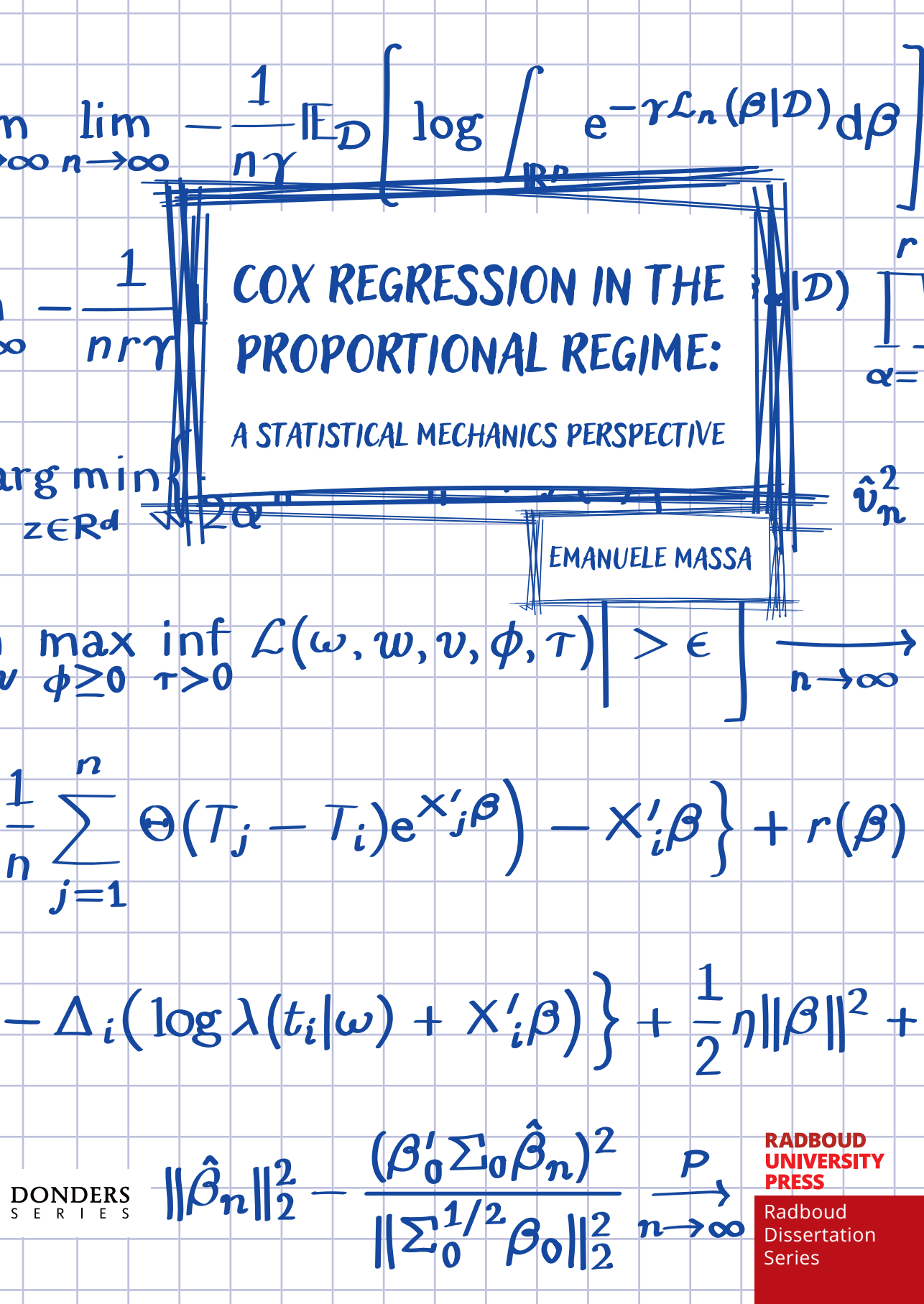


DONDERS SERIES

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Cox regression in the proportional regime: a statistical mechanics perspective

by

Emanuele Massa

Emanuele Massa

Cox regression in the proportional regime: a statistical mechanics perspective

Radboud Dissertation Series

ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Proefschrift AIO/Guus Gijben
Cover image: Guntra Laivacuma
Printing: DPN Rikken/Pumbo
ISBN: 9789465150857
DOI: 10.54195/9789465150857

Free download at: https://doi.org/10.54195/9789465150857

RADBOUD
UNIVERSITY
PRESS

Cox regression in the proportional regime: a statistical mechanics perspective

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 16 september 2025
om 16:30 uur precies

door

Emanuele Massa
geboren op 19 september 1995
te Genua, Italië

Promotoren

Prof. dr. A.C.C. Coolen
Prof. dr. C.B. Roes

Copromotor

Dr. M.A. Jonker

Manuscriptcommissie

Prof. dr. H.J. Kappen
Dr. H. Battey (Imperial College London, Verenigd Koninkrijk)
Prof. dr. M. Opper (Technische Universität Berlin, Duitsland)

Dedication

A mamma Patrizia e papá Luca, per avermi dato l' opportunita' di studiare e per il loro incondizionato e sempre presente supporto. A mio fratello Federico perché sai sempre come farmi ridere e tirarmi su di morale. Anche se questi anni ci hanno messo a dura prova, siamo sempre la famiglia Massa, e sono molto fiero di ció.

Aan mijn partner Manon, jouw liefde en voortdurende steun gaven mij de kracht om door te gaan. Zonder jou zou dit proefschrift er niet zijn. Ik kijk uit naar onze toekomst samen, ik weet zeker dat wij samen alles kunnen doen.

Contents

Chapter 1

Introduction

Survival analysis is a branch of statistics dealing with time to event data. A time-to-event data point is a measurement of the amount of time that is passed between a predefined time origin and the event of interest. To be concrete, the predefined time origin might be the time at which a patient is admitted in a medical study, or when a user subscribes to a video streaming platform the first time. The event of interest can be cancer relapse, or when the user terminates his/her subscription. Survival analysis might be defined succinctly as the statistical analysis of survival data, i.e. the collection of methods used to conduct inference and making predictions with survival data. These methods find application in several fields ranging from epidemiology, to finance and engineering. For this reason, the user/ patient in the previous example is generically referred to as the subject. In general, the scientist involved in the analysis of time to event data is interested in understanding which "features" of a subject (and eventually how) influence its survival time. Having an idea of the relationship between the covariates (i.e. the "features") and survival time allows performing patient screening and, eventually, aid decision-making for treatment. Furthermore, recent years have seen a surge of interest in the concept of personalized or precision medicine. The core idea is to gather many covariates for each subject, e.g. clinical information and genetic characteristics, and then use this massive amount of data to discover patterns, i.e. understand which covariates are important to predict the survival time of patients. The fundamental belief is that this might eventually lead to a better understanding of the mechanisms underlying diseases and allow for personalized treatment. However, traditional regression methods in survival analysis are not suited to cope with such "high dimensional" data.

The scope of this thesis is to understand the limitations of regression methods in Survival analysis in the modern high dimensional regime, eventually proposing novel strategies to improve the latter.

1.1 A short introduction to Survival Analysis

Time to event data are generally censored, i.e. the information over certain subjects is incomplete. There exist (at least) three type of censoring known respectively as left, interval and right censoring [40]. When it is only known that a subject experienced the event before a certain time, we say that the subject is left censored; e.g. the time at which a patient is diagnosed might (most likely) not coincide with the onset time of the disease. A subject is interval censored if it is only known that she experienced the event in a certain time interval; e.g. when a patient is regularly monitored by a physician, the relapse of an acute medical condition is only known to have happened in between the penultimate and last check-ups.

Finally, a patient is right censored when it is only known that she experienced the event after a certain time (if at all); for instance, if death is the event of interest and a patient under study survives until the end of the study. In this thesis, we will only deal with right censoring, being the case encountered most often in applications. The discussion might be extended "mutati mutandis" to other censoring types. We will further assume that the mechanism underlying censoring is uninformative, i.e. the event "Alice is censored" does not depend on Alice's covariates. The assumption of uninformative censoring is mathematically convenient, but might not always be realistic. In that case, competing risks models [48] might be considered, but these are not treated in the present thesis.

1.1.1 Homogeneous Populations

In order to introduce the mathematical notation relevant for Survival Analysis, we start by considering the case of a homogeneous population. An homogeneous population might be thought of as made up of a priori "identical" subjects. Perhaps the simplest example is that of a batch of e.g. springs produced in the same factory, all in the same manner. If we are interested in testing the mechanical strain resistance of the springs, we might run an experiment where we stretch the springs until they break. In this case the survival time is the positive stretching force applied on the string, when the latter breaks. It is clear that not all the springs will break at the same value of the applied force, but there will rather be a distribution of forces. Models used for heterogeneous populations are introduced as a generalization of the homogenous ones in the next section. Mathematically, a subject is taken to be a realization of a random variable T, the time to event, which follows an unknown law, also known as probability distribution. This associates to each event $T \in (t, t + \mathrm{d}t)$ a probability $f(t)\mathrm{d}t$. One of the aim of survival analysis, in this context, is to infer the shape of f, i.e. the law of the population, from a sample $T_1, \ldots, T_n$ (a finite set of measurements), i.e. the observed time to event for n subjects.
In order to make inference and/or predict the survival time of future patients, we usually need to postulate a statistical model. A statistical model encodes our assumptions regarding the data generating process. In practice, this amounts to assuming that the data generating distribution belongs to a certain class of functions or that it can be well approximated by a function in the latter. Often, Survival analysis models are specified in terms of the hazard rate function λ. This is the probability per unit time at which a subject experiences the event of interest, given that he/she has survived up to t. Mathematically,

$$\lambda(t)\ \mathrm{d}t := P\Big[T \in (t, t + dt)\Big|T > t\Big]\ . \tag{1.1}$$

Knowledge of λ fixes the density f too. Introducing the survival function

$$S(t) := P\Big[T > t\Big]\ , \tag{1.2}$$

we notice that by definition

$$f(t)\mathrm{d}t = P\Big[T \in (t, t + dt)\Big] = P\Big[T \in (t, t + dt)\Big|T > t\Big]P\Big[T > t\Big] = \lambda(t)\ S(t)\ \mathrm{d}t\ , \tag{1.3}$$

on the other hand

$$f(t)\mathrm{d}t = P\Big[T \geq t\Big] - P\Big[T \geq t + dt\Big] = -\frac{\mathrm{d}}{\mathrm{d}t}S(t)\ \mathrm{d}t \tag{1.4}$$

which together imply

$$-\frac{\mathrm{d}}{\mathrm{d}t}S(t) = \lambda(t)\ S(t)\ , \tag{1.5}$$

which is solved by

$$S(t) = \mathrm{e}^{-\Lambda(t)}, \quad \Lambda(t) := \int_0^t \lambda(t)\mathrm{d}t\ , \tag{1.6}$$

since $S(0) = 1$ by definition of survival function. Hence, the probability density for a positive random variable T can always be written as

$$f(t) = \lambda(t)\mathrm{e}^{-\Lambda(t)}, \quad t \in \mathbb{R}^+ \to \lambda(t) \in \mathbb{R}^+\ . \tag{1.7}$$

Consider the case where all n subjects under investigation have independently experienced the event of interest before the end of the study, then the likelihood of the sample reads

$$\prod_{i=1}^{n} f(T_i) = \prod_{i=1}^{n} \lambda(T_i)\mathrm{e}^{-\Lambda(T_i)}\ . \tag{1.8}$$

In practice, however, some subjects will be censored. When a subject is right censored, for instance, the typical format in which data are stored is the following: for each subject a time $T \in \mathbb{R}_+$ and an indicator $\Delta \in \{0, 1\}$ are recorded (a value $\Delta = 1$ indicates that the subject experienced the event while under observation and 0 otherwise). Pragmatically one might describe the data generating process by introducing two fictitious random variables, the latent event time Y and the latent censoring time C and define the observable T, Δ as follows

$$T = \min\{Y, C\} \qquad \Delta = \mathbf{1}\Big[Y \leq C\Big]\ . \tag{1.9}$$

The latent event time Y is the time at which the subject experience the event under study, i.e. Y follows the model (1.7), which is only observable if the subject has not yet been censored, i.e. if $\Delta = 1$, meaning $Y < C$. The probability of the event $T \in (t, t + \mathrm{d}t), \Delta = 1$ under the model (1.9) can then be easily deduced,

$$\begin{aligned} P\Big[T \in (t, t+\mathrm{d}t), \Delta = 1\Big] &= P\Big[Y \in (t, t+\mathrm{d}t), Y \leq C\Big] = \\ &= P\Big[C > t + \mathrm{d}t\Big] P\Big[Y \in (t, t+\mathrm{d}t)\Big] = \lambda(t)\mathrm{e}^{-\Lambda(t)} S_C(t+\mathrm{d}t)\mathrm{d}t \end{aligned} \tag{1.10}$$

where $S_C(t) = P\Big[C > t\Big]$ is the survival function of C. The case $\Delta = 0$ can be worked out similarly and gives

$$P\Big[T \in (t, t+\mathrm{d}t), \Delta = 0\Big] = \mathrm{e}^{-\Lambda(t+\mathrm{d}t)} f_C(t)\mathrm{d}t \tag{1.11}$$

where $f_c(t)$ is the density of C. This gives the probability density

$$f(t, \Delta) = \lambda(t)^{\Delta}\mathrm{e}^{-\Lambda(t)} \times S_C(t)^{\Delta} f_C(t)^{1-\Delta}\ . \tag{1.12}$$

Since we are only interested in fitting the value of λ, the additional contribution coming from the distribution of the censoring time is generally neglected, as it appears as an additive constant in the log-likelihood.

1.1.2 Heterogeneous Populations

As already anticipated in the introduction of this chapter, it is often of interest to understand how the covariates of the subject are related to the survival time. In medicine, for example, the risk of prostate cancer might increase with the age of the patient and its smoking habits. This motivates the introduction of survival regression models that try to capture the relationship between the event time t of the subject and his/her characteristics $\mathbf{X}$. This is equivalent to specify an hazard rate function $\lambda(.|\mathbf{X})$ which depends not only on time, but also on the covariates $\mathbf{X} \in \mathbb{R}^p$. Different functional forms for the hazard $\lambda(.|\mathbf{X})$ or assumptions on how $\mathbf{X}$ enters in $\lambda(.|\mathbf{X})$ yield different models.
The class of Proportional Hazards (PH) models, for instance, postulates that $\mathbf{X}$ appears in the hazard rate only via a so-called risk score $\exp\{\mathbf{X}'\boldsymbol{\beta}\}$, which multiplies the common baseline hazard rate λ

$$\lambda(t|\mathbf{X}) = \lambda(t)\exp\{\mathbf{X}'\boldsymbol{\beta}\} \ . \tag{1.13}$$

Above we have indicated with $'$ the scalar product of two vectors, i.e. for $\boldsymbol{a}, \boldsymbol{v} \in \mathbb{R}^d$, $\boldsymbol{a}'\boldsymbol{v} := \sum_{k=1}^d a_k v_k$. This implies that the ratio of the hazards of two patients depends only on the difference in risk score which is constant over time

$$\log \frac{\lambda(t|\mathbf{X})}{\lambda(t|\mathbf{Z})} = \mathbf{X}'\boldsymbol{\beta} - \mathbf{Z}'\boldsymbol{\beta} \tag{1.14}$$

and hence the name, since the hazards for different patients are proportional (at all times with the same constant).
Another example is the class of Accelerated Failure Time (AFT) models,

$$\lambda(t|\mathbf{X}) = \lambda(t\exp\{\mathbf{X}'\boldsymbol{\beta}\})\exp\{\mathbf{X}'\boldsymbol{\beta}\} \ . \tag{1.15}$$

For a complete overview of different survival models we refer to [48]. In this thesis, we will focus on fully-parametric and semi-parametric models.
Fully parametric models assume a parametric form for the hazard rate function λ, e.g. the Weibull model, which is the only model in both PH and AFT classes, reads

$$\lambda_{\boldsymbol{\theta}}(t|\mathbf{X}) = \rho \mathrm{e}^{\phi} t^{\rho-1} \mathrm{e}^{\mathbf{X}'\boldsymbol{\beta}}, \quad \boldsymbol{\theta} := (\rho, \phi, \boldsymbol{\beta}) \in \mathbb{R}^+ \times \mathbb{R} \times \mathbb{R}^p \ . \tag{1.16}$$

But making assumptions on the baseline hazard might result in a mis-specified model, hence practictioners generally prefer the Cox semi-parametric PH model [25], which does not make any assumption on the baseline hazard, but assumes the PH form. The latter can be informally regarded as a parametric model with the function $t \to \lambda_0(t)$ being an "infinite dimensional" parameter to be inferred from the data, i.e.

$$\lambda_{\boldsymbol{\theta}}(t|\mathbf{x}) = \lambda_0(t)\mathrm{e}^{\mathbf{X}'\boldsymbol{\beta}}, \quad \boldsymbol{\theta} = (\lambda_0, \boldsymbol{\beta}) \in \mathcal{F} \times \mathbb{R}^p \tag{1.17}$$

with $\mathcal{F}$ is the set of all mappings f from $\mathbb{R}^+$ to $\mathbb{R}^+$, such that $\exp\big\{-\int_0^{+\infty} f(t)\mathrm{d}t\big\} = 0$.
Alternatives to the Cox semi-parametric PH model are flexible parametric models for the hazard rate, such as the Piece-Wise Exponential model [35]. In this case the hazard is assumed to be piece-wise constant, with jumps at $d+1$ knots $\tau_1, \ldots, \tau_{d+1}$, this gives

$$\lambda_{\boldsymbol{\theta}}(t|\mathbf{X}) = \sum_{k=1}^{d} \exp(\omega_k)\mathbf{1}[\tau_k < t < \tau_{k+1}]\mathrm{e}^{\mathbf{X}'\boldsymbol{\beta}}, \quad \boldsymbol{\theta} := (\omega_1, \ldots, \omega_d, \boldsymbol{\beta}) \in \mathbb{R}^d \times \mathbb{R}^p \ ,$$

where $\exp(\omega_k)$ is the value of the hazard in the interval (τ_k, τ_{k+1}). This is a simple case of a parametrization in terms of a linear combination of basis spline functions [8].

Having briefly illustrated a restricted subset of the models available in the analysis of time to event data, we give a short introduction to estimation via the maximum likelihood method.

1.2 A primer on frequentist estimation

A statistical regression model is a collection of conditional distributions of the response Y given the covariates $\mathbf{X}$, i.e.

$$Y|\mathbf{X} \sim f(.|\mathbf{X}'\boldsymbol{\beta}, \boldsymbol{\sigma}), \tag{1.18}$$

with the functional form f fixed and where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the regression parameters and $\boldsymbol{\sigma} \in \mathbb{R}^d$ are the nuisance parameters. To be clear, in Linear regression Y is the continuous value to be predicted, whilst in Survival analysis $Y = (\Delta, T)$, an array containing as first component the event indicator and as second component the event time. Each value of the parameters $\boldsymbol{\beta}, \boldsymbol{\sigma}$ identifies a conditional distribution that might describe the empirical data.

Once a model is chosen, one wonders : i) if there is a "best" conditional distribution in our putative set and ii) how do we select it. Generally, the most widely adopted method to select a conditional distribution is the Maximum Likelihood (ML) one, originally introduced by Fisher [34, 2] (although other methods like e.g. the method of moments exists and are used in some contexts). This prescribes to select the values of $\boldsymbol{\beta}, \boldsymbol{\sigma}$ that maximize the likelihood of the data

$$\prod_{i=1}^{n} f(Y_i|\mathbf{X}_i'\boldsymbol{\beta}, \boldsymbol{\sigma}), \tag{1.19}$$

or, as is done in practice more often, by minimizing minus the logarithm of the latter (rescaled by the sample size for better numerical behaviour)

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\sigma}) := -\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|\mathbf{X}_i'\boldsymbol{\beta}, \boldsymbol{\sigma}) \ , \tag{1.20}$$

where ℓ stands for loss. The value of $\boldsymbol{\beta}, \boldsymbol{\sigma}$ where ℓ_n attains its minimum is called the Maximum Likelihood Estimator (since there the Likelihood is maximal)

$$\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\sigma}}_n := \arg\min_{\boldsymbol{\beta},\boldsymbol{\sigma}} \{\ell_n(\boldsymbol{\beta}, \boldsymbol{\sigma})\} \ . \tag{1.21}$$

This choice for $\boldsymbol{\beta}, \boldsymbol{\sigma}$ is optimal in the sense that it minimizes the "distance" (which we shall soon define appropriately) between the conditional distribution of the data under the model (1.18) and the empirical density of the data

$$f_n(y, \mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} \delta(y - Y_i)\delta(\mathbf{x} - \mathbf{X}_i), \tag{1.22}$$

where δ is the Dirac's delta, which assigns a probability mass of $1/n$ to each point in the training data-set. This "distance" is quantified by the Kullback-Liebler(KL) divergence [24]. Given two distributions $f, g : \boldsymbol{\Omega} \subseteq \mathbb{R}^\ell \to \mathbb{R}^+$ (almost everywhere non-vanishing over $\boldsymbol{\Omega}$), the KL divergence is defined as

$$D_{KL}[f, g] := \int_{\boldsymbol{\Omega}\subseteq\mathbb{R}^\ell} f(\mathbf{x}) \log\left[\frac{f(\mathbf{x})}{g(\mathbf{x})}\right] \mathrm{d}\mathbf{x} \ . \tag{1.23}$$

The Jensen inequality [71] can be used to show that the quantity above is non-negative and vanishes if and only if the f coincides with g almost everywhere, i.e. on all subsets of $\boldsymbol{\Omega}$ with non-vanishing f probability density. However, D_{KL} is not symmetric and does not satisfy

the triangle inequality, thus it is formally not a distance and hence the name divergence. In our setting, we have

$$\begin{aligned} D_{KL}[f_n, f] &= -\int f_n(y, \mathbf{x}) \log \frac{f(y|\mathbf{x}'\boldsymbol{\beta}, \boldsymbol{\sigma})}{f_n(y, \mathbf{x})} \mathrm{d}y\mathrm{d}\mathbf{x} = \\ &= \int f_n(y, \mathbf{x}) \log f(y|\mathbf{x}'\boldsymbol{\beta}, \boldsymbol{\sigma}) \mathrm{d}y\mathrm{d}\mathbf{x} - H[f_n] = \qquad (1.24) \\ &= \ell_n(\boldsymbol{\beta}, \boldsymbol{\sigma}) + H[f_n] \qquad (1.25) \end{aligned}$$

where $H[f_n]$ is the Shannon Entropy of the discrete distribution f_n [24], which does not depend on the parameters of the model. Hence, the Maximum Likelihood method selects the value of $(\boldsymbol{\beta}, \boldsymbol{\sigma})$ that makes our model closest to the empirical density of the data f_n. However, one would rather like to minimize the divergence $D_{KL}[f_0, f]$ between the model (1.18) and the *actual* density from which the data are (assumed to be) "pulled", which we call f_0.

The Maximum Likelihood (ML) methodology was developed under the scenario where the number of covariates p is much smaller than the total number of patients recorded n. In this case, it is possible to show that the ML estimator posses several desirable properties like consistency and asymptotic normality and is even optimal in a certain sense, as it saturates the Cramer-Rao bound [11, 84]. This scaling regime was realistic during the last century, when the number of recordings per patient (the number of covariates) counted few units. Although the "classical" setting with $p \ll n$ remains relevant in a number of applications, for instance when there is experimental control of covariates. Nowadays, the scenarios "small n, large p" and "large n, large p", where the number of recordings per patient p is much larger than, or comparable to, the total number of patients recorded n, are becoming increasingly common. In this modern setting, which is referred to as "high dimensional regime", the classical methodology breaks down: the ML estimator might not be well-defined or may not even exist; when it does exist, it is frequently biased and has a large variance, making it an unreliable tool for inference and prediction.

While the main focus of the thesis is on survival analysis models, it is easier to showcase some of the aforementioned "overfitting problems", arising when p is proportional to n, in the context of the linear regression model. For this model the ML estimator is available in closed form, i.e. can be written explicitly as a function of the data, hence the main concepts can be easily illustrated in this framework. Survival analysis models, in contrast, are non-linear and there exists no closed form for the ML estimator. We must then resort to more complicated mathematical methods, which might hinder the exposition of the core ideas at this stage, but will be timely addressed in the following chapters of the thesis.

1.3 What is the problem with high dimensional data?

Consider data generated according to a simple linear model,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n) \ , \qquad (1.26)$$

with $\mathbf{X} \in \mathbb{R}^{n \times p}$ fixed and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ fixed.
To make an example, suppose we model the height of the firstborn child (at adulthoood) as a function of the height of the father and of the mother. That is, we suppose that the height of the firstborn child is a linear combination of the height of the parents. To do this, we use data from a registry containing the measured height of 1000 firstborn sons or daughters,

together with the height of the parents. Sex might be an important factor here, as generally men are taller than women, so we include a binary variable indicating the gender of the child. In this (very simplistic) model $p = 3$, $n = 1000$ and the design matrix $\mathbf{X}$ is 1000×3. When $p < n$ the ML estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}_0$ exists and can be computed explicitly as

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big\{\frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\Big\} = \big(\mathbf{X}'\mathbf{X}\big)^{-1}\mathbf{X}'\mathbf{Y} \ , \tag{1.27}$$

provided $\mathbf{X}'\mathbf{X}$ is non-singular, i.e. all its eigenvalues are strictly positive. Substituting the expression for the linear model, we obtain

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + \tilde{\boldsymbol{\epsilon}}, \qquad \tilde{\boldsymbol{\epsilon}} := \big(\mathbf{X}'\mathbf{X}\big)^{-1}\mathbf{X}'\boldsymbol{\epsilon} \ , \tag{1.28}$$

hence $\hat{\boldsymbol{\beta}}_n$ is unbiased, since $\tilde{\boldsymbol{\epsilon}}$ has a null expectation by (1.26). The Mean Squared Error (MSE) reads

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2\Big] = \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\|\tilde{\boldsymbol{\epsilon}}\|_2^2\Big] = \sigma^2\mathrm{Tr}\Big(\big(\mathbf{X}'\mathbf{X}\big)^{-1}\Big) = \frac{p}{n}\ \sigma^2\ \frac{1}{p}\mathrm{Tr}\big(\hat{\boldsymbol{\Sigma}}_n^{\ -1}\big) \tag{1.29}$$

where the expectation is conditional on the covariates, which are regarded as fixed, and we defined $\hat{\boldsymbol{\Sigma}}_n := \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'/n$, the empirical covariance matrix. Notice that

$$\lambda_{\max}^{-1}(\hat{\boldsymbol{\Sigma}}_n) \le \frac{1}{p}\mathrm{Tr}\big(\hat{\boldsymbol{\Sigma}}_n^{-1}\big) \le \lambda_{\min}^{-1}(\hat{\boldsymbol{\Sigma}}_n), \tag{1.30}$$

where $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_n), \lambda_{\max}(\hat{\boldsymbol{\Sigma}}_n)$ are, respectively, the smallest and largest eigenvalue of $\hat{\boldsymbol{\Sigma}}_n$. Hence, the MSE is controlled by the ratio p/n, the minimum eigenvalue of the empirical covariance matrix $\lambda_{\min}^{-1}(\hat{\boldsymbol{\Sigma}}_n)$ and by the conditioning number $c(\hat{\boldsymbol{\Sigma}}_n) := \lambda_{\max}(\hat{\boldsymbol{\Sigma}}_n)/\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_n)$. Whilst the prediction error is

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\|_2^2\Big] = \frac{1}{n}\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\boldsymbol{\epsilon}'\mathbf{P}\boldsymbol{\epsilon}\Big] = \frac{p}{n}\sigma^2, \quad \mathbf{P} := \mathbf{X}\big(\mathbf{X}'\mathbf{X}\big)^{-1}\mathbf{X} \tag{1.31}$$

since $\mathbf{P}$ is a projection matrix of rank p.
When $p \ll n$ and provided $p\lambda_{\min}^{-1}(\hat{\boldsymbol{\Sigma}}_n)\sigma^2 = o(n)$, the estimation error, quantified by the MSE, and the prediction error tend asymptotically to 0 as n diverges to infinity. In particular, the fact that the MSE vanishes asymptotically implies (via Chebishev inequality) that the estimator $\hat{\boldsymbol{\beta}}_n$ is consistent for the *whole* vector $\boldsymbol{\beta}_0$, i.e.

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 \xrightarrow[n\to\infty]{P} 0 \ . \tag{1.32}$$

In words, as n increases we estimate better and better the target $\boldsymbol{\beta}_0$.

This picture changes as soon as p is comparable (not even bigger than) n, i.e. when p grows proportionally to n ($p < n$), such that $p = p_n = \zeta n,\ \zeta \in (0, 1)$. In this case,

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\frac{1}{n}\|\mathbf{X}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\|_2^2\Big] = \zeta\sigma^2 \tag{1.33}$$

and

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2\Big] = \zeta\delta\lambda_{\min}^{-1}(\hat{\boldsymbol{\Sigma}}_n), \qquad \delta \in [1/c(\hat{\boldsymbol{\Sigma}}_n), 1] \tag{1.34}$$

which does *not* vanish as $n \to \infty$. Hence, the consistency of $\hat{\boldsymbol{\beta}}_n$ is lost, and the prediction error does not vanish, even asymptotically. The situation further complicates when $p > n$.

In that case, the empirical covariance matrix $\hat{\mathbf{\Sigma}}_n$ is singular, and the Maximum Likelihood estimator is ill-defined, since there exist infinite minimizers of (1.27). Notice, however, that we have, for any $\mathbf{v}$ independent from the data $\mathbf{Y}, \mathbf{X}$

$$\mathbf{v}'\hat{\boldsymbol{\beta}}_n = \mathbf{v}'\boldsymbol{\beta}_0 + \varepsilon, \quad \varepsilon := \mathbf{v}'\big(\mathbf{X}'\mathbf{X}\big)^{-1}\mathbf{X}'\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, v_n^2), \tag{1.35}$$

where

$$v_n^2 := \frac{1}{n}\mathbf{v}'\hat{\mathbf{\Sigma}}_n^{-1}\mathbf{v}, \quad \frac{1}{n}\lambda_{\max}^{-1}(\hat{\mathbf{\Sigma}}_n)\|\mathbf{v}\|_2^2 \leq v_n^2 \leq \frac{1}{n}\lambda_{\min}^{-1}(\hat{\mathbf{\Sigma}}_n)\|\mathbf{v}\|_2^2 \ . \tag{1.36}$$

Thus, provided $\lambda_{\min}^{-1}(\hat{\mathbf{\Sigma}}_n)\|\mathbf{v}\|_2^2 = o(n)$, we have L_2 consistency

$$\|\mathbf{v}'\hat{\boldsymbol{\beta}}_n - \mathbf{v}'\boldsymbol{\beta}_0\|_2^2 \xrightarrow[n\to\infty]{P} 0 \ . \tag{1.37}$$

This means that the estimation error for each component is still "small", however, when adding all the contributions up, the vector difference remains finite even in the limit, since the number of components p grows with n proportionally.

The impossibility of estimating consistently a large number of parameters *simultaneously* is a manifestation of the so-called "curse of dimensionality". This "problem" is known in the statistical community since the 60', starting from the works of Kolmogorov and Huber, notably [86, 43, 31]. In a certain sense, we are "asking too much" : when p is large and comparable to, or larger than, n there is not enough data to meaningfully estimate *all* the parameters of the model. Put it in other words, if p is of order n, minimizing the Kullback-Liebler divergence (1.23) between the model (1.18) and the empirical density leads to a poor estimate of $(\boldsymbol{\beta}, \boldsymbol{\sigma})$. The value of the parameters $(\boldsymbol{\beta}, \boldsymbol{\sigma})$ that are selected leads to a conditional distribution that "mimics excessively" the empirical density, in jargon it is said that the model "over-fits" [3]. This can be easily seen by computing the estimator for the noise variance

$$\hat{\boldsymbol{\sigma}}_n^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2 = \frac{1}{n}\boldsymbol{\epsilon}'(\boldsymbol{I}_n - \mathbf{P})\boldsymbol{\epsilon} \ . \tag{1.38}$$

Its expectation equals (again, always for $p < n$)

$$\mathbb{E}\Big[\hat{\boldsymbol{\sigma}}_n^2\Big] = \frac{1}{n}\mathbb{E}\Big[\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2\Big] = \frac{1}{n}\mathbb{E}\Big[\boldsymbol{\epsilon}'(\boldsymbol{I}_n - \mathbf{P})\boldsymbol{\epsilon}\Big] = (1 - p/n)\sigma^2 \ . \tag{1.39}$$

It can be shown via standard concentration inequalities [31, 85] that (1.38) concentrates around its average for large n (1.39), i.e. the fluctuations around the mean are of order $1/n$. Hence, as the ratio $\zeta = p/n$ increases towards one, the variance is progressively under-estimated, eventually reaching zero at $\zeta = 1$. This shows that the model increasingly "interprets noise as signal", leading to interpolation when $\zeta = 1$ where the inferred noise variance $\hat{\boldsymbol{\sigma}}_n^2$ is zero (with high probability) and the Mean Squared Error of $\hat{\boldsymbol{\beta}}_n$ equals σ^2.

The lack of consistency prevents the study of the asymptotics of MAP estimators via the classical statistical framework, as in e.g. [33], since Taylor expansions around the true value are not valid. A possible way to restore consistency is to reduce the number of covariates included in the regression model, i.e. reduce p. This is the subject of the next subsection.

1.4 Feature selection and data reduction

Feature selection and data reduction are distinct collections of methods which aim at selecting a subset of the original covariates/features in order to avoid the problems that we have

discussed in the previous section by effectively reducing p, the number of covariates included in the regression model. The main distinction between feature selection and data reduction is in how the data are used to make this selection. Feature (or model) selection makes use of the response data $\mathbf{Y} = (Y_1, \ldots, Y_n)$, whilst data reduction does not. The aims are also intrinsically different : feature or model selection aims at selecting a target "true" model that generated the data, whilst data reduction only aims at reducing the number of covariates.

1.4.1 Feature selection

Suppose that originally we are given p "raw" covariates. A priori there exist 2^p models, one for each possible subset of covariates (not accounting for the interactions between variables!),

$$\mathcal{M}_{\boldsymbol{\tau}} := \{\boldsymbol{\beta}_0 \ : \ \beta_{0,k} = 0, \text{ if } \tau_k = 0\}, \ \boldsymbol{\tau} \in \{0,1\}^p \ . \tag{1.40}$$

In a high dimensional setting, brute force or "all subset" model selection by searching in the set above is computationally intractable since, as noted above, the cardinality of this set is exponential in p. Alternatively, one can restrict the set of models to

$$\mathcal{M}_k := \{\boldsymbol{\beta}_0 \ : \ \beta_{0,k+1} = \beta_{0,k+2} = \cdots = \beta_{0,p} = 0\}, \ k \in \{1, \ldots, p\} \ , \tag{1.41}$$

which is a sequence of subsets, i.e. $M_k \subset M_{k+1}$. In this case, one is required to compare p different models at most, which is more computationally tractable. However, compared to all subset selection, the ordering (or labelling) of the various features matters. The true model might include, say features $1, 2, 10$, which is a subset of cardinality 3, but is only included in $\mathcal{M}_{10}$ [11]. There exist testing procedures that implement this idea in practice, like stepwise forward (or backward) selection [13] and minimization of model selection criteria such as AIC [1] and BIC [72], to name a few. Step-wise selection is advocated to be unstable, as small perturbation of the data set can lead to wildly different estimated models [13]. Furthermore, statistical inference after having selected a model based on the response data should account for the model selection procedure : confidence statements computed "as if" the selected model were the true one are too narrow. Likewise, goodness of fit or prediction metrics will be optimistically biased. However, besides very simple models, the computation of the confidence intervals is far from trivial [40, 50]. Nevertheless, a reduction of the number of covariates to be included in the model might be necessary in applications, for instance when the sample size is not sufficiently large. For this reason, post model selection inference is still an active research topic. In this context, the Cox reduction method [26] has been recently proposed as an approximation to all subset selection in the high dimensional regime. The authors proposed the following procedure. Initially, the covariates are arranged on a cube $s \times s \times s = p$ (without loss on generality as one can just add zeros to make a perfect cube). This cube can be traversed in $3s^2$ ways, each time collecting s features. The idea then is to run $3s^2$ regressions over s variables (so that each feature is regressed 3 times with other $s-1$ different covariates) and retain those features that are deemed significant (up to a certain user defined confidence level via a statistical test) two or more times (on the total of 3, for each covariate). This procedure is iterated, (if necessary one arranges the features in a square, rather than a cube and proceeds similarly), until $10, 20$ covariates are left. Then one can proceed to check if the remaining features are dependent and eventually eliminate those that are strongly correlated. Finally, once a sufficienlty small number of covariates is isolated, one can run an "all subset regression" in order to select the models that are consistent with the data. Recently, a theoretical analysis of this model selection process have been carried

out in [52]. There, the authors point out some weaknesses and propose improvements of the original approach of [26]. In particular, a strength of the method is the fact that it should be consistent if the procedure is run several times with different randomization. An inconsistency points out that the conclusion drawn by the method might not be accurate.

1.4.2 Data reduction

Data-reduction techniques, conversely, aim at reducing the number of predictors/covariates to be included in the regression model, in an "un-supervised" manner, i.e. without using the responses. Doing so guarantees to not alter the statistical inference task that follows [40]. Perhaps the simplest instance of data reduction is Principal Components (PC). Consider the sample scatter matrix $S_n := \mathbf{X}'\mathbf{X}$, this is a symmetric positive semi-definite matrix in $\mathbb{R}^{p\times p}$, hence it can be diagonalized as

$$S = \mathbf{X}'\mathbf{X} = \mathbf{O}'\mathbf{\Lambda}\mathbf{O}, \tag{1.42}$$

with $\mathbf{O} \in \mathbb{R}^{p\times p}$ an orthogonal matrix and $\mathbf{\Lambda}$ a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_p$, i.e. $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$. By its definition, there can be at most $\min\{p, n\}$ non-zero eigenvalues. The eigenvalues can always be ordered as $\lambda_1 > \lambda_2 > \cdots > \lambda_p$: the first principal component is the eigenvector corresponding to the first eigenvalue (the largest), and so on and so forth. Retaining only a subset of the principal components, say those corresponding to eigenvalues larger than a certain (user specified) threshold, leads to a reduced, or more parsimonious, model. However, it remains the problem of how to choose in practice the threshold, or equivalently how many principal components one should include. After having reduced the original covariate vector, retaining, say, the first $k < p$ Principal Components, one can conduct ordinary regression. This is generally known as Principal Component Analysis (PCA). There exist several more involved generalization of PCA. We refer to [41] for an exhaustive presentation, which is not the scope of this manuscript. Alternative strategies, that do not rely on Principal Components, have also been proposed in the literature [40]. In redundancy analysis, for instance, the idea is that some columns of $\mathbf{X}$ are redundant, i.e. can be predicted from the knowledge of the remaining columns. The method proceed to eliminate features that can be predicted sufficiently accurately (which is quantified by a user provided threshold), by a possibly non-linear relationship involving the remaining predictors. More generally, methods aiming at extracting a reduced representation of the features, from knowledge of the features only, go under the umbrella of unsupervised learning algorithms [41].

Instead of reducing the number of covariates/features included in the regression model, an alternative way to restore consistency is to incorporate previous knowledge in the regression problem. This is the topic of the next section.

1.5 Prior beliefs, Bayesian statistics and consistency in the high dimensional regime

Previous or "prior" knowledge is essentially incorporated as *unverifiable* assumptions on the data generating process. This is often done by adding a regularizer, also-called a penalization function, to the objective function (1.20) that is to be minimized in order to estimate the parameters of our model. The regularizer acts by discouraging *a priori* unlikely values for the parameters of the model, effectively penalizing the latter, hence the alternative name.

Such a procedure might be perceived to be "ad hoc", or "artificial", however, it has a "natural" interpretation in the Bayesian formulation of statistics.

In Bayesian statistics, after a likelihood function (or equivalently a model) is assumed, the central object of interest is the *posterior* distribution of the model parameters given the data

$$p(\boldsymbol{\beta}, \boldsymbol{\sigma}|\mathcal{D}) \propto \left\{ \prod_{i=1}^{n} f(T_i|\mathbf{X}_i'\boldsymbol{\beta}, \boldsymbol{\sigma}) \right\} \pi(\boldsymbol{\beta}, \boldsymbol{\sigma}), \tag{1.43}$$

where π is the prior density function. Mathematically, the prior function incorporates the previous knowledge of the scientist by assigning more weight (density) to values of the parameters that are *believed* (by all scientist with the same prior π) to be more likely "a priori", i.e. before seeing the data. The posterior function might be interpreted as the updated belief of the scientist (with the prior π) over the parameters $(\boldsymbol{\beta}, \boldsymbol{\sigma})$ given the data (the empirical evidence). At difference with the Frequentist framework, where we select a parameter value from model (1.18), also-called a point estimate [11], in Bayesian statistics we obtain a distribution over all the conditional distributions (identified by the values of the parameters, once a likelihood is assumed) in the model. This distribution is subjective, i.e. depends on the prior. This means that all scientists with the same prior will agree on the conclusion, i.e. will have the same posterior, but this is not true if the priors are different. However, one might argue that also in the Frequentist approach the scientist chooses a model, based on e.g. affinity with the data and computational convenience, and this point is necessarily subjective. Hence, all the scientist that agree on the same set of assumptions will finally draw the same conclusion. Eventually, in a fully Bayesian approach, we should consider also a prior over the different models adopted in the analysis [37].

In general, however, obtaining a closed form expression for the posterior is impossible (computationally hard). When the number of parameters included in the model is reasonably contained, there exist numerical ways of sampling from the posterior (1.43), e.g. via the Monte-Carlo Markov Chain (MCMC) method, or via the Gibbs sampling procedure [37]. In the modern high dimensional regime these methods are computationally slower [9, 18] and the posterior density might be replaced, for computational convenience, with a Dirac delta at the posterior mode (the value of $\boldsymbol{\beta}, \boldsymbol{\sigma}$ where the posterior is maximal), also known as the Maximum A Posterior (MAP) estimator. By its definition the MAP estimator solves

$$\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\sigma}}_n := \arg\min_{\boldsymbol{\beta}, \boldsymbol{\sigma}} \{\mathcal{H}_n(\boldsymbol{\beta}, \boldsymbol{\sigma})\} \tag{1.44}$$

$$\mathcal{H}_n(\boldsymbol{\beta}, \boldsymbol{\sigma}) := n\ell_n(\boldsymbol{\beta}, \boldsymbol{\sigma}) + \mathrm{r}(\boldsymbol{\beta}, \boldsymbol{\sigma}) \ , \tag{1.45}$$

where r is called the regularization function

$$\mathrm{r}(\boldsymbol{\beta}, \boldsymbol{\sigma}) := -\log \pi(\boldsymbol{\beta}, \boldsymbol{\sigma}) \tag{1.46}$$

To summarize, the Maximum A Posteriori estimator can be seen as a generalization of the Maximum Likelihood estimator, where additional previous knowledge is incorporated through the regularization function r (the minus logarithm of the prior). Since the regularization acts as a penalty for those values of $\boldsymbol{\beta}, \boldsymbol{\sigma}$ which have a small prior probability (values that are assumed to be unlikely a priori), (1.44) is sometimes called Penalized or Regularized Maximum Likelihood estimator. Different priors enforce different beliefs on the data generating process, which result in different properties of the resulting MAP estimator.

Priors that enforce the belief in an underlying sparse model, have received much attention in the last two decades. A model is sparse if the number of covariates that are correlated with

the outcome is "small". To make the discussion clear, suppose that we have data generated from the linear model (1.26). Formally, we can define the support of $\boldsymbol{\beta}_0$ as the set of non-null components, i.e. $\text{supp}(\boldsymbol{\beta}_0) = \big\{\mu \in \{1, \dots, p\} \ : \ |\beta_{0,\mu}| > 0\big\}$, where we indicated with $\beta_{0,\mu}$ the μ-th component of $\boldsymbol{\beta}_0$. Then the model is sparse if $\text{supp}(\boldsymbol{\beta}_0)$ has a small cardinality $s_0 := |\text{supp}(\boldsymbol{\beta}_0)| \ll p$. The Least Absolute Shrinkage and Selection Operator (LASSO) regularization introduced in [83]

$$\text{r}(\boldsymbol{\beta}) = \alpha \|\boldsymbol{\beta}\|_1 = \alpha \sum_{k=1}^{p} |\beta_k| \tag{1.47}$$

which corresponds to a Laplace (or double exponential) prior, is an important, but by no means the only, example. For the linear regression model, the corresponding MAP estimator is defined as

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \Big\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \Big\} \tag{1.48}$$

where α is a free parameter that tunes the amount of regularization. Unfortunately, $\hat{\boldsymbol{\beta}}_n$ in (1.48) cannot be solved in closed form. However, it has been shown in e.g. [15, 10], that for α in a suitable range $O(\sqrt{\log(p)/n})$, the Mean Square Error converges to zero in probability as n diverges

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 \xrightarrow[n\to\infty]{P} 0 \ , \tag{1.49}$$

under the assumption that the underlying true signal is sufficiently sparse, i.e.

$$\|\boldsymbol{\beta}_0\|_1 = o\big(\sqrt{\log(p)/n}\big) \ , \tag{1.50}$$

where we used the little o notation of Landau [84, 15], provided that the compatibility condition

$$\Big\| \frac{\boldsymbol{\beta}_0' \hat{\boldsymbol{\beta}}_n}{\|\boldsymbol{\beta}_0\|_2^2} \boldsymbol{\beta}_0 - \boldsymbol{\beta}_0 \Big\|_2^2 \leq \hat{\boldsymbol{\beta}}_n' \hat{\boldsymbol{\Sigma}}_n \hat{\boldsymbol{\beta}}_n \frac{s_0}{\phi_0^2} \tag{1.51}$$

is satisfied for some $\phi_0 > 0$. The condition above is essentially a constraint on the correlations between the covariates. Informally, the covariates should not be too correlated [15]. In practice, however, α is generally chosen in a data-driven manner, for instance by cross validation, or variants of the latter, which aims at optimal prediction (which can be actually measured), which in turn is often in conflict with variable selection. As a consequence, cross validation methodologies tend to select a value of α that is smaller than the one prescribed by the theory in order to achieve consistency [62, 15]. Furthermore, the validity of (1.51) is in general difficult to check, in practice, as $\boldsymbol{\beta}_0$ is unknown. Finally, while a "parsimonious" model is often appealing in terms of interpretability, the assumption of a sparse model is *not* directly verifiable and might not be always empirically justified [31]. In fact, it can be shown that the Lasso or other typically used sparsity inducing penalties might return estimators having very large MSE, when the sparsity assumption does not hold [51].

To summarize, a vast theoretical literature has established the LASSO regularization and its variants as valuable tools in high dimensional statistics, since when properly tuned they can restore consistency, if the underlying $\boldsymbol{\beta}_0$ is sufficiently sparse. However, in the *proportional regime*, where the number of covariates associated with the outcome (s_0) is proportional to the number of samples (n), i.e. $s_0 = \nu p, \nu \in (0, 1)$ with the number of covariates $p = \zeta n, \ \zeta \in (0, \infty)$, consistency is lost, even when imposing a properly tuned regularization. During the last two decades, different theoretical methods have been explored in order to establish the asymptotic behaviour of MAP estimators in this regime. The realization that

heuristic methods from statistical physics are well suited to cope with optimization problems in this "inconsistency" regime can be traced back to [58] and more specifically to the study of learning in neural networks in its early days [36].
In the next section, we aim at introducing this alternative language or notation, with emphasis on the application to statistical models.

1.6 Asymptotic behaviour of MAP estimators in the proportional regime via statistical physics

An optimization problem can be viewed as a statistical physics' problem, via the analogy that "a physical system tends to occupy the minimal energy configuration as the temperature tends toward zero". This seemingly trivial observation is at the base of successful ideas like simulated annealing [49]. Simulated annealing is a randomized algorithm which aims at obtaining a solution of an optimization problem, for instance the Maximum A Posteriori estimator (1.44). This is achieved by sampling configurations $\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\sigma} \in \mathbb{R}^d$ according to the Boltzmann probability density

$$p_\gamma(\boldsymbol{\beta}, \boldsymbol{\sigma}|\mathcal{D}) = \frac{1}{Z_n(\gamma|\mathcal{D})} \mathrm{e}^{-\gamma \mathcal{H}_n(\boldsymbol{\beta}, \boldsymbol{\sigma}|\mathcal{D})}, \quad Z_n(\gamma|\mathcal{D}) = \int \mathrm{e}^{-\gamma \mathcal{H}_n(\boldsymbol{\beta}, \boldsymbol{\sigma}|\mathcal{D})} \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\boldsymbol{\sigma} \tag{1.52}$$

with $\mathcal{H}_n(.|\mathcal{D})$ defined as in (1.20) and where $\mathcal{D}$ indicates the data-set, i.e. $\mathcal{D} := \{(Y_1, \mathbf{X}_1), \ldots, (Y_n, \mathbf{X}_n)\}$. For the more statistically oriented reader, the expression (1.52) might appear very similar to the posterior distribution of Bayesian Statistics [55, 65]. Indeed, the latter is recovered when $\gamma = 1$. As the temperature vanishes, i.e. in the limit $\gamma \to \infty$ (and with a sufficiently smooth annealing schedule), the sampled configurations "will be close" to $\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\sigma}}_n$. This is because as $\gamma \to \infty$, the Boltzmann density (1.52) is "strongly peaked" around the solution of the optimization problem defining the MAP estimator (1.44). So we see that the present formulation interpolates elegantly between the Bayesian formulation and the Frequentist one. However, one might question if there is any practical advantage of such a formulation. We shall soon see that there is indeed an advantage in this reformulation of the problem. Densities like the one in (1.52) are the "bread and butter" of the statistical physics of disordered systems, where one would identify:

- $\mathcal{D}$ as the quenched disorder
- $\mathcal{H}(\boldsymbol{\beta}, \boldsymbol{\sigma}|\mathcal{D})$ as the Hamiltonian of a fictitious physical system with degrees of freedom coinciding with the parameters of the model, i.e. with $\boldsymbol{\beta}, \boldsymbol{\sigma}$
- $Z_n(\gamma|\mathcal{D})$ as the partition function (a.k.a. the evidence in Bayesian terminology) .

Powerful methods like the replica and the cavity method have been invented to deal with such problems in the last century [59, 75, 69]. Indeed, the so-called "statistical mechanics approach" has been extremely successful when applied to modelling in several cross-disciplinary scientific areas involving some kind of random optimization, e.g. neural networks and machine learning [73, 32, 29, 20, 54, 38], random satisfiability problems [57, 60, 67], economy [20, 16], information processing [28, 5, 47, 79, 87] and more recently high dimensional statistics [27, 6, 45, 46, 63, 21].
In statistical physics, the central quantity of interest is the free energy density

$$f_n(\gamma|\mathcal{D}) := -\frac{1}{n\gamma} \log Z_n(\gamma|\mathcal{D}) \ . \tag{1.53}$$

When $\gamma = 1$ this is exactly the logarithm of the Bayesian's evidence (per sample), i.e. the normalization constant of the posterior (which is a function of the observed data). In the limit $\gamma \to \infty$, this quantity is related to the optimal value of the optimization problem defining the MAP estimator. This can be seen via an application of the Laplace approximation method

$$-\frac{1}{\gamma}\log\int_{\Omega\subseteq\mathbb{R}^d}\exp\{-\gamma h(\mathbf{x})\}\,\mathrm{d}\mathbf{x}\underset{\gamma\to\infty}{\sim}\min_{\mathbf{x}\in\Omega}\{h(\mathbf{x})\}+O(1/\gamma)\;, \tag{1.54}$$

where the right-hand side is an asymptotic expansion in γ with reminder of order $1/\gamma$ [61, 66]. In the limit, we obtain an identity

$$\lim_{\gamma\to\infty} f_n(\gamma|\mathcal{D}) = \frac{1}{n}\min_{\boldsymbol{\beta},\boldsymbol{\sigma}\in\mathbb{R}^p\times\mathbb{R}^d}\Big\{\mathcal{H}_n(\boldsymbol{\beta},\boldsymbol{\sigma}|\mathcal{D})\Big\}\;. \tag{1.55}$$

The Laplace's identity above can be interpreted as a formal statement of the intuition behind simulated annealing: all the points that are not the one(s) at which the Hamiltonian attains its minimum, have an exponentially smaller probability of being observed, where the exponent is proportional to the inverse temperature γ that is diverging.

Asymptotic theory makes statements on the *average behaviour* of an estimator, e.g. what is its MSE, or what is the plug in prediction error, under an assumed "true" model and other regularity conditions over the data generating process [11]. Suppose in fact that we are interested in computing the expected value of the optimum of the objective function under the *joint distribution of covariates and responses*, i.e.

$$\mathcal{E}_n = \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\min_{\boldsymbol{\beta},\boldsymbol{\sigma}\in\mathbb{R}^p\times\mathbb{R}^d}\Big\{\mathcal{H}_n(\boldsymbol{\beta},\boldsymbol{\sigma}|\mathcal{D})\Big\}\Big]\;. \tag{1.56}$$

In this case, even when the samples in $\mathcal{D}$ are i.i.d., $\min\ \mathcal{H}_n(\boldsymbol{\beta},\boldsymbol{\sigma}|\mathcal{D})$ is not a sum of i.i.d. random variables and computing its expectation is non-trivial. In the statistical mechanics formulation, we have

$$\mathcal{E}_n = \lim_{\gamma\to\infty} f_n(\gamma),\quad f_n(\gamma) := -\frac{1}{n\gamma}\mathbb{E}_{\mathcal{D}}\Big[\log Z_n(\gamma|\mathcal{D})\Big]\;, \tag{1.57}$$

where we have *heuristically* exchanged the expectation with the limit. However, $\mathcal{E}_n$ is still non-trivial to compute, as computing $Z_n(\gamma|\mathcal{D})$ at fixed $\mathcal{D}$ is already a challenge in its own. It is here that the advantage of the statistical mechanics calculation bears fruit : the replica method and its relative, the cavity method, have been devised to circumvent exactly this problem.

The replica method builds on the following relation

$$\mathbb{E}_X\Big[\log X\Big] = \lim_{r\to 0}\frac{1}{r}\log\mathbb{E}_X\Big[X^r\Big]\;, \tag{1.58}$$

which can be justified rigorously for a positive random variable (a random variable that assumes strictly positive values with probability one). Applied to our problem at hand, this leads to

$$f_n(\gamma) = \lim_{r\to 0} f_n^{(r)}(\gamma),\qquad f_n^{(r)}(\gamma) := -\frac{1}{\gamma n r}\log\mathbb{E}_{\mathcal{D}}\Big[Z_n^r(\gamma|\mathcal{D})\Big]\;. \tag{1.59}$$

So far the procedure is perfectly "respectable", in the sense that if $Z_n(\gamma|\mathcal{D})$ exists, then it can be justified rigorously under mild assumptions on [64]. However, computing arbitrary real moments of $Z_n(\gamma|\mathcal{D})$ is yet a daunting task in many cases. For this reason, the replica

method proceeds by computing $f_n^{(r)}(\gamma)$ *approximately* to leading order in n, "only" for r *integer*. This gives a formal expression for $f^{(r)}(\gamma)$

$$f^{(r)}(\gamma) = \lim_{n\to\infty} f_n^{(r)}(\gamma) \ , \tag{1.60}$$

when r is an *integer*. In physics $f^{(r)}(\gamma)$ is known as the "replicated" free energy density, since

$$Z_n^r(\gamma|\mathcal{D})) = \int \mathrm{e}^{-\gamma \sum_{\alpha=1}^r \mathcal{H}_n(\boldsymbol{\beta},\boldsymbol{\sigma}|\mathcal{D})} \prod_{\alpha=1}^r \mathrm{d}\boldsymbol{\beta}_\alpha \ \mathrm{d}\boldsymbol{\sigma}_\alpha \tag{1.61}$$

is the partition function of a system of r independent replicas of the system, at a fixed realization of the disorder $\mathcal{D}$ (a fixed realization of the data-set). Finally, the value of $f(\gamma)$ is obtained by taking the limit $r \to 0$ of the formal expression for $f^{(r)}(\gamma)$ valid at integer r. The procedure can be compactly summarized as

$$\lim_{n\to\infty} \mathcal{E}_n = \lim_{\gamma\to\infty} \lim_{r\to 0} \lim_{n\to\infty} f_n^{(r)}(\gamma) \ . \tag{1.62}$$

Notice that we have exchanged : i) the limit $n \to \infty$ with $\gamma \to \infty$ and ii) the limit $n \to \infty$ with $r \to 0$.

The replica and the cavity method give access to the asymptotic value of $\mathcal{E}_n$ and test functions of the estimators $\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\sigma}}_n$. However, because of their *heuristic* nature, neither of the methods are able to clarify what is the type of convergence in the previous sentence. The latter might, in some cases, be obtained via a rigorous approach. A massive amount of work, first in the field of mean field spin glass models [77, 78, 19], established the rigorous validity of various heuristic results obtained via the replica method, by means of the interpolation method, introduced originally in [39], and then later extended in [4] for applications to information theory and inference. A different "algorithmic" method, is the one introduced originally in [12], later adopted to establish the rigorous formulation of the Approximate Message Passing framework [5, 28]. This method differs from the Gaussian Interpolation method previously mentioned, and is based on the analysis of the Thouless Anderson Palmer (TAP) equations, which were originally introduced in [80]. Recently, particularly in the field of inference and high dimensional statistics, the Convex Min Max Gaussian theorem [82, 81, 62, 54] has emerged as an alternative to prove the replica formulas in the zero temperature limit and when the original Hamiltonian is a convex function of the degrees of freedom.
For these reasons, the belief of the statistical mechanics community is that the replica method leads to the "correct answer". However, the question "why does the replica calculation recipe work?" remains, as yet, unanswered [64].
We would however highlight some points in favor of the heuristic statistical mechanics procedure :

- the "power" of the replica method is that it allows to tackle any problem that can be re-written in the statistical mechanics formulation (via a Boltzmann-Gibbs measure), e.g. optimization problems, in a unified manner following a well-defined recipe ;
- the proof of the results obtained via the replica method have been possible because of the knowledge acquired via the replica method in the first place. There are recent exceptions [82], which however recover the same results;
- when possible, proving the results obtained via the replica (or the cavity) method requires a non-trivial amount of pages of work, which easily exceeds the number of pages of a replica derivation [64] .

Furthermore, the emerging mathematical theory is, at least in the opinion of the author, beautiful and somewhat "exotic": the free energy $f_n^{(r)}(\gamma)$ is defined over the space of configurations of r replicas and we are "taking the limit" $r \to 0$. It is here that the "magic" of the replica method is believed to lie, i.e. in the parametrization of this dependency. For this step there exist several, in fact infinite, possible choices of increasing complexity [59, 29, 30]. The simplest parametrization, is the so-called replica symmetric ansatz. In this thesis we will only deal with the latter, as it has been shown to lead to the correct result, when the Hamiltonian is a convex function of the degrees of freedom.
However, the steps of the "recipe" somewhat obscure the underlying approximation taken. In this regard the cavity method [59, 56], is more "direct", since it hinges on explicit approximations and assumptions. This is one of the reasons that motivated the author to explore also this alternative heuristic, which provides a complementary view of the subject.

1.7 Aim and structure of this thesis

The first chapter of this thesis focus on Generalized Linear Models (GLMs) and can be considered as a preparation to the following chapters, which conversely are devoted to Survival analysis models.

The goal of chapter 2 is to study an ad hoc penalization that can be tuned in order to obtain an asymptotically unbiased MAP estimator. Limiting our-selves to quadratic penalties for mathematical simplicity, we propose to use a "covariant" prior, also known as uniform shrinkage prior in Bayesian literature [23]. This is a generalization of the well known ridge penalization [42], where the matrix of the quadratic form defining the regularization is the empirical covariance matrix (in place of the identity as for ridge). The net effect is to return an estimator which is uniformly shrunk with respect to the ML one. The proposed methodology is shown to return asymptotically unbiased estimators, provided the regularization strength is chosen appropriately (as a function of $\|\boldsymbol{\beta}_0\|_2$) and $p < n$. The precise value of the regularization strength can be chosen by solving a set of non-linear equations, which is derived by setting up a statistical mechanics theory and carrying out the calculations via the cavity method. This gives me the opportunity to introduce the statistical mechanics approach to learning and optimization [32], which will also be a key ingredient in many chapters. The performance of the proposed estimator is tested numerically, confirming the theoretical findings.

Chapter 3 deals with the Cox semi-parametric Proportional Hazards model, which is arguably the most famous model used for survival data. Previous articles [21, 74, 22] attempted to construct a statistical mechanics theory for this model, but were unable to: i) account for the semi-parametric nature of the Cox model, ii) account for censoring. My first contribution in this chapter is to overcome these limitations, namely to devise an efficient numerical scheme to solve the replica symmetric equations of the model in question directly, without the need of the variational approximation (which might not always be accurate) and to include censoring in the theory. The inclusion of censoring complicates the picture painted in the previous studies in a non-trivial manner, both numerically and theoretically. In fact, when observations are censored, the ML estimator might not exist [17, 44, 88]. When the ML estimator exists, we confirm through numerical simulations that the theory is able to predict accurately the behaviour of the estimator. Starting from the considerations of the first chapter, we then devise a procedure to solve, approximately, the replica symmetric equations starting from the data, i.e. without the need to know the true data generating process (of course under assumptions on the latter). This is vital in order to compute the de-

biasing factors from the theory, in practice. The methodology has been tested on synthetic data, showing that the estimators obtained in this manner are effectively un-biased.

In chapter 4, it was the intention of the author to "put on firmer ground" at least some of the findings of chapter 3, in order to complement the insight gained from the statistical physics approach. Tackling directly the Cox regression model is a highly non-trivial mathematical challenge. For this reason, I focused instead on the asymptotic behaviour of the Piece-Wise Exponential model, which can be rigorously established by means of the Convex Gaussian Min Max theorem [82]. The Piece-Wise Exponential model is a "flexible" parametric PH model, where the base hazard rate is parametrized by a linear combination of piece-wise functions defined over user-specified intervals. The Cox model can be interpreted as a special case of the model considered in this section, when the intervals are fixed according to a specific data-driven manner first proposed by Breslow [14, 53]. However, here, the proof hinges on the fact that the number of intervals is finite, i.e. does not grow with the sample size (n), hence our results are not directly applicable to the Cox model.

We extend, in chapter 5, the replica theory for the Cox model by considering the regularized Cox regression, allowing for an arbitrary regularization function when the covariates are sampled from a multivariate standard normal distribution. Hence, the first contribution of this chapter is to precisely quantify the behaviour of the Regularized Maximum Partial Likelihood Estimator as function of the distribution of the entries of $\boldsymbol{\beta}_0$. The second contribution of this chapter is to propose a variant of the Generalized Approximate Message Passing algorithm [28, 70] to compute the MAP estimator for the Cox regression model. The update equations of this algorithm suggest the computation of a local field that can then be used to estimate all the order parameters of the theory from the data, i.e. without the need of solving the RS equations of the theory which require the perfect knowledge of the data generating process (which is not available in practice). This strategy is similar, but not identical, to the Adaptive TAP or Expectation Consistent [68] inspired procedure adopted in [76] for the LASSO in linear regression models. We show via numerical experiments that this strategy is correct and easy to implement.

In the last chapter we show that, when the covariant prior is used as a regularizer and $\zeta < 1$, one can estimate all the unknowns of the theory, including the signal strength and the nuisance parameters for any positive definite population covariance matrix $\boldsymbol{\Sigma}_0$. Thanks to recent advancements in the field [7], the Replica Symmetric order parameters can be measured directly from the data, without solving the Replica Symmetric equations of the theory both for Generalized Linear Models (GLMs) and Cox regression model. Second, when the model is a parametric GLM, we show that it is possible to easily estimate the signal strength and other nuisance parameters, directly from the data by solving a set of moment matching equations. A more involved procedure is proposed and tested for the Cox semi-parametric model, building on the ideas proposed in chapter two, showing via numerical experiments that the resulting estimators are approximately unbiased.

Bibliography

[1] H Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] J Aldrich. R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162 – 176, 1997.

[3] MA Babyak. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, 66:411–421, 2004.

[4] J Barbier and N Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability Theory and Related Fields*, 174:1133–1185, 2018.

[5] M Bayati and A Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[6] M Bayati and A Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

[7] PC Bellec. Observable adjustments in single-index models for regularized m-estimators, 2024.

[8] A Bender, A Groll, and F Scheipl. A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, 18(3-4):299–321, 2018.

[9] A Beskos and A Stuart. Computational complexity of metropolis-hastings methods in high dimensions. In Pierre L' Ecuyer and Art B. Owen, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2008*, pages 61–71, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[10] PJ Bickel, Y Ritov, and AB Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009.

[11] F Bijma, MA Jonker, and AW van der Vaart. *An Introduction to Mathematical Statistics*. Amsterdam University Press, 2017.

[12] E Bolthausen. An iterative construction of solutions of the tap equations for the sherrington–kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[13] L Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383, 1996.

[14] N Breslow. Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99, 1974.

[15] P Bühlmann and S van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg, 2011.

[16] D Challet, M Marsili, and R Zecchina. Statistical mechanics of systems with heterogeneous agents: Minority games. *Phys. Rev. Lett.*, 84:1824–1827, Feb 2000.

[17] MH Chenm, JG Ibrahim, and QM Shao. Maximum likelihood inference for the cox regression model with applications to missing covariates. *Journal of Multivariate Analysis*, 100(9):2018–2030, 2009.

[18] S Chewi, C Lu, K Ahn, X Cheng, T Le Gouic, and P Rigollet. Optimal dimension dependence of the metropolis-adjusted langevin algorithm. In *Annual Conference Computational Learning Theory*, 2020.

[19] P Contucci and C Giardinà. *Perspectives on Spin Glasses*. Perspectives on Spin Glasses. Cambridge University Press, 2013.

[20] ACC Coolen. *The Mathematical Theory of Minority Games: Statistical mechanics of interacting agents*. Oxford Finance Series. OUP Oxford, 2005.

[21] ACC Coolen, JE Barrett, P Paga, and CJ Perez-Vicente. Replica analysis of overfitting in regression models for time-to-event data. *Journal of Physics A: Mathematical and Theoretical*, 50(37):375001, aug 2017.

[22] ACC Coolen, M Sheikh, A Mozeika, F Aguirre-Lopez, and F Antenucci. Replica analysis of overfitting in generalized linear regression models. *Journal of Physics A: Mathematical and Theoretical*, 53(36):365001, aug 2020.

[23] JB Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(3):311–335, 12 1983.

[24] TM Cover and JA Thomas. *Elements of Information Theory*. Wiley, 2012.

[25] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[26] DR Cox and HS Battey. Large numbers of explanatory variables, a semi-descriptive analysis. *Proceedings of the National Academy of Sciences*, 114(32):8592–8595, 2017.

[27] D Donoho and A Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969, 2016.

[28] DL Donoho, A Maleki, and A Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[29] V Dotsenko. *An Introduction to the Theory of Spin Glasses and Neural Networks*. Notes in Physics Series. World Scientific, 1994.

[30] V Dotsenko. *Introduction to the Replica Theory of Disordered Statistical Systems*. Aléa-Saclay. Cambridge University Press, 2001.

[31] N El Karoui, D Bean, PJ Bickel, C Lim, and B Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[32] A Engel and C Broeck. *Statistical Mechanics of Learning*. Statistical Mechanics of Learning. Cambridge University Press, 2001.

[33] L Fahrmeir and H Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

[34] RA Fisher and EJ Russell. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

[35] M Friedman. Piecewise Exponential Models for Survival Data with Covariates. *The Annals of Statistics*, 10(1):101 – 113, 1982.

[36] E Gardner and B Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, jan 1988.

[37] A Gelman, JB Carlin, HS Stern, DB Dunson, A Vehtari, and DB Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[38] F Gerace, F Krzakala, B Loureiro, L Stephan, and L Zdeborová. Gaussian universality of perceptrons with random labels. *Phys. Rev. E*, 109:034305, Mar 2024.

[39] F Guerra and FL Toninelli. The thermodynamic limit in mean field spin glass models. *Communications in Mathematical Physics*, 230:71–79, 2002.

[40] FE Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. Springer International Publishing, 2015.

[41] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009.

[42] AE Hoerl and RW Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[43] PJ Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799 – 821, 1973.

[44] M Jacobsen. Existence and unicity of mles in discrete exponential family distributions. *Scandinavian Journal of Statistics*, pages 335–349, 1989.

[45] A Javanmard and A Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[46] A Javanmard and A Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593 – 2622, 2018.

[47] Y Kabashima, T Wadayama, and T Tanaka. A typical reconstruction limit for compressed sensing based on lp-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003, sep 2009.

[48] JD Kalbfleisch and RL Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley, 2011.

[49] S Kirkpatrick, CD Gelatt, and MP Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[50] H Leeb and BM Pötscher. Model selection and inference: facts and fiction. *Econometric Theory*, 21(1):21–59, 2005.

[51] H Leeb and BM Pötscher. Sparse estimators and the oracle property, or the return of hodges' estimator. *Journal of Econometrics*, 142(1):201–211, 2008.

[52] RM Lewis and HS Battey. Cox reduction and confidence sets of models: a theoretical elucidation. *arXiv preprint arXiv:2302.12627*, 2023.

[53] DY Lin. On the breslow estimator. *Lifetime data analysis*, 13:471–480, 2007.

[54] B Loureiro, C Gerbelot, H Cui, S Goldt, F Krzakala, M Mézard, and L Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114001, nov 2022.

[55] DJC MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[56] M Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A: Mathematical and General*, 22(12):2181, jun 1989.

[57] M Mézard and A Montanari. *Information, Physics, and Computation*. Oxford Graduate Texts. OUP Oxford, 2009.

[58] M Mézard and G Parisi. Replicas and optimization. *Journal de Physique Lettres*, 46(17):771–778, 1985.

[59] M Mézard, G Parisi, and MA Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[60] M Mézard and R Zecchina. Random k-satisfiability problem: From an analytic solution to an efficient algorithm. *Phys. Rev. E*, 66:056126, Nov 2002.

[61] PD Miller. *Applied Asymptotic Analysis*. Graduate studies in mathematics. American Mathematical Society, 2006.

[62] L Miolane and A Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313 – 2335, 2021.

[63] A Montanari. *Mean field asymptotics in high-dimensional statistics: from exact results to efficient algorithms*, pages 2973–2994.

[64] A Montanari and S Sen. A friendly tutorial on mean-field spin glass techniques for non-physicists, 2024.

[65] KP Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press, 2012.

[66] JD Murray. *Asymptotic Analysis*. Applied Mathematical Sciences. Springer New York, 2012.

[67] CM Olivier, R Monasson, and R Zecchina. Statistical mechanics methods and phase transitions in optimization problems. *Theoretical Computer Science*, 265(1):3–67, 2001. Phase Transitions in Combinatorial Problems.

[68] M Opper, O Winther, and MJ Jordan. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6(12), 2005.

[69] G Parisi. The order parameter for spin glasses: a function on the interval 0-1. *Journal of Physics A: Mathematical and General*, 13(3):1101, mar 1980.

[70] S Rangan. Generalized approximate message passing for estimation with random linear mixing. *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172, 2010.

[71] W. Rudin. *Real and Complex Analysis*. Higher Mathematics Series. McGraw-Hill Education, 1987.

[72] G Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978.

[73] HS Seung, H Sompolinsky, and N Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45:6056–6091, Apr 1992.

[74] M Sheikh and ACC Coolen. Analysis of overfitting in the regularized cox model. *Journal of Physics A: Mathematical and Theoretical*, 52(38):384002, aug 2019.

[75] D Sherrington and S Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35:1792–1796, Dec 1975.

[76] T Takahashi and Y Kabashima. A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(7):073405, 2018.

[77] M Talagrand. *Mean Field Models for Spin Glasses: Volume I: Basic Examples*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer Berlin Heidelberg, 2010.

[78] M Talagrand. *Mean Field Models for Spin Glasses: Volume II: Advanced Replica-Symmetry and Low Temperature*. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer Berlin Heidelberg, 2011.

[79] T Tanaka. A statistical-mechanics approach to large-system analysis of cdma multiuser detectors. *IEEE Transactions on Information Theory*, 48(11):2888–2910, 2002.

[80] DJ Thouless, PW Anderson, and RG Palmer. Solution of 'solvable model of a spin glass'. *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics*, 35(3):593–601, 1977.

[81] C Thrampoulidis, E Abbasi, and B Hassibi. Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[82] C Thrampoulidis, S Oymak, and B Hassibi. The gaussian min-max theorem in the presence of convexity, 2015.

[83] R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[84] AW van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000.

[85] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[86] MJ Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[87] L Zdeborová and F Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[88] X Zhang, H Zhou, and H Ye. A modern theory for high-dimensional cox regression models. *arXiv preprint arXiv:2204.01161*, 2022.

Chapter 2

Penalization-induced shrinking without rotation in high dimensional GLM regression: a cavity analysis

This chapter is published in Journal of Physics A : Mathematical and Theoretical, https://iopscience.iop.org/article/10.1088/1751-8121/aca4ab/meta. The notation in this chapter is different from the notation in the introduction. The dot product between two vectors $\boldsymbol{a}, \boldsymbol{v} \in \mathbb{R}^d$ is indicated with a $\cdot$, i.e $\boldsymbol{a} \cdot \boldsymbol{v} = \sum_{k=1}^{d} a_k \; v_k$.

2.1 Abstract

In high dimensional regression, where the number of covariates is of the order of the number of observations, ridge penalization is often used as a remedy against overfitting. Unfortunately, for correlated covariates such regularisation typically induces in generalized linear models not only shrinking of the estimated parameter vector, but also an unwanted *rotation* relative to the true vector. We show analytically how this problem can be removed by using a generalization of ridge penalization, and we analyse the asymptotic properties of the corresponding estimators in the high dimensional regime, using the cavity method. Our results also provide a quantitative rationale for tuning the parameter controlling the amount of shrinking. We compare our theoretical predictions with simulated data and find excellent agreement.

2.2 Introduction

2.2.1 Setting and motivation

In statistical regression the goal is to understand the relationship between a response $T \in \mathbb{R}$ and a set of covariates $\mathbf{X} \in \mathbb{R}^p$, given n previously recorded observations. In generalized linear models (GLM) our hypotheses are encoded in a statistical model

$$p(T|\mathbf{X} \cdot \boldsymbol{\beta}, \boldsymbol{\sigma}) \tag{2.1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the regression parameters and $\boldsymbol{\sigma} \in \mathbb{R}^d$ are the nuisance parameters. For simplicity we assume that the covariates follow a multivariate normal distribution

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_0) \tag{2.2}$$

and that our model is correctly specified, so our n observations $\{(T_1, \mathbf{X}_1), \ldots, (T_n, \mathbf{X}_n)\}$ are actually generated according to (2.1), i.e. $p(T_i|\mathbf{X}_i \cdot \boldsymbol{\beta}_0, \boldsymbol{\sigma}_0)$ for some "true" unknown values $\boldsymbol{\beta}_0$ and $\boldsymbol{\sigma}_0$ that we would like to estimate. The parameter estimators $\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\sigma}}_n$ are found by optimizing an objective function constructed from the observations available

$$\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\sigma}}_n := \arg\max_{\boldsymbol{\beta},\boldsymbol{\sigma}} \Big\{ l_n(\boldsymbol{\beta}, \boldsymbol{\sigma}) \Big\} \tag{2.3}$$

$$l_n(\boldsymbol{\beta}, \boldsymbol{\sigma}) = \sum_{i=1}^{n} \log p(T_i|\mathbf{X}_i \cdot \boldsymbol{\beta}, \boldsymbol{\sigma}) + \pi(\boldsymbol{\beta}, \boldsymbol{\sigma}) \ . \tag{2.4}$$

In (2.4) the first term on the right hand side is the log-likelihood of the observations under the model (2.1) and π is the penalization function. This term is usually added in order to enforce a constraint by Lagrange multipliers.

In the high dimensional regime, the number of covariates p in the model is large and proportional to the number of observations n, so that as $n, p \to \infty$ the ratio $\zeta := p/n$ is fixed. The true regression parameter vector $\boldsymbol{\beta}_0$ is assumed to have a finite modulus $S_0 := \|\boldsymbol{\beta}_0\|$ that does not depend on p nor n. A common choice for $\pi(\boldsymbol{\beta}, \boldsymbol{\sigma})$ is the so-called ridge penalty

$$\pi_{\text{ridge}}(\boldsymbol{\beta}, \boldsymbol{\sigma}|\Delta) = -\frac{1}{2}\Delta\boldsymbol{\beta} \cdot \boldsymbol{\beta} \tag{2.5}$$

which forces the estimator $\hat{\boldsymbol{\beta}}_n$ to take values closer to the origin. The effect is more pronounced for larger values of Δ. Ridge penalization is often employed to prevent overfitting, and thereby improve the prediction performance of the fitted model [14]. Overfitting causes the Maximum Likelihood (ML) estimator $\hat{\boldsymbol{\beta}}_n^{ML}$ of GLMs to lie typically in the same direction of $\boldsymbol{\beta}_0$, but with modulus larger than that of $\boldsymbol{\beta}_0$ [5, 20], and ridge penalization counter-acts this phenomenon. When the covariates are correlated, however, the effect of the penalty (2.5) is to return an estimator that is both a shrunken and rotated version of the ML estimator: $\hat{\boldsymbol{\beta}}_n$ differs on average both in magnitude and in *direction* from $\boldsymbol{\beta}_0$ [5, 26]. In fact, ridge penalization was originally introduced for the purpose of obtaining a lower variance of $\hat{\boldsymbol{\beta}}_n$ by "trading" variance for bias [15]. While the amount of rotation induced by the penalization is in principle predictable [5, 26], computing the latter is in practice hard to do, because it requires the computation of averages over the spectrum of the population covariance matrix $\mathbf{A}_0$, which is unknown and not easy to estimate [16, 10]. Furthermore, while the Mean Squared Error of the estimator might be lower than the Maximum Likelihood one, it is evident that one would still report a value of $\hat{\boldsymbol{\beta}}_n$ that is on average not even in the direction of $\boldsymbol{\beta}_0$. This represents a major problem in inference if one wants to use penalization as a remedy against overfitting. These considerations leads us to look for a generalization of the ridge penalty such that the resulting estimator: 1) is geometrically unbiased, that is, it lies typically in the direction of $\boldsymbol{\beta}_0$, and 2) has an amplitude tuned by a penalization parameter, so that by selecting the proper value of the latter we can make $\hat{\boldsymbol{\beta}}_n$ unbiased.

2.2.2 Covariant penalties

In absence of the ridge penalization (i.e. in ML regression) one can formally map the problem for correlated covariates to one where the transformed covariates are uncorrelated. This is possible because the ML estimator

$$\hat{\boldsymbol{\beta}}_n^{ML}\Big(\boldsymbol{\beta}_0, \{T_i, \mathbf{X}_i\}_{i=1}^n\Big) := \arg\max_{\boldsymbol{\beta}} \left(\max_{\boldsymbol{\sigma}} \Big\{ \sum_{i=1}^{n} \log p(T_i|\mathbf{X}_i \cdot \boldsymbol{\beta}, \boldsymbol{\sigma}) \Big\} \right) \tag{2.6}$$

has the following covariant property: if we rotate and rescale our covariates according to the linear transformation $\boldsymbol{\mathcal{X}}_i := \mathbf{M}^{-1}\mathbf{X}_i$, then $\hat{\boldsymbol{\beta}}_n^{ML}$ satisfies

$$\hat{\boldsymbol{\beta}}_n^{ML}\Big(\boldsymbol{\beta}_0, \{T_i, \mathbf{X}_i\}_{i=1}^n\Big) \;\; := \mathbf{M}^{-1}\hat{\boldsymbol{\beta}}_n^{ML}\Big(\mathbf{M}\boldsymbol{\beta}_0, \{T_i, \boldsymbol{\mathcal{X}}_i\}_{i=1}^n\Big) \tag{2.7}$$

The ML estimator $\hat{\boldsymbol{\beta}}_n^{ML}$ transforms under rotation and rescaling of the covariates in the exact opposite way, so that the scalar product $\mathbf{X}_i \cdot \hat{\boldsymbol{\beta}}_n$ does not change. Because the quantity $\boldsymbol{\beta}_0$ that we want to estimate is a fixed vector in $\mathbb{R}^p$, a linear transformation executed on the underlying basis will change its components, hence we must have $\mathbf{M}\boldsymbol{\beta}_0$ in the right hand side. Upon transforming $\boldsymbol{\mathcal{X}}_i := \mathbf{A}_0^{-1/2}\mathbf{X}_i$, where $\mathbf{A}_0$ the covariance matrix of $\mathbf{X}_i$, the transformed covariates will be uncorrelated, $\boldsymbol{\mathcal{X}}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and

$$\hat{\boldsymbol{\beta}}_n^{ML}\Big(\boldsymbol{\beta}_0, \{T_i, \mathbf{X}_i\}_{i=1}^n\Big) \;\; \stackrel{d}{=} \;\; \mathbf{A}_0^{-1/2}\hat{\boldsymbol{\beta}}_n^{ML}\Big(\mathbf{A}_0^{1/2}\boldsymbol{\beta}_0, \{T_i, \boldsymbol{\mathcal{X}}_i\}_{i=1}^n\Big) \; . \tag{2.8}$$

Since (2.7) does not hold in general for the ridge penalized estimator (only under rotations), the correlations cannot be "transformed out" as in (2.8). This motivates our interest in *covariant* generalizations of ridge penalization, such that the resulting estimator $\hat{\boldsymbol{\beta}}_n$ would satisfy (2.7). For simplicity and analytical tractability, we consider quadratic penalties. A first candidate is

$$\pi_{\mathrm{O}}(\boldsymbol{\beta}|\eta, \mathbf{A}_0) = -\frac{1}{2}\eta\boldsymbol{\beta} \cdot \mathbf{A}_0\boldsymbol{\beta} \tag{2.9}$$

which we call "oracle" penalty, as it requires the knowledge of the population covariance matrix $\mathbf{A}_0$. For uncorrelated covariates, (2.9) reduces to the ridge penalty. For correlated covariates, (2.9) shrinks differently in different directions, depending on the covariate correlations. While in some cases the population under investigation is sufficiently well characterized, in most cases the matrix $\mathbf{A}_0$ is not known. A possible "quick and dirty" alternative is to use the empirical correlation matrix as an estimate of $\mathbf{A}_0$, giving

$$\pi_{\mathrm{E}}(\boldsymbol{\beta}, \boldsymbol{\sigma}|\tau) = -\frac{1}{2}\tau\boldsymbol{\beta} \cdot \Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i\Big)\boldsymbol{\beta} \; . \tag{2.10}$$

We refer to (2.10) as the "empirical" penalty. We know from Random Matrix Theory (RMT) that the sample covariance matrix is not a good estimator of the population covariance matrix in the high dimensional regime [17], so there we have no guarantee that the estimators obtained using (2.9) and (2.10) will have similar properties. Furthermore, the sample covariance matrix develops eigenvalues approaching zero [17, 9] as $\zeta \to 1$, and (2.10) will not define a well posed optimization problem. On the other hand, if $\mathbf{A}_0$ is non-singular, the oracle penalty defines a well posed optimization problem for all $\zeta > 0$.

2.2.3 Previous work

The idea of biasing an estimator to decrease its Mean Squared Error, i.e. shrinking, goes back to [27]. Ridge penalization is the simplest implementation of this idea, equivalent to a constrained optimization where the length of the maximizer is fixed [15]. Our goal here differs from the original application of shrinking in that we seek to make $\hat{\boldsymbol{\beta}}_n$ unbiased. The geometrical properties of the objective function play a role in determining the distribution of the estimator $\hat{\boldsymbol{\beta}}_n$. We will show that our penalization functions (2.9,2.10) give shrunken and geometrically unbiased estimators. Their origin can be traced back to the Bayesian interpretation of uniform shrinkage [6, 7]; here we provide another interpretation of uniform

shrinkage priors. We will show that the asymptotic properties of the resulting estimators can be computed analytically, under simple assumptions about the data generating process, for arbitrary generalized linear models (GLM). The covariant penalties (2.9,2.10) provide a natural route to asymptotically unbiased estimators in GLM inference, together with our asymptotic analytical results to select an appropriate value of the penalization parameter, even when the sample size is modest.

Several papers in literature deal with the asymptotic behaviour of M-estimators, i.e. those derived by optimizing a sample average of an objective function (see [12, 5, 31, 18]). While the techniques used vary, e.g. Random Matrix Theory, Cavity Method and Replica Method, the conclusions are always expressed via self consistent equations that give the bias and variance of the estimators for the regression parameters. The replica approach [5, 4, 26, 19] appears the most flexible, as it does not assume any specific form of the utility function and (unlike other approaches) allows one to find also the asymptotic expected estimators of nuisance parameters. On the other hand, the replica approach tends to obscure some assumptions needed for its results to hold. We show in 2.6.3 that the Replica Symmetric (RS) equations can be obtained also via another method inspired by the statistical physics analogy with optimization [22]: the cavity method [23, 21]. We extend the results obtained with the replica method [5] to incorporate the penalties (2.9,2.10), and identify the key assumptions needed to reach the result obtained using replicas. Our results confirm again that the cavity method is complementary to the replica method [21]. However, "it is always easier to descend a mountain than to climb it", and the replica method is always our first step. Alternative heuristic approaches, based on Random Matrix Theory, already appeared in the statistical literature too [12]. Several rigorous proofs have also been carried out. In [19], the authors proved rigorous results for robust linear M- estimators. In [20] the asymptotic behaviour of the estimators in a linear model with normally distributed errors and fixed variance was established. Most notably in [18] the authors have rigorously established the validity of the RS equations for non-linear models that depend on the linear predictor, but do not involve nuisance parameters. Their work is general enough to allow for a general penalization and indirectly includes the case where the nuisance parameters are taken to be fixed and known. With this in mind our argument in 2.6.3 might prove useful in establishing a future proof of these results when also the nuisance parameters are inferred. Ultimately, all methods (including ours) rely on the concentration of measure phenomenon, i.e. that specific random variables fluctuate with very low probability around a deterministic value [30, 11, 1, 3], and hence we support our assumptions with simulations.

2.2.4 Aim and structure of the paper

In this paper we study the asymptotic properties of the Penalized Maximum Likelihood (PML) estimator, mathematically equivalent to the Maximum A Posteriori (MAP) estimator, obtained by maximizing the objective function

$$\begin{aligned} l_n(\boldsymbol{\beta}, \boldsymbol{\sigma} | \eta', \tau', \mathbf{A}_0) &= \sum_{i=1}^{n} \log p(T_i | \mathbf{X}_i \cdot \boldsymbol{\beta}, \boldsymbol{\sigma}) + \\ &- \frac{1}{2} p \boldsymbol{\beta} \cdot \Big(\tau' \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i + \eta' \mathbf{A}_0 \Big) \boldsymbol{\beta} \ . \end{aligned} \tag{2.11}$$

The penalization combines (2.9) and (2.10), with re-scaled penalization parameters $\eta = p\eta'$ and $\tau = p\tau'$ such that penalization and log-likelihood scale in the same manner with n in

the high dimensional asymptotic limit [26]. We aim to compare the asymptotic properties of the estimators derived by using the two different penalizations (2.9, 2.10), so we will only be interested in the cases where either η or τ equal zero. In particular we want to answer the following questions: do the covariant penalizations (2.9, 2.10) lead to a geometrically unbiased PML estimator for $\boldsymbol{\beta}_0$, what are the properties of the PML estimator in the high dimensional regime, and how do these depend on the penalization parameters?

The article is divided in four sections. In section 2.3 we present our results concerning the asymptotic properties of the estimator in the high dimensional regime, and show that the PML estimator $\hat{\boldsymbol{\beta}}_n$ is typically oriented in the same direction as $\boldsymbol{\beta}_0$. In section 2.4 we compare our theoretical predictions with simulated data for two prototypical models: the Logit regression model for binary data, and the Weibull Proportional Hazard model for skewed data. We conclude in section 2.5 with a discussion of our results and their applicability, and we identify future directions of investigation.

2.3 The Covariantly Penalized Maximum Likelihood estimator

In this section we give an explicit expression for the PML/MAP estimator, and derive its properties in the high dimensional limit $n, p \to \infty$, with fixed ratio $p/n = \zeta > 0$.

2.3.1 Representation of $\hat{\boldsymbol{\beta}}_n$

The estimator $\hat{\boldsymbol{\beta}}_n$ obtained by maximizing (2.11) satisfies (2.7) and (2.8). Hence we can simply study the properties of the following estimator, because the PML/MAP estimator with correlated covariates is simply a rotated and re-scaled version of it:

$$\hat{\boldsymbol{\beta}}_n^{\sim} := \hat{\boldsymbol{\beta}}_n(\boldsymbol{\beta}_0^{\sim}, \{T_i, \boldsymbol{\mathcal{X}}_i\}) \text{ with } \boldsymbol{\mathcal{X}}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}),\ \boldsymbol{\beta}_0^{\sim} := \mathbf{A}_0^{1/2}\boldsymbol{\beta}_0 \ . \tag{2.12}$$

Following a similar strategy as in [12], we argue in 2.6.1 that this estimator admits the following representation

$$\hat{\boldsymbol{\beta}}_n^{\sim} \stackrel{d}{=} K_n \boldsymbol{\beta}_0^{\sim} + V_n \mathbf{U} \tag{2.13}$$

where

$$K_n := (\boldsymbol{\beta}_0^{\sim} \cdot \hat{\boldsymbol{\beta}}_n^{\sim})/\|\boldsymbol{\beta}_0^{\sim}\|^2 \tag{2.14}$$

$$V_n^2 := \hat{\boldsymbol{\beta}}_n^{\sim} \cdot \hat{\boldsymbol{\beta}}_n^{\sim} - K_n^2 \tag{2.15}$$

with $\mathbf{U}$ independent from K_n and V_n, uniformly distributed on the unit sphere in the $p-1$ dimensional subspace of $\mathbb{R}^p$ that is orthogonal to $\boldsymbol{\beta}_0^{\sim}$. Using (2.8) and (2.13) we have

$$\hat{\boldsymbol{\beta}}_n \stackrel{d}{=} K_n \boldsymbol{\beta}_0 + V_n \mathbf{A}_0^{-1/2}\mathbf{U} \ . \tag{2.16}$$

The PML/MAP estimator consists of two geometrically independent contributions: a signal component in the direction of $\boldsymbol{\beta}_0$, and an error component which is independent from and orthogonal to the former. The expected value of the error component over the possible datasets is zero. Representation (2.16) shows explicitly that the covariance matrix $\mathbf{A}_0$ multiplies the error term $\mathbf{U}$ and enters otherwise in the factor $\|\boldsymbol{\beta}_0^{\sim}\| := \|\mathbf{A}_0^{1/2}\boldsymbol{\beta}_0\|$ only. Furthermore, (2.16) tells us that the distribution of the PML/MAP estimator is fully determined by K_n and V_n. These determine the amplitude of the two independent components of $\hat{\boldsymbol{\beta}}_n$: K_n

gives the (random) inflation factor with respect to the true value, and V_n is the (random) factor controlling the size of the error term. In the statistical physics of disordered systems [23, 25, 22] K_n and V_n are called "overlaps". It is appealing that these quantities, familiar from replica calculations, also have a geometric interpretation for finite n.

With (2.16) we can compute and understand the expected properties of the estimator $\hat{\boldsymbol{\beta}}_n$. For instance, the Mean Square Error (MSE) of the estimator $\hat{\boldsymbol{\beta}}_n$ takes the form

$$\mathbb{E}\big[\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2\big] \;=\; \mathbb{E}\big[\|\hat{\boldsymbol{\beta}}_n - \mathbb{E}[\hat{\boldsymbol{\beta}}_n]\|^2\big] + \|\mathbb{E}[\hat{\boldsymbol{\beta}}_n] - \boldsymbol{\beta}_0\|^2 \,. \tag{2.17}$$

Where we used $\mathbb{E}[]$ to indicate the expectation with respect to the data generating distribution. As everybody knows it consists of two distinct terms, namely the squared bias

$$\|\mathbb{E}[\hat{\boldsymbol{\beta}}_n] - \boldsymbol{\beta}_0\|^2 = (\mathbb{E}[K_n] - 1)^2 \|\boldsymbol{\beta}_0\|^2 \tag{2.18}$$

and the variance

$$\mathbb{E}\big[\|\hat{\boldsymbol{\beta}}_n - \mathbb{E}[\hat{\boldsymbol{\beta}}_n]\|^2\big] = \big(\mathbb{E}[K_n^2] - \mathbb{E}[K_n]^2\big)\|\boldsymbol{\beta}_0\|^2 + \mathbb{E}[V_n^2]\mathbb{E}[A_p^2] \tag{2.19}$$

where we defined

$$A_p^2 := \mathbf{U} \cdot \mathbf{A}_0^{-1}\mathbf{U} \tag{2.20}$$

and used the fact that $\mathbf{U}$ is orthogonal to $\boldsymbol{\beta}_0$ and independent of V_n (see 2.6.1). According to (2.19), the moments of the generalized overlaps K_n, V_n weigh the contribution of the true signal strength $\|\boldsymbol{\beta}_0\|$ and of the fluctuations $\mathbb{E}[A_p^2]$ to the MSE.

2.3.2 The PML/MAP estimator in the high dimensional limit

The next step is to assume that K_n, V_n and the estimators for the nuisance parameters $\hat{\boldsymbol{\sigma}}_n$ (as well as A_p), should concentrate, as $n, p \to \infty$ with $\zeta := p/n$ fixed, around deterministic values. These latter values are expected to satisfy a set of self consistent equations. In this section we display these self-consistent asymptotic equations. The derivation based on cavity arguments is presented in 2.6.3. We introduce some variations of the standard cavity argument, leading to a faster derivation and allowing us to tackle also non-linear objective functions and nuisance parameters, thus we believe this derivation to be of interest on its own. We have obtained the same results by means of the replica method, as an independent verification.

Result 1 (Replica Symmetric (RS) equations)**.** *Suppose we are given n i.i.d. observations $\{(T_i, \mathbf{X}_i)\}_{i=1}^n$, generated as $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_0)$ and $T_i|\mathbf{X}_i \sim p(T_i|\mathbf{X}_i \cdot \boldsymbol{\beta}_0, \boldsymbol{\sigma}_0)$. Under regularity conditions[1] over the objective function (2.11) as $n, p \to \infty$ with fixed ratio $\zeta = p/n$, the random variables K_n, V_n and $\hat{\boldsymbol{\sigma}}_n$ concentrate around deterministic values. The latter, which we indicate with $w_\star/S$, $v_\star$ and $\boldsymbol{\sigma}_\star$ respectively, are obtained, together with $u_\star$, by solving the so-called Replica Symmetric (RS) equations*

$$\zeta v^2 = \mathbb{E}_{T,Q,Z_0}\Big[\big(\xi - vQ - wZ_0\big)^2\Big] \tag{2.21}$$

$$v\Big(1 + \zeta(u^2\eta' - 1)\Big) = \mathbb{E}_{T,Q,Z_0}\Big[\frac{\partial}{\partial Q}\xi\Big] \tag{2.22}$$

$$w\zeta = \mathbb{E}_{T,Q,Z_0}\Big[\xi\frac{\partial}{\partial Z_0}\log p(T|SZ_0, \boldsymbol{\sigma}_0)\Big] \tag{2.23}$$

$$0 = \mathbb{E}_{T,Q,Z_0}\Big[\nabla_{\boldsymbol{\sigma}}\rho(T|\xi, \boldsymbol{\sigma})\Big] \tag{2.24}$$

In the equations above $S^2 := \lim_{p\to\infty} \boldsymbol{\beta}_0 \cdot \mathbf{A}_0\boldsymbol{\beta}_0 = \lim_{p\to\infty} \|\boldsymbol{\beta}_0^{\sim}\|^2$, $Z_0, Q \sim \mathcal{N}(0,1)$, *with* $Z_0 \perp Q^2$ *and* $T|Z_0 \sim p(.|SZ_0, \boldsymbol{\sigma}_0)$. *We also used the short hand* $\xi := \xi(vQ + wZ_0, u, \boldsymbol{\sigma}, T, \tau')$ *to denote the quantity*

$$\xi(x, u, \boldsymbol{\sigma}, T, \tau') := \arg\min_y \Big\{ \frac{1}{2}\Big(\frac{y-x}{u}\Big)^2 - \rho(T|y, \boldsymbol{\sigma}) \Big\} \tag{2.25}$$

known as the proximal mapping of the function $-u\rho(T|., \boldsymbol{\sigma})$, *where*

$$\rho(T|y, \boldsymbol{\sigma}) := \log p(T|y, \boldsymbol{\sigma}) - \frac{1}{2}\zeta\tau' y^2 \ . \tag{2.26}$$

As a consequence of the above result, the asymptotic properties of the estimators $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\sigma}}_n$ depend on the RS order parameters $v_\star, w_\star, \boldsymbol{\sigma}_\star$. For instance, the asymptotic squared bias of $\hat{\boldsymbol{\beta}}_n$ reads

$$\lim_{\substack{n,p\to\infty \\ p/n=\zeta}} \|\mathbb{E}[\hat{\boldsymbol{\beta}}_n] - \boldsymbol{\beta}_0\|^2 = (w_\star/S - 1)^2 S_0^2 \tag{2.27}$$

where we have defined $S_0^2 := \lim_{p\to\infty} \|\boldsymbol{\beta}_0\|^2$. The variability of $\hat{\boldsymbol{\beta}}_n$ around its expected value takes asymptotically the following simple form:

$$\lim_{\substack{n,p\to\infty \\ p/n=\zeta}} \mathbb{E}\big[\|\hat{\boldsymbol{\beta}}_n - \mathbb{E}[\hat{\boldsymbol{\beta}}_n]\|^2\big] = \lim_{\substack{n,p\to\infty \\ p/n=\zeta}} \mathbb{E}[V_n^2]\mathbb{E}[A_p^2] = v_\star^2\alpha^2 \tag{2.28}$$

where

$$\alpha^2 := \lim_{p\to\infty} \mathbb{E}[A_p^2] = \lim_{p\to\infty} \frac{1}{p}\mathrm{Tr}(\mathbf{A}_0^{-1}) \ . \tag{2.29}$$

The derivation of (2.29) in found in 2.6.2. One observes that the fluctuations of $\hat{\boldsymbol{\beta}}_n$ are asymptotically controlled only by the trace of the inverse of the covariance matrix. Furthermore, equation (2.28) gives the statistical meaning of the order parameter $v_\star$. The latter is zero in the low-dimensional regime, i.e. for $\zeta = 0$, where the number of components of $\boldsymbol{\beta}$ is finite and the scale of the fluctuations of each component is $\sqrt{n}$ by standard asymptotic theory [31]. In the high dimensional regime $\zeta > 0$ this is no longer true [12, 5, 4, 19, 26]. Combining (2.27) with (2.28), we obtain an expression for the (asymptotic) MSE of $\hat{\boldsymbol{\beta}}_n$

$$\begin{aligned} &\lim_{\substack{n,p\to\infty \\ p/n=\zeta}} \mathbb{E}\big[\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|^2\big] = \\ &\quad = \lim_{\substack{n,p\to\infty \\ p/n=\zeta}} \mathbb{E}\big[\|\hat{\boldsymbol{\beta}}_n - \mathbb{E}[\hat{\boldsymbol{\beta}}_n]\|^2\big] + \|\mathbb{E}[\hat{\boldsymbol{\beta}}_n] - \boldsymbol{\beta}_0\|^2 \\ &\quad v_\star^2\alpha^2 + S_0^2\big(w_\star/S - 1\big)^2 \ . \end{aligned} \tag{2.30}$$

It is important to note that the data set enters only in three scalar quantities, namely α, S and S_0, but only S determines the solution of the RS equations.

Since $(u_\star, v_\star, w_\star, \boldsymbol{\sigma}_\star)$ depend implicitly on the regularization parameters τ' and η', this is also true for the asymptotic bias and variance (and thus the MSE) of $\hat{\boldsymbol{\beta}}_n$, and for the asymptotic properties of the nuisance parameter estimators $\hat{\boldsymbol{\sigma}}_n$. One can therefore, in principle, define the optimal value of the regularization parameters by imposing an additional condition. In inference the focus is usually on obtaining an unbiased estimator for the regression parameters $\hat{\boldsymbol{\beta}}_n$, possibly with minimum variance. In this case one can require that the asymptotic bias be zero, or equivalently that $w_\star/S = 1$ and solve for the value of the penalization parameter that satisfies this condition.

[1] We do not provide a rigorous proof, but our heuristic derivation relies on Laplace integration and assumes that the overlaps concentrate or equivalently are self averaging. We refer to 2.6.3 for more details.

[2] With $A \perp B$ we indicate that A is statistically independent of B.

2.4 Application to selected regression models

In this section we compare the solution of the RS equations against simulated data for two popular regression models, namely the Logit regression model for binary data and the Weibull proportional hazard model for time to event data. This confirms that asymptotically the overlaps fluctuate increasingly tightly around the values predicted by the theory. We also show that the theory describes accurately how the properties of the PML/MAP estimator depend on the regularization parameters, and that the RS equations can be used to select the parameter values that gives unbiased estimators.

2.4.1 Simulation protocol

The data used in this section were generated with the open source programming language R [2]. The covariates in the simulations are always taken to be $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}_0)$. In each simulation, with given p, the covariance matrix $\mathbf{A}_0$ is generated randomly as follows. First we generate a random orthogonal matrix $\mathbf{O}$ in $\mathbb{R}^{p\times p}$ with the R-package `pracma` [2]. We then sample the p eigenvalues of $\mathbf{A}_0$ from a uniform distribution with support in $[0.1, 10.0]$ and store them in a diagonal matrix $\mathbf{\Lambda}_0$. Finally $\mathbf{A}_0 = \mathbf{O} \cdot \mathbf{\Lambda}_0 \mathbf{O}$. The true value of the regression parameters $\boldsymbol{\beta}_0$ is fixed to $\boldsymbol{\beta}_0 = \mathbf{e}_1$, the first basis vector of $\mathbb{R}^p$, for simplicity. All elements of the covariance matrix $\mathbf{A}_0$ are rescaled by a common factor in order to enforce that $S = \|\mathbf{A}_0^{1/2}\boldsymbol{\beta}_0\| = 1$. The present definitions clearly differ from the simpler orthonormal design case $\mathbf{X} = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. We assume that there is no model mis-specification, so the responses are generated according to the same regression model that is used to construct the likelihood. The generalized overlaps are computed from their definitions. For K_n we build on (2.16), so that asymptotically

$$K_n = \hat{\boldsymbol{\beta}}_n \cdot \boldsymbol{\beta}_0 / \|\boldsymbol{\beta}_0\|^2 \tag{2.31}$$

We compute V_n approximately using (2.28)

$$V_n = \hat{\boldsymbol{\beta}}_n \Big(\mathbf{I} - \frac{\boldsymbol{\beta}_0 \boldsymbol{\beta}_0}{\|\boldsymbol{\beta}_0\|^2} \Big) \cdot \hat{\boldsymbol{\beta}}_n / \alpha \ . \tag{2.32}$$

Where $\hat{\boldsymbol{\beta}}_n$ is the maximizer of the objective function (2.11). Although the theory was derived in the asymptotic limit, we carried out our simulations at the rather modest sample size of $n = 200$. This is done on purpose, in order to show that the high dimensional asymptotic theory is suitable in practice to understand even the typical properties of estimators obtained from small sample size datasets.

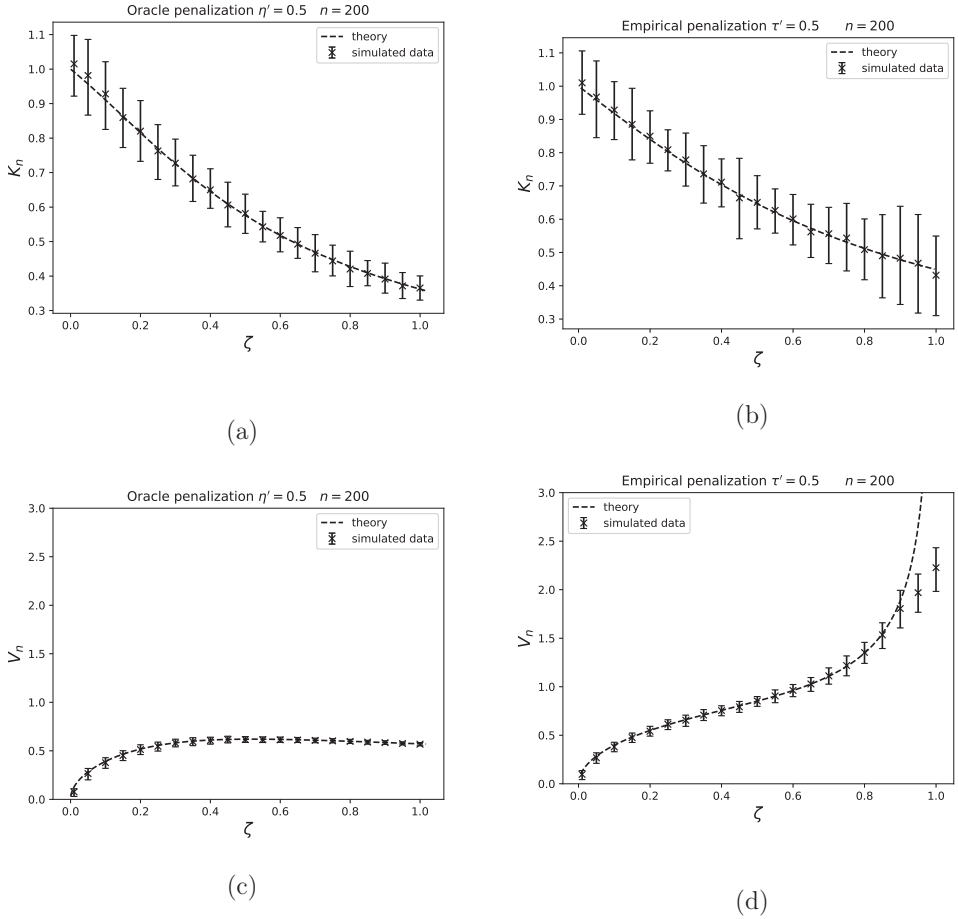


Figure 2.1: Simulated data for the Logit regression model, tested against the solution of the RS equations ($n = 200$, $S = 1.0$). The crosses represent sample averages, and the error bars represent the first and third sample quartiles based on 500 realizations of the PML/MAP estimator. Dashed curves: solution of the RS equations. Left: results for the covariant 'oracle' regularizer. Right: results for the covariant 'empirical' regularizer.

2.4.2 Logit regression model

The Logit regression model for a binary response $T \in \{-1, 1\}$ is defined as

$$p(T|\mathbf{X} \cdot \boldsymbol{\beta}) = \frac{\mathrm{e}^{T(\mathbf{X} \cdot \boldsymbol{\beta})}}{2\cosh(\mathbf{X} \cdot \boldsymbol{\beta})} \ . \tag{2.33}$$

It is the simplest non-linear regression model $p(T|\mathbf{X} \cdot \boldsymbol{\beta}, \boldsymbol{\sigma})$ that cannot be rewritten in the form $p(T - \mathbf{X} \cdot \boldsymbol{\beta})$, and probably the most widely used model in applications.

It is known that for the model (2.33) the MLE is biased for $\zeta > 0$. Furthermore, this bias diverges at a critical value of ζ, as was shown in previous studies, with a sharp phase transition [5, 26, 29] beyond which the MLE does not exist. Figure 2.1 shows that in PML

regression for both covariant regularizers (2.9) (oracle) and (2.10) (empirical) the overlaps fluctuate around average values that are correctly predicted by the theory. The theoretical curves represented by dashed lines are the solution of the RS equations (see 2.6.4 for the derivation), reading

$$\frac{\zeta\nu^2}{(1-\tau\mu^2)^2} = \mu^4\ \mathbb{E}_{T,Q,Z_0}\left[\left(T - \tau\frac{\nu Q + \omega Z_0}{1-\tau\mu^2} - \tanh(x_\star)\right)^2\right] \tag{2.34}$$

$$\zeta\Big(1 - \frac{\mu^2\eta'}{1-\tau'\zeta\mu^2}\Big) = 1 - (1-\tau'\zeta\mu^2)\mathbb{E}_{T,Q,Z_0}\left[\frac{\cosh^2(x_\star)}{u^2+\cosh^2(x_\star)}\right] \tag{2.35}$$

$$\zeta\omega/S = (1-\tau\mu^2)\mathbb{E}_{T,Q,Z_0}\Big[x\Big(T - \tanh(SZ_0)\Big)\Big] \tag{2.36}$$

where $Z_0, Q \sim \mathcal{N}(0,1)$, $Q \perp Z_0$, $T|Z_0 \sim \exp(TSZ_0)/\big(2\cosh(SZ_0)\big)$ and where $x_\star$ is the solution of the transcendental equation $x = a - b\tanh(x)$, in which

$$a = \nu Q + \omega Z_0 + \mu^2 T, \qquad b = \mu^2 \tag{2.37}$$

$$\mu^2 = u^2/(1+\tau'\zeta u^2), \qquad \nu = v/(1+\tau'\zeta u^2), \qquad \omega = w/(1+\tau'\zeta u^2)\ . \tag{2.38}$$

The above coupled equations can be solved numerically by fixed point iteration, once S and the regularization parameter (τ' or η') are specified.

For ζ close to one, the sample covariance matrix becomes singular and the empirical penalization fails to regularize the optimization problem, leading to discrepancy between theory and simulated data (Figure 1*d*). The estimator obtained with the oracle penalization is slightly more biased than the one obtained with the empirical penalization, but has a smaller (and bounded) variance. Note that, away from $\zeta = 1$, the fluctuations of the estimators obtained with the empirical penalty (Figure 1*d*) are of the same order as those found with oracle penalization (Figure 1*c*).

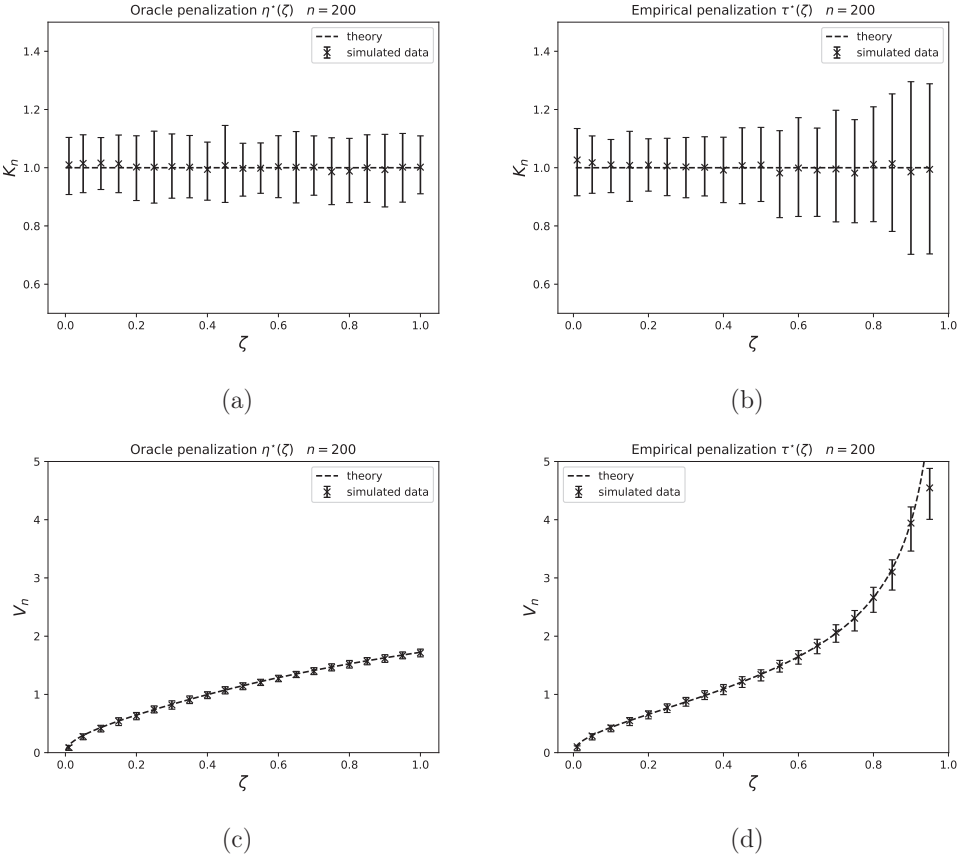


Figure 2.2: Simulation data for the Logit regression model along the zero bias line (for n=200, $S = 1$), with optimal oracle (left) or empirical (right) covariant penalization. The penalization parameters now depend on ζ in such a way that the asymptotic bias is zero, i.e. $w_\star/S = 1.0$. This is confirmed in Figures 2*a* and 2*b*. The estimator fluctuations are shown in Figures 2*c* and 2*d*. The error bars represent the first and third sample quartiles. Dashed lines give the theoretical predictions (solutions of the RS equations). In this case the asymptotic MSE as given in (2.30) will be equal to $v^2\alpha^2$, as $w_\star/S = 1$, where α^2 is defined in (2.29).

The special values of η' and τ' that would make the estimator $\hat{\boldsymbol{\beta}}_n$ strictly unbiased in both direction and amplitude, which we will denote by $\eta^\star$ and $\tau^\star$, can be computed as a function of ζ and S by solving the RS equations, subject to the condition $w/S = \omega/S(1-\tau'\zeta\mu^2)^{-1} = 1$. This constraint gives

$$\zeta = \mathbb{E}_{T,Q,Z_0}\big[x(T - \tanh(SZ_0))\big] \ . \tag{2.39}$$

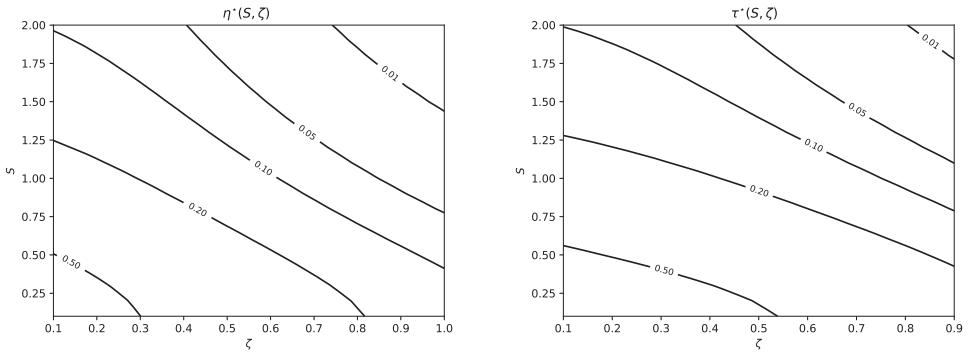
Equation (2.35) can then be used to compute τ' or η'. For $\eta' = 0$ one obtains

$$\tau^\star = \frac{\zeta - \chi}{1-\chi}\,\frac{1}{\mu^2\zeta} \qquad \text{with} \qquad \chi = \mathbb{E}_{T,Q,Z_0}\Big[\frac{u^2}{u^2 + \cosh^2(x_\star)}\Big] \ . \tag{2.40}$$

Whilst for $\tau' = 0$ one finds the prescription

$$\eta^\star = (\zeta - \chi)/\mu^2\zeta \ . \tag{2.41}$$

Equations (2.40,2.41) are added to the RS equations in place of (2.36), upon which the combined system is once more solved by fixed point iteration. The surfaces $\tau^\star(S,\zeta)$ (empirical covariant regularization) or $\eta^\star(S,\zeta)$ (oracle covariant regularization) define the values of the regularization parameters that should be used to obtain an unbiased estimator (that is, given S).



(a) (b)

Figure 2.3: Contour plots of the surfaces $\tau^\star(S,\zeta)$ and $\eta^\star(S,\zeta)$. These are the regularizion parameter values of the covariant penalizations that make the estimator $\hat{\boldsymbol{\beta}}_n$ unbiased. Left: oracle penalization. Right: empirical penalization. Both values are seen to decrease with increasing S and ζ.

In Figure 2.2 we show the result of carrying out regression following the latter protocols and compare the values computed from simulations with the theoretical curves. We observe satisfactory agreement, with the simulation averages almost identical to the theoretical predictions. The improvements over ML estimator are clear: the estimator is unbiased (for known S), and has a finite variance. It is also clear that V_n, measuring the typical fluctuations around the average, gets larger as ζ increases, especially for the estimator obtained with the empirical penalization (Figure 2*d*). Nevertheless, it is remarkable and of practical relevance that, according to Figures 2*c* and 2*d*, for sufficiently small values of ζ (e.g. $\zeta < 0.4$) the sample to sample fluctuations of the estimator obtained with empirical penalization is comparable to the one obtained with oracle penalization. This implies that for such ζ values, in absence of $\mathbf{A}_0$ one can safely use empirical penalization. In fact in the un-biased case the MSE (2.17) is given by the variance (2.19) which is equal to $\alpha^2 v^2$ asymptotically (2.28), where α (2.29) is a constant that is fixed for a particular data set. Hence one can deduce the asymptotic performance of $\hat{\boldsymbol{\beta}}_n$, in terms of MSE directly from Figure 2*d* and conclude accordingly as above.

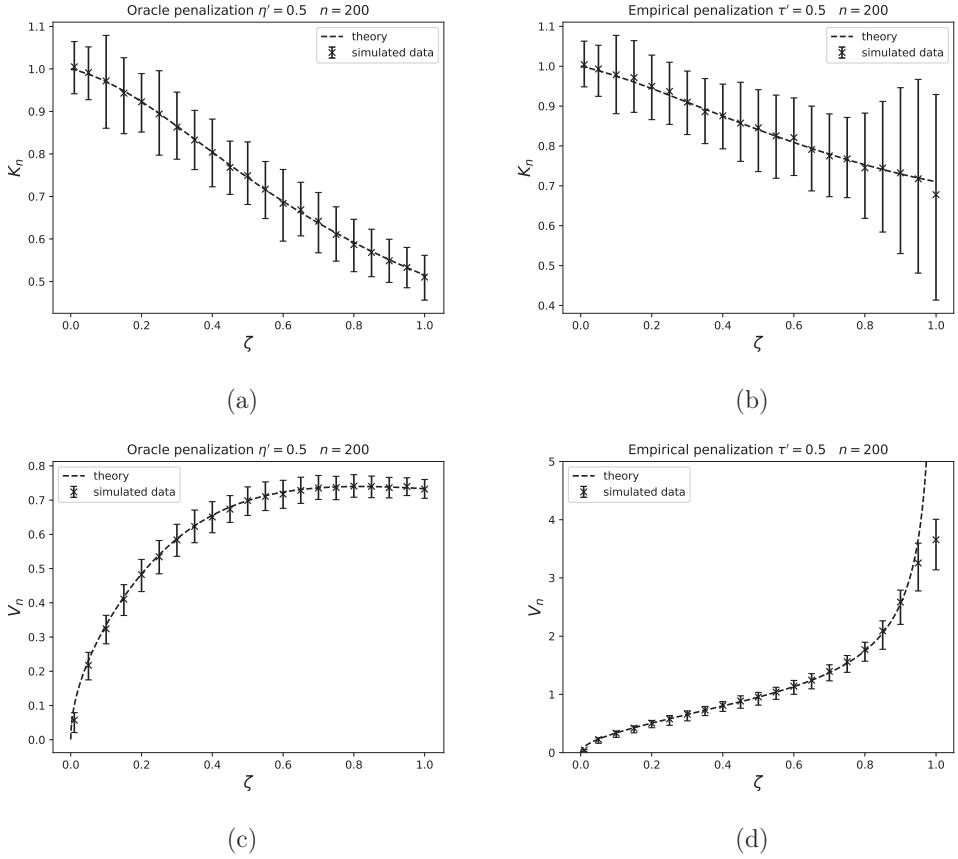


Figure 2.4: Simulation data for the Weibull proportional hazards model (with n=200, $S = 1$), with oracle (left) or empirical (right) covariant penalization. The error bars represent the first and third sample quartiles, while the dashed line is the curve obtained by solving the RS equations. Dashed lines give the theoretical predictions.

It is interesting to inspect the dependence on the global parameters S and ζ of the special values for $\tau^\star(S,\zeta)$ and $\eta^\star(S,\zeta)$ that lead to unbiased $\hat{\boldsymbol{\beta}}_n$. These values are shown as contour plots in the (S,ζ) plane in Figure 2.3. We see that the de-biasing regularization parameters decrease with both S and ζ. Whilst the decrease with S is intuitive (it is easier to estimate a powerful signal than a weak one, hence there is less need for correction), the decrease with ζ is not. On the other hand, our theory predicts that $\hat{\boldsymbol{\beta}}_n \cdot \mathbf{A}_0 \hat{\boldsymbol{\beta}}_n$ and $\sum_{i=1}^n \hat{\boldsymbol{\beta}}_n \cdot \mathbf{X}_i \mathbf{X}_i \hat{\boldsymbol{\beta}}_n / n$ both increase with ζ. It is hence not unreasonable that the optimal penalization parameter gets smaller for larger ζ, to prevent penalization from contributing the leading term in the objective function (which would force $\hat{\boldsymbol{\beta}}_n$ excessively towards the origin). One would expect to find that $\eta^\star$ or $\tau^\star$ goes to zero with ζ. This is actually not needed as the penalization parameters are multiplied by p (see 2.11), hence when ζ "goes" to zero, the same is true for $p = \zeta n$. In (2.11) the amount of penalization is then tuned by $\zeta\tau'$ (or $\zeta\eta'$) and the penalty amounts to a small contribution in the objective function, no matter what finite value of τ' or η', which are indeed found to be non vanishing, but finite. Notice finally, that while it is

possible to make the estimator $\hat{\boldsymbol{\beta}}_n$ strictly unbiased, this may not always be meaningful. It is well known that in some cases a biased estimator with lower variance might outperform an unbiased one; for an unbiased estimator with excessive variance the signal may get lost in a sea of fluctuations. Once $v_\star$ and α are estimated, one has access to the variance of $\hat{\boldsymbol{\beta}}_n$ (2.28). Alternative one can control the bias-variance trade-off manually via the solution of the RS equations.

2.4.3 The Weibull proportional hazards model

In Proportional hazards models, the conditional density of the response $T \geq 0$, given the covariates, is of the form

$$p(T|\mathbf{X}\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) = h(T|\boldsymbol{\sigma})\mathrm{e}^{\mathbf{X}\cdot\boldsymbol{\beta}-H(T|\boldsymbol{\sigma})\exp(\mathbf{X}\cdot\boldsymbol{\beta})} \tag{2.42}$$

where $h(.|.)$ is the base hazard function, and $H(T|\boldsymbol{\sigma}) := \int_0^T \mathrm{d}t'\ h(t'|\boldsymbol{\sigma})$ is the integrated base hazard rate. A widely employed parametric survival model is the Weibull proportional hazards model, for which the integrated base hazard rate has the form[3]

$$H(T|\phi,\sigma) = \mathrm{e}^{\phi/\sigma}T^{1/\sigma} \tag{2.43}$$

In 2.6.5 we show that for this model the RS equations can be written in the form

$$\frac{\nu^2\zeta}{(1-\tau'\zeta\mu^2)^2} = \mathbb{E}_{Z,Q,Z_0}\Big[\Big(\mu^2 - W\Big(Z^{\sigma_0/\sigma}\mu^2\mathrm{e}^{\mu^2+(\omega-S\frac{\sigma_0}{\sigma})Z_0+\nu Q+(\phi-\phi_0)/\sigma}\Big) + -\frac{\tau'\zeta\mu^2}{1-\tau'\zeta\mu^2}(\nu Q+\omega Z_0)\Big)^2\Big] \tag{2.44}$$

$$\frac{1}{1-\tau'\zeta\mu^2}\Big(1-\zeta\big(1-\frac{\mu^2\eta'}{1-\tau'\zeta\mu^2}\big)\Big) = \tag{2.45}$$

$$1-\mathbb{E}_{Z,Q,Z_0}\left[\frac{W\Big(Z^{\sigma_0/\sigma}\mu^2\mathrm{e}^{\mu^2+(\omega-S\frac{\sigma_0}{\sigma})Z_0+\nu Q+(\phi-\phi_0)/\sigma}\Big)}{1+W\Big(Z^{\sigma_0/\sigma}\mu^2\mathrm{e}^{\mu^2+(\omega-S\frac{\sigma_0}{\sigma})Z_0+\nu Q+(\phi-\phi_0)/\sigma}\Big)}\right]$$

$$\frac{\omega}{S} = \Big(1-\tau'\mu^2-\frac{\eta'\mu^2}{1-\tau'\mu^2\zeta}\Big)\frac{\sigma_0}{\sigma} \tag{2.46}$$

$$\mu^2 = \mathbb{E}_{Z,Q,Z_0}\Big[W\Big(Z^{\sigma_0/\sigma}\mu^2\mathrm{e}^{\mu^2+(\omega-S\frac{\sigma_0}{\sigma})Z_0+\nu Q+(\phi-\phi_0)/\sigma}\Big)\Big] \tag{2.47}$$

$$\frac{\sigma}{\sigma_0} = \tag{2.48}$$

$$\mathbb{E}_{Z,Q,Z_0}\Big[\Big(\log Z - SZ_0\Big)\Big(\frac{W\Big(Z^{\sigma_0/\sigma}\mu^2\mathrm{e}^{\mu^2+(\omega-S\frac{\sigma_0}{\sigma})Z_0+\nu Q+(\phi-\phi_0)/\sigma}\Big)}{\mu^2}-1\Big)\Big]$$

where $W(x)$ is Lambert's W-function, $Z\sim\mathrm{Exp}(1)$, $Q,Z_0\sim\mathcal{N}(0,1)$ with $Q\perp Z_0$ and μ,ω,ν are related to u,w,v as in (2.38). The above equations can be again solved by fixed point iteration. Similar to the Logit model, we see in Figure 2.4 that the overlaps fluctuate closely around the theoretical prediction (i.e. the solution of the RS equations, displayed as a dashed line). The estimator obtained with oracle penalization is again more biased with respect to the one obtained with the empirical penalization, but fluctuates less from sample to sample.

[3]This is not the usual parametrization of the Weibull base hazard rate, which can be recovered by setting $\lambda=\exp(\phi)$ and $\rho=1/\sigma$.

These fluctuations of the two penalization formulae are of the same order of magnitude for sufficiently small ζ.

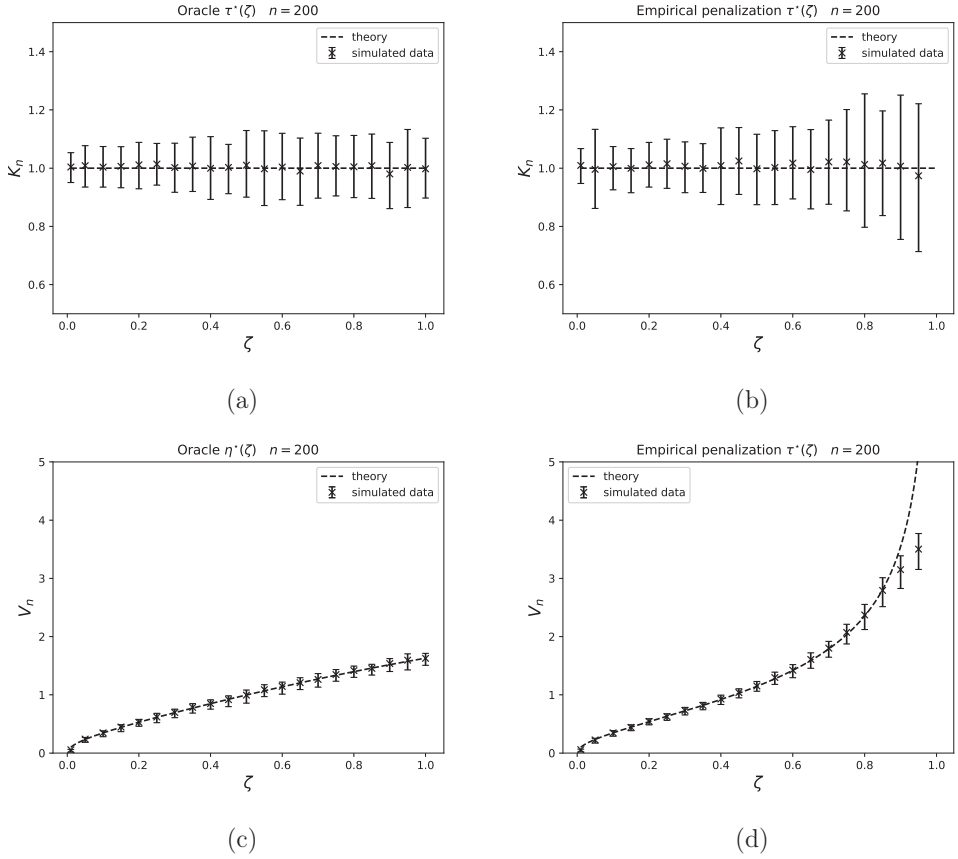


Figure 2.5: Simulation data for the Weibull proportional hazards model along the zero bias line (for n=200, $S = 1$), with optimal oracle (left) or empirical (right) covariant penalization. The penalization parameters now depend on ζ in such a way that the asymptotic bias is zero, i.e. $w_\star/S = 1.0$. This is confirmed in Figures 5a and 5b. The estimator fluctuations are shown in Figures 6c and 6d. The error bars represent the first and third sample quartiles. Dashed lines give the theoretical predictions. In this case the asymptotic MSE as given in (2.30) will be equal to $v^2\alpha^2$, as $w_\star/S = 1$, where α^2 is defined in (2.29).

To obtain unbiased estimators we impose the condition $w/S = \omega/S(1 - \tau\zeta\mu^2)^{-1} = 1$, which leads to an equation either for $\tau^\star$ or for $\eta^\star$, to be substituted for (2.46). Setting $\eta' = 0$ we obtain

$$\tau^\star = \frac{1 - \sigma/\sigma_0}{1 - \zeta\sigma/\sigma_0} \tag{2.49}$$

while setting $\tau' = 0$ we obtain

$$\eta^\star = \mu^{-2}(1 - \sigma/\sigma_0) \ . \tag{2.50}$$

We show the result of using the protocols $\tau^\star(\zeta, S)$ and $\eta^\star(\zeta, S)$, at fixed $S = 1$, in Figure 2.5. We see that K_n now fluctuates around the value one, as desired, indicating unbiased inference.

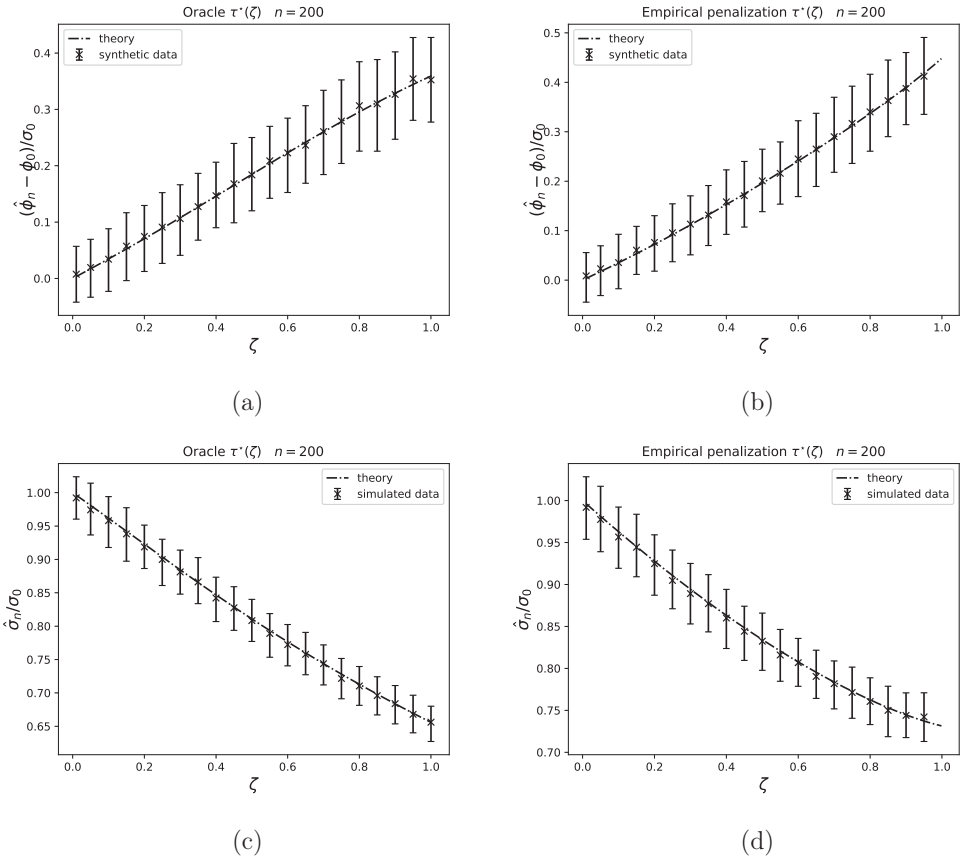


Figure 2.6: Simulated data for the Weibull proportional hazards model (n=200, $S = 1$). The error bars represents the first and third sample quartiles, while the dashed line is the curve obtained by solving the RS equations. The penalization parameters are obtained by solving the RS equations, upon requiring $\hat{\boldsymbol{\beta}}_n$ be unbiased. It is evident that the nuisance parameters are biased and that the quantities displayed fluctuate closely around the solution of RS equations.

As we already argued for the Logit model, the performance of the estimator $\hat{\boldsymbol{\beta}}_n$ obtained with the empirical penalization at $\tau^\star$ can be deduce directly from Figure 5d because the asymptotic MSE is directly proportional to v via a constant α that depends only on the correlations between covariates.

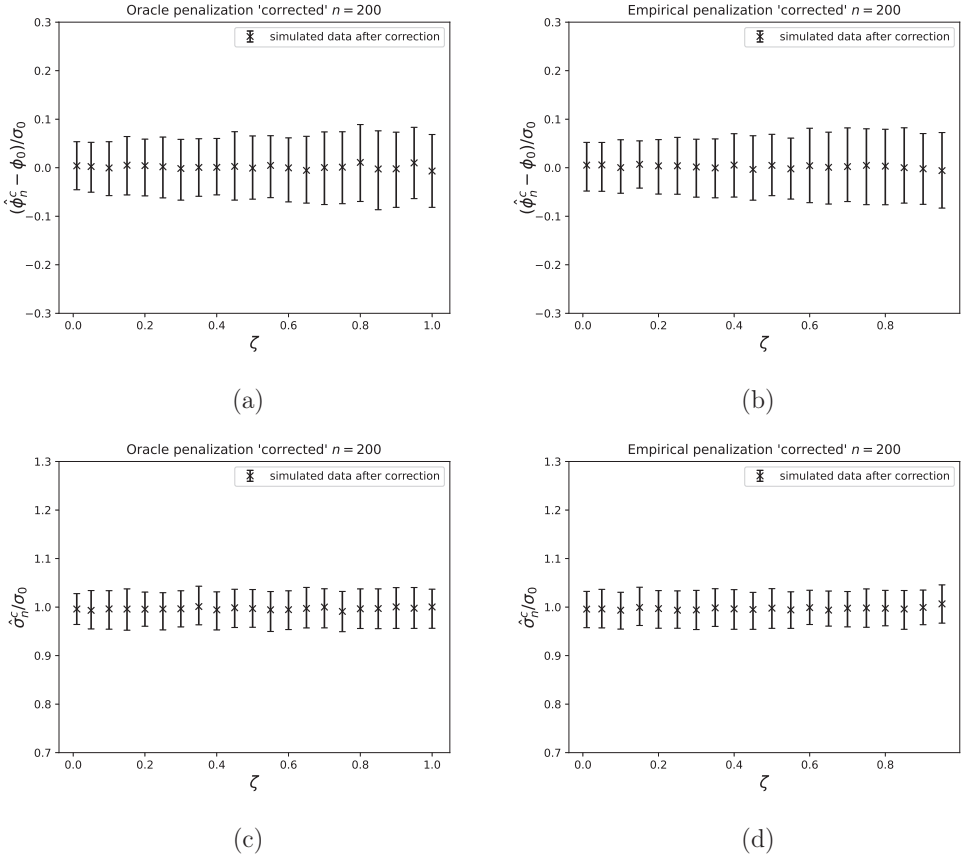


Figure 2.7: Simulated data for the Weibull proportional hazards model ($n = 200$, $S = 1$) after application of regularization parameters computed with the RS equations upon demanding unbiased nuisance parameter estimators. Error bars represent the first and third sample quartiles, while the dashed line is the theoretical prediction. Figures 8a and 8c show that $(\hat{\phi}_n^c - \phi_0)/\hat{\sigma}_0$ fluctuates around zero, while Figures 8c and 8d show that $\hat{\sigma}_n^c/\sigma_0$ fluctuates around one. Hence the estimators $\hat{\phi}_n^c$ and $\hat{\sigma}_n^c$ are indeed unbiased.

In contrast to the Logit model, the Weibull model has further nuisance parameters, namely ϕ and σ. It should be emphasised that, while $\hat{\boldsymbol{\beta}}_n$ is unbiased, the nuisance parameters estimators are not, as one can see in Figure 2.6. While in inference the focus is usually on the regression parameters, it might also be of interest to have unbiased estimators for the nuisance parameters; for instance in probabilistic prediction. We could in principle use our equations to predict for which τ' or η' the nuisance parameters estimators will be unbiased, but a more efficient and general alternative is to use the solution of the RS equation to compute a de-biasing factor, similar to the one employed in [19]. In the present case the RS equations can be rewritten in terms of $(\phi - \phi_0)/\sigma_0$ and σ/σ_0. The latter combinations turn out to depend only on ζ, τ' or η', and S. Since the value of the estimator will be close to

the solution of the RS equations, we may write

$$\frac{\hat{\phi}_n - \phi_0}{\sigma_0} = h(\zeta, S) + o_n(1), \qquad \frac{\hat{\sigma}_n}{\sigma_0} = g(\zeta, S) + o_n(1) \tag{2.51}$$

where $h(.,.)$ and $g(.,.)$ are expressions obtained by solving the RS equations. One subsequently solves these two equations for ϕ_0 and σ_0, leading to new unbiased estimators $\hat{\phi}_n^c, \hat{\sigma}_n^c$. In Figure 2.7 it can be observed that this procedure indeed leads to estimators for the nuisance parameters that are unbiased.

2.5 Discussion and conclusion

Ridge regression, a popular inference approach in the high dimensional regime where $\zeta = p/n$ is finite and fixed, leads in generalized linear models typically to rotated estimators of the association parameters $\boldsymbol{\beta} \in \mathbf{R}^p$. This prompted us to search for generalizations of the ridge penalty, for which the corresponding estimator will be unbiased. We focused on two elliptical generalizations, of the form $\boldsymbol{\beta} \cdot \mathbf{A}\boldsymbol{\beta}$. An 'oracle' penalization, defined by $\mathbf{A} = \mathbf{A}_0$, uses the population covariance matrix $\mathbf{A}_0$ of the covariates (information that is not always available). An 'empirical' penalization uses instead the sample covariance matrix $\mathbf{A} = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i / n$. Both are covariant under linear transformations of the covariates. We derive an explicit expression for the distribution of the PML/MAP estimator $\hat{\boldsymbol{\beta}}_n$, obtained by optimizing the corresponding objective functions, under the assumption of Gaussian covariates. This estimator is for both covariant priors shown to be typically aligned with the true parameter vector $\boldsymbol{\beta}_0$, and its distribution depends solely on two scalar quantities K_n and V_n, and on $\mathbf{A}_0$. The quantities K_n and V_n are finite n equivalents of the replica symmetric (RS) order parameters of [19, 5, 4]. In the high dimensional regime they fluctuate weakly around deterministic values. Under the assumption that they are asymptotically self-averaging, we show how their values can be computed from a small system of self-consistent equations, the RS equations. Deriving the RS equations alternatively via the cavity method clarifies the relevant underlying assumptions. The RS equations enable one to control the bias-variance trade-off; one can select the value of the regularization parameter by fixing the value of bias that is deemed to be acceptable for the problem at hand, and compute the RS curves to predict the associated variance. Our equations can also be used to give explicit expressions for covariant regularization parameters such that not only the direction but also the length of the estimator is correct, i.e. such that the estimators will be strictly unbiased.

We tested our theoretical results against the parameter statistics computed from inferences with simulated data, and found excellent agreement. We worked out the theory for two popular generalized linear models: the Logit model for binary classification, and the Weibull proportional hazards model for survival analysis. For both applications, the solution of the RS equations correctly predicts the behaviour of the estimators obtained from simulated data, even for relatively small sample sizes. The PML/MAP estimators obtained with the oracle penalization achieve a lower variance than the ones computed using the empirical penalty, which can be explained in terms of the bias-variance trade-off. When ζ is close to one, i.e where the number of covariates is almost equal to the number of samples, the sample covariance matrix may develop zero eigenvalues, leading to pathological estimators. Sufficiently below $\zeta = 1$ zero eigenvalues will not occur, and the empirical penalization is an effective, practical and easy to implement method for achieving unbiased estimators. Oracle penalization obviously does not have this problem, and can be used even for $\zeta > 1$ (provided

$\mathbf{A}_0$ has full rank). This also suggests an alternative third route, where one estimates the population covariance matrix based on previous data, with methods such as [16], to be used as a proxy for $\mathbf{A}_0$ in the oracle penalization (a possible subject of future investigation).

While the RS equations can be solved for any given value of $S = |\mathbf{A}_0^{1/2}\boldsymbol{\beta}_0|$, in practice one does not know this value. For a significant class of models S is in fact found to drop out of the RS equations, and for those models where this does not happen there are transparent approaches to estimate S. Still this is a subject hat warrants further research. It might also be interesting to analyse what happens when the covariant regularizers are used in a model mis-match setting (see e.g. [1]), which in principle can be addressed in the present formulation.

Our results are derived under the assumption of Gaussian distributed covariates. The results concerning the first two moments of the estimator $\hat{\boldsymbol{\beta}}_n$ should however hold more generally, as suggested by the work of e.g. [11, 4, 26, 5, 19]. The underlying reason is that even for non-Gaussian covariates, the quantities $\mathbf{X} \cdot \boldsymbol{\beta}$ may asymptotically have Gaussian statistics due to the Central Limit Theorem (under certain conditions). See also [13, 28, 18]. The universality of the asymptotic behaviour of the objective function has been proven for the machine learning setting in [24]. It would be interesting to study whether in arbitrary generalized linear models this universality would extend to the asymptotic properties of the estimators, at least concerning the first two moments.

To conclude, we present covariant generalizations of ridge penalization in high-dimensional regression, such that the parameter estimators no longer exhibit the undesirable rotation that plagues conventional ridge regularization (especially for correlated data). For arbitrary generalized linear models the corresponding inference process is described accurately by an RS theory (derived alternatively via the cavity method), even for modest sample sizes. This theory can be used e.g. to tune the bias-variance trade-off, or to predict which values of the covariant regularization parameters should be used in practice to obtain strictly unbiased estimators.

Bibliography

[1] J Barbier, WK Chen, D Panchenko, and M Sáenz. Performance of bayesian linear regression in a model with mismatch, 2021.

[2] HW Borchers. Package 'pracma'. *Practical numerical math functions, version*, 2(5), 2019.

[3] S Boucheron, G Lugosi, and P Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[4] ACC Coolen, JE Barrett, P Paga, and CJ Perez-Vicente. Replica analysis of overfitting in regression models for time-to-event data. *Journal of physics A: mathematical and theoretical*, 50(37):375001, 2017.

[5] ACC Coolen, M Sheikh, A Mozeika, F Aguirre-Lopez, and F Antenucci. Replica analysis of overfitting in generalized linear regression models. *Journal of Physics A: Mathematical and Theoretical*, 53(36):365001, 2020.

[6] JB Copas. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(3):311–335, 12 1983.

[7] JB Copas. Using regression models for prediction: shrinkage and regression to the mean. *Statistical methods in medical research*, 6(2):167–183, 1997.

[8] RM Corless, GH Gonnet, DEG Hare, DJ Jeffrey, and DE Knuth. On the lambert w function. *Advances in Computational mathematics*, 5:329–359, 1996.

[9] R Couillet and M Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2014.

[10] N El Karoui. Spectrum estimation for large dimensional covariance matrices using random matrix theory. *The Annals of Statistics*, 36(6):2757 – 2790, 2008.

[11] N El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170:95 – 175, 2017.

[12] N El Karoui, PJ Bean, Dand Bickel, C Lim, and B Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[13] S Goldt, B Loureiro, G Reeves, F Krzakala, M Mézard, and L Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

[14] FE Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics. Springer International Publishing, 2015.

[15] AE Hoerl and RW Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[16] O Ledoit and M Wolf. Spectrum estimation: A unified framework for covariance matrix estimation and pca in large dimensions. *Journal of Multivariate Analysis*, 139:360–384, 2015.

[17] G Livan, M Novaes, and P Vivo. *Introduction to Random Matrices: Theory and Practice*. SpringerBriefs in Mathematical Physics. Springer International Publishing, 2018.

[18] B Loureiro, C Gerbelot, H Cui, S Goldt, F Krzakala, M Mézard, and L Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114001, nov 2022.

[19] E Massa, M Jonker, and A Coolen. Correction of overfitting bias in regression models, 2023.

[20] E Massa, MA Jonker, and ACC Coolen. Penalization-induced shrinking without rotation in high dimensional glm regression: a cavity analysis. *Journal of Physics A: Mathematical and Theoretical*, 55(48):485002, 2022.

[21] M Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A: Mathematical and General*, 22(12):2181, 1989.

[22] M Mézard and G Parisi. Replicas and optimization. *Journal de Physique Lettres*, 46(17):771–778, 1985.

[23] M Mézard, G Parisi, and MA Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[24] A Montanari and BN Saeed. Universality of empirical risk minimization. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4310–4312. PMLR, 02–05 Jul 2022.

[25] G Parisi. The mean field theory of spin glasses: the heuristic replica approach and recent rigorous results. *Letters in Mathematical Physics*, 88(1):255–269, 2009.

[26] M Sheikh and ACC Coolen. Analysis of overfitting in the regularized cox model. *Journal of Physics A: Mathematical and Theoretical*, 52(38):384002, 2019.

[27] C Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206, 1956.

[28] S Stergiopoulos. *Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems*. CRC Press Revivals. CRC Press, 2017.

[29] P Sur and EJ Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[30] M Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

[31] AW Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

2.6 Appendix

2.6.1 Representation of the PML/MAP estimator

The PML/MAP estimator for the regression parameters is defined by the following optimization problem

$$\begin{aligned}&\hat{\boldsymbol{\beta}}_n^{\sim} = \\ &= \arg\max_{\boldsymbol{\beta}}\Big(\max_{\boldsymbol{\sigma}}\Big\{\sum_{i=1}^{n}\log p(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) - \frac{p}{2}\boldsymbol{\beta}\cdot\Big(\tau'\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\mathcal{X}}_i\boldsymbol{\mathcal{X}}_i + \eta'\boldsymbol{I}\Big)\boldsymbol{\beta}\Big\}\Big)\end{aligned} \tag{2.52}$$

where

$$T_i|\boldsymbol{\mathcal{X}}_i \sim p(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}_0^{\sim},\boldsymbol{\sigma}_0) \tag{2.53}$$

Now consider any rotation $\mathbf{R}_0$ in $\mathbb{R}^p$ around $\boldsymbol{\beta}_0$. From property (2.7) in the main text and the fact that the conditional distribution of the response T_i depends only on the projection $\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}_0^{\sim}$,

$$T_i|\boldsymbol{\mathcal{X}}_i = T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}_0^{\sim} \tag{2.54}$$

we obtain

$$\hat{\boldsymbol{\beta}}_n^{\sim} = \hat{\boldsymbol{\beta}}_n(\{T_i, \boldsymbol{\mathcal{X}}_i\}, \boldsymbol{\beta}_0^{\sim}) = \mathbf{R}_0\hat{\boldsymbol{\beta}}_n(\{T_i, \mathbf{R}_0\boldsymbol{\mathcal{X}}_i\}, \boldsymbol{\beta}_0^{\sim}) \stackrel{d}{=} \mathbf{R}_0\hat{\boldsymbol{\beta}}_n(\{T_i, \boldsymbol{\mathcal{X}}_i\}, \boldsymbol{\beta}_0^{\sim}) \tag{2.55}$$

where we have denoted equality in distribution with $\stackrel{d}{=}$. The first equality follows because $\boldsymbol{\beta}_0^{\sim} = \mathbf{R}_0\boldsymbol{\beta}_0^{\sim}$, and the last equality follows because $\mathbf{O}\boldsymbol{\mathcal{X}}_i$ has the same distribution as $\boldsymbol{\mathcal{X}}_i$ for any rotation $\mathbf{O}$ since $\boldsymbol{\mathcal{X}}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ (hence also for any $\mathbf{R}_0$). To understand the implications of (2.55), note that we can always decompose $\hat{\boldsymbol{\beta}}_n^{\sim}$ into a component along $\boldsymbol{\beta}_0^{\sim}$ and a component in the subspace of $\mathbb{R}^p$ orthogonal to $\boldsymbol{\beta}_0^{\sim}$

$$\hat{\boldsymbol{\beta}}_n^{\sim} = \hat{\boldsymbol{\beta}}_{n,\|}^{\sim} + \hat{\boldsymbol{\beta}}_{n,\perp}^{\sim} \qquad \hat{\boldsymbol{\beta}}_{n,\|}^{\sim} \cdot \hat{\boldsymbol{\beta}}_{n,\perp}^{\sim} = 0 \tag{2.56}$$

with

$$\hat{\boldsymbol{\beta}}_{n,\|}^{\sim} := \boldsymbol{\beta}_0^{\sim}(\boldsymbol{\beta}_0^{\sim} \cdot \hat{\boldsymbol{\beta}}_n^{\sim})/\|\boldsymbol{\beta}_0^{\sim}\|^2 \qquad \hat{\boldsymbol{\beta}}_{n,\perp}^{\sim} := \Big(\boldsymbol{I} - \boldsymbol{\beta}_0^{\sim}\boldsymbol{\beta}_0^{\sim}/\|\boldsymbol{\beta}_0^{\sim}\|^2\Big)\hat{\boldsymbol{\beta}}_n^{\sim} \ . \tag{2.57}$$

Then equation (2.55) implies that

$$\mathbf{R}_0\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim} \stackrel{d}{=} \hat{\boldsymbol{\beta}}_{n,\perp}^{\sim} \tag{2.58}$$

because $\hat{\boldsymbol{\beta}}_{n,\|}^{\sim}$ is not affected by any $\mathbf{R}_0$ as it is parallel to $\boldsymbol{\beta}_0^{\sim}$. Hence all the values of $\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim}$ that have the same length, i.e. that lie on a sphere in the subspace of $\mathbb{R}^p$ orthogonal to $\boldsymbol{\beta}_0^{\sim}$, must have the same probability density. In turn this means that the length of $\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim}$ must be independent from its direction: conditional on the length, the direction is uniformly distributed over a sphere in the above mentioned subspace. This means that, conditional on $\hat{\boldsymbol{\beta}}_{n,\|}^{\sim}$, we have the following representation for $\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim}$

$$\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim} = \|\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim}\|\frac{\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim}}{\|\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim}\|} = \|\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim}\|\mathbf{U} \tag{2.59}$$

with $\mathbf{U}$ uniformly distributed on the unit sphere in the $p-1$ dimensional subspace $\mathcal{S}_{p-2}$ of $\mathbb{R}^p$ that is orthogonal to $\boldsymbol{\beta}_0^{\sim}$. Now note that

$$\|\hat{\boldsymbol{\beta}}_{n,\perp}^{\sim}\| = \sqrt{\|\hat{\boldsymbol{\beta}}_n^{\sim}\|^2 - \|\hat{\boldsymbol{\beta}}_{n,\|}^{\sim}\|^2} = \sqrt{\|\hat{\boldsymbol{\beta}}_n^{\sim}\|^2 - (\boldsymbol{\beta}_0^{\sim} \cdot \hat{\boldsymbol{\beta}}_n^{\sim})^2/\|\boldsymbol{\beta}_0^{\sim}\|^2} \ . \tag{2.60}$$

So finally we obtain the representation

$$\hat{\boldsymbol{\beta}}_n^{\sim} = \frac{\boldsymbol{\beta}_0^{\sim} \cdot \hat{\boldsymbol{\beta}}_n^{\sim}}{\|\boldsymbol{\beta}_0^{\sim}\|^2}\boldsymbol{\beta}_0^{\sim} + \sqrt{\|\hat{\boldsymbol{\beta}}_n^{\sim}\|^2 - \frac{(\boldsymbol{\beta}_0^{\sim} \cdot \hat{\boldsymbol{\beta}}_n^{\sim})^2}{\|\boldsymbol{\beta}_0^{\sim}\|^2}}\mathbf{U} \ . \tag{2.61}$$

2.6.2 The value of α^2

To compute the limit defining α^2 in (2.29), we note that

$$\begin{aligned} \mathbb{E}[A_p^2] &= \mathbb{E}\Big[\mathbf{U} \cdot \mathbf{A}_0^{-1}\mathbf{U}\Big] = \mathbb{E}\Big[\frac{\mathbf{Z} \cdot \mathbf{P}\mathbf{A}_0^{-1}\mathbf{P}\mathbf{Z}}{\mathbf{Z} \cdot \mathbf{P}\mathbf{Z}}\Big] = \\ &= \frac{1}{p-1}\Big(\mathrm{Tr}(\mathbf{A}_0^{-1}) - \frac{\boldsymbol{\beta}_0^{\sim} \cdot \mathbf{A}_0^{-1}\boldsymbol{\beta}_0^{\sim}}{\|\boldsymbol{\beta}_0^{\sim}\|^2}\Big) + o_n(1) \ . \end{aligned} \tag{2.62}$$

In (2.62), to obtain the third equality we used the fact that

$$\mathbf{U} \stackrel{d}{=} \frac{\mathbf{P}\mathbf{Z}}{\|\mathbf{P}\mathbf{Z}\|} \qquad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad \mathbf{P} := \boldsymbol{I} - \frac{\boldsymbol{\beta}_0^{\sim}\boldsymbol{\beta}_0^{\sim}\cdot}{\|\boldsymbol{\beta}_0^{\sim}\|^2} \ . \tag{2.63}$$

For the last equality we used the fact that quadratic forms like $\mathbf{Z}\cdot\mathbf{MZ}/p$ concentrate around their expected value $\mathrm{Tr}(\mathbf{M})$ in the asymptotic high dimensional limit [12, 3]

$$\mathbf{Z}\cdot\mathbf{MZ}/p = \mathrm{Tr}(\mathbf{M})/p + o_n(1)\ . \tag{2.64}$$

At this point we see that by definition of $\boldsymbol{\beta}_0^{\sim} := \mathbf{A}_0^{1/2}\boldsymbol{\beta}_0$

$$\frac{\boldsymbol{\beta}_0^{\sim}\cdot\mathbf{A}_0^{-1}\boldsymbol{\beta}_0^{\sim}}{\|\boldsymbol{\beta}_0^{\sim}\|^2} = \frac{\boldsymbol{\beta}_0\cdot\boldsymbol{\beta}_0}{\boldsymbol{\beta}_0\cdot\mathbf{A}_0\boldsymbol{\beta}_0} \tag{2.65}$$

and

$$\lambda_{\max}^{-1}(\mathbf{A}_0) \leq \frac{\boldsymbol{\beta}_0\cdot\boldsymbol{\beta}_0}{\boldsymbol{\beta}_0\cdot\mathbf{A}_0\boldsymbol{\beta}_0} \leq \lambda_{\min}^{-1}(\mathbf{A}_0) \tag{2.66}$$

thus

$$\lambda_{\max}^{-1}(\mathbf{A}_0)/(p-1) \leq \left|\alpha_p^2 - \mathrm{Tr}(\mathbf{A}_0^{-1})/(p-1)\right| \leq \lambda_{\min}^{-1}(\mathbf{A}_0)/(p-1)\ . \tag{2.67}$$

If $\lambda_{min}(\mathbf{A}_0) = O(p^{\alpha})$, with $\alpha > -1$, then $\lim_{p\to\infty}\lambda_{\min}^{-1}(\mathbf{A}_0)/(p-1) = 0$. Hence we conclude that

$$\alpha^2 := \lim_{p\to\infty}\mathbb{E}[A_p^2] = \lim_{p\to\infty}\frac{1}{p}\mathrm{Tr}(\mathbf{A}_0^{-1})\ . \tag{2.68}$$

2.6.3 Derivation of the RS equations with the Cavity method

In this section we derive the Replica symmetric equations (2.21, 2.22, 2.23, 2.24). We first establish the asymptotic properties of the PML/MAP estimator of the regression parameters $\hat{\boldsymbol{\beta}}_n$ for a fixed value of the nuisance parameters $\boldsymbol{\sigma}$. We subsequently show that, given our assumptions and approximations, this is sufficient to establish also the asymptotic behaviour of $\hat{\boldsymbol{\sigma}}_n$. We will use several times the notation $o_n(1)$ to indicate a reminder that goes to zero as n diverges.

Analysis via statistical physics

We exploit the analogy between optimization and statistical physics, whereby variables over which an optimization is carried out are interpreted as degrees of freedom of particles, with a Hamiltonian that is minus the objective function to be optimized. The solution of the optimization problem then equals the ground state of the Hamiltonian. Here we aim to understand the properties of the PML/MAP estimator obtained with $\{T_i, \boldsymbol{\mathcal{X}}_i\}_{i=1}^n$ where $\boldsymbol{\mathcal{X}}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $T_i|\boldsymbol{\mathcal{X}}_i \sim p(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}_0^{\sim}, \boldsymbol{\sigma}_0)$ and $\boldsymbol{\beta}_0^{\sim} := \mathbf{A}_0^{1/2}\boldsymbol{\beta}_0$. So we consider the Hamiltonian

$$\mathcal{H}_n(\boldsymbol{\beta}, \boldsymbol{\sigma}) := -l_n(\boldsymbol{\beta}, \boldsymbol{\sigma}) = \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{n}\rho(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}, \boldsymbol{\sigma})\ . \tag{2.69}$$

Where

$$\rho(T|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}, \boldsymbol{\sigma}) := \log p(T|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}, \boldsymbol{\sigma}) - \frac{1}{2}\zeta\tau'(\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta})^2 \tag{2.70}$$

The statistical properties of the physical system at temperature $1/\gamma$ follow from the Gibbs measure

$$\boldsymbol{\beta}|\{T_i, \boldsymbol{\mathcal{X}}_i\}_{i=1}^n \sim p(\boldsymbol{\beta}) = \mathrm{e}^{-\gamma\mathcal{H}_n(\boldsymbol{\beta},\boldsymbol{\sigma})}/Z_n(\gamma) \tag{2.71}$$

where the partition function ensures the normalization, i.e.

$$Z_n(\gamma) = \int \mathrm{d}\boldsymbol{\beta}\ \mathrm{e}^{-\gamma\mathcal{H}_n(\boldsymbol{\beta},\boldsymbol{\sigma})}\ . \tag{2.72}$$

Under regularity conditions on the Hamiltonian function, and for a well behaved functions f, the Laplace argument implies that

$$f(\hat{\boldsymbol{\beta}}_n^{\sim}) = \lim_{\gamma\to\infty} \int \mathrm{d}\boldsymbol{\beta}\ f(\boldsymbol{\beta}) \frac{\mathrm{e}^{-\gamma\mathcal{H}_n(\boldsymbol{\beta},\boldsymbol{\sigma})}}{Z_n(\gamma)} = \lim_{\gamma\to\infty} \langle f(\boldsymbol{\beta})\rangle \tag{2.73}$$

where

$$\hat{\boldsymbol{\beta}}_n^{\sim} := \arg\min_{\boldsymbol{\beta}} \Big\{ \mathcal{H}_n(\boldsymbol{\beta},\boldsymbol{\sigma}) \Big\} = \arg\max_{\boldsymbol{\beta}} \Big\{ l_n(\boldsymbol{\beta},\boldsymbol{\sigma}) \Big\} \ . \tag{2.74}$$

We are interested in the typical properties of (functions of) the PML estimator, hence we study averages over all possible realizations of the data-set:

$$\mathbb{E}\Big[f(\hat{\boldsymbol{\beta}}_n^{\sim}) \Big] = \lim_{\gamma\to\infty} \mathbb{E}\Big[\langle f(\boldsymbol{\beta})\rangle \Big] \ . \tag{2.75}$$

We switched the limit and the expectation over the data-set, which is valid under regularity conditions on f. The right hand side of (2.75) has the structure of the computation of a disorder-average observable f in a spin glass model [23, 21, 25], with the role of disorder played by the data set $\{T_i, \boldsymbol{\mathcal{X}}_i\}_{i=1}^n$. This enables the use of tools from the physics of disordered systems.

Application of the Cavity method

We next use the cavity method to derive the asymptotic properties of the PML/MAP estimator. Consider a sufficiently well behaved function f that depends only on the linear predictor $\boldsymbol{\mathcal{X}}_i \cdot \boldsymbol{\beta}$, the response T_i of the ith observation and the (for now fixed) nuisance parameters $\boldsymbol{\sigma}$. By noting that

$$\mathcal{H}_{n,p}(\boldsymbol{\beta},\boldsymbol{\sigma}) := \mathcal{H}_{(i),p}(\boldsymbol{\beta},\boldsymbol{\sigma}) - \rho(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) \ . \tag{2.76}$$

We can write

$$\begin{aligned} \langle f(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma})\rangle &= \frac{\int \mathrm{d}\boldsymbol{\beta}\ f(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) \mathrm{e}^{\gamma\rho(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) - \gamma\mathcal{H}_{(i)}(\boldsymbol{\beta},\boldsymbol{\sigma})}}{\int \mathrm{d}\boldsymbol{\beta}\ \mathrm{e}^{\gamma\rho(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) - \gamma\mathcal{H}_{(i)}(\boldsymbol{\beta},\boldsymbol{\sigma})}} = \\ &\frac{\Big\langle f(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) \exp\Big\{ \gamma\rho(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) \Big\}\Big\rangle_{(i)}}{\Big\langle \exp\Big\{ \gamma\rho(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) \Big\}\Big\rangle_{(i)}} \ . \end{aligned} \tag{2.77}$$

Here $\langle\rangle_{(i)}$ refers to an average under the Gibbs measure with Hamiltonian $\mathcal{H}_{(i),p}(\boldsymbol{\beta},\boldsymbol{\sigma})$. The computation of (2.77) is difficult because we need the distribution of the linear predictor $Y_i := \boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}$ under the cavity measure. The first two moments of Y_i are

$$\langle Y_i\rangle = \boldsymbol{\mathcal{X}}_i\cdot\langle\boldsymbol{\beta}\rangle_{(i)}, \qquad \langle Y_i^2\rangle = \boldsymbol{\mathcal{X}}_i\cdot\langle\boldsymbol{\beta}\boldsymbol{\beta}\rangle_{(i)}\boldsymbol{\mathcal{X}}_i \ . \tag{2.78}$$

Now standard results in the theory of concentration of measure [12, 3] tell us that

$$\begin{aligned} \mathcal{U}_{(i)}^2 &= \langle Y_i^2\rangle - \langle Y_i\rangle^2 = \boldsymbol{\mathcal{X}}_i\cdot\Big(\langle\boldsymbol{\beta}\boldsymbol{\beta}\rangle_{(i)} - \langle\boldsymbol{\beta}\rangle_{(i)}\langle\boldsymbol{\beta}\rangle_{(i)}\Big)\boldsymbol{\mathcal{X}}_i = \\ &= \mathrm{Tr}\Big(\langle\boldsymbol{\beta}\boldsymbol{\beta}\rangle_{(i)} - \langle\boldsymbol{\beta}\rangle_{(i)}\langle\boldsymbol{\beta}\rangle_{(i)}\Big) + o_n(1) = \\ &= \langle\boldsymbol{\beta}\cdot\boldsymbol{\beta}\rangle_{(i)} - \langle\boldsymbol{\beta}\rangle_{(i)}\cdot\langle\boldsymbol{\beta}\rangle_{(i)} + o_n(1) \ . \end{aligned} \tag{2.79}$$

We note that $\mathcal{U}^2_{(i)}$ is exactly the nonlinear susceptibility under the cavity Gibbs measure. We assume that the cavity distribution of Y_i is Gaussian, so

$$Y_i = \boldsymbol{\mathcal{X}}_i \cdot \langle \boldsymbol{\beta} \rangle_{(i)} + \mathcal{U}_{(i)} Z_i, \qquad Z_i \sim \mathcal{N}(0,1) \tag{2.80}$$

with $Z_i \perp T_i, \boldsymbol{\mathcal{X}}_i$. Remembering that $T_i | \boldsymbol{\mathcal{X}}_i \sim p(T_i | \boldsymbol{\mathcal{X}}_i^T \boldsymbol{\beta}_0^\sim, \boldsymbol{\sigma}_0)$, so the distribution of the disorder depends on $\boldsymbol{\mathcal{X}}_i \cdot \boldsymbol{\beta}_0^\sim$ which is correlated with $\boldsymbol{\mathcal{X}}_i \cdot \langle \boldsymbol{\beta} \rangle_{(i)}$, it is possible to separate $\boldsymbol{\mathcal{X}}_i \cdot \langle \boldsymbol{\beta} \rangle_{(i)}$ into two independent contributions

$$\begin{gathered} \boldsymbol{\mathcal{X}}_i \cdot \langle \boldsymbol{\beta} \rangle_{(i)} = \boldsymbol{\mathcal{X}}_i \cdot \Big(\boldsymbol{I} - \frac{\boldsymbol{\beta}_0^\sim \boldsymbol{\beta}_0^\sim}{\|\boldsymbol{\beta}_0^\sim\|^2} \Big) \langle \boldsymbol{\beta} \rangle_{(i)} + \frac{\boldsymbol{\mathcal{X}}_i \cdot \boldsymbol{\beta}_0^\sim}{\|\boldsymbol{\beta}_0^\sim\|} \frac{\boldsymbol{\beta}_0^\sim \cdot \langle \boldsymbol{\beta} \rangle_{(i)}}{\|\boldsymbol{\beta}_0^\sim\|} = \\ V_{(i)} Q_i + W_{(i)} Z_{0,i}, \qquad Q_i \perp Z_{0,i}, \quad Q_i, Z_{0,i} \sim \mathcal{N}(0,1) \end{gathered} \tag{2.81}$$

involving the generalized overlaps under the leave i-th out measure

$$W_{(i)} := \frac{\boldsymbol{\beta}_0^\sim \cdot \langle \boldsymbol{\beta} \rangle_{(i)}}{\|\boldsymbol{\beta}_0^\sim\|} \tag{2.82}$$

$$V_{(i)} := \Big\| \Big(\boldsymbol{I} - \frac{\boldsymbol{\beta}_0^\sim \boldsymbol{\beta}_0^\sim}{\|\boldsymbol{\beta}_0^\sim\|^2} \Big) \langle \boldsymbol{\beta} \rangle_{(i)} \|^2 = \| \langle \boldsymbol{\beta} \rangle_{(i)} \|^2 - W^2_{(i)} \ . \tag{2.83}$$

These are independent of $T_i, \boldsymbol{\mathcal{X}}_i$, hence

$$Y_i = W_{(i)} Z_{0,i} + V_{(i)} Q_i + \mathcal{U}_{(i)} Z_i \tag{2.84}$$

and

$$\begin{gathered} \langle f(T_i | \boldsymbol{\mathcal{X}}_i \cdot \boldsymbol{\beta}, \boldsymbol{\sigma}) \rangle = \\ \frac{\mathbb{E}_{Z_i} \Big[f(T_i | W_{(i)} Z_{0,i} + V_{(i)} Q_i + \mathcal{U}_{(i)} Z_i, \boldsymbol{\sigma}) \mathrm{e}^{\gamma \rho(T_i | W_{(i)} Z_{0,i} + V_{(i)} Q_i + U_{(i)} Z_i, \boldsymbol{\sigma})} \Big]}{\mathbb{E}_{Z_i} \Big[\mathrm{e}^{\gamma \rho(T_i | W_{(i)} Z_{0,i} + V_{(i)} Q_i + \mathcal{U}_{(i)} Z_i, \boldsymbol{\sigma})} \Big]} \ . \end{gathered} \tag{2.85}$$

Where we indicated with $\mathbb{E}_{Z_i}[]$ the expectation over the Gaussian random variable Z_i. This can be rewritten via a simple change of variable, $\xi := W_{(i)} Z_{0,i} + V_{(i)} Q_i + \mathcal{U}_{(i)} Z_i$, as

$$\begin{aligned} \langle f(T_i | \boldsymbol{\mathcal{X}}_i \cdot \boldsymbol{\beta}, \boldsymbol{\sigma}) \rangle &= \mathbb{E}_{\xi_i} \Big[f(T_i, \xi_i, \boldsymbol{\sigma}) \Big] = \\ &= \frac{\int \mathrm{d}\xi \ f(T_i, \xi, \boldsymbol{\sigma}) \exp \Big\{ - \frac{1}{2} \Big(\frac{\xi - W_{(i)} Z_{0,i} + V_{(i)} Q_i}{\mathcal{U}_{(i)}} \Big)^2 + \gamma \rho(T_i, \xi, \boldsymbol{\sigma}) \Big\}}{\int \mathrm{d}\xi \ \exp \Big\{ - \frac{1}{2} \Big(\frac{\xi - W_{(i)} Z_{0,i} + V_{(i)} Q_i}{\mathcal{U}_{(i)}} \Big)^2 + \gamma \rho(T_i, \xi, \boldsymbol{\sigma}) \Big\}} \end{aligned} \tag{2.86}$$

The disorder is now contained only in the leave-one-out generalized overlaps $W_{(i)}, V_{(i)}, \mathcal{U}_{(i)}$. It seems reasonable to approximate the overlaps computed under the full Gibbs measure by their leave-one-out approximations for each i,

$$W_n := \frac{\boldsymbol{\beta}_0^\sim \cdot \langle \boldsymbol{\beta} \rangle}{\|\boldsymbol{\beta}_0^\sim\|} = W_{(i)} + o_n(1) \tag{2.87}$$

$$V_n := \Big\| \Big(\boldsymbol{I} - \frac{\boldsymbol{\beta}_0^\sim \boldsymbol{\beta}_0^\sim}{\|\boldsymbol{\beta}_0^\sim\|^2} \Big) \langle \boldsymbol{\beta} \rangle \Big\|^2 = V_{(i)} + o_n(1) \tag{2.88}$$

$$\mathcal{U}_n := \langle \boldsymbol{\beta} \cdot \boldsymbol{\beta} \rangle - \langle \boldsymbol{\beta} \rangle \cdot \langle \boldsymbol{\beta} \rangle = \mathcal{U}_{(i)} + o_n(1) \tag{2.89}$$

because, in the limit of large n, disregarding one data point should not influence the inference results. This, in turn, means that one can alternatively use the approximation

$$W_n := \frac{1}{n}\sum_{i=1}^{n} W_{(i)} + o_n(1), \qquad V_n := \frac{1}{n}\sum_{i=1}^{n} V_{(i)} + o_n(1) \tag{2.90}$$

$$\mathcal{U}_n := \frac{1}{n}\sum_{i=1}^{n} \mathcal{U}_{(i)} + o_n(1)\ . \tag{2.91}$$

Which shows that it is plausible that W_n, V_n and $\mathcal{U}_n$ concentrate around deterministic values, or are self averaging in physical jargon, which can be stated as

$$W_n := w + o_n(1), \qquad V_n := v + o_n(1), \qquad \mathcal{U}_n := \tilde{u} + o_n(1)\ . \tag{2.92}$$

The remaining task is to determine the values $(w, v, \tilde{u})$, which according to the above arguments should satisfy a set of coupled self-consistency equations.

Self-consistent equations for the overlaps

Given the Gibbs measure

$$\boldsymbol{\beta}|\{T_i, \boldsymbol{\mathcal{X}}_i\} \sim \frac{1}{Z_n} \mathrm{e}^{\gamma\left(\sum_{i=1}^{n} \rho(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma}) - \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2\right)} \tag{2.93}$$

it is easy to derive the following equation via integration by parts in $\boldsymbol{\beta}$:

$$\gamma\eta\langle\boldsymbol{\beta}\rangle = \gamma\sum_{i=1}^{n} \boldsymbol{\mathcal{X}}_i\langle\varphi(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma})\rangle \tag{2.94}$$

By taking the scalar product with $\boldsymbol{\beta}_0^{\sim}$ and dividing by $\|\boldsymbol{\beta}_0^{\sim}\| = S$, we get

$$\gamma\eta\boldsymbol{\beta}_0^{\sim}\cdot\langle\boldsymbol{\beta}\rangle/\|\boldsymbol{\beta}_0^{\sim}\| = \gamma\sum_{i=1}^{n} \boldsymbol{\beta}_0^{\sim}\cdot\boldsymbol{\mathcal{X}}_i/\|\boldsymbol{\beta}_0^{\sim}\|\langle\varphi(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma})\rangle\ . \tag{2.95}$$

Using the concentration/self averaging hypothesis (2.92) and taking the expectation over the disorder (i.e the data-set), we then obtain a self consistent equation for w

$$\begin{aligned} w &= \mathbb{E}\Big[\boldsymbol{\beta}_0^{\sim}\cdot\langle\boldsymbol{\beta}\rangle\Big]/S + o_n(1) \\ &= \frac{1}{\eta}\sum_{i=1}^{n}\mathbb{E}\Big[\langle\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}_0^{\sim}\varphi(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma})\rangle\Big]/S + o_n(1) \\ &= \frac{n}{\eta}\mathbb{E}\Big[Z_0\mathbb{E}_\xi[\varphi(T|\xi,\boldsymbol{\sigma})]\Big] + o_n(1) \end{aligned} \tag{2.96}$$

with

$$\mathbb{E}_\xi\Big[f(\xi)\Big] := \frac{\int \mathrm{d}\xi\ \mathrm{e}^{-\frac{1}{2}\left(\frac{\xi - vQ - wZ_0}{\tilde{u}}\right)^2 + \gamma\rho(T|\xi,\boldsymbol{\sigma})} f(\xi)}{\int \mathrm{d}\xi\ \mathrm{e}^{-\frac{1}{2}\left(\frac{\xi - vQ - wZ_0}{\tilde{u}}\right)^2 + \gamma\rho(T|\xi,\boldsymbol{\sigma})}}\ . \tag{2.97}$$

We will now drop the reminders $o_n(1)$ to ease the notation, bearing in mind that these will be negligible as n diverges, so long as the overlaps concentrate, or are self averaging in physical jargon. We see that

$$\gamma\mathbb{E}_\xi\Big[\varphi(T|\xi,\boldsymbol{\sigma})\Big] = \mathbb{E}_\xi\Big[\Big(\frac{\xi - vQ - wZ_0}{\tilde{u}^2}\Big)\Big]\ . \tag{2.98}$$

This implies

$$w = \frac{n}{\gamma\eta\tilde{u}^2}\Big(\mathbb{E}\Big[Z_0\mathbb{E}_\xi[\xi]\Big] - w\Big) \tag{2.99}$$

and thus

$$w(1+\gamma\tilde{u}^2\eta/n) = \mathbb{E}\Big[Z_0\mathbb{E}_\xi[\xi]\Big] \ . \tag{2.100}$$

Similarly

$$\begin{aligned}
\gamma\eta(v^2+w^2+\tilde{u}^2) &= \gamma\eta\mathbb{E}\Big[\langle\boldsymbol{\beta}\cdot\boldsymbol{\beta}\rangle\Big] = \gamma\sum_{i=1}^{n}\mathbb{E}\Big[\langle\boldsymbol{\beta}\cdot\boldsymbol{\mathcal{X}}_i\varphi(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma})\rangle\Big] + \gamma p \\
&= \gamma p + \gamma n\mathbb{E}\Big[\mathbb{E}_\xi[\xi\varphi(T|\xi,\boldsymbol{\sigma})]\Big] \\
&= \gamma(p-n) + n\mathbb{E}\Big[\mathbb{E}_\xi[\xi(\xi - vQ - wZ_0)]\Big]/\tilde{u}^2
\end{aligned} \tag{2.101}$$

where we have used

$$\gamma\mathbb{E}\Big[\mathbb{E}_\xi[\xi\varphi(T|\xi,\boldsymbol{\sigma})]\Big] = \mathbb{E}\Big[\mathbb{E}_\xi[\xi(\xi - vQ - wZ_0)]\Big]/\tilde{u}^2 - 1 \ . \tag{2.102}$$

Also

$$\begin{aligned}
\gamma\eta(v^2+w^2) &= \gamma\eta\mathbb{E}\Big[\langle\boldsymbol{\beta}\rangle\cdot\langle\boldsymbol{\beta}\rangle\Big] = \gamma\eta(v^2+w^2) \\
&= \gamma\sum_{i=1}^{n}\mathbb{E}\Big[\langle\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta}\rangle\langle\langle\varphi(T_i|\boldsymbol{\mathcal{X}}_i\cdot\boldsymbol{\beta},\boldsymbol{\sigma})\rangle\Big] = \gamma n\mathbb{E}\Big[\mathbb{E}_\xi[\xi]\mathbb{E}_\xi[\varphi(T|\xi,\boldsymbol{\sigma})]\Big] \\
&= \frac{n}{\tilde{u}^2}\mathbb{E}\Big[\mathbb{E}_\xi[\xi]\mathbb{E}_\xi[\xi - vQ - wZ_0]\Big] \ .
\end{aligned} \tag{2.103}$$

Now take the difference of (2.101) and (2.103) to find

$$\gamma\eta\tilde{u}^2 = p - n + n\mathbb{E}\Big[\mathbb{E}_\xi[\xi^2] - \mathbb{E}_\xi[\xi]^2\Big]/\tilde{u}^2 = p - n + n\mathbb{E}\Big[Q\mathbb{E}_\xi[\xi]\Big]/v \tag{2.104}$$

dividing by n and rearranging the terms gives

$$v(1-\zeta) + v\gamma\tilde{u}^2\eta/n = \mathbb{E}\Big[Q\mathbb{E}_\xi[\xi]\Big] \ . \tag{2.105}$$

We obtain another self consistent equation by substitution

$$\begin{aligned}
\eta\tilde{u}^2\gamma(v^2+w^2)/n &= \mathbb{E}\Big[\mathbb{E}_\xi[\xi]^2\Big] - v\mathbb{E}\Big[Q\mathbb{E}_\xi[\xi]\Big] - w\mathbb{E}\Big[Z_0\mathbb{E}_\xi[\xi]\Big] \\
&= \mathbb{E}\Big[\mathbb{E}_\xi[\xi^2]\Big] - v^2\Big(1-\zeta+\frac{\gamma\tilde{u}^2\eta}{n}\Big) - w^2(1+\frac{\gamma\tilde{u}^2\eta}{n}) \ .
\end{aligned} \tag{2.106}$$

Rearranging terms leads to

$$\mathbb{E}\Big[\mathbb{E}_\xi[\xi]^2\Big] = v^2(1-\zeta) + w^2 + 2\tilde{u}^2\gamma(v^2+w^2)\eta/n \ . \tag{2.107}$$

In summary we have obtained the following approximate equations

$$\mathbb{E}\Big[\mathbb{E}_\xi[\xi]^2\Big] = v^2(1-\zeta) + w^2 + 2\tilde{u}^2\gamma(v^2+w^2)\eta'\zeta \tag{2.108}$$

$$v(1-\zeta) + v\gamma\tilde{u}^2\eta'\zeta = \mathbb{E}\Big[Q\mathbb{E}_\xi[\xi]\Big] \tag{2.109}$$

$$w(1+\gamma\tilde{u}^2\eta'\zeta) = \mathbb{E}\Big[Z_0\mathbb{E}_\xi[\xi]\Big] \tag{2.110}$$

where we used the scaling $\eta = \eta' p$ as in the main text. These equations are valid under our stated assumptions and approximations, and apply so far to any inverse temperature γ. Their solution gives us the values $\tilde{u}_\star, v_\star, w_\star$ as functions of γ, $\boldsymbol{\sigma}$, S and ζ. The value of $\boldsymbol{\sigma}$ though, is generally unknown and must be estimated from the data.

Equations for the nuisance parameters

Following a strategy similar to the one followed so far, we assume also the estimator $\hat{\boldsymbol{\sigma}}_n$ for the nuisance parameters to assume deterministic values in the asymptotic limit $n, p \to \infty \quad p/n = \zeta$. Under the assumption of concentration of the overlaps (2.92), we expect that the internal energy per data point

$$F_n(\boldsymbol{\sigma}) = \Big\langle \frac{1}{n} \sum_{i=1}^{n} \rho(T_i | \boldsymbol{\mathcal{X}}_i \cdot \boldsymbol{\beta}, \boldsymbol{\sigma}) - \frac{1}{2} \eta' \zeta \|\boldsymbol{\beta}\|^2 \Big\rangle \tag{2.111}$$

will be self averaging. i.e. concentrate on a deterministic function $f(\boldsymbol{\sigma})$:

$$F_n(\boldsymbol{\sigma}) = f(\boldsymbol{\sigma}) + o_n(1) \ . \tag{2.112}$$

Using (2.86) for the cavity average, together with (2.92) for the overlap values, gives

$$F_n(\boldsymbol{\sigma}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\xi \Big[\rho(T_i | \xi, \boldsymbol{\sigma}) \Big] - \frac{1}{2} \eta (w^2 + v^2 + \tilde{u}^2) + o_n(1) \ . \tag{2.113}$$

By the law of large numbers we expect that also this quantity will concentrate on

$$f(\boldsymbol{\sigma}) = \mathbb{E}_{T,Q,Z_0} \Big[\mathbb{E}_\xi \Big[\rho(T | \xi, \boldsymbol{\sigma}) \Big] \Big] - \frac{1}{2} \eta (w^2 + v^2 + \tilde{u}^2) \ . \tag{2.114}$$

Hence the estimator for the nuisance parameter $\hat{\boldsymbol{\sigma}}_n$, the value that maximizes F_n, will be close to the deterministic value that maximizes f. The latter solves

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\sigma}} f(\boldsymbol{\sigma}) = \frac{\partial f(\boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} + \frac{\partial f}{\partial w} \frac{\partial w}{\partial \boldsymbol{\sigma}} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial \boldsymbol{\sigma}} + \frac{\partial f}{\partial \tilde{u}} \frac{\partial \tilde{u}}{\partial \boldsymbol{\sigma}} = 0 \ . \tag{2.115}$$

It can be verified by direct calculation that $\frac{\partial}{\partial w} f, \frac{\partial}{\partial v} f$ and $\frac{\partial}{\partial \tilde{u}} f$ vanish at the solution of the self consistent equations for the overlaps (as expected). Hence the deterministic value that maximizes f at inverse temperature γ, is the solution of the following equation (where again we dropped the reminder because negligible in the large n limit), which must be solved simultaneously with the self consistent equations for the overlaps:

$$\mathbb{E}_{T,Q,Z_0} \Big[\mathbb{E}_\xi \big[\nabla_{\boldsymbol{\sigma}} \rho(T | \xi, \boldsymbol{\sigma}) \big] \Big] = \mathbf{0} \ . \tag{2.116}$$

The zero temperature limit

What remains is to take the limit $\gamma \to \infty$ in our equations to recover the overlaps appearing in the representation of the MAP/PML estimator $\tilde{u}_\star, v_\star, w_\star$ and the deterministic values of the nuisance parameters estimator $\boldsymbol{\sigma}_\star$. Since the distribution of the PML/MAP estimator depends on

$$K_n = \mathbb{E} \Big[\frac{\boldsymbol{\beta}_0^\sim \cdot \hat{\boldsymbol{\beta}}_n^\sim}{\|\boldsymbol{\beta}_0^\sim\|^2} \Big] + o_n(1) \qquad V_n^2 = \mathbb{E} \Big[\hat{\boldsymbol{\beta}}_n^\sim \cdot \hat{\boldsymbol{\beta}}_n^\sim \Big] - K_n^2 + o_n(1) \tag{2.117}$$

where

$$\mathbb{E} \Big[\frac{\boldsymbol{\beta}_0^\sim \cdot \hat{\boldsymbol{\beta}}_n^\sim}{\|\boldsymbol{\beta}_0^\sim\|} \Big] = \lim_{\gamma \to \infty} \mathbb{E} \Big[\frac{\boldsymbol{\beta}_0^\sim \cdot \langle \boldsymbol{\beta} \rangle}{\|\boldsymbol{\beta}_0^\sim\|} \Big] = \lim_{\gamma \to \infty} w \tag{2.118}$$

$$\mathbb{E} \Big[\hat{\boldsymbol{\beta}}_n^\sim \cdot \hat{\boldsymbol{\beta}}_n^\sim \Big] = \lim_{\gamma \to \infty} \mathbb{E} \Big[\langle \boldsymbol{\beta} \cdot \boldsymbol{\beta} \rangle \Big] = \lim_{\gamma \to \infty} (v^2 + w^2 + \tilde{u}^2) \tag{2.119}$$

together with the value $\boldsymbol{\sigma}_\star$ that minimizes the internal energy (2.114) at zero temperature, we must compute the limit $\gamma \to \infty$ of equations (2.108, 2.109, 2.110, 2.116).
To do so, we consider the general expression

$$\lim_{\gamma\to\infty} \mathbb{E}_\xi[f(\xi)] = \lim_{\gamma\to\infty} \frac{\int \mathrm{d}\xi \; \mathrm{e}^{-\frac{1}{2}\left(\frac{\xi - vQ - wZ_0}{\tilde{u}}\right)^2 + \gamma\rho(T|\xi,\boldsymbol{\sigma})} f(\xi)}{\int \mathrm{d}\xi \; \mathrm{e}^{-\frac{1}{2}\left(\frac{\xi - vQ - wZ_0}{\tilde{u}}\right)^2 + \gamma\rho(T|\xi,\boldsymbol{\sigma})}} \; . \tag{2.120}$$

We assume the scaling $\tilde{u}^2 = u^2/\gamma$ suggested by equations (2.108,2.109,2.110), where $\tilde{u}^2$ appears always multiplied by γ. We can now proceed via the Laplace argument:

$$\lim_{\gamma\to\infty} \frac{\int \mathrm{d}\xi \; \mathrm{e}^{-\gamma\left(\frac{1}{2}\frac{(\xi - vQ - wZ_0)^2}{u^2} - \rho(T|\xi,\boldsymbol{\sigma})\right)} f(\xi)}{\int \mathrm{d}\xi \; \mathrm{e}^{-\gamma\left(\frac{1}{2}\frac{(\xi - vQ - wZ_0)^2}{u^2} - \rho(T|\xi,\boldsymbol{\sigma})\right)}} = f(\xi_\star) \tag{2.121}$$

with $\xi_\star$ denoting the proximal mapping of the function $\rho(T|.,\boldsymbol{\sigma})$, defined as

$$\xi_\star = \arg\min_\xi \left\{ \frac{1}{2}\frac{(\xi - vQ - wZ_0)^2}{u^2} - \rho(T|\xi,\boldsymbol{\sigma}) \right\} \tag{2.122}$$

(as always under appropriate regularity conditions on f). It is now easy to take the limit in equations (2.108, 2.109, 2.110, 2.116), obtaining

$$\begin{aligned}
\mathbb{E}\left[\xi_\star^2\right] &= v^2(1-\zeta) + w^2 + 2u^2(v^2+w^2)\eta'\zeta & (2.123)\\
v(1-\zeta) + vu^2\eta'\zeta &= \mathbb{E}\left[Q\xi_\star\right] & (2.124)\\
w(1+u^2\eta'\zeta) &= \mathbb{E}\left[Z_0\xi_\star\right] & (2.125)\\
\nabla_{\boldsymbol{\sigma}}\rho(T|\xi_\star,\boldsymbol{\sigma}) &= \mathbf{0} & (2.126)
\end{aligned}$$

Some additional algebraic steps then result in equations (2.21, 2.22, 2.23) in the main text.

2.6.4 Logit regression model

Working out the RS equations

The Logit regression model is defined by

$$T|\mathbf{X} \sim \frac{\mathrm{e}^{T(\mathbf{X}\cdot\boldsymbol{\beta})}}{2\cosh(\mathbf{X}\cdot\boldsymbol{\beta})} \tag{2.127}$$

where $T \in \{-1,+1\}$. The log-likelihood is $\log p(T|\mathbf{X}\cdot\boldsymbol{\beta}) = T(\mathbf{X}\cdot\boldsymbol{\beta}) - \log[2\cosh(\mathbf{X}\cdot\boldsymbol{\beta})]$, so the first RS equation (2.34) follows by substitution into inside formula (2.21) of

$$\xi_\star = \nu Q + \omega Z_0 + \mu^2 T - \mu^2 \tanh(x) \tag{2.128}$$

This gives

$$\frac{\nu^2\zeta}{(1-\tau'\zeta\mu^2)^2} = \mu^4 \; \mathbb{E}_{Y,Q,Z_0}\left[\left(T - \tau\frac{\nu Q + \omega Z_0}{1-\tau\mu^2} - \tanh(x_\star)\right)^2\right] \; . \tag{2.129}$$

For equation (2.35) we need the derivative $\partial\xi_\star/\partial Q = \partial x_\star/\partial Q$, which is obtained by differentiating the equation that defines $x_\star$:

$$\frac{\partial\xi_\star}{\partial Q} = \nu - \frac{\mu^2}{\cosh^2(x_\star)}\frac{\partial\xi_\star}{\partial Q} \tag{2.130}$$

with solution

$$\frac{\partial \xi_\star}{\partial Q} = \frac{\nu \cosh^2(x_\star)}{\mu^2 + \cosh^2(x_\star)} \ . \tag{2.131}$$

Substitution into (2.22), followed by division by the common factor ν, leads after some simple rewriting to

$$\zeta\Big(1 - \eta'\mu^2/(1 - \tau'\zeta\mu^2)\Big) = \mathbb{E}_{T,Q,Z_0}\Big[\frac{u^2}{u^2 + \cosh^2(x_\star)}\Big] \ . \tag{2.132}$$

Next we compute

$$\frac{\partial}{\partial Z_0} \log p(T|SZ_0) = S\Big(T - \tanh(SZ_0)\Big) \tag{2.133}$$

and insert the result into

$$\frac{\zeta\omega}{1 - \tau'\zeta\mu^2} = \mathbb{E}_{T,Q,Z_0}\Big[\xi_\star \frac{\partial}{\partial Z_0} \log p(T|SZ_0)\Big] \tag{2.134}$$

to give

$$\zeta\omega/S\zeta = (1 - \tau'\zeta\mu^2)\mathbb{E}_{T,Q,Z_0}\big[(x - \phi)\big(T - \tanh(SZ_0)\big)\big] \ . \tag{2.135}$$

We note that

$$\mathbb{E}_{T,Q,Z_0}[T] = \mathbb{E}_{T,Z_0}[\tanh(SZ_0)] \tag{2.136}$$

so

$$\zeta\omega/S\zeta = (1 - \tau'\zeta\mu^2)\mathbb{E}_{T,Q,Z_0}\big[x\big(T - \tanh(SZ_0)\big)\big] \ . \tag{2.137}$$

Numerical solution of the RS equations

The numerical solution of the RS equations is obtained by means of fixed point iteration. The system (2.34,2.35,2.36) can be seen as a dynamical mapping for $\mathbf{x} = (\zeta, \nu, \omega)$:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}; \mu^2, S) \ . \tag{2.138}$$

The solution of (2.34,2.35,2.36) corresponds to a fixed point of the dynamical system, which is found by iterating the mapping above until some convergence criterion is met. In our case we chose to set the tolerance at $\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \leq 1 \cdot 10^{-10}$. We see that

$$\mathbb{E}_{T,Q,Z_0}\Big[g(T,Q,Z_0)\Big] = \mathbb{E}_{Q,Z_0}\Big[\sum_{t=-1,1} p(t|SZ_0)g(t,Q,Z_0)\Big] \tag{2.139}$$

so we just need to compute two Gaussian integrals. We computed the mapping $\mathbf{f}$ via Gauss-Hermite quadratures, at order 40 to achieve a reasonably high accuracy (since the functions appearing inside the integrals are not simple polynomials of a fixed degree).

2.6.5 Weibull Proportional Hazards model

Working out the RS equations

We first derive the RS equations for an arbitrary parametric Proportional Hazards model, for which the conditional response density given the covariates is

$$p(t|\mathbf{X} \cdot \boldsymbol{\beta}, \boldsymbol{\sigma}) = h(t|\boldsymbol{\sigma})\mathrm{e}^{\mathbf{X}^T\boldsymbol{\beta}}\mathrm{e}^{-H(t|\boldsymbol{\sigma})\mathrm{e}^{\mathbf{X}^T\boldsymbol{\beta}}} \tag{2.140}$$

with $t \geq 0$ and $H(t|\boldsymbol{\sigma}) = \int_0^t \mathrm{d}t'\ h(t'|\boldsymbol{\sigma})$. Equation (2.25) new reads

$$\mu^{-2}(\xi_\star - \nu Q - \omega Z_0) = 1 - H(T|\boldsymbol{\sigma})\mathrm{e}^{\xi_\star} \tag{2.141}$$

with the following solution, involving Lambert's W-function [8]:

$$\xi_\star = \nu Q + \omega Z_0 + \mu^2 - W\big(H(T|\boldsymbol{\sigma})\mu^2 \mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big) \tag{2.142}$$

By substituting (2.142) into (2.21) and writing (u, v, w) in terms of (μ, ν, ω) one obtains

$$\begin{aligned} &\frac{\zeta\nu^2}{(1-\tau\mu^2)^2} = \\ &= \mathbb{E}_{T,Q,Z_0}\Big[\Big(\mu^2 - W\Big(H(T|\boldsymbol{\sigma})\mu^2\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\Big) - \frac{\tau\mu^2}{1-\tau\mu^2}(\nu Q + \omega Z_0)\Big)^2\Big] \ . \end{aligned} \tag{2.143}$$

To obtain equation (2.22) we need $\partial\xi_\star/\partial Q$. This is computed by the inverse function rule

$$\frac{\partial\xi_\star}{\partial Q} = \nu\left(1 - \frac{W\big(H(T|\boldsymbol{\sigma})\mu^2\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big)}{1 + W\big(H(T|\boldsymbol{\sigma})\mu^2\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big)}\right) \ . \tag{2.144}$$

After substituting and dividing for the common factor ν we obtain

$$\begin{aligned} &\frac{1}{1-\tau'\zeta\mu^2}\left(1 - \zeta\Big(1 - \frac{\mu^2\eta'}{1-\tau'\zeta\mu^2}\Big)\right) \\ &\quad = 1 - \mathbb{E}_{T,Q,Z_0}\Big[\frac{W\big(H(T|\boldsymbol{\sigma})\mu^2\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big)}{1 + W\big(H(T|\boldsymbol{\sigma})\mu^2\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big)}\Big] \ . \end{aligned} \tag{2.145}$$

For the third RS equation (2.23) we need to compute

$$\frac{\partial}{\partial Z_0}\log p(T|Z_0, \phi_0, \sigma_0) = S - SH(T|SZ_0, \phi_0, \sigma_0)\mathrm{e}^{SZ_0} \ . \tag{2.146}$$

Substitution into (2.23) gives

$$\begin{aligned} &\frac{\omega\zeta}{1-\tau\mu^2} = \\ &\quad S\ \mathbb{E}_{T,Q,Z_0}\Big[W\big(H(T|\boldsymbol{\sigma})\mu^2\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big)\Big(H(T|SZ_0, \phi_0, \sigma_0)\mathrm{e}^{SZ_0} - 1\Big)\Big] \ . \end{aligned} \tag{2.147}$$

Note that the distribution (2.140) of $T|Z_0$ has the property that upon changing to the new variable

$$Z = H(T|\boldsymbol{\sigma}_0)\mathrm{e}^{SZ_0} \tag{2.148}$$

one will have $Z \sim Exp(1)$, $T = H^{-1}(Z\mathrm{e}^{-SZ_0}|\boldsymbol{\sigma}_0)$ and

$$H(T|\boldsymbol{\sigma}) = H(H^{-1}(Z\mathrm{e}^{-SZ_0}|\boldsymbol{\sigma}_0)|\boldsymbol{\sigma}) \ . \tag{2.149}$$

Hence the first three RS equations derived above (2.143,2.145,2.147) can be rewritten as

$$\frac{\zeta\nu^2}{(1-\tau\zeta\mu^2)^2} =$$
$$\mathbb{E}_{Z,Q,Z_0}\Big[\Big(\mu^2 - W\big(\mu^2 H(H^{-1}(Ze^{-SZ_0}|\boldsymbol{\sigma}_0)|\sigma)\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big)\Big)^2\Big] \tag{2.150}$$
$$\frac{1}{1-\tau'\zeta\mu^2}\Big(1-\zeta\Big(1-\frac{\mu^2\eta'}{1-\tau\zeta\mu^2}\Big)\Big) =$$
$$1-\mathbb{E}_{T,Q,Z_0}\left[\frac{W\big(\mu^2 H(H^{-1}(Ze^{-SZ_0}|\boldsymbol{\sigma}_0)|\sigma)\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big)}{1+W\big(\mu^2 H(H^{-1}(Ze^{-SZ_0}|\boldsymbol{\sigma}_0)|\sigma)\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big)}\right] \tag{2.151}$$
$$\frac{\omega}{1-\tau\zeta\mu^2}\zeta =$$
$$S\mathbb{E}_{T,Q,Z_0}\Big[W\big(\mu^2 H(H^{-1}(Ze^{-SZ_0}|\boldsymbol{\sigma}_0)|\sigma)\mathrm{e}^{\mu^2+\nu Q+\omega Z_0}\big)\Big)(Z-1)\Big]\ . \tag{2.152}$$

To work out the RS equations for the nuisance parameters (2.24) we must choose an explicit parametric form for the base hazard rate. Upon assuming the integrated base hazard to have the Weibull parametrization,

$$H(T|\phi,\sigma) = T^{1/\sigma}\mathrm{e}^{\phi/\sigma} \tag{2.153}$$

we obtain

$$H(H^{-1}(Ze^{-SZ_0}|\phi_0,\sigma_0)|\phi,\sigma) = Z^{\sigma_0/\sigma}\mathrm{e}^{-SZ_0\sigma_0/\sigma+(\phi-\phi_0)/\sigma}\ . \tag{2.154}$$

Substituting (2.154) into (2.143, 2.145) then directly leads to equations (2.44, 2.45) in the main text. We simplify (2.147) via integration by parts, use (2.145), and after some minor algebraic simplifications obtain (2.46). For the remaining equations one has to compute

$$\frac{\partial}{\partial\phi}\log p(T|\xi_*,\phi,\sigma) =$$
$$-\frac{1}{\sigma}-\frac{1}{\sigma\mu^2}W\big(Z^{\sigma_0/\sigma}\mu^2\mathrm{e}^{\mu^2+(\omega-S\frac{\sigma_0}{\sigma})Z_0+\nu Q+(\phi-\phi_0)/\sigma}\big) \tag{2.155}$$
$$\frac{\partial}{\partial\sigma}\log p(T|\xi_*,\phi,\sigma) =$$
$$-\frac{1}{\sigma}-\frac{1}{\sigma}\Big(\frac{\phi-\phi_0}{\sigma}+\frac{\sigma_0}{\sigma}\log Z-\frac{\sigma_0}{\sigma}SZ_0\Big)$$
$$\times\left(\frac{1}{\mu^2}W\big(Z^{\sigma_0/\sigma}\mu^2\mathrm{e}^{\mu^2+(\omega-S\frac{\sigma_0}{\sigma})Z_0+\nu Q+(\phi-\phi_0)/\sigma}\big)-1\right)\ . \tag{2.156}$$

Upon taking the expectation and setting the previous two derivatives to zero we obtain the remaining RS equations (2.47, 2.48)

$$\mu^2 = \mathbb{E}_{Z,Q,Z_0}\Big[W\Big(\mu^2\mathrm{e}^{\mu^2}Z^{\sigma_0/\sigma}\mathrm{e}^{(\omega-S\frac{\sigma_0}{\sigma})Z_0+\nu Q}\mathrm{e}^{(\phi-\phi_0)/\sigma}\Big)\Big] \tag{2.157}$$
$$\frac{\sigma}{\sigma_0} = \tag{2.158}$$
$$\mathbb{E}_{Z,Q,Z_0}\Big[\Big(\log Z - SZ_0\Big)\Big(\frac{1}{\mu^2}W\Big(\mu^2\mathrm{e}^{\mu^2}Z^{\sigma_0/\sigma}\mathrm{e}^{(\omega-S\frac{\sigma_0}{\sigma})Z_0+\nu Q}\mathrm{e}^{(\phi-\phi_0)/\sigma}\Big)-1\Big)\Big]\ .$$

Numerical solution of the RS equations

The numerical solution of the RS equations is again obtained via fixed point iteration. The system (2.44,2.45,2.46,2.47,2.48) is interpreted as the fixed-point condition of a dynamical mapping of $\mathbf{x} = (\zeta, \nu, \omega, \mu, \sigma/\sigma_0)$:

$$\mathbf{x}_t = \mathbf{f}(\mathbf{x}_{t-1}; (\phi - \phi_0)/\sigma, S) \ . \tag{2.159}$$

The mapping is iterated until some convergence criterion is met, here chosen to be $\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \leq 1 \cdot 10^{-10}$. We see that all the RS equations involve terms of the form

$$\mathbb{E}_{Z,Q,Z_0}\Big[g(Z, Q, Z_0)\Big] \tag{2.160}$$

involving two Gaussian and one exponential integral. We computed these via Gauss-Hermite and Gauss-Laguerre quadratures, to order 40 to achieve sufficient accuracy.

Chapter 3

Replica analysis of overfitting in regression models for time to event data: the impact of censoring

This chapter is published in Journal of Physics A : Mathematical and Theoretical, `https://iopscience.iop.org/article/10.1088/1751-8121/ad2e40/meta`. The notation in this chapter is different from the notation in the introduction. The dot product between two vectors $\boldsymbol{a}, \boldsymbol{v} \in \mathbb{R}^d$ is indicated with a $\cdot$, i.e $\boldsymbol{a} \cdot \boldsymbol{v} = \sum_{k=1}^{d} a_k \, v_k$, instead of $'$. Furthermore the covariates are indicated with the bold letter $\mathbf{z}$, instead of $\mathbf{X}$.

3.1 Abstract

We use statistical mechanics techniques, viz. the replica method, to model the effect of censoring on overfitting in Cox's proportional hazards model, the dominant regression method for time-to-event data. In the overfitting regime, Maximum Likelihood (ML) parameter estimators are known to be biased already for small values of the ratio of the number of covariates over the number of samples. The inclusion of censoring was avoided in previous overfitting analyses for mathematical convenience, but is vital to make any theory applicable to real-world medical data, where censoring is ubiquitous. Upon constructing efficient algorithms for solving the new (and more complex) Replica Symmetric (RS) equations and comparing the solutions with numerical simulation data, we find excellent agreement, even for large censoring rates. We then address the practical problem of using the theory to correct the biased ML estimators without knowledge of the data-generating distribution. This is achieved via a novel numerical algorithm that self-consistently approximates all relevant parameters of the data generating distribution while simultaneously solving the RS equations. We investigate numerically the statistics of the corrected estimators, and show that the proposed new algorithm indeed succeeds in removing the bias of the ML estimators, for both the association parameters and for the cumulative hazard.

3.2 Introduction

Inference relies on the foundations provided by classical statistical theory, which was developed for use in settings where the number p of covariates is small compared to the number n of observations [21, 10]. The standard Maximum Likelihood (ML) method for estimating model

parameters indeed fails in the high-dimensional regime where $p = \mathcal{O}(n)$ [7, 6, 2, 18, 19, 11], due to overfitting. Overfitting is the phenomenon that data noise is misinterpreted as signal, leading to biased parameter estimators with large sample to sample fluctuations. An overfitting model will predict outcomes for training examples well, but will fail in predicting outcomes for new data. Early strategies to mitigate overfitting include leaving out covariates (with the risk of overlooking relevant predictive information), penalization (equivalent to adding a parameter prior in Bayesian language), and shrinking parameter estimates after model fitting. The penalization weight or shrinking factor are typically estimated via bootstrapping or cross-validation, which forces one to sacrifice some of the training data; this makes the overfitting even worse. Moreover, penalization and shrinking may at best repair the overfitting-induced bias in the length of the parameter vector, but not the bias in its direction (which will appear as soon as the covariates are correlated [3]).

In order to use existing regression models also in the overfitting regime, it is vital to correctly quantify and undo the effects of overfitting, both for inference and for prediction purposes. As a consequence, in recent years we have seen increased research efforts aimed at extending the classical inference theory to the so-called proportional asymptotic regime, characterized by taking the limit $n \to \infty$ while keeping the ratio $\zeta = p/n$ fixed. Several mathematical methods from the domains of the statistical physics of disordered system [2, 11, 3, 13], computer science [19, 22] and statistics [7, 6, 18], have by now been applied successfully in this latter regime to model the statistics of ML and Maximum A Posteriori Probability (MAP) estimators.

The condition $p \ll n$ for ML/MAP inference to be used safely is especially problematic in post-genome medicine: we can now routinely measure very many variables per patient and want to use these to develop personalized therapies, but overfitting prevents us from doing so. For time-to-event data, the most common type in medicine, the main regression tools are variations on the model of Cox [5]. Overfitting in this model was analysed in [3, 16] as a statistical mechanics problem, via the replica method [15]. A later study [22] used the Convex Gaussian Min-max theorem [20], with the logarithm of Cox's partial likelihood as utility function, to study overfitting in the association parameters. The latter route avoids the use of replicas, but unlike [2, 16], cannot model overfitting in the base hazard rate.

Unlike some applied statistical studies on the subject, see e.g. [14, 1], all theoretical studies of overfitting based on the replica or cavity approach have so far assumed for simplicity that the data were not subject to censoring. Censoring means that samples are lost to observation prior to events occurring. It is ubiquitous in medicine, since medical studies are always of finite duration and also since patients often fail to return to hospital appointments for unknown reasons. Before the modern overfitting analysis methods for the proportional asymptotic regime, and their corresponding overfitting bias decontamination formulae, can be applied in the real world, it is hence vital that the theory of [2, 16] is extended to include censoring. That is the first aim of this paper.

We generalise the theory of[2, 16] by allowing for arbitrary types of non-informative censoring, carry out a replica analysis in the regime $p, n \to \infty$ with fixed $\zeta = p/N$, and derive the extended RS equations. Methodologically we also improve upon the approach in [2, 16] by: (i) using an more compact form of the replica approach, (ii) avoiding the variational approximation for the base hazard rate, and (iii) constructing novel and more powerful numerical algorithms for solving and inverting the RS equations, including in cases where one does not know anything about the data generating distribution, resulting in precise algorithms for decontaminating estimators of association parameters and the integrated base hazard for overfitting-induced bias.

This paper is organized as follows. In Section 3.3 we define our notation and describe briefly the Cox model. We explain the results obtained via the replica method (whose derivation is relegated to an Appendix) and the physical interpretation of the RS order parameters in Section 3.4, and show in Section 3.5 how this interpretation inspires an efficient numerical method for solving the RS equations. In Section 3.6 we test the predictions of the theory against numerical simulations, and find perfect agreement. We construct in Section 3.7 a new algorithm for simultaneously inferring relevant characteristics of the data generating distribution and solving the RS equations, leading to a realistic and practical tool for correcting the biased ML estimates of association parameters and the nuisance parameters (i.e. the integrated base hazard) in the overfitting regime, even in the presence of censoring. We conclude in Section 3.8 with a discussion of our results and future research.

3.3 Definitions

In time to event data, each observation ideally reports the time T at which the subject experiences the event under investigation and the covariate vector $\mathbf{z} = (z_1, \ldots, z_p) \in \mathbb{R}^p$, i.e. the list of characteristics of the subject measured at time zero. All covariate vectors are assumed to have been drawn independently from some distribution $p(\mathbf{z})$. Since any non-zero average will drop out of our formulae, we can always take $p(\mathbf{z})$ to have average zero. However, in virtually all real-world time-to-event data sets observations are censored, i.e. for some subjects we have a missing or partial observation of their event times. For instance, we might only know that an event occurred after or before a certain time point, or within a specific interval (called right, left and interval censoring, respectively). In what follows we focus on right censoring, the most common form of censoring in medical applications. In practice this amounts to assuming that when collecting data we actually observe

$$t = \min\{T, C\}. \tag{3.1}$$

The random variable C models the censoring time, drawn randomly from an a priori unknown distribution $p_c(x)$, and we are given a binary variable that indicates whether the subject has at time t experienced the event ($\Delta = 1$) or has been censored ($\Delta = 0$):

$$\Delta = \begin{cases} 1 & \text{if } T < C \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

Note that we can always write $p_c(x) = \lambda_c(x)\mathrm{e}^{-\Lambda_c(x)}$, where $\lambda_c(x)$ is the censoring rate and $\Lambda_c(x) = \int_0^x \mathrm{d}x'\ \lambda_c(x')$. The Cox semi-parametric model [5] assumes that the time at which a subject experiences the event is distributed according to

$$p_T(x|\mathbf{z}) = \lambda_0(x)\mathrm{e}^{\boldsymbol{\beta}_0\cdot\mathbf{z} - \Lambda_0(x)\exp(\boldsymbol{\beta}_0\cdot\mathbf{z})}, \tag{3.3}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, $\lambda_0(x)$ is the true (non-negative) hazard rate, and $\Lambda_0(x) = \int_0^x \mathrm{d}x'\ \lambda_0(x')$ is the true cumulative hazard. The censoring times C and the event times T are assumed to be statistically independent, i.e. censoring time is said to be non-informative, and the joint distribution of observed times $t \geq 0$ and event type indicators $\Delta \in \{0, 1\}$, given covariates $\mathbf{z}$, can then be written as

$$p(t, \Delta|\boldsymbol{\beta}_0\cdot\mathbf{z}) = \big(\lambda_0(t)\mathrm{e}^{\boldsymbol{\beta}_0\cdot\mathbf{z}}\big)^{\Delta}\big(\lambda_c(t)\big)^{1-\Delta}\mathrm{e}^{-\Lambda_0(t)\exp(\boldsymbol{\beta}_0\cdot\mathbf{z})-\Lambda_c(t)}\ . \tag{3.4}$$

The parameters one seeks to infer are the true vector $\boldsymbol{\beta}_0$ and the true function $\lambda_0(t)$. It follows from (3.4) upon computing the log-likelihood density for a data-set of i.i.d. observations $\mathcal{D} = \{(t_i, \Delta_i, \mathbf{z}_i)\}_{i=1}^n$, that ML inference will give the estimators

$$(\hat{\boldsymbol{\beta}}, \hat{\lambda})_n = \text{argmax}_{\boldsymbol{\beta},\lambda}\, L_n\big(\boldsymbol{\beta}, \lambda|\mathcal{D}\big) \tag{3.5}$$

$$L_n\big(\boldsymbol{\beta}, \lambda|\mathcal{D}\big) = \sum_{i=1}^n \Big\{\Delta_i\big(\log\lambda(t_i) + \boldsymbol{\beta}\cdot\mathbf{z}_i\big) - \Lambda(t_i)\mathrm{e}^{\boldsymbol{\beta}\cdot\mathbf{z}_i}\Big\}\,. \tag{3.6}$$

The problem (3.5) can be reduced by first maximizing over the hazard rate. Setting the functional derivative with respect to λ to zero gives an equation that can be solved for $\lambda(t)$, leading to the so-called Breslow estimator [5]:

$$\hat{\lambda}_n(t|\boldsymbol{\beta}) = \frac{1}{n}\sum_{k=1}^n \frac{\Delta_k\delta(t-t_k)}{n^{-1}\sum_{j=1}^n\theta(t_j-t_k)\mathrm{e}^{\boldsymbol{\beta}\cdot\mathbf{z}_j}}\,, \tag{3.7}$$

where $\theta(x)$ denotes the step-function ($\theta(x>0)=1$ and $\theta(x<0)=0$). Substituting (3.7) into (3.6) and disregarding terms that are independ of $\boldsymbol{\beta}$ we obtain

$$\hat{\boldsymbol{\beta}}_n = \text{argmax}_{\boldsymbol{\beta}}\, L_n(\boldsymbol{\beta}|\mathcal{D}) \tag{3.8}$$

$$L_n(\boldsymbol{\beta}|\mathcal{D}) = \sum_{i=1}^n \Delta_i\Big\{\boldsymbol{\beta}\cdot\mathbf{z}_i - \log\Big(\frac{1}{n}\sum_{j=1}^n\theta(t_j-t_i)\mathrm{e}^{\boldsymbol{\beta}\cdot\mathbf{z}_j}\Big)\Big\}\,. \tag{3.9}$$

The right-hand side of (3.9) is the logarithm of the famous Cox partial likelihood [5].

3.4 Typical behaviour of the Maximum (Partial) Likelihood estimator

Computing the estimator $\hat{\boldsymbol{\beta}}_{\rm ML}$ in (3.8) is equivalent to finding the ground state of a fictitious physical system with Hamiltonian $\mathcal{H}_n(\boldsymbol{\beta}|\mathcal{D}) = -L_n(\boldsymbol{\beta}|\mathcal{D})$, in which the association parameters $\boldsymbol{\beta}$ act as the degrees of freedom, and the data-set $\mathcal{D}$ plays the role of quenched disorder. Statistical physics thus allows us to compute the average over the disorder (i.e. the data $\mathcal{D}$) of the log-partial likelihood, giving the typical behaviour of the ML estimator, in the proportional asymptotic regime as

$$\begin{aligned}\lim_{n,p\to\infty,\ \zeta=p/n}\Big\langle\frac{1}{n}L_n(\hat{\boldsymbol{\beta}}_{\rm ML}|\mathcal{D})\Big\rangle_{\mathcal{D}} &= -\lim_{n,p\to\infty\ \zeta=p/n}\frac{1}{n}\Big\langle\min_{\boldsymbol{\beta}}\,\mathcal{H}_n(\boldsymbol{\beta}|\mathcal{D})\Big\rangle_{\mathcal{D}}\\ &= \lim_{n,p\to\infty,\ \zeta=p/n}\lim_{\gamma\to\infty}\frac{1}{n\gamma}\Big\langle\log\int\mathrm{d}\boldsymbol{\beta}\;\mathrm{e}^{-\gamma\mathcal{H}_n(\boldsymbol{\beta}|\mathbf{T},\mathbf{z})}\Big\rangle_{\mathcal{D}}\,.\end{aligned} \tag{3.10}$$

In fact, for reasons that will become clear, we insert a constant regularization factor (whose impact will be zero due to the limit $\gamma\to\infty$), and replace (3.10) by

$$\begin{aligned}&\lim_{n,p\to\infty,\ \zeta=p/n}\Big\langle\frac{1}{n}L_n(\hat{\boldsymbol{\beta}}_{\rm ML}|\mathcal{D})\Big\rangle_{\mathcal{D}} =\\ &\qquad\lim_{n,p\to\infty,\ \zeta=p/n}\lim_{\gamma\to\infty}\frac{1}{n\gamma}\Big\langle\log\int\mathrm{d}\boldsymbol{\beta}\;\mathrm{e}^{\frac{1}{2}p\log p-\gamma\mathcal{H}_n(\boldsymbol{\beta}|\mathbf{T},\mathbf{z})}\Big\rangle_{\mathcal{D}}\,.\end{aligned} \tag{3.11}$$

The standard procedure to carry out the disorder average in parametric regression models via the replica method (which goes back to [9]) builds on the assumption that the log-likelihood

is a sum of n independent terms, each depending on a single observation $(t_i, \Delta_i, \mathbf{z}_i)$. For the present Hamiltonian $\mathcal{H}_n(\boldsymbol{\beta}|\mathcal{D})$ this is not the case, so we need an intermediate step. We introduce the empirical distribution

$$P_n(t, \Delta, h|\boldsymbol{\beta}, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \delta(t - t_i)\delta_{\Delta,\Delta_i}\delta(h - \boldsymbol{\beta}\cdot\mathbf{z}_i) \tag{3.12}$$

and observe that we can write the Hamiltonian (i.e. minus the log-partial likelihood) as the following functional:

$$\mathcal{H}\Big[P_n(.|\boldsymbol{\beta}, \mathcal{D})\Big] = n\mathcal{E}\Big[P_n(.|\boldsymbol{\beta}, \mathcal{D})\Big], \tag{3.13}$$

$$\mathcal{E}\Big[P(.)\Big] = \tag{3.14}$$

$$\int \mathrm{d}t\mathrm{d}h\ P(t, 1, h)\left[\log\sum_{\Delta'}\int \mathrm{d}h'\mathrm{d}t'\ \theta(t' - t)\mathrm{e}^{h'}P(t', \Delta', h') - h\right].$$

We can now write (3.11) in a form where the disorder average effectively is computed over the density of states associated with (3.12). Upon introducing suitable delta-functionals one then achieves factorisation of the disorder average over the samples i in the data set $\mathcal{D}$. The full replica derivation is given in 3.9.1 for completeness, and gives upon making the so-called replica-symmetric (RS) ansatz:

$$\lim_{\substack{n,p\to\infty\\ \zeta=p/n}} \left\langle \frac{1}{n}\mathcal{H}_n(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}|\mathcal{D})\right\rangle_{\mathcal{D}} = \operatorname*{extr}_{u,v,w} \mathcal{F}(u, v, w) = \mathcal{F}(u_\star, v_\star, w_\star) \tag{3.15}$$

with

$$\mathcal{F}(u_\star, v_\star, w_\star) =$$

$$= \int \mathrm{D}y\mathrm{D}z \sum_{\Delta}\int \mathrm{d}t\ p(t, \Delta|Sy)\Big[\Lambda(t)\mathrm{e}^{\xi_\star(t,\Delta,y,z)} - \Delta\xi_\star(t, \Delta, y, z)\Big], \tag{3.16}$$

$$p(t, \Delta|Sy) = \big(\mathrm{e}^{Sy}\lambda_0(t)\big)^{\Delta}\lambda_c^{1-\Delta}(t)\mathrm{e}^{-\Lambda_0(t)\exp(Sy)-\Lambda_c(t)}, \tag{3.17}$$

$$\xi(t, \Delta, y, z) = wy + vz + u^2\Delta - W\big(u^2\mathrm{e}^{u^2\Delta+wy+vz}\Lambda(t)\big), \tag{3.18}$$

and

$$\Lambda(s) = \int \mathrm{D}y \sum_{\Delta\in\{0,1\}} \times$$

$$\times \int \mathrm{d}t\ p(t, \Delta|Sy)\left[\frac{\Delta\theta(s - t)}{\int \mathrm{D}y'\mathrm{D}z' \sum_{\Delta'}\int_t^\infty \mathrm{d}t'\ p(t', \Delta'|Sy')\mathrm{e}^{\xi_\star(t',\Delta',y',z')}}\right].$$

Here we used the standard short-hand $\mathrm{D}y = (2\pi)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}y^2}\mathrm{d}y$, and $W(x)$ is the Lambert W-function [4], defined by the equation $W(x)\exp[W(x)] = x$ for all x. The only condition on $p(\mathbf{z})$ needed in the derivation of the above RS equations is that for $p \to \infty$ the true and inferred linear predictors $\boldsymbol{\beta}\cdot\mathbf{z}$ acquire Gaussian statistics[1]. The extremum $(u_\star, v_\star, w_\star)$ in

[1]This will be trivially true if the covariate distribution $p(\mathbf{z})$ is itself Gaussian, but also if the conditions for the central limit theorem to hold apply (i.e. if the components of $\mathbf{z}$ not too strongly correlated, and the components of $\boldsymbol{\beta}$ are not too dissimilar in their scaling with n; see [8]).

(3.15) is found to satisfy the so-called RS equations:

$$\zeta v = \int \mathrm{D}y\mathrm{D}z\ z \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|Sy)\ W(u^2\Lambda(t)\mathrm{e}^{\Delta u^2+vz+wy}), \tag{3.19}$$

$$0 = \int \mathrm{D}y\mathrm{D}z\ y \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|Sy)\ \Big[W(u^2\Lambda(t)\mathrm{e}^{\Delta u^2+vz+wy}) - \Delta u^2\Big], \tag{3.20}$$

$$\zeta v^2 = \int \mathrm{D}y\mathrm{D}z \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|Sy)\Big[W(u^2\Lambda(t)\mathrm{e}^{\Delta u^2+vz+wy}) - \Delta u^2\Big]^2. \tag{3.21}$$

These can be written in alternative forms, for instance by using identities such as $W(u^2\Lambda(t)\mathrm{e}^{\Delta u^2+vz+wy}) = u^2\Lambda(t)\mathrm{e}^{\xi(t,\Delta,y,z)} = \Delta u^2 + vz + wy - \xi(t,\Delta,y,z)$, $W'(x) = W(x)/x[1+W(x)]$, and via integration by parts over z in (3.19), giving

$$\zeta v^2 = \int \mathrm{D}y\mathrm{D}z \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|y)\Big[\xi(t,\Delta,y,z) - vz - wy\Big]^2, \tag{3.22}$$

$$1-\zeta = \int \mathrm{D}y\mathrm{D}z \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|y)\Big[\frac{1}{1+u^2\Lambda(t)\mathrm{e}^{\xi(t,\Delta,y,z)}}\Big], \tag{3.23}$$

$$w = \int \mathrm{D}y\mathrm{D}z \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|y)\ y\xi(t,\Delta,y,z). \tag{3.24}$$

For the distribution $\mathcal{P}(t,\Delta,h)$ and the functions $\Lambda(t)$ and $\mathcal{S}(t)$ we find in RS ansatz the following expressions (see 3.9.1):

$$\mathcal{P}(t,\Delta,h) = \int \mathrm{D}y\mathrm{D}z\ p(t,\Delta|Sy)\ \delta\big[h - \xi(t,\Delta,y,z)\big], \tag{3.25}$$

$$\Lambda(t) = \int \mathrm{d}t'\mathrm{d}h'\ \frac{\theta(t-t')\mathcal{P}(t',1,h')}{\mathcal{S}(t')}, \tag{3.26}$$

$$\mathcal{S}(t) = \sum_{\Delta'} \int \mathrm{d}t'\mathrm{d}h'\ \theta(t'-t)\mathrm{e}^{h'}\mathcal{P}(t',\Delta',h'). \tag{3.27}$$

We note that an alternative way to derive equations (3.19,3.20,3.21) (and a corresponding equation for $\Lambda(t)$) would have been to make appropriate choices such as $y \to (t,\Delta)$ and $\theta \to \{\lambda_0(t), \lambda_C(t)\}$ in the RS formulae of [3]. However, that alternative route would not have generated the powerful expression (3.25) found with the present approach. By retracing its derivation and making simple adaptations, (3.25) can be generalized to

$$\mathcal{P}(t,\Delta,h_0,h) = \frac{\mathrm{e}^{-\frac{1}{2}h_0^2/S^2}}{S\sqrt{2\pi}} p(t,\Delta|h_0) \int \mathrm{D}z\ \delta\big[h - \xi(t,\Delta,y,z)\big], \tag{3.28}$$

where

$$\mathcal{P}(t,\Delta,h_0,h) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Big\langle \delta(t-t_i)\delta_{\Delta,\Delta_i}\delta(h_0 - \boldsymbol{\beta}_0\cdot\mathbf{z}_i)\delta(h - \boldsymbol{\beta}\cdot\mathbf{z}_i)\Big\rangle_{\mathcal{D}} \tag{3.29}$$

(with the standard convention that $\zeta = p/n$ is kept fixed in the limits $n, p \to \infty$). This result shows very transparently the impact of overfitting on the inferred risk factors $\boldsymbol{\beta}\cdot\mathbf{z}_i$,

relative to their true values $\boldsymbol{\beta}_0 \cdot \mathbf{z}_i$.

At this point it is helpful to write $w = S\kappa$. The interpretation of the values $(\kappa_\star, v_\star)$ as solved from the RS equations above can be inferred directly from the results in [3]:

$$\kappa_\star = \lim_{p,n\to\infty} \left\langle \hat{\kappa}_n \right\rangle_{\mathcal{D}}, \qquad \hat{\kappa}_n = \frac{\boldsymbol{\beta}_0 \cdot \boldsymbol{A}\hat{\boldsymbol{\beta}}_n}{\boldsymbol{\beta}_0 \cdot \boldsymbol{A}\boldsymbol{\beta}_0} \tag{3.30}$$

$$v_\star^2 = \lim_{p,n\to\infty} \left\langle \hat{v}_n^2 \right\rangle_{\mathcal{D}}, \qquad \hat{v}_n^2 = \hat{\boldsymbol{\beta}}_n \cdot \boldsymbol{A}\hat{\boldsymbol{\beta}}_n - \hat{\kappa}_n^2 \boldsymbol{\beta}_0 \cdot \boldsymbol{A}\boldsymbol{\beta}_0 \tag{3.31}$$

with $p/n = \zeta$ fixed as $p, n \to \infty$. The importance of $(\kappa_\star, v_\star)$ derives from the facts that (asymptotically) the asymptotic distribution of $\hat{\boldsymbol{\beta}}_n$ depends only these two quantities [13], and that the overfitting induced inference bias and noise can be expressed in terms of $\kappa_\star$ and $v_\star$, respectively. The function $\Lambda(t)$ has the following interpretation:

$$\Lambda(t) = \lim_{p,n\to\infty} \int_0^t \mathrm{d}t' \left\langle \hat{\lambda}_{\mathrm{ML}}(t') \right\rangle_{\mathcal{D}} \tag{3.32}$$

with $p/n = \zeta$ fixed, and with the Breslow estimator as given in (3.7).

3.5 Numerical solution of RS equations

For fully parametric GLM models in the overfitting regime the RS equations can always be solved numerically in a relatively straightforward iterative manner by evaluating the relevant integrals via Gaussian quadratures. Here, for the Cox model with censoring, we have the added complication of the functional order parameter $\Lambda(t)$, which suggests we take a different approach inspired by population dynamics algorithms and our interpretation (3.25,3.26,3.27). For any given values of the scalar order parameters (u, v, w), and given values of $S = |\boldsymbol{A}^{1/2}\boldsymbol{\beta}_0|$, we can generate $m \gg 1$ samples from the distribution

$$\mathcal{P}(t, \Delta, y, z|u, v, w) \;=\; (2\pi)^{-1} \mathrm{e}^{-\frac{1}{2}(y^2+z^2)} p(t, \Delta|Sy), \tag{3.33}$$

resulting in $\{(t_\ell, \Delta_\ell, y_\ell, z_\ell)\}_{\ell=1}^m$. We then define for each (u, v, w, Λ) the quantities $\xi_\ell(u, v, w, \Lambda) = \xi(t_\ell, \Delta_\ell, y_\ell, z_\ell|u, v, w, \Lambda)$, computed via (3.18), and the estimator $\tilde{\Lambda}_m(t)$:

$$\tilde{\Lambda}_m(t|u, v, w, \Lambda) \;=\; \sum_{\ell=1}^m \frac{\theta(t - t_\ell)\Delta_\ell}{\sum_{k=1}^m \theta(t_k - t_\ell)\mathrm{e}^{\xi_k(u,v,w,\Lambda)}} \tag{3.34}$$

For sufficiently large population size m, the iterative algorithm defined by the mapping $\Lambda(.) \to \Lambda'(.) = \tilde{\Lambda}_m(.|u, v, w, \Lambda)$ will now by construction have as its fixed-point the solution of (3.25,3.26,3.27). Furthermore, it is relatively easy to invert numerically equation (3.23) and write u for fixed (v, w, Λ) as a function of ζ, for instance via Newton's method, so that ζ can always be chosen as the independent parameter of our RS equations (as it is always

known). We can then solve the following surrogate set of RS equations:

$$\zeta v_m^2 = \frac{1}{m}\sum_{\ell=1}^{m}\Big(\xi_\ell - v_m z_\ell - w_m y_\ell\Big)^2, \tag{3.35}$$

$$1-\zeta = \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{1+u_m^2\tilde{\Lambda}_m(t_\ell)\mathrm{e}^{\xi_\ell}}, \tag{3.36}$$

$$w_m = \frac{1}{m}\sum_{\ell=1}^{m} y_\ell\xi_\ell, \tag{3.37}$$

$$\tilde{\Lambda}_m(t_i) = \sum_{\ell=1}^{m}\frac{\theta(t_i-t_\ell)\Delta_\ell}{\sum_{j=1}^{m}\theta(t_j-t_\ell)\mathrm{e}^{\xi_j}}, \tag{3.38}$$

where now

$$\xi_i = u^2\Delta_i + v_m z_i + w_m y_i - W\big(u_m^2\tilde{\Lambda}_m(t_i)\mathrm{e}^{u_m^2\Delta_i+v_m z_i+w_m y_i}\big). \tag{3.39}$$

(e.g. via damped fixed point iteration). When simulating the data, $\Lambda_0(.)$ is known and, empirically, we find that substituting equation (3.37) with its equivalent version, obtained by Gaussian integration by parts of (3.24),

$$w_m = S\frac{1}{m}\sum_{\ell=1}^{m}\Big(u_m^2\Delta_\ell - W(u_m^2\tilde{\Lambda}_m(t_\ell)\mathrm{e}^{\Delta_\ell u_m^2+v_m z+w_m y})\Big)\Big(\Delta_\ell - \Lambda_0(t_\ell)\mathrm{e}^{Sy_\ell}\Big)/\zeta \tag{3.40}$$

leads to much faster convergence with smaller population size and hence the latter is adopted instead of (3.37).

3.6 Simulations tests of RS theory

When some observations are censored, it is known that in the proportional asymptotic regime the ML estimator $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ does not always exists [17], and a recent paper [22] established that, asymptotically and under the assumption of Gaussian covariates, $\hat{\boldsymbol{\beta}}_n$ undergoes a sharp phase transition at some critical value ζ_c. Here we focus on the region $\zeta < \zeta_c$ where the ML estimator does exist. Since one obviously cannot solve the RS equations for all possible choices of the hazard function $\lambda_0(.)$, the censoring rate $\lambda_c(.)$ and the true association amplitude $S = |\boldsymbol{A}^{1/2}\boldsymbol{\beta}_0|$, we have in our simulations chosen a censoring distribution $p_c(t)$ that is uniform between 0 and $t_{\max}$, reflecting the realistic scenario of a clinical trial of duration $t_{\max}$ where patients are recruited at constant rate for the trial duration. Different choices of S, $t_{\max}$ and true hazard rate $\lambda_0(.)$ will lead to different expected fractions $\langle\Delta\rangle$ of true (non-censored) events:

$$\begin{aligned}\langle\Delta\rangle &= \int \mathrm{D}y\int_0^\infty \mathrm{d}t\; p(t,1|Sy) = -\int \mathrm{D}y\int_0^\infty \mathrm{d}t\; \mathrm{e}^{-\Lambda_c(t)}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-\Lambda_0(t)\exp(Sy)}\\ &= 1-\int_0^{t_{\max}}\frac{\mathrm{d}t}{t_{\max}}\int \mathrm{D}y\; \mathrm{e}^{-\Lambda_0(t)\exp(Sy)}\end{aligned} \tag{3.41}$$

To test the predictions of our RS equations we generated 500 survival data-sets $\mathcal{D}$ of size $n = |\mathcal{D}| = 400$, with uniformly distributed censoring on the interval $[0, t_{\max}]$, using cumulative hazards of the Log-logistic $\Lambda_0(t) = \log(1+t^2)$ and Weibull $\Lambda_0(t) = \frac{1}{2}t^2$ form. We used uncorrelated unit-variance Gaussian covariates, and true association vector $\boldsymbol{\beta}_0 = \hat{\mathbf{e}}_1$ (i.e.

the first unit vector), hence $S = 1$. We then applied Cox's ML regression protocol to infer the associations and the cumulative hazard (via Breslow's formula), giving the estimators $\hat{\boldsymbol{\beta}}_n$ and $\hat{\Lambda}_n(.)$. From $\hat{\boldsymbol{\beta}}_n$ we subsequently computed the overfitting markers $(\hat{\kappa}_n, \hat{v}_n)$ in (3.30,3.31).

Below we show results obtained when $t_{\max} = 4$, corresponding to $\langle\Delta\rangle \approx 0.6$, implying that around 40% of the events are censored, for both choices of $\Lambda_0(.)$. This relatively large censored fraction was deliberately chosen to confirm that the new theory is capable of predicting the behaviour of the various estimators, despite the complication posed by censoring. Further evidence in the form of additional simulations for different values of $t_{\max}$ and hence of $\langle\Delta\rangle$ is presented in 3.9.3. There we considered $t_{\max} = 2$ and $t_{\max} = 6$, corresponding to $\langle\Delta\rangle \approx 40\%$ and 70%, respectively. In real medical data sets the fraction of censored events depends heavily on the kind of risk under study and on the duration of the clinical study. For example, a fast acting risk will lead to small numbers of censored individuals, since most will experience the primary event before the end of the study.

In Figure 3.1 we plot the inferred cumulative hazards $\hat{\Lambda}_n(.)$ versus the true cumulative hazards $\Lambda_0(.)$, which can be done unambiguously since both are always non-decreasing functions of time. By the nature of the Breslow estimator (3.7), these curves take the form of 'staircase' functions. We also show in the same panels the RS prediction of the relation between the two quantities, as red dashed lines. We conclude from Figure 1 that in all cases the solution of the RS equations correctly predict the typical values of $\hat{\Lambda}_n(.)$. In Figure 3.2 we compare the replica predictions $(\kappa_\star, v_\star)$ (solid lines) with the corresponding measurements $(\hat{\kappa}_n, \hat{v}_n)$ (markers with error bars), for different values of $\zeta = p/n \in [0, 0.5]$. Markers give the averages over the 500 data set realizations; error bars indicate the standard deviations. Again we observe excellent agreement between the RS theory and the simulations, in spite of the relatively small sample size $n = 400$. The latter feature is important for applications of the theory, since real data sets in survival analysis indeed often have sizes of that order.

3.7 De-biasing protocols for overfitted ML estimators

Given the agreement between theory and simulations, we now explore the potential of using our theory as a systematic tool with which to decontaminate ML estimators and build *asymptotically unbiased* estimators $\tilde{\boldsymbol{\beta}}_n$ and $\tilde{\Lambda}_n(.)$. Two obstacles appear initially to hinder direct application of our RS equations for bias decontamination. First, our equations involve the amplitude S (which is not directly accessible), Second, bias decontamination requires inverting the complex relation between the inferred and true integrated hazards. We have already constructed an algorithm for computing $\hat{\Lambda}_n(.)$ in terms of $\Lambda_0(.)$, but now we need to compute $\Lambda_0(.)$ given $\hat{\Lambda}_n(.)$. In this section we remove both obstacles, and show that the RS theory can indeed be used for creating accurate unbiased estimators in the overfitting regime.

3.7.1 Derivation of a practical algorithm for de-biasing

Given the assumptions of our theory, we may write the marginal outcome probability $p(t, \Delta) = \int \mathrm{d}\mathbf{z}\; p(\mathbf{z}) p(t, \Delta | \boldsymbol{\beta}_0 \cdot \mathbf{z})$ as

$$p(t, \Delta) \;=\; \lambda_0^{\Delta}(t) \lambda_c^{1-\Delta}(t) \mathrm{e}^{-\Lambda_c(t)} \int \mathrm{D}y\; \mathrm{e}^{\Delta S y - \Lambda_0(t) \exp(S y)}. \tag{3.42}$$

Hence the log-marginal likelihood density for n i.i.d. observations $\{(t_i, \Delta_i)\}$ reads

$$\ell_n(S, \lambda, \lambda_c) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\Delta_i \log\lambda(t_i) + (1-\Delta_i)\log\lambda_c(t_i) - \Lambda_c(t_i) + \log\phi_{\Delta_i}\big(\Lambda(t_i), S\big)\Big\}, \tag{3.43}$$

where we introduced the function

$$\phi_\Delta(x, s) = \int \mathrm{D}y\ \mathrm{e}^{\Delta s y - x\exp(sy)}. \tag{3.44}$$

We next compute the ML estimators for (λ, λ_c) by maximization of (3.43). Taking the functional derivative with respect to $\lambda(.)$ gives, after standard manipulations, the following estimator for the hazard rate:

$$\tilde\lambda(t|S) = \sum_{i=1}^{n}\frac{\Delta_i\delta(t-t_i)}{\sum_{j=1}^{n}\theta(t_j-t_i)\big[\phi_{\Delta_j+1}(\Lambda(t_j),S)/\phi_{\Delta_j}(\Lambda(t_j),S)\big]}. \tag{3.45}$$

Upon integrating both sides over t, and replacing in the right-hand side the (as yet unknown) value of $\Lambda(t_i)$ by the estimator $\tilde\Lambda_n(t|S) = \int_0^t \mathrm{d}t'\ \tilde\lambda_n(t'|S)$ one then finds the following simple fixed-point equation for $\tilde\Lambda_n(.|S)$, that does not involve any parameter that could be susceptible to overfitting:

$$\tilde\Lambda(t|S) = \sum_{i=1}^{n}\frac{\Delta_i\theta(t-t_i)}{\sum_{j=1}^{n}\theta(t_j-t_i)\big[\phi_{\Delta_j+1}(\tilde\Lambda(t_j),S)/\phi_{\Delta_j}(\tilde\Lambda(t_j),S)\big]}. \tag{3.46}$$

The result of solving this fixed-point equation by (damped) iteration is remarkably accurate, as will be shown below. Repeating the same steps for the censoring rate $\lambda_c(.)$ gives the unbiased ML estimator

$$\tilde\Lambda_c(t) = \sum_{i=1}^{n}\frac{(1-\Delta_i)\theta(t-t_i)}{\sum_{j=1}^{n}\theta(t_j-t_i)}. \tag{3.47}$$

One could in principle attempt to determine S in the same way: extremize (3.43) with respect to S, and solve the resulting equation simultaneously with (3.46) for $\tilde\Lambda(.|S)$ and S. Unfortunately, the resulting equations admit the trivial solution $S = 0$, describing the trivial situation where the outcomes (t, Δ) are independent of the covariates. Thus we need an independent extra equation to determine the value of S, which must be added to and solved simultaneously with the RS order parameter equations and (3.46). We first note that (3.30,3.31) imply

$$S^2 = \frac{1}{\kappa_\star^2}\Big(\lim_{n,p\to\infty}\big\langle\hat{\boldsymbol{\beta}}_n\cdot\boldsymbol{A}\hat{\boldsymbol{\beta}}_n\big\rangle_{\mathcal{D}} - v_\star^2\Big) \tag{3.48}$$

with $p/n = \zeta$ fixed as $p, n \to \infty$. If the covariate correlation matrix $\boldsymbol{A}$ is known, the above could serve as the desired extra equation. If the covariate correlation matrix is not known, as would generally be the case, we can use the following result, which exploits the fact that in regression the n quantities $\hat{\boldsymbol{\beta}}_{\rm ML}\cdot\mathbf{z}_i$ are always observable:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\beta}}_{\rm ML}\cdot\mathbf{z}_i)^2 \ = \ (1-\zeta)v_\star^2 + w_\star^2. \tag{3.49}$$

This identity is derived in 3.9.2. It uses explicitly the alternative form of the replica calculation followed in the present paper, from which followed equation (3.25) (a result that could have been but was not derived in earlier papers). In contrast to (3.48), (3.49) can always be used as a additional equation with which to determine S.

The final result is the following algorithm, that seeks to combine optimally the RS theory with the available data $\mathcal{D}$ and the biased ML estimators. After measuring the left-hand side of (3.49), one solves numerically the four scalar equations (3.22,3.23,3.24,3.49) simultaneously for (u, v, w, S), e.g. via (damped) fixed-point iteration, with the short-hands (3.18) for the function $\xi(t, \Delta, y, z)$ and (3.17) for $p(t, \Delta|Sy)$. The censoring rate $\lambda_c(t)$ in (3.17) is estimated by the time derivative of (3.47), and for each value of S, the base hazard $\lambda_0(t)$ by the time derive of the result of iterating (3.46) to a fixed-point. The various Gaussian integrals can be done via Gauss-Hermite quadrature. The resulting new and unbiased estimator for the association parameters is according to (3.30) then given by $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\rm ML}/\kappa_\star$, where $\kappa_\star = w_\star/S$.

3.7.2 Simulation tests of the proposed new estimators

In Figure 3.3 we plot both the uncorrected and the de-biased estimator for the cumulative hazard, for simulations of survival data with $n = 400$ and around 40% of censoring events. These data show convincingly that the ML estimator $\hat{\Lambda}_n(.)$ (shown in black) is quite biased, with average values as predicted by the RS theory (dashed yellow line), but that the corrected estimator $\tilde{\Lambda}(.)$ (shown in blue) when plotted against the true cumulative hazard is centered around the red dashed line (the diagonal, corresponding to unbiased estimation). This is remarkable, given that our de-biasing algorithm does not require any knowledge of the data generating process, apart from assuming the form of a semi-parametric proportional hazard model (3.4). As expected, the corrected estimator exhibits a larger variance than the ML estimator (reflecting the general and well-known bias-variance trade-off in inference). The increasing variance of both estimators at large times is simply a finite size effect: already at $t = 2.5$ the survival probability is below 0.1, so only few subjects are still alive and able to generate events that may contribute to hazard rate inference. Similarly we show in Figures 3.4, 3.5, and 3.6 the differences between the ML estimators and the new de-biased estimators for the amplitude S (the 'signal strength' of the true associations), the nonzero and zero components of the true association vector respectively. Additional simulations for different values of the censoring rate have been carried out, and the corresponding results can be found in 3.9.3.

Our simulations support the conclusion that the new decontamination algorithm proposed above indeed leads to virtually unbiased estimators for $\Lambda_0(.)$, S , and $\boldsymbol{\beta}_0$, even for relatively small data sets and in the presence of significant levels of censoring. This is a nontrivial result, since we have only required that there is no model mismatch, i.e. that the data were generated from a Cox model, and that the distribution of the risk factors $\boldsymbol{\beta}_0 \cdot \mathbf{z}_i$ will asymptotically be Gaussian. In particular, the algorithm does not require a priori knowledge of any of the parameters of the data generating process.

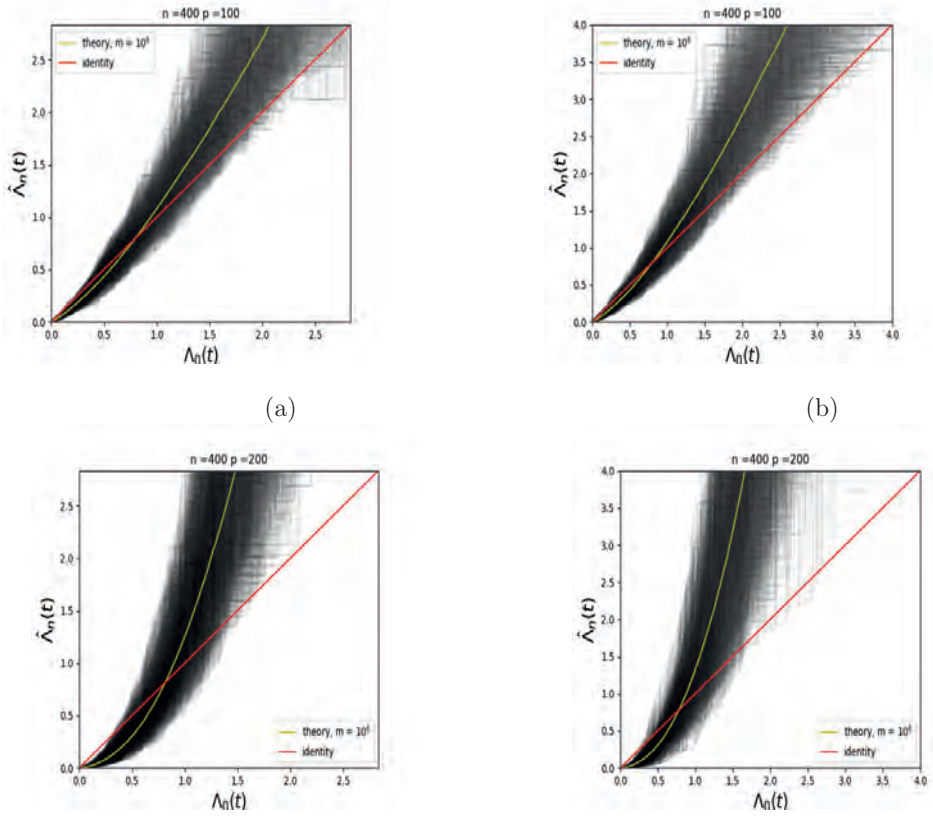


Figure 3.1: Comparison of theoretical predictions and simulation data, for Cox's survival analysis model with uniform censoring on the interval $[0, 4]$, giving around 40% censoring events, and data set size $n = 400$. Top row: $p = 100$ (so $\zeta = 0.25$); bottom row: $p = 200$ (so $\zeta = 0.5$). Panels (a,c): cumulative hazard $\Lambda_0(t) = \log(1+t^2)$. Panels (b,d): cumulative hazard $\Lambda_0(t) = \frac{1}{2}t^2$. In all panels we plot with black lines the Breslow estimator $\hat{\Lambda}_n(.)$ versus the true cumulative hazard $\Lambda_0(.)$, for 500 independent simulations with distinct data realizations. Yellow curves show the predictions of the RS theory (solved with populations of size $m = 10^5$), and red curves indicate the diagonal (that would have been found for perfect regression). Further details are given in the main text.

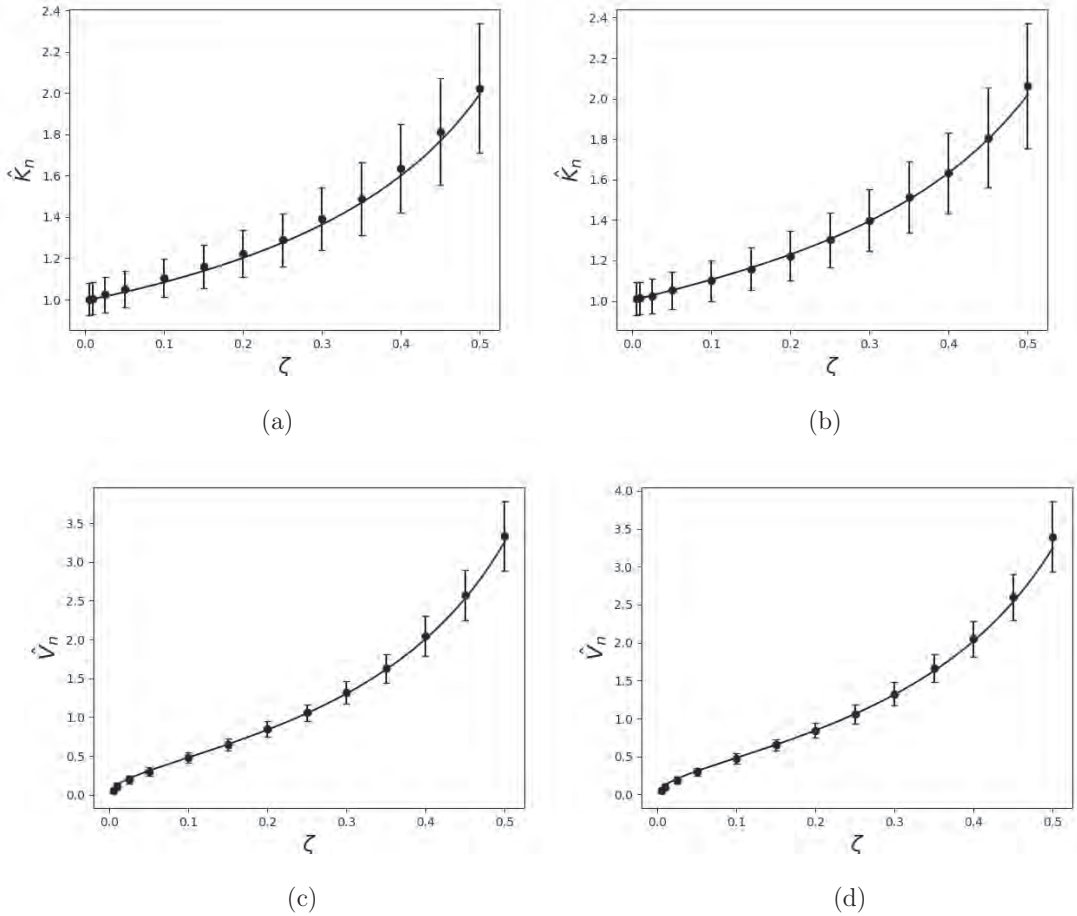


Figure 3.2: Comparison of theoretical predictions and simulation data, for Cox's survival analysis model with uniform censoring on the interval $[0, 4]$, giving around 40% censoring events, and data set size $n = 400$. Top row: overfitting-induced bias factor $\hat{\kappa}_n$ versus ζ; bottom row: overfitting-induced noise amplitude $\hat{v}_n$ versus ζ. Panels (a,c): data for cumulative hazard $\Lambda_0(t) = \log(1 + t^2)$. Panels (b,d): data for cumulative hazard $\Lambda_0(t) = \frac{1}{2}t^2$. In all panels we plot with markers and error bars the simulation results, averaged over 500 independent simulations with distinct data realizations, and with solid lines the predictions $\kappa_\star$ and $v_\star$ of the RS theory (solved with populations of size $m = 10^5$).

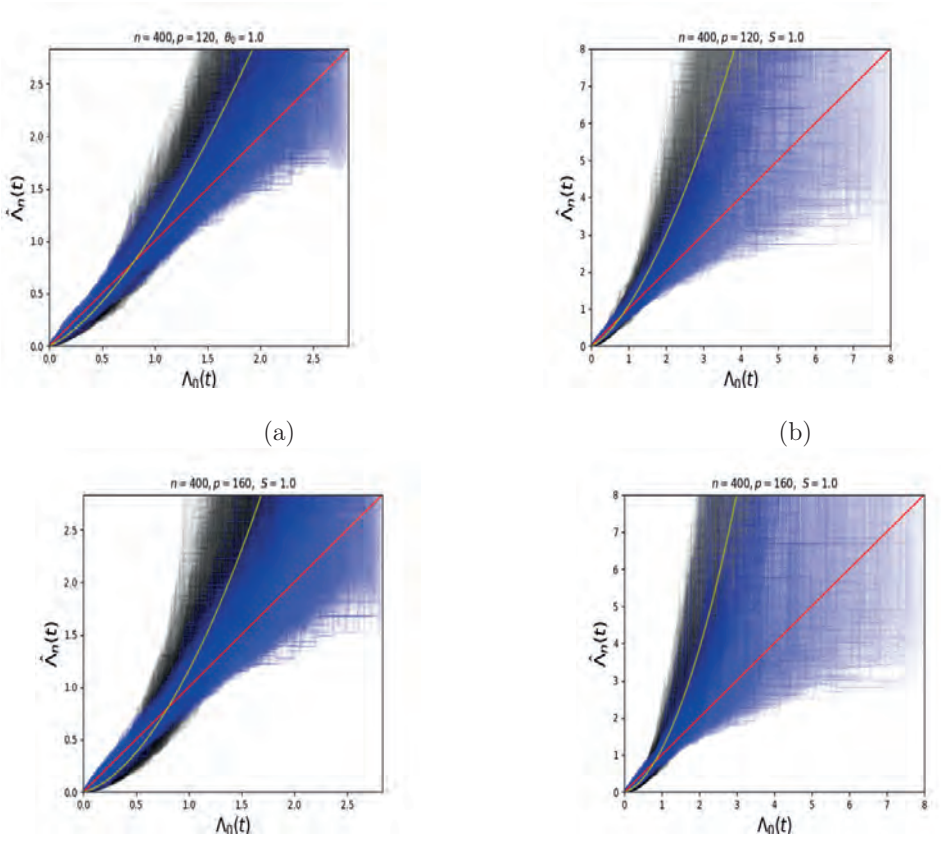
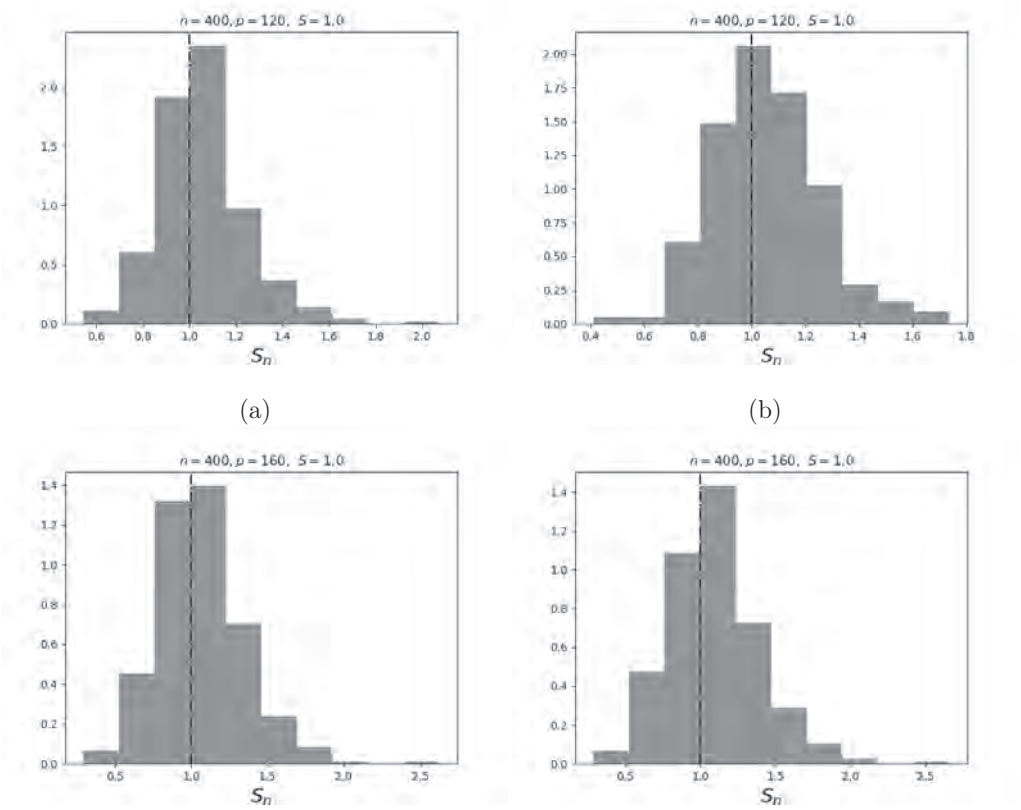


Figure 3.3: Comparison of uncorrected and corrected integrated hazard rate estimators derived from simulation data, for Cox's survival analysis model with uniform censoring on the interval $[0, 4]$, giving around 40% censoring events, and data set size $n = 400$. In all cases $\boldsymbol{\beta}_0 = \hat{\mathbf{e}}_1$, so $S = 1$. Top row: $p = 120$ (so $\zeta = 0.3$); bottom row: $p = 200$ (so $\zeta = 0.4$). Panels (a,c): cumulative hazard $\Lambda_0(t) = \log(1+t^2)$. Panels (b,d): cumulative hazard $\Lambda_0(t) = \frac{1}{2}t^2$. In all panels we plot with black lines the Breslow estimator $\hat{\Lambda}_n(.)$ versus the true cumulative hazard $\Lambda_0(.)$, for 500 independent simulations with distinct data realizations. Yellow curves show the predictions of the RS theory (solved with populations of size $m = 10^5$), and red curves indicate the diagonal (that would have been found for perfect regression). We also show in blue the *de-biased estimator* $\tilde{\Lambda}(.)$ for the cumulative hazard versus the true cumulative hazard $\Lambda_0(.)$, for the same 500 simulations.



(a) (b)

(c) (d)

Figure 3.4: Distributions of the values of the effective signal strength $S = |\boldsymbol{A}^{1/2}\boldsymbol{\beta}_0|$ inferred by our de-biasing algorithm, for the same data as those used in Figure 3. In all panels the maximum of the histogram corresponds to the true value $S = 1$, in spite of the relatively modest data set size (around 240 non-censored events).

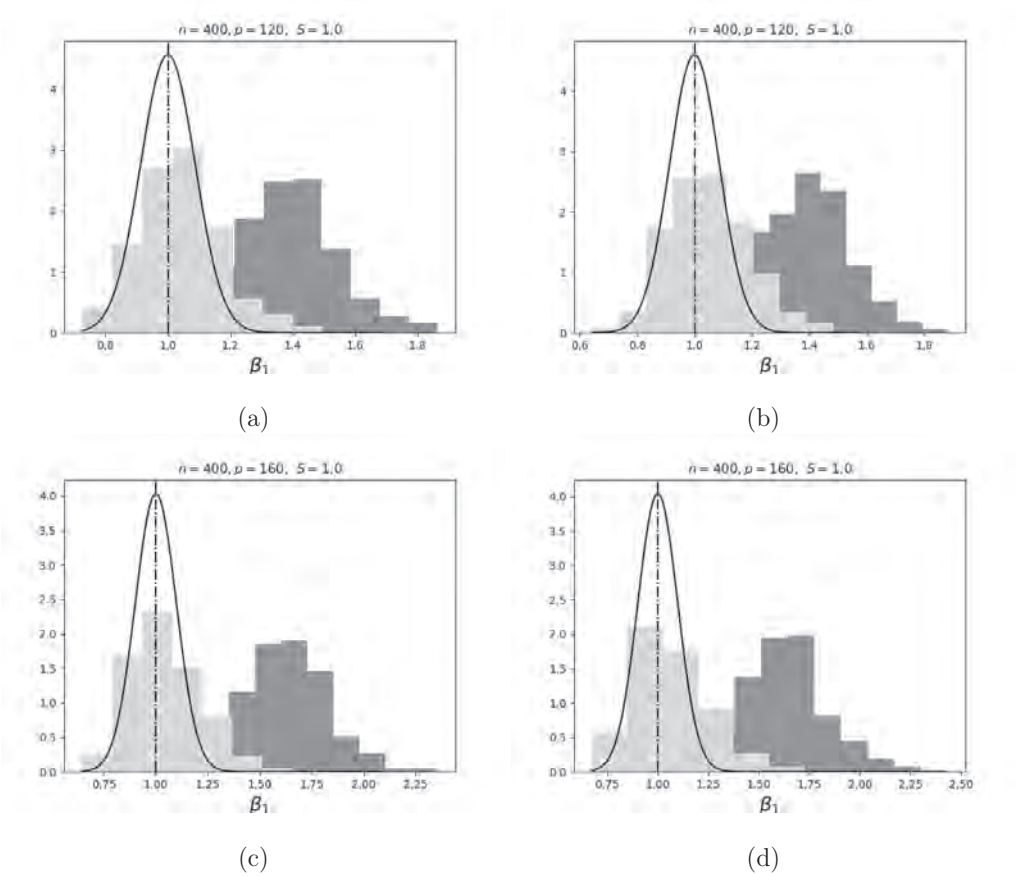


Figure 3.5: Distributions of the values of the non-zero component $\hat{\mathbf{e}}_1 \cdot \boldsymbol{\beta}$ of the association vector, both for the standard ML Cox regression estimator (dark grey) and for the new estimator inferred by our de-biasing algorithm (light grey), for the same data as those used in Figure 3. We also show as a vertical dashed line the location of the true value $\hat{\mathbf{e}}_1 \cdot \boldsymbol{\beta}_0 = 1$, and as a solid curve the predicted Gaussian asymptotic distribution of the new estimator, with average 1 and variance $v_\star^2/\kappa_\star^2 p$. We observe that the new estimator indeed removed successfully the overfitting-induced bias of the ML one, while its distribution exhibits finite size effects.

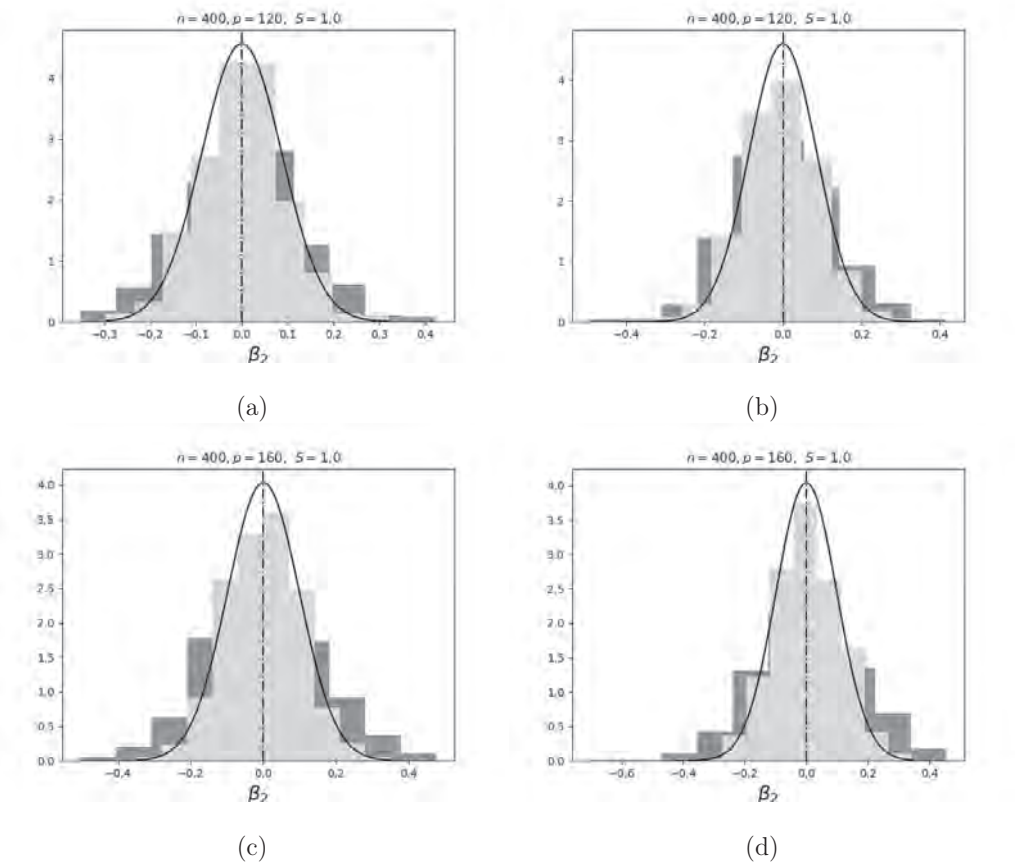


Figure 3.6: Distributions of the values of the zero component $\hat{\mathbf{e}}_2 \cdot \boldsymbol{\beta}$ of the association vector, both for the standard ML Cox regression estimator (dark grey) and for the new estimator inferred by our de-biasing algorithm (light grey), for the same data as those used in Figure 3. We also show as a vertical dashed line the location of the true value $\hat{\mathbf{e}}_2 \cdot \boldsymbol{\beta}_0 = 0$, and as a solid curve the predicted Gaussian asymptotic distribution of the new estimator, here with zero average and variance $v_\star^2/\kappa_\star^2 p$. Here, as expected, already the ML estimator was unbiased, since the theory predicts the overfitting-induced bias to take the form of a multiplicative factor $\kappa_\star$ (which would here just multiply the number zero).

3.8 Conclusion

In this paper we have extended previous analytical studies on overfitting in high dimensional Cox regression for time-to-event data [2, 3, 16], by including censoring. Censoring is a necessary ingredient for any survival analysis theory if its results are to be applied to real medical data (the main application area of survival analysis). While still using the replica method as our main tool, in addition to the inclusion of censoring we have here also chosen a different route to carry out the analysis compared to [2, 3, 16], which enabled us to derive an explicit formula for the inferred distribution of risk factors. This latter formula, in turn, paved the way for the creation of new algorithms for applications of the replica-symmetric (RS) theory, that previously required additional knowledge and/or additional approximations. All our new theoretical results and predictions are validated using numerical simulations.

We consider the main deliverables of the present paper to be the following three. First, we now have an accurate theory with which to understand quantitatively the impact of censoring on overfitting phenomena in the Cox model, so that replica-based overfitting analysis can now be applied to realistic real-world problems in medicine (i.e. to data from clinical studies of finite as opposed to infinite duration). Second, we are now able to solve directly and accurately the RS equations for the inferred cumulative hazard, instead of using the variational approximations proposed in [2, 3, 16]. Thirdly, we have been able to construct from the present theory a practical and accurate algorithm with which to decontaminate inferences for overfitting-induced bias, that does not require knowledge of *any* of the parameters of the true data generating distribution. The latter algorithm combines in a very effective way the RS equations with those quantities that are measurable in ML regression, and infers the effective signal strength $S = |\boldsymbol{A}^{1/2}\boldsymbol{\beta}_0|$ as a by-product.

The results presented here thus have not only a theoretical but also a very practical value. The availability of tools for de-biasing the important estimators of the Cox model implies that one can now achieve significantly improved outcome predictions in realistic clinical scenarios, compared to what would have been achievable with ML regression alone, either by using more covariates for prediction (i.e. increasing $\zeta = p/n$ by increasing p), or by using fewer patients to do so (i.e. increasing $\zeta = p/n$ by decreasing n), or by allowing for increased levels of censoring without detrimental consequences. To the best of our knowledge, the presently proposed bias decontamination algorithm, that corrects both association parameters and nuisance parameters (i.e. the cumulative hazard in the Cox model), is the first of its kind.

We envisage future work to involve application of the new estimator decontamination algorithm to real (partially censored) survival data, finding an expression or equation for the transition point ζ_c of the Cox model in the presence of censoring(via the statistical mechanics route), and working out the RS theory upon including ridge penalization in the Cox formulae for survival analysis with censoring. We also envisage generalizing the new decontamination algorithm to more general GLM regression models, where the efficient computation of the signal strength S, or other possible parameters of the true data generating model, and of the nuisance parameters, have in the past always required unwelcome approximations that we expect will now no longer be necessary.

Bibliography

[1] F Bertrand and M Maumy-Bertrand. Fitting and cross-validating cox models to censored big data with missing values using extensions of partial least squares regression

models. *Frontiers in big Data*, 4:684794, 2021.

[2] ACC Coolen, JE Barrett, P Paga, and CJ Perez-Vicente. Replica analysis of overfitting in regression models for time-to-event data. *Journal of physics A: mathematical and theoretical*, 50(37):375001, 2017.

[3] ACC Coolen, M Sheikh, A Mozeika, F Aguirre-Lopez, and F Antenucci. Replica analysis of overfitting in generalized linear regression models. *Journal of Physics A: Mathematical and Theoretical*, 53(36):365001, 2020.

[4] RM Corless, GH Gonnet, DEG Hare, DJ Jeffrey, and DE Knuth. On the lambert w function. *Advances in Computational mathematics*, 5:329–359, 1996.

[5] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[6] D Donoho and A Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969, 2016.

[7] N El Karoui, PJ Bean, Dand Bickel, C Lim, and B Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[8] W Feller. *An Introduction to Probability Theory and Its Applications, Volume 1.* An Introduction to Probability Theory and Its Applications. Wiley, 1968.

[9] E Gardner and B Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, jan 1988.

[10] JD Kalbfleisch and RL Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley, 2011.

[11] B Loureiro, C Gerbelot, H Cui, S Goldt, F Krzakala, M Mézard, and L Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114001, nov 2022.

[12] E Massa, M Jonker, and A Coolen. Correction of overfitting bias in regression models, 2023.

[13] E Massa, MA Jonker, and ACC Coolen. Penalization-induced shrinking without rotation in high dimensional glm regression: a cavity analysis. *Journal of Physics A: Mathematical and Theoretical*, 55(48):485002, 2022.

[14] SF McGough, D Incerti, S Lyalina, R Copping, B Narasimhan, and R Tibshirani. Penalized regression for left-truncated and right-censored survival data. *Statistics in Medicine*, 40(25):5487–5500, 2021.

[15] M Mézard, G Parisi, and MA Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[16] M Sheikh and ACC Coolen. Analysis of overfitting in the regularized cox model. *Journal of Physics A: Mathematical and Theoretical*, 52(38):384002, 2019.

[17] MJ Silvapulle and J Burridge. Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(1):100–106, 1986.

[18] P Sur and EJ Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[19] C Thrampoulidis, E Abbasi, and B Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[20] C Thrampoulidis, S Oymak, and B Hassibi. A tight version of the gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

[21] AW Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[22] X Zhang, H Zhou, and H Ye. A modern theory for high-dimensional cox regression models. *arXiv preprint arXiv:2204.01161*, 2022.

3.9 Appendix

3.9.1 Replica derivation

We start from formulae (3.12,3.13,3.14) for the Hamiltonian, being minus the partial log-likelihood of the Cox model. The relevant information on the typical estimators $(\hat{\boldsymbol{\beta}},\hat{\lambda})_{\rm ML}$ follows from the asymptotic disorder-averaged free energy density in the $\gamma\to\infty$ limit, viz. from $f=\lim_{\gamma\to\infty}f(\gamma)$, where

$$f(\gamma) = -\lim_{n\to\infty}\frac{1}{n\gamma}\Big\langle \log Z_n(\gamma,\mathcal{D})\Big\rangle_{\mathcal{D}}, \tag{3.50}$$

$$Z_n(\gamma,\mathcal{D}) = \int {\rm d}\boldsymbol{\beta}\ {\rm e}^{\frac{1}{2}p\log p-\gamma\mathcal{H}[P_n(.|\boldsymbol{\beta},\mathcal{D})]}. \tag{3.51}$$

To compute the average of the logarithm we use the replica method, giving

$$f(\gamma) = -\lim_{n\to\infty}\lim_{r\to 0}\frac{1}{n\gamma r}\log\Big\langle [Z_n(\gamma,\mathcal{D})]^r\Big\rangle_{\mathcal{D}}. \tag{3.52}$$

The replicated free energy density

In order to compute $\big\langle [Z_n(\gamma,\mathcal{D})]^r\big\rangle_{\mathcal{D}}$ we first write $Z_n(\gamma,\mathcal{D})$ in a more convenient form. We introduce the functional delta distribution

$$\delta\Big[\mathcal{P}(.)-P_n(.|\boldsymbol{\beta},\mathcal{D})\Big] = \int\mathcal{D}\hat{\mathcal{P}}\ {\rm e}^{{\rm i}n\sum_\Delta\int{\rm d}t{\rm d}h\ \hat{\mathcal{P}}(t,\Delta,h)\left[\mathcal{P}(t,\Delta,h)-P_n(t,\Delta,h|\boldsymbol{\beta},\mathcal{D})\right]} \tag{3.53}$$

in which $\mathcal{D}\hat{\mathcal{P}}$ and $\mathcal{D}\mathcal{P}$ are normalized such that $\int\mathcal{D}\mathcal{P}\ \delta\big[\mathcal{P}(.)\big]=1$. We can now write

$$\begin{aligned} Z_n(\gamma,\mathcal{D}) &= \int\mathcal{D}\hat{\mathcal{P}}\mathcal{D}\mathcal{P}\ \mathrm{e}^{n\left\{\mathrm{i}\sum_\Delta\int\mathrm{d}t\mathrm{d}h\ \hat{\mathcal{P}}(t,\Delta,h)\mathcal{P}(t,\Delta,h)-\gamma\mathcal{E}[\mathcal{P}(.)]\right\}} \\ &\quad\times\int\mathrm{d}\boldsymbol{\beta}\ \mathrm{e}^{\frac{1}{2}p\log p-\mathrm{i}\sum_{i=1}^n\hat{\mathcal{P}}(t_i,\Delta_i,\boldsymbol{\beta}\cdot\mathbf{z}_i)}. \end{aligned} \tag{3.54}$$

For integer r we can next do the average over the data $\mathcal{D}$:

$$\big\langle[Z_n(\gamma,\mathcal{D})]^r\big\rangle_{\mathcal{D}}= \tag{3.55}$$

$$\begin{aligned} &\int\Big[\prod_{\alpha=1}^r\mathcal{D}\hat{\mathcal{P}}_\alpha\mathcal{D}\mathcal{P}_\alpha\Big]\ \mathrm{e}^{n\sum_{\alpha=1}^r\left\{\mathrm{i}\sum_\Delta\int\mathrm{d}t\mathrm{d}h\ \hat{\mathcal{P}}_\alpha(t,\Delta,t)\mathcal{P}_\alpha(t,\Delta,t)-\gamma\mathcal{E}[\mathcal{P}_\alpha(.)]\right\}} \\ &\times\int\Big[\prod_{\alpha=1}^r\mathrm{d}\boldsymbol{\beta}_\alpha\Big]\mathrm{e}^{\frac{1}{2}rp\log p}\Big[\sum_\Delta\int\mathrm{d}\mathbf{z}\mathrm{d}t\ p(t,\Delta|\boldsymbol{\beta}_0\cdot\mathbf{z})p(\mathbf{z})\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^r\hat{\mathcal{P}}_\alpha(t,\Delta,\boldsymbol{\beta}_\alpha\cdot\mathbf{z})}\Big]^n \\ &=\int\Big[\prod_{\alpha=1}^r\mathcal{D}\hat{\mathcal{P}}_\alpha\mathcal{D}\mathcal{P}_\alpha\Big]\ \mathrm{e}^{n\sum_{\alpha=1}^r\left\{\mathrm{i}\sum_\Delta\int\mathrm{d}t\mathrm{d}h\ \hat{\mathcal{P}}_\alpha(t,\Delta,h)\mathcal{P}_\alpha(t,\Delta,h)-\gamma\mathcal{E}[\mathcal{P}_\alpha(.)]\right\}} \\ &\times\int\Big[\prod_{\alpha=1}^r\mathrm{d}\boldsymbol{\beta}_\alpha\Big]\mathrm{e}^{\frac{1}{2}rp\log p}\Big[\sum_\Delta\int\mathrm{d}t\mathrm{d}\mathbf{x}\ W(\mathbf{x}|\{\boldsymbol{\beta}\})p(t,\Delta|x_0)\mathrm{e}^{-\mathrm{i}\sum_\alpha\hat{\mathcal{P}}_\alpha(t,\Delta,x_\alpha)}\Big]^n \end{aligned} \tag{3.56}$$

in which $\mathbf{x}=(x_0,\ldots,x_r)\in\mathbb{R}^{r+1}$, $p(\mathbf{z})$ is the distribution from which the n covariate vectors $\mathbf{z}_i$ were drawn, $p(t,\Delta|\boldsymbol{\beta}_0\cdot\mathbf{z})$ is defined in (3.4), and

$$W(\mathbf{x}|\{\boldsymbol{\beta}\}) = \int\mathrm{d}\mathbf{z}\ p(\mathbf{z})\prod_{\alpha=0}^r\delta\Big[x_\alpha-\boldsymbol{\beta}_\alpha\cdot\mathbf{z}\Big] \tag{3.57}$$

The integrand in the last line of (3.56) depends on the vectors $\boldsymbol{\beta}_\alpha$ only via the linear predictors $x_\alpha=\boldsymbol{\beta}_\alpha\cdot\mathbf{z}$. We assume that the distribution $W(\mathbf{x}|\{\boldsymbol{\beta}\})$ of these predictors is Gaussian, either because $p(\mathbf{z})$ is Gaussian, or because for $p\to\infty$ the central limit theorem will apply. Since $\int\mathrm{d}\mathbf{z}\ p(\mathbf{z})\mathbf{z}=\mathbf{0}$, we may now write, with $\boldsymbol{C}=\{C_{\alpha\rho}\}$:

$$W(\mathbf{x}|\{\boldsymbol{\beta}\})=\big[(2\pi)^{r+1}\mathrm{det}\boldsymbol{C}[\{\boldsymbol{\beta}\}]\big]^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}\mathbf{x}\cdot\boldsymbol{C}^{-1}[\{\boldsymbol{\beta}\}]\mathbf{x}} \tag{3.58}$$

$$\forall\alpha,\rho=0\ldots r:\qquad C_{\alpha\rho}[\{\boldsymbol{\beta}\}]=\int\mathrm{d}\mathbf{z}\ p(\mathbf{z})(\boldsymbol{\beta}_\alpha\cdot\mathbf{z})(\boldsymbol{\beta}_\rho\cdot\mathbf{z})=\boldsymbol{\beta}_\alpha\cdot\boldsymbol{A}\boldsymbol{\beta}_\rho \tag{3.59}$$

To describe the generation or explanation of finite event times, the linear predictors will always have to remain finite. Hence for realistic data sets the matrix elements $C_{\alpha\rho}$ must remain of order $O(1)$ even in the limit $p\to\infty$[2]. If we now also insert the integral $1=\int\mathrm{d}\boldsymbol{C}\ \delta[\boldsymbol{C}-\boldsymbol{C}[\{\boldsymbol{\beta}\}]]$, and write the δ-function in integral form, we find:

$$\big\langle[Z_n(\gamma,\mathcal{D})]^r\big\rangle_{\mathcal{D}} = \int\mathrm{d}\boldsymbol{C}\mathrm{d}\hat{\boldsymbol{C}}\Big[\prod_{\alpha=1}^r\mathcal{D}\hat{\mathcal{P}}_\alpha\mathcal{D}\mathcal{P}_\alpha\Big]\mathrm{e}^{-nr\Psi\left[\{\mathcal{P}_\alpha,\hat{\mathcal{P}}_\alpha\},\boldsymbol{C},\hat{\boldsymbol{C}}\right]+\mathcal{O}(\log n)} \tag{3.60}$$

[2]In earlier papers [2, 3, 16, 13] the definition (3.57) involved $\boldsymbol{\beta}_\alpha\cdot\mathbf{z}/\sqrt{p}$ instead of $\boldsymbol{\beta}_\alpha\cdot\mathbf{z}$, so the components of $\boldsymbol{\beta}_\alpha$ would typically be of order $\mathcal{O}(1)$. The present choice links more directly to the standard ML algorithms, but is the reason why in (3.11) the regularizer constant $\frac{1}{2}p\log p$ was required.

where, using $p=\zeta n$,

$$-r\Psi\big[\{\mathcal{P}_\alpha,\hat{\mathcal{P}}_\alpha\},\boldsymbol{C},\hat{\boldsymbol{C}}\big]=\mathrm{i}\zeta\mathrm{Tr}(\hat{\boldsymbol{C}}\boldsymbol{C})+\varphi[\boldsymbol{C},\{\hat{\mathcal{P}}_\alpha\}]+\zeta\phi[\hat{\boldsymbol{C}}]$$
$$+\mathrm{i}\sum_{\alpha=1}^{r}\sum_{\Delta}\int\mathrm{d}t\mathrm{d}h\;\hat{\mathcal{P}}_\alpha(t,\Delta,h)\mathcal{P}_\alpha(t,\Delta,h)-\gamma\sum_{\alpha=1}^{r}\mathcal{E}[\mathcal{P}_\alpha(.)] \tag{3.61}$$

with

$$\varphi[\boldsymbol{C},\{\hat{\mathcal{P}}_\alpha\}] = \log\left[\sum_{\Delta}\int\mathrm{d}t\mathrm{d}\mathbf{x}\;p(t,\Delta|x_0)\frac{\mathrm{e}^{-\frac{1}{2}\mathbf{x}\cdot\boldsymbol{C}^{-1}\mathbf{x}-\mathrm{i}\sum_\alpha\hat{\mathcal{P}}_\alpha(t,\Delta,x_\alpha)}}{\left[(2\pi)^{r+1}\mathrm{det}\boldsymbol{C}\right]^{\frac{1}{2}}}\right] \tag{3.62}$$

$$\phi[\hat{\boldsymbol{C}}] = \frac{1}{p}\log\int\Big[\prod_{\alpha=1}^{r}\mathrm{d}\boldsymbol{\beta}_\alpha\Big]\;\mathrm{e}^{\frac{1}{2}rp\log p-\mathrm{i}p\sum_{\alpha\rho=0}^{r}\hat{C}_{\alpha\rho}\boldsymbol{\beta}_\alpha\cdot\boldsymbol{A}\boldsymbol{\beta}_\rho} \tag{3.63}$$

It now follows, after exchanging the limits $n\to\infty$ and $r\to 0$, that the asymptotic disorder-average free energy density (3.52) can be computed via steepest descent:

$$f(\gamma) = \frac{1}{\gamma}\lim_{r\to 0}\mathrm{extr}_{\{\mathcal{P}_\alpha,\hat{\mathcal{P}}_\alpha\},\boldsymbol{C},\hat{\boldsymbol{C}}}\;\Psi\big[\{\mathcal{P}_\alpha,\hat{\mathcal{P}}_\alpha\},\boldsymbol{C},\hat{\boldsymbol{C}}\big]. \tag{3.64}$$

Saddle point integration

We first derive the stationary conditions for (3.61) with respect to the functions $\hat{\mathcal{P}}_\alpha$ and $\mathcal{P}_\alpha$, via functional differentiation and using (3.14). This gives, respectively,

$$\mathcal{P}_\alpha(t,\Delta,h)=\frac{\int\mathrm{d}\mathbf{x}\;\delta[h-x_\alpha]p(t,\Delta|x_0)\mathrm{e}^{-\frac{1}{2}\mathbf{x}\cdot\boldsymbol{C}^{-1}\mathbf{x}-\mathrm{i}\sum_{\rho=1}^{r}\hat{\mathcal{P}}_\rho(t,\Delta,x_\rho)}}{\sum_{\Delta'}\int\mathrm{d}t'\mathrm{d}\mathbf{x}\;p(t',\Delta'|x_0)\mathrm{e}^{-\frac{1}{2}\mathbf{x}\cdot\boldsymbol{C}^{-1}\mathbf{x}-\mathrm{i}\sum_{\rho=1}^{r}\hat{\mathcal{P}}_\rho(t',\Delta',x_\rho)}} \tag{3.65}$$

$$\mathrm{i}\hat{\mathcal{P}}_\alpha(t,\Delta,h)=\gamma\Delta\left[\log\left(\sum_{\Delta'}\int\mathrm{d}t'\mathrm{d}h'\;\theta(t'-t)\mathrm{e}^{h'}\mathcal{P}_\alpha(t',\Delta',h')\right)-h\right]$$
$$+\gamma\mathrm{e}^{h}\int\mathrm{d}t'\mathrm{d}h'\left(\frac{\theta(t-t')\mathcal{P}_\alpha(t',1,h')}{\sum_{\Delta''}\int\mathrm{d}t''\mathrm{d}h''\;\theta(t''-t')\mathrm{e}^{h''}\mathcal{P}_\alpha(t'',\Delta'',h'')}\right). \tag{3.66}$$

Combination of (3.7) with (3.12) enables us to recognize in (3.66) the replicated cumulative hazard function,

$$\Lambda_\alpha(t) = \int\mathrm{d}t'\mathrm{d}h'\;\frac{\theta(t-t')\mathcal{P}_\alpha(t',1,h')}{\mathcal{S}_\alpha(t')} \tag{3.67}$$

$$\mathcal{S}_\alpha(t) = \sum_{\Delta'}\int\mathrm{d}t'\mathrm{d}h'\;\theta(t'-t)\mathrm{e}^{h'}\mathcal{P}_\alpha(t',\Delta',h') \tag{3.68}$$

with which we may write (3.66) more compactly as

$$\mathrm{i}\hat{\mathcal{P}}_\alpha(t,\Delta,h) = \gamma\Big(\Delta[\log\mathcal{S}_\alpha(t)-h]+\mathrm{e}^{h}\Lambda_\alpha(t)\Big). \tag{3.69}$$

Insertion into (3.65) then simplifies the latter to

$$\mathcal{P}_\alpha(t,\Delta,h)= \tag{3.70}$$
$$\frac{\int\mathrm{d}\mathbf{x}\;\delta[h-x_\alpha]p(t,\Delta|x_0)\mathrm{e}^{-\frac{1}{2}\mathbf{x}\cdot\boldsymbol{C}^{-1}\mathbf{x}-\gamma\sum_{\rho=1}^{r}\left(\Delta[\log\mathcal{S}_\rho(t)-x_\rho]+\exp(x_\rho)\Lambda_\rho(t)\right)}}{\sum_{\Delta'}\int\mathrm{d}t'\mathrm{d}\mathbf{x}\;p(t',\Delta'|x_0)\mathrm{e}^{-\frac{1}{2}\mathbf{x}\cdot\boldsymbol{C}^{-1}\mathbf{x}-\gamma\sum_{\rho=1}^{r}\left(\Delta'[\log\mathcal{S}_\rho(t')-x_\rho]+\exp(x_\rho)\Lambda_\rho(t')\right)}}$$

Upon using also a modest amount of foresight we transform $\mathrm{i}\hat{\boldsymbol{C}} = \frac{1}{2}\boldsymbol{D}$. We then find that our saddle point problem (3.64) can be written as follows:

$$f(\gamma) \;=\; \frac{1}{\gamma}\lim_{r\to 0}\mathrm{extr}_{\boldsymbol{C},\boldsymbol{D},\{\mathcal{P}_\alpha\}}\;\tilde{\Psi}[\boldsymbol{C},\boldsymbol{D},\{\mathcal{P}_\alpha\}], \tag{3.71}$$

where

$$-r\tilde{\Psi}[\ldots] = \frac{1}{2}\zeta\mathrm{Tr}(\boldsymbol{D}\boldsymbol{C}) + \tilde{\varphi}[\boldsymbol{C}] + \zeta\tilde{\phi}[\boldsymbol{D}] - \frac{1}{2}\log\det\boldsymbol{C} - \frac{1}{2}(r+1)\log(2\pi)$$
$$+\gamma\sum_{\alpha=1}^{r}\Big\{\sum_{\Delta}\int \mathrm{d}t\mathrm{d}h\;\mathcal{P}_\alpha(t,\Delta,h)\Big(\Delta[\log\mathcal{S}_\alpha(t)-h]+\mathrm{e}^{h}\Lambda_\alpha(t)\Big) - \mathcal{E}[\mathcal{P}_\alpha(.)]\Big\}$$

with

$$\tilde{\phi}[\boldsymbol{D}] = \lim_{p\to\infty}\frac{1}{p}\log\int\Big[\prod_{\alpha=1}^{r}\mathrm{d}\boldsymbol{\beta}_\alpha\Big]\;\mathrm{e}^{\frac{1}{2}rp\log p-\frac{1}{2}p\sum_{\alpha\rho=0}^{r}D_{\alpha\rho}\boldsymbol{\beta}_\alpha\cdot\boldsymbol{A}\boldsymbol{\beta}_\rho} \tag{3.72}$$

$$\tilde{\varphi}[\boldsymbol{C}] = \tag{3.73}$$
$$\log\left[\sum_{\Delta}\int\mathrm{d}t\mathrm{d}\mathbf{x}\;p(t,\Delta|x_0)\mathrm{e}^{-\frac{1}{2}\mathbf{x}\cdot\boldsymbol{C}^{-1}\mathbf{x}-\gamma\sum_{\alpha=1}^{r}\left(\Delta[\log\mathcal{S}_\alpha(t)-x_\alpha]+\exp(x_\alpha)\Lambda_\alpha(t)\right)}\right]$$

We have not used (3.70) to derive (3.72), only (3.69). Hence in (3.71) the functions $\{\mathcal{P}_\alpha\}$ indeed still have the status of independent variational parameters, and functional differentiation of (3.72) with respect to $\mathcal{P}_\alpha$ must therefore reproduce (3.70).

The replica symmetry (RS) ansatz

We now make the so-called replica symmetric (RS) ansatz. Replica symmetry should affect only the nonzero replica labels $\alpha = 1\ldots r$, so for the present order parameters $\boldsymbol{C}$ and $\boldsymbol{D}$ the RS ansatz takes the following form, $\forall\alpha\in\{1,\ldots,r\}$:

$$C_{\alpha 0} = C_{0\alpha} = c_0, \quad C_{\alpha\nu} = C_{\nu\alpha} = c + (C-c)\delta_{\alpha\nu} \tag{3.74}$$
$$D_{\alpha 0} = D_{0\alpha} = \hat{m}, \quad D_{\alpha\nu} = D_{\nu\alpha} = -\hat{q} + (\hat{\rho}+\hat{q})\delta_{\alpha\nu} \tag{3.75}$$
$$\mathcal{P}_\alpha(t,\Delta,h) = \mathcal{P}(t,\Delta,h), \quad \Lambda_\alpha(t) = \Lambda(t), \quad \mathcal{S}_\alpha(t) = \mathcal{S}(t) \tag{3.76}$$

The RS form of $\boldsymbol{C}$ implies the same for its inverse, where we define $(\boldsymbol{C}^{-1})_{00} = \tilde{\mu}$ and $\forall\alpha\in\{1,\ldots,r\}$

$$(\boldsymbol{C}^{-1})_{\alpha 0} = (\boldsymbol{C}^{-1})_{0\alpha} = \tilde{m}, \tag{3.77}$$
$$(\boldsymbol{C}^{-1})_{\alpha\nu} = (\boldsymbol{C}^{-1})_{\nu\alpha} = \tilde{q} + (\tilde{\rho}-\tilde{q})\delta_{\alpha\nu} \tag{3.78}$$

After some algebra one then finds that

$$\tilde{\mu} \;=\; \frac{C+c(r-1)}{C_{00}(C+c(r-1)) - rc_0^2} \tag{3.79}$$
$$\tilde{m} \;=\; \frac{c_0}{rc_0^2 - C_{00}(C+c(r-1))} \tag{3.80}$$
$$\tilde{q} \;=\; \frac{1}{C-c}\,\frac{c_0^2 - cC_{00}}{C_{00}(C+c(r-1)) - rc_0^2} \tag{3.81}$$
$$\tilde{\rho} \;=\; \frac{1}{C-c} + \tilde{q}\;. \tag{3.82}$$

With the RS ansatz we can simplify the functions (3.72,3.73). We start with $\tilde{\phi}[\boldsymbol{D}]$, using the short-hand $S^2 = C_{00} = \boldsymbol{\beta}_0 \cdot \boldsymbol{A}\boldsymbol{\beta}_0$:

$$
\begin{aligned}
&\tilde{\phi}_{\rm RS}[\boldsymbol{D}] = \\
&\lim_{p\to\infty}\Bigg\{\frac{1}{2}rp\log p - \frac{1}{2}D_{00}\boldsymbol{\beta}_0\cdot\boldsymbol{A}\boldsymbol{\beta}_0 - \frac{r}{2p}\log{\rm Det}\boldsymbol{A} \\
&+\frac{1}{p}\log\int\Big[\prod_{\alpha=1}^{r}{\rm d}\boldsymbol{\beta}_\alpha\Big]\ {\rm e}^{-p\hat{m}\sum_{\alpha=1}^r\boldsymbol{\beta}_0\cdot\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\beta}_\alpha+\frac{1}{2}p\hat{q}[\sum_{\alpha=1}^r\boldsymbol{\beta}_\alpha]^2-\frac{1}{2}p(\hat{\rho}+\hat{q})\sum_{\alpha=1}^r(\boldsymbol{\beta}_\alpha)^2}\Bigg\} \\
&= \lim_{p\to\infty}\Bigg\{\frac{1}{2}rp\log p - \frac{1}{2}D_{00}\boldsymbol{\beta}_0\cdot\boldsymbol{A}\boldsymbol{\beta}_0 - \frac{r}{2p}\log{\rm Det}\boldsymbol{A} \\
&+\frac{1}{p}\log\Big[\prod_{\mu=1}^{p}{\rm D}x_\mu\Big]\Big[\prod_{\mu=1}^{p}\int{\rm d}\beta\ {\rm e}^{\beta[x_\mu\sqrt{p\hat{q}}-p\hat{m}(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\beta}_0)_\mu]-\frac{1}{2}p(\hat{\rho}+\hat{q})\beta^2}\Big]^r\Bigg\} \\
&= -\frac{1}{2}D_{00}S^2 + \frac{1}{2}r\Big(\frac{\hat{m}^2S^2+\hat{q}}{\hat{q}+\hat{\rho}} - \frac{1}{p}\log{\rm Det}\boldsymbol{A} - \log(\hat{\rho}+\hat{q}) - \log(2\pi)\Big) + \mathcal{O}(r^2)
\end{aligned}
\tag{3.83}
$$

Next we apply the RS ansatz to $\tilde{\varphi}[\boldsymbol{C}]$:

$$
\begin{aligned}
&{\rm e}^{\tilde{\varphi}_{\rm RS}[\boldsymbol{C}]} = \\
&\sum_{\Delta\in\{0,1\}}\int{\rm d}t{\rm d}\mathbf{x}\ p(t,\Delta|x_0){\rm e}^{\Delta\gamma\sum_{\alpha=1}^r x_\alpha-\gamma\Lambda(t)\sum_{\alpha=1}^r\exp(x_\alpha)-r\gamma\Delta\log\mathcal{S}(t)} \\
&\times{\rm e}^{-\frac{1}{2}\tilde{\mu}x_0^2-\tilde{m}x_0\sum_{\alpha=1}^r x_\alpha-\frac{1}{2}\tilde{q}[\sum_{\alpha=1}^r x_\alpha]^2-\frac{1}{2}(\tilde{\rho}-\tilde{q})\sum_{\alpha=1}^r x_\alpha^2} \\
= &\sum_{\Delta\in\{0,1\}}\int{\rm D}z\int{\rm d}t{\rm d}x_0\ {\rm e}^{-\frac{1}{2}\tilde{\mu}x_0^2}p(t,\Delta|x_0){\rm e}^{-r\gamma\Delta\log\mathcal{S}(t)} \\
&\times\Big[\int{\rm d}x\ {\rm e}^{x(\Delta\gamma+{\rm i}z\sqrt{\tilde{q}}-\tilde{m}x_0)-\gamma\Lambda(t)\exp(x)-\frac{1}{2}(\tilde{\rho}-\tilde{q})x^2}\Big]^r \\
= &\Big(\frac{2\pi}{\tilde{\mu}}\Big)^{\frac{1}{2}}\Bigg\{1 - r\gamma\int{\rm D}y_0\int{\rm d}t\ p(t,1|\frac{y_0}{\sqrt{\tilde{\mu}}})\log\mathcal{S}(t) + \frac{1}{2}r\log\Big(\frac{2\pi}{\tilde{\rho}-\tilde{q}}\Big) + \mathcal{O}(r^2) \\
&+r\sum_{\Delta\in\{0,1\}}\int{\rm D}y_0{\rm D}z\int{\rm d}t\ p(t,\Delta|\frac{y_0}{\sqrt{\tilde{\mu}}}) \\
&\times\log\int{\rm D}x\ {\rm e}^{x[\Delta\gamma+{\rm i}z\sqrt{\tilde{q}}-(\tilde{m}/\sqrt{\tilde{\mu}})y_0]/\sqrt{\tilde{\rho}-\tilde{q}}-\gamma\Lambda(t)\exp(x/\sqrt{\tilde{\rho}-\tilde{q}})}
\end{aligned}
\tag{3.84}
$$

Hence

$$
\begin{aligned}
&\tilde{\varphi}_{\rm RS}[\boldsymbol{C}] = \frac{1}{2}\log\Big(\frac{2\pi}{\tilde{\mu}}\Big) + \\
&-r\gamma\int{\rm D}y_0\int{\rm d}t\ p(t,1|\frac{y_0}{\sqrt{\tilde{\mu}}})\log\mathcal{S}(t) + \frac{1}{2}r\log\Big(\frac{2\pi}{\tilde{\rho}-\tilde{q}}\Big) + \mathcal{O}(r^2) \\
&+r\sum_{\Delta\in\{0,1\}}\int{\rm D}y_0{\rm D}z\int{\rm d}t\ p(t,\Delta|\frac{y_0}{\sqrt{\tilde{\mu}}}) \\
&\times\log\int{\rm D}x\ {\rm e}^{x[\Delta\gamma+{\rm i}z\sqrt{\tilde{q}}-(\tilde{m}/\sqrt{\tilde{\mu}})y_0]/\sqrt{\tilde{\rho}-\tilde{q}}-\gamma\Lambda(t)\exp(x/\sqrt{\tilde{\rho}-\tilde{q}})}
\end{aligned}
\tag{3.85}
$$

We then find, using identities such as $\mathrm{Tr}(\boldsymbol{DC}) = D_{00}S^2 + r(2c_0\hat{m} + c\hat{q} + C\hat{\rho}) + \mathcal{O}(r^2)$, $\log \mathrm{Det}\boldsymbol{C} = \log S^2 + r[\log(C-c) + (c - c_0^2/S^2)/(C-c)] + \mathcal{O}(r^2)$ (obtained by diagonalizing the RS form of $\boldsymbol{C}$), as well as $\tilde{\mu}S^2 = 1 + rc_0^2/S^2(C-c) + \mathcal{O}(r^2)$, $\tilde{m} = -c_0/S^2(C-c) + \mathcal{O}(r)$, $\tilde{q} = (c_0^2/S^2 - c)/(C-c)^2 + \mathcal{O}(r)$, and $\tilde{\rho} - \tilde{q} = (C-c)^{-1} + \mathcal{O}(r)$, that the RS ansatz implies the following for expression (3.72) (modulo irrelevant constants):

$$
\begin{aligned}
&-\lim_{r\to 0}\tilde{\Psi}_{\rm RS}[\ldots] = \\
&-\frac{1}{2}\frac{c}{C-c} + \frac{1}{2}\zeta\Big(2\hat{m}c_0 + \hat{q}c + \hat{\rho}C + \frac{\hat{m}^2S^2 + \hat{q}}{\hat{q}+\hat{\rho}} - \log(\hat{\rho}+\hat{q})\Big) \\
&+\sum_{\Delta}\int \mathrm{D}y_0\mathrm{D}z \int \mathrm{d}t\; p(t,\Delta|Sy_0)\times \\
&\times \log\int \mathrm{D}x\; \mathrm{e}^{\left[\Delta\gamma\sqrt{C-c} + z\frac{\sqrt{c-c_0^2/S^2}}{\sqrt{C-c}} + \frac{y_0c_0}{S\sqrt{C-c}}\right]x - \gamma\Lambda(t)\exp(x\sqrt{C-c})} \\
&+\gamma\Bigg\{\sum_{\Delta\in\{0,1\}}\int \mathrm{d}t\mathrm{d}h\; \mathcal{P}(t,\Delta,h)\Big(\Delta[\log\mathcal{S}(t) - h] + \mathrm{e}^h\Lambda(t)\Big) - \mathcal{E}[\mathcal{P}(.)] \\
&-\int \mathrm{D}y_0\int \mathrm{d}t\; p(t,1|Sy_0)\log\mathcal{S}(t)\Bigg\}\;.
\end{aligned}
\tag{3.86}
$$

We can now extremize $\Psi_{\rm RS}[\ldots]$ over $(\hat{m},\hat{q},\hat{\rho})$, as demanded by (3.71), giving

$$
\hat{m} = -\frac{c_0}{S^2(C-c)}, \qquad \hat{q} = \frac{c - c_0^2/S^2}{(C-c)^2}, \qquad \hat{\rho} + \hat{q} = \frac{1}{C-c}
\tag{3.87}
$$

Our various formulae takes more compact forms upon introducing the three short-hands $u = \sqrt{\gamma(C-c)}$, $v = \sqrt{c - c_0^2/S^2}$ and $w = c_0/S$, which we know from earlier studies to have finite $\gamma\to\infty$ limits. We may then write

$$
\hat{m} = -\frac{\gamma w}{Su^2}, \qquad \hat{q} = \frac{\gamma^2 v^2}{u^4}, \qquad \hat{\rho} + \hat{q} = \frac{\gamma}{u^2}
\tag{3.88}
$$

and

$$
\begin{aligned}
&-\lim_{\gamma\to\infty}\frac{1}{\gamma}\lim_{r\to 0}\tilde{\Psi}_{\rm RS}[\ldots] = -\frac{1}{2u^2}[w^2 + (1-\zeta)v^2] + \sum_{\Delta}\int \mathrm{D}y_0\mathrm{D}z\mathrm{d}t\; p(t,\Delta|Sy_0)\times \\
&\times \lim_{\gamma\to\infty}\frac{1}{\gamma}\log\int \mathrm{d}x\; \mathrm{e}^{\gamma\left[-\frac{1}{2}u^2x^2 + [\Delta u^2 + vz + wy_0]x - \Lambda(t)\exp(xu^2)\right]} \\
&+\sum_{\Delta\in\{0,1\}}\int \mathrm{d}t\mathrm{d}h\; \mathcal{P}(t,\Delta,h)\Big(\Delta[\log\mathcal{S}(t) - h] + \mathrm{e}^h\Lambda(t)\Big) - \mathcal{E}[\mathcal{P}(.)] \\
&-\int \mathrm{D}y_0\int \mathrm{d}t\; p(t,1|Sy_0)\log\mathcal{S}(t)\;.
\end{aligned}
\tag{3.89}
$$

We may hence now write $f_{\rm RS} = \lim_{\gamma\to\infty} f_{\rm RS}(\gamma)$ as follows:

$$
\begin{aligned}
f_{\rm RS} = {\rm extr}_{u,v,w,\mathcal{P}}\Bigg\{ & \frac{1}{2u^2}[w^2{+}(1{-}\zeta)v^2] - \sum_{\Delta\in\{0,1\}}\int {\rm D}y{\rm D}z{\rm d}t\ p(t,\Delta|Sy)\times \\
& \times\ {\rm max}_x\Big[-\frac{1}{2}u^2x^2 + (\Delta u^2{+}vz{+}wy)x - \Lambda(t){\rm e}^{xu^2}\Big] + \\
& - \sum_{\Delta\in\{0,1\}}\int {\rm d}t{\rm d}h\ \mathcal{P}(t,\Delta,h)\Big(\Delta[\log\mathcal{S}(t)-h] + {\rm e}^h\Lambda(t)\Big) - \mathcal{E}[\mathcal{P}(.)] \\
& + \int {\rm D}y\int {\rm d}t\ p(t,1|Sy)\log\mathcal{S}(t)\Bigg\}
\end{aligned}
\tag{3.90}
$$

with $\mathcal{E}[\mathcal{P}(.)]$ as defined by (3.14), and

$$
\Lambda(t) = \int {\rm d}t'{\rm d}h'\ \frac{\theta(t-t')\mathcal{P}(t',1,h')}{\mathcal{S}(t')} \tag{3.91}
$$

$$
\mathcal{S}(t) = \sum_{\Delta'}\int {\rm d}t'{\rm d}h'\ \theta(t'-t){\rm e}^{h'}\mathcal{P}(t',\Delta',h') \tag{3.92}
$$

The maximization over x in (3.90) is achieved for $x = u^{-2}\xi(t,\Delta,y,z)$, where

$$
\xi(t,\Delta,y,z) = \Delta u^2{+}vz{+}wy - W\Big(u^2\Lambda(t){\rm e}^{\Delta u^2+vz+wy}\Big) \tag{3.93}
$$

with Lambert's function $W(z)$, the inverse of the function $g(x) = x{\rm e}^x$. Note that expression (3.93) can be interpreted in terms of a Moreau envelope, and is in other approaches to overfitting indeed often found via that route [11]. Hence we now arrive at

$$
\begin{aligned}
f_{\rm RS} = {\rm extr}_{u,v,w,\mathcal{P}}\Bigg\{ & \frac{1}{2u^2}[w^2{+}(1{-}\zeta)v^2] + \frac{1}{u^2}\sum_{\Delta\in\{0,1\}}\int {\rm D}y{\rm D}z{\rm d}t\ p(t,\Delta|Sy)\times \\
& \left[\frac{1}{2}\xi^2(t,\Delta,y,z) - (\Delta u^2{+}vz{+}wy)\xi(t,\Delta,y,z) + u^2\Lambda(t){\rm e}^{\xi(t,\Delta,y,z)}\right] \\
& - \sum_{\Delta\in\{0,1\}}\int {\rm d}t{\rm d}h\ \mathcal{P}(t,\Delta,h)\Big(\Delta[\log\mathcal{S}(t)-h] + {\rm e}^h\Lambda(t)\Big) - \mathcal{E}[\mathcal{P}(.)] \\
& + \int {\rm D}y\int {\rm d}t\ p(t,1|Sy)\log\mathcal{S}(t)\Bigg\},
\end{aligned}
\tag{3.94}
$$

which can be written in alternative ways, using identities such as $z{\rm e}^{-W(z)} = W(z)$.

Replica symmetric order parameter equations

We next work out the RS form of equation (3.70) in the limit $r \to 0$, using the same manipulations as in working out the previous function $\tilde{\varphi}_{\rm RS}[\boldsymbol{C}]$, giving

$$\begin{aligned}
&\mathcal{P}(t,\Delta,h)\lim_{r\to 0}\Bigg[\int\!{\rm D}y{\rm D}z\ p(t,\Delta|\frac{y}{\sqrt{\tilde{\mu}}}){\rm e}^{-r\gamma\Delta\log\mathcal{S}(t)}\ \frac{\sqrt{\tilde{\rho}-\tilde{q}}}{\sqrt{2\pi}}{\rm e}^{-\frac{1}{2}(\tilde{\rho}-\tilde{q})h^2}\\
&\times{\rm e}^{[\Delta\gamma+{\rm i}z\sqrt{\tilde{q}}-\frac{\tilde{m}y}{\sqrt{\tilde{\mu}}}]h-\gamma\Lambda(t)\exp(h)}\Big(\int\!{\rm D}x\ {\rm e}^{[\Delta\gamma+{\rm i}z\sqrt{\tilde{q}}-\frac{\tilde{m}y}{\sqrt{\tilde{\mu}}}]\frac{x}{\sqrt{\tilde{\rho}-\tilde{q}}}-\gamma\Lambda(t)\exp(\frac{x}{\sqrt{\tilde{\rho}-\tilde{q}}})}\Big)^{r-1}\Bigg]\\
&\times\Bigg[\sum_{\Delta'\in\{0,1\}}\int\!{\rm d}t'\int\!{\rm D}y{\rm D}z\ p(t',\Delta'|\frac{y}{\sqrt{\tilde{\mu}}}){\rm e}^{-r\gamma\Delta'\log\mathcal{S}(t')}\\
&\times\Big(\int\!{\rm D}x\ {\rm e}^{[\Delta'\gamma+{\rm i}z\sqrt{\tilde{q}}-\frac{\tilde{m}y}{\sqrt{\tilde{\mu}}}]\frac{x}{\sqrt{\tilde{\rho}-\tilde{q}}}-\gamma\Lambda(t')\exp(\frac{x}{\sqrt{\tilde{\rho}-\tilde{q}}})}\Big)^{r}\Bigg]^{-1}\\
&=\lim_{r\to 0}\int\!{\rm D}y{\rm D}z\ p(t,\Delta|\frac{y}{\sqrt{\tilde{\mu}}})\frac{\frac{\sqrt{\tilde{\rho}-\tilde{q}}}{\sqrt{2\pi}}{\rm e}^{-\frac{1}{2}(\tilde{\rho}-\tilde{q})h^2+[\Delta\gamma+{\rm i}z\sqrt{\tilde{q}}-\frac{\tilde{m}y}{\sqrt{\tilde{\mu}}}]h-\gamma\Lambda(t)\exp(h)}}{\int\!{\rm D}x\ {\rm e}^{[\Delta\gamma+{\rm i}z\sqrt{\tilde{q}}-\frac{\tilde{m}y}{\sqrt{\tilde{\mu}}}]\frac{x}{\sqrt{\tilde{\rho}-\tilde{q}}}-\gamma\Lambda(t)\exp(\frac{x}{\sqrt{\tilde{\rho}-\tilde{q}}})}}\Bigg]\\
&=\int\!{\rm D}y{\rm D}z\ p(t,\Delta|Sy)\frac{{\rm e}^{\frac{\gamma}{u^2}\left[-\frac{1}{2}h^2+(\Delta u^2+vz+wy)h-u^2\Lambda(t){\rm e}^h\right]}}{\int\!{\rm d}h'\ {\rm e}^{\frac{\gamma}{u^2}\left[-\frac{1}{2}h'^2+(\Delta u^2+vz+wy)h'-u^2\Lambda(t){\rm e}^{h'}\right]}}
\end{aligned}\tag{3.95}$$

Upon taking the limit $\gamma\to\infty$ we then find the following remarkably simple result, in which the function $\xi(\ldots)$ is as given by equation (3.93):

$$\lim_{\gamma\to\infty}\mathcal{P}(t,\Delta,h)\ \ =\ \ \int\!{\rm D}y{\rm D}z\ p(t,\Delta|Sy)\ \delta\big[h-\xi(t,\Delta,y,z)\big].\tag{3.96}$$

From this one then obtains the RS expressions of $\Lambda(t)$ and $\mathcal{S}(t)$, via (3.91, 3.92).

Finally we need to work out the scalar order parameter equations for the trio (u,v,w), by executing the extremization demanded in expression (3.94) for the free energy density. These equations are seen to take the form $\partial\mathcal{F}(u,v,w)/\partial u=\partial\mathcal{F}(u,v,w)/\partial v=\partial\mathcal{F}(u,v,w)/\partial w=0$, with $\xi(\ldots)$ as given in (3.93), where

$$\begin{aligned}
\mathcal{F}(u,v,w)&=\frac{1}{2u^2}[w^2+(1-\zeta)v^2]+\frac{1}{u^2}\sum_{\Delta\in\{0,1\}}\int\!{\rm D}y{\rm D}z{\rm d}t\ p(t,\Delta|Sy)\times\\
&\times\Big[\frac{1}{2}\xi^2(t,\Delta,y,z)-(\Delta u^2+vz+wy)\xi(t,\Delta,y,z)+u^2\Lambda(t){\rm e}^{\xi(t,\Delta,y,z)}\Big]
\end{aligned}\tag{3.97}$$

By using identities such as $W(x)\exp[W(x)]=x$, one finds that
$u^2\Lambda(t)\exp[\xi(t,\Delta,y,z)]=W(u^2\Lambda(t){\rm e}^{\Delta u^2+vz+wy})$, which enables us to simplify $\mathcal{F}(u,v,w)$ to

$$\begin{aligned}
\mathcal{F}(u,v,w)=\frac{1}{u^2}\Bigg\{&\frac{1}{2}[w^2+(1-\zeta)v^2]+\sum_{\Delta}\int\!{\rm D}y{\rm D}z{\rm d}t\ p(t,\Delta|Sy)\times\\
&\Big[\frac{1}{2}W^2(u^2\Lambda(t){\rm e}^{\Delta u^2+vz+wy})+W(u^2\Lambda(t){\rm e}^{\Delta u^2+vz+wy})-\frac{1}{2}(\Delta u^2+vz+wy)^2\Big]\Bigg\}
\end{aligned}\tag{3.98}$$

From this expression one then derives directly the final order parameter equations for (u, v, w), in which only the equation for u involves some further manipulations:

$$\frac{\partial \mathcal{F}}{\partial v} = 0:$$
$$\zeta v = \int \mathrm{D}y\mathrm{D}z\ z \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|Sy)\ W(u^2\Lambda(t)\mathrm{e}^{\Delta u^2+vz+wy}) \tag{3.99}$$
$$\frac{\partial \mathcal{F}}{\partial w} = 0:$$
$$0 \ = \int \mathrm{D}y\mathrm{D}z\ y \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|Sy)\ \Big[W(u^2\Lambda(t)\mathrm{e}^{\Delta u^2+vz+wy}) - \Delta u^2\Big] \tag{3.100}$$
$$\frac{\partial \mathcal{F}}{\partial u} = 0:$$
$$\zeta v^2 = \int \mathrm{D}y\mathrm{D}z \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|Sy)\Big[W(u^2\Lambda(t)\mathrm{e}^{\Delta u^2+vz+wy}) - \Delta u^2\Big]^2 \tag{3.101}$$

$$\mathcal{F}(u_\star, v_\star, w_\star) =$$
$$= \int \mathrm{D}y\mathrm{D}z \sum_{\Delta} \int \mathrm{d}t\ p(t,\Delta|Sy)\Big[\Lambda(t)\mathrm{e}^{\xi_\star(t,\Delta,y,z)} - \Delta\xi_\star(t,\Delta,y,z)\Big]\ . \tag{3.102}$$

3.9.2 Variance of inferred risk factors

In this Appendix we derive equation (3.49) for the variance of the inferred risk factors $\hat{\boldsymbol{\beta}}_{\rm ML}\cdot\mathbf{z}_i$, which is an important tool with which to remove the need for knowing the amplitude S in the construction of unbiased estimators. It exploits our result (3.25) for the asymptotic form of the distribution (3.12) (which is the type of quantity that within the replica theory is assumed to be self-averaging[3]):

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (\hat{\boldsymbol{\beta}}_{\rm ML}\cdot\mathbf{z}_i)^2 &= \lim_{n\to\infty} \sum_{\Delta\in\{0,1\}} \int \mathrm{d}t\mathrm{d}h\ \big\langle P_n(t,\Delta,h|\hat{\boldsymbol{\beta}}_{\rm ML},\mathcal{D})\big\rangle_{\mathcal{D}} h^2 \\
&= \sum_{\Delta\in\{0,1\}} \int \mathrm{d}t\mathrm{d}h\ \mathcal{P}(t,\Delta,h) h^2 \\
&= \sum_{\Delta\in\{0,1\}} \int \mathrm{D}y\mathrm{D}z\mathrm{d}t\ p(t,\Delta|Sy)\xi^2(t,\Delta,y,z).
\end{aligned} \tag{3.103}$$

[3]This assumption is validated by the accuracy of the predictions of the RS equations, as confirmed repeatedly in previous studies [2, 3, 13, 12], and also again in our present numerical simulations.

Upon using (3.18) we can thus write

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\beta}}_{\rm ML}\cdot\mathbf{z}_i)^2 =$$
$$\sum_{\Delta}\int {\rm D}y{\rm D}z{\rm d}t\ p(t,\Delta|Sy)\Big[wy+vz+u^2\Delta - W\Big(u^2{\rm e}^{u^2\Delta+wy+vz}\Lambda(t)\Big)\Big]^2$$
$$= v^2+w^2+\int {\rm D}y{\rm D}z\sum_{\Delta}\int {\rm d}t\ p(t,\Delta|Sy)\Big[u^2\Delta - W\Big(u^2{\rm e}^{u^2\Delta+wy+vz}\Lambda(t)\Big)\Big]^2$$
$$-2w\int {\rm D}y{\rm D}z\sum_{\Delta}\int {\rm d}t\ p(t,\Delta|Sy)\Big[yW\Big(u^2{\rm e}^{u^2\Delta+wy+vz}\Lambda(t)\Big) - \Delta u^2 y\Big]$$
$$-2v\int {\rm D}y{\rm D}z\sum_{\Delta}\int {\rm d}t\ p(t,\Delta|Sy)\ zW\Big(u^2{\rm e}^{u^2\Delta+wy+vz}\Lambda(t)\Big). \tag{3.104}$$

Upon using the order parameter equations (3.19,3.20,3.21) we can now simplify the right-hand side of the latter expression, and find the remarkably simple but powerful result

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\beta}}_{\rm ML}\cdot\mathbf{z}_i)^2 = (1-\zeta)v_\star^2+w_\star^2\ . \tag{3.105}$$

3.9.3 Additional simulations

This Appendix describes tests of the theory against simulations for further choices of the censoring rate, i.e. for different values of for different values of $\langle\Delta\rangle$. In 3.9.3 we compare the measurements in additional simulations with the solutions of the RS equations, as in section 3.6. In 3.9.3 we apply the algorithm of section 3.7 for different values of $\langle\Delta\rangle$, to investigate the robustness of the proposed algorithm. The data are generated as explained in the main text, i.e. the cumulative hazard is either of the Weibull form $\Lambda_0(t)=\frac{1}{2}t^2$ or of the Log-logistic form $\Lambda_0(t)=\log(1+t^2)$, the true signal strength is $|\boldsymbol{\beta}_0|=1$, and the censoring is uniformly distributed in the interval $[0,t_{\rm max}]$. In order to vary the fraction of censored individuals we tune the end-of-trial time $t_{\rm max}$. In particular, we consider the values $t_{\rm max}=2$ and $t_{\rm max}=6$.

In all cases it can be seen that the RS theory predicts the typical values of the estimators very accurately (even for $n=400$).

Comparison of theory and simulations

In Figures 3.7 and 3.8 we compare the predicted values of the overlaps $(\kappa_\star,v_\star)$, obtained by solving the RS equations, against the values $(\hat{\kappa}_n,\hat{v}_n)$ in simulations, for different values of ζ and for $t_{\rm max}\in\{2,6\}$ (corresponding to $\langle\Delta\rangle\approx 40\%$ and 70%, respectively). We also show in (3.9) the comparison between the Breslow estimator $\hat{\Lambda}_n(.)$ (solid black lines) against the solution of the RS equations (yellow solid line) when $\zeta=0.3$ at $t_{\rm max}=2.0$ (top row) and when $\zeta=0.4$ at $t_{\rm max}=6$ (bottom row). As a reference we plot the identity line (solid red), which corresponds to perfect recovery.

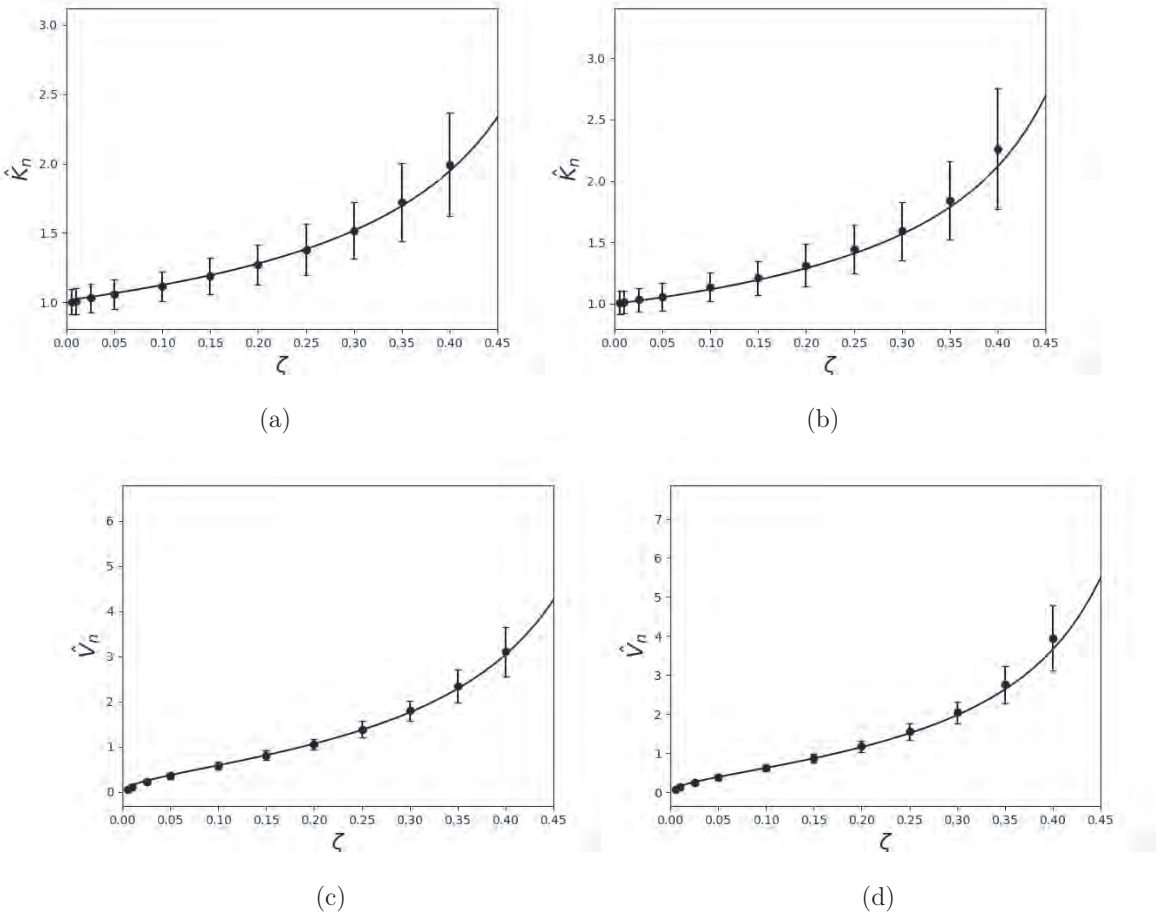


Figure 3.7: Comparison of theoretical predictions and simulation data, for Cox's survival analysis model with uniform censoring on the interval $[0, 2]$ ($\langle\Delta\rangle \approx 40\%$) and data set size $n = 400$. Top row: overfitting-induced bias factor $\hat{\kappa}_n$ versus ζ; bottom row: overfitting-induced noise amplitude $\hat{v}_n$ versus ζ. Panels (a,c): data for cumulative hazard $\Lambda_0(t) = \log(1 + t^2)$. Panels (b,d): data for cumulative hazard $\Lambda_0(t) = \frac{1}{2}t^2$. In all panels we plot with markers and error bars the simulation results, averaged over 500 independent simulations with distinct data realizations, and with solid lines the predictions $\kappa_\star$ and $v_\star$ of the RS theory (solved with populations of size $m = 10^5$).

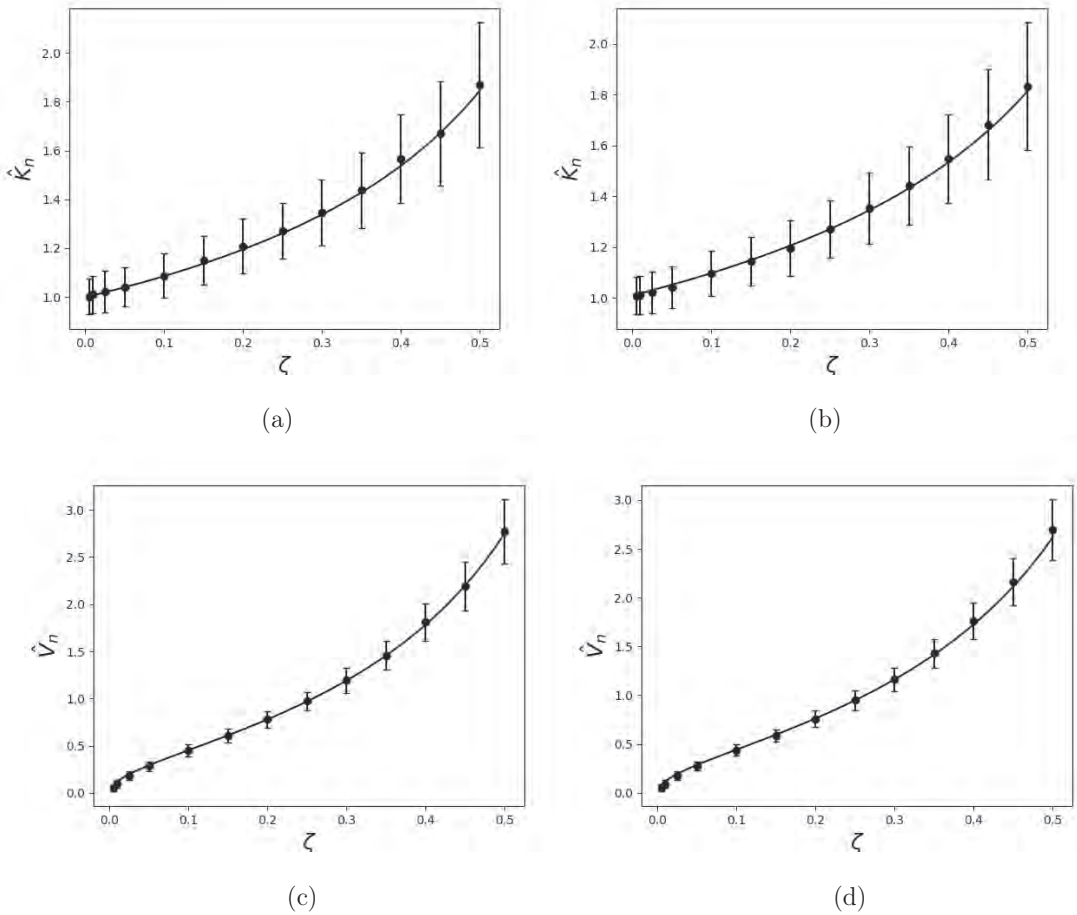


Figure 3.8: Comparison of theoretical predictions and simulation data, for Cox's survival analysis model with uniform censoring on the interval $[0, 6]$ ($\langle\Delta\rangle \approx 70\%$) and data set size $n = 400$. Top row: overfitting-induced bias factor $\hat{\kappa}_n$ versus ζ; bottom row: overfitting-induced noise amplitude $\hat{v}_n$ versus ζ. Panels (a,c): data for cumulative hazard $\Lambda_0(t) = \log(1+t^2)$ ($\langle\Delta\rangle \approx 50\%$). Panels (b,d): data for cumulative hazard $\Lambda_0(t) = \frac{1}{2}t^2$ ($\langle\Delta\rangle \approx 40\%$). In all panels we plot with markers and error bars the simulation results, averaged over 500 independent simulations with distinct data realizations, and with solid lines the predictions $\kappa_\star$ and $v_\star$ of the RS theory (solved with populations of size $m = 10^5$).

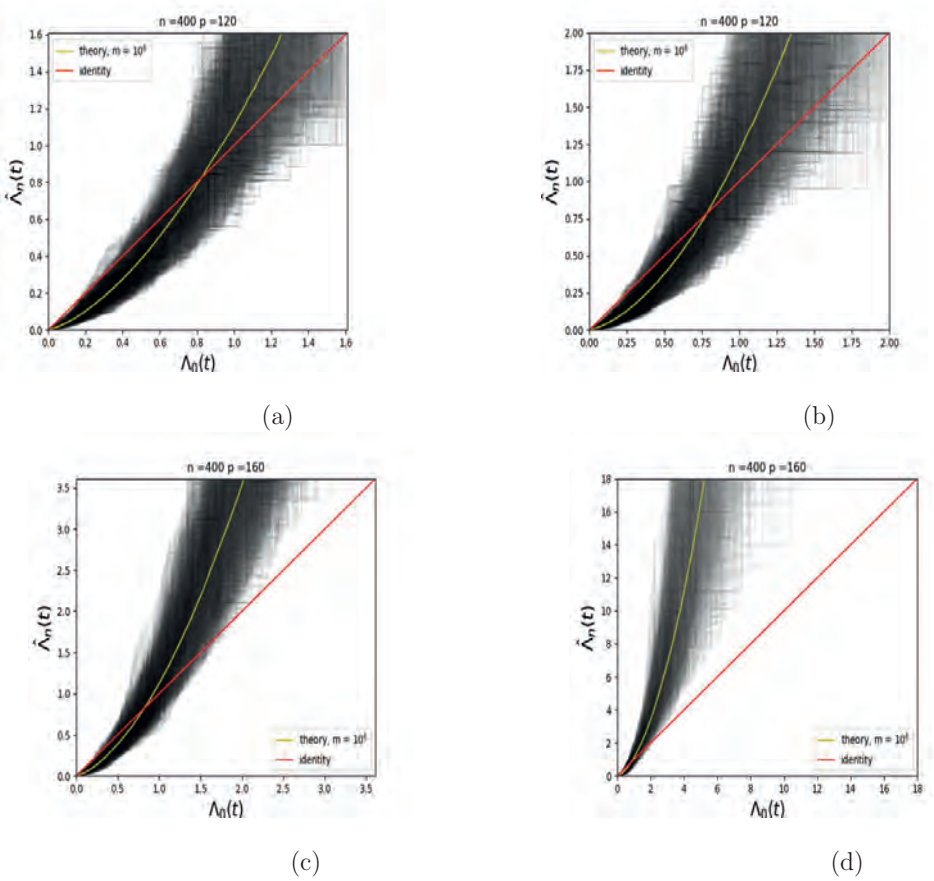


Figure 3.9: Comparison of theoretical predictions and simulation data, for Cox's survival analysis model with uniform censoring on the interval $[0, t_{\max}]$ and data set size $n = 400$. Top row: $p = 120$ ($\zeta = 0.3$), $t_{\max} = 2.0$ ($\langle\Delta\rangle \approx 40\%$); bottom row : $p = 160$ ($\zeta = 0.4$), $t_{\max} = 6.0$ ($\langle\Delta\rangle \approx 70\%$). Panels (a,c): cumulative hazard $\Lambda_0(t) = \log(1+t^2)$. Panels (b,d): cumulative hazard $\Lambda_0(t) = \frac{1}{2}t^2$. In all panels we plot with black lines the Breslow estimator $\hat{\Lambda}_n(.)$ versus the true cumulative hazard $\Lambda_0(.)$, for 500 independent simulations with distinct data realizations. Yellow curves show the predictions of the RS theory (solved with populations of size $m = 10^5$), and red curves indicate the diagonal (that would have been found for perfect regression).

Simulations test for the proposed decontamination algorithm

In Figure 3.10 we compare the Breslow estimator $\hat{\Lambda}_n(.)$ (solid black lines) with its decontamined version $\tilde{\Lambda}_n(.)$ (blue solid lines), when $\zeta = 0.3$ at $t_{\max} = 2$ (top row) and when $\zeta = 0.4$ at $t_{\max} = 6$ (bottom row). As a reference we also plot the diagonal (solid red), which corresponds to perfect estimation. We also show in Figure 3.11 the histograms of the estimator S_n for S, as obtained via the decontamination algorithm, together with $\tilde{\Lambda}_n(.), \tilde{\kappa}_n$, and $\tilde{v}_n$. The true value $S = 1$ is plotted as dashed vertical line in black. Finally we compare the histograms of the ML estimators $\mathbf{e}_1 \cdot \hat{\boldsymbol{\beta}}_n$ and $\mathbf{e}_2 \cdot \hat{\boldsymbol{\beta}}_n$ against their corrected values $\mathbf{e}_1 \cdot \tilde{\boldsymbol{\beta}}_n$

and $\mathbf{e}_2 \cdot \tilde{\boldsymbol{\beta}}_n$, with $\tilde{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_n/\kappa_n$, and we superimpose the normal density with mean zero and variance $v_\star^2/\kappa_\star^2$ (the prediction of the RS theory). Since here $\boldsymbol{\beta}_0 = \mathbf{e}_1$, the true values are $\mathbf{e}_1 \cdot \boldsymbol{\beta}_0 = 1$ and $\mathbf{e}_2 \cdot \boldsymbol{\beta}_0 = 0$.

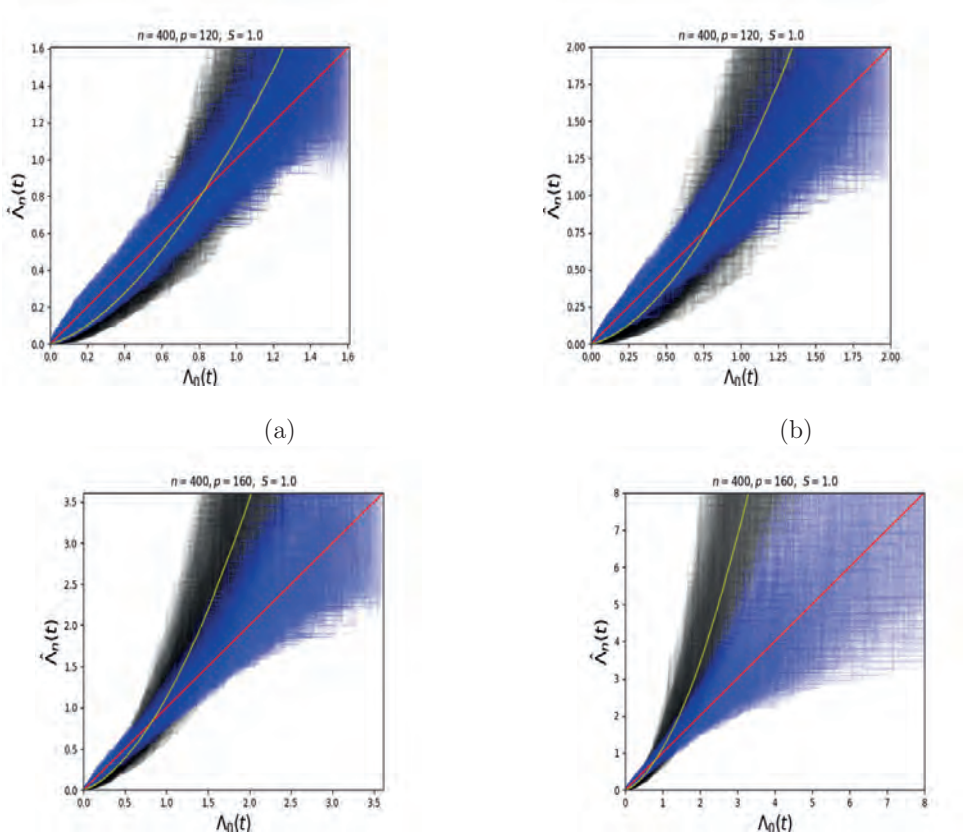


Figure 3.10: Comparison of uncorrected and corrected integrated hazard rate estimators derived from simulation data, for Cox's survival analysis model with uniform censoring on the interval $[0, t_{\max}]$and data set size $n = 400$. In all cases $\boldsymbol{\beta}_0 = \hat{\mathbf{e}}_1$, so $S = 1$. Top row: $p = 120$ (so $\zeta = 0.3$), $t_{\max} = 2.0$; bottom row: $p = 160$ (so $\zeta = 0.4$), $t_{\max} = 6.0$. Panels (a,c): cumulative hazard $\Lambda_0(t) = \log(1+t^2)$. Panels (b,d): cumulative hazard $\Lambda_0(t) = \frac{1}{2}t^2$. In all panels we plot with black lines the Breslow estimator $\hat{\Lambda}_n(.)$ versus the true cumulative hazard $\Lambda_0(.)$, for 500 independent simulations with distinct data realizations. Yellow curves show the predictions of the RS theory (solved with populations of size $m = 10^5$), and red curves indicate the diagonal (that would have been found for perfect regression). We also show in blue the *de-biased estimator* $\tilde{\Lambda}(.)$ for the cumulative hazard versus the true cumulative hazard $\Lambda_0(.)$, for the same 500 simulations.

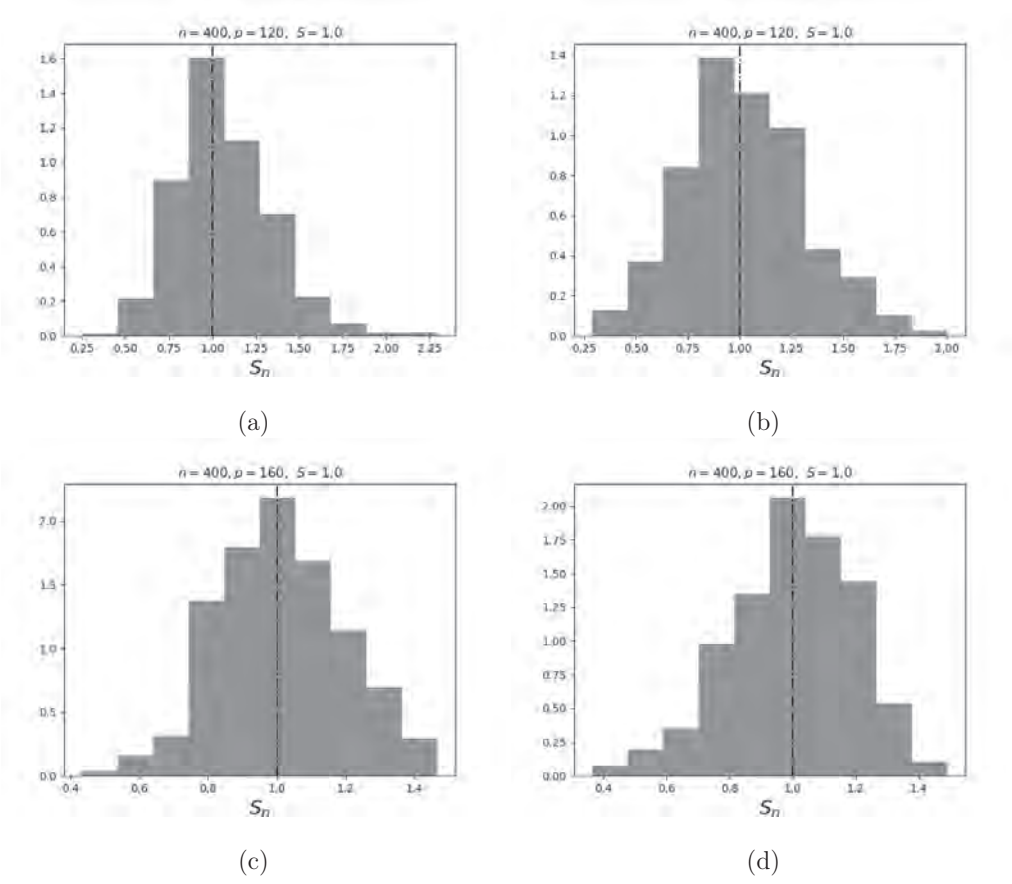


Figure 3.11: Distributions of the values of the effective signal strength $S = |\boldsymbol{A}^{1/2}\boldsymbol{\beta}_0|$ inferred by our de-biasing algorithm, for the same data as those used in Figure (3.10,3.12,3.13). Top row: $p = 120$ (so $\zeta = 0.3$), $t_{\max} = 2.0$; bottom row: $p = 160$ (so $\zeta = 0.4$), $t_{\max} = 6.0$. In all panels the maximum of the histogram corresponds to the true value $S = 1$, in spite of the relatively modest data set size (around 240 non-censored events).

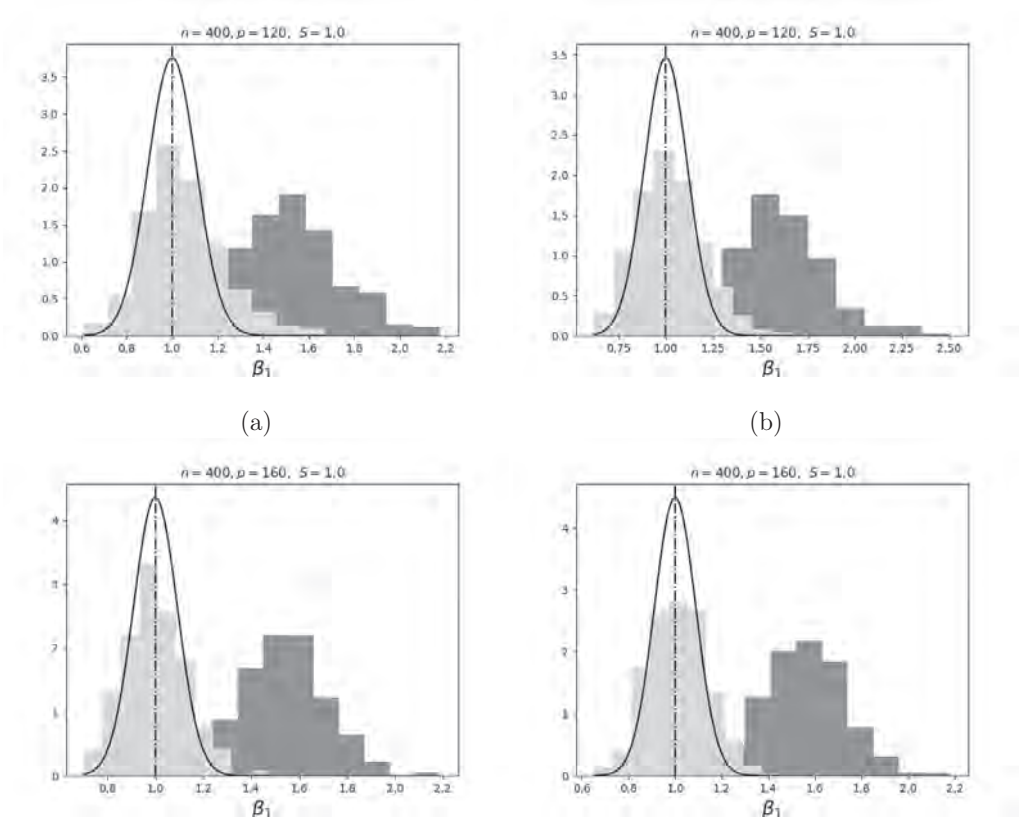
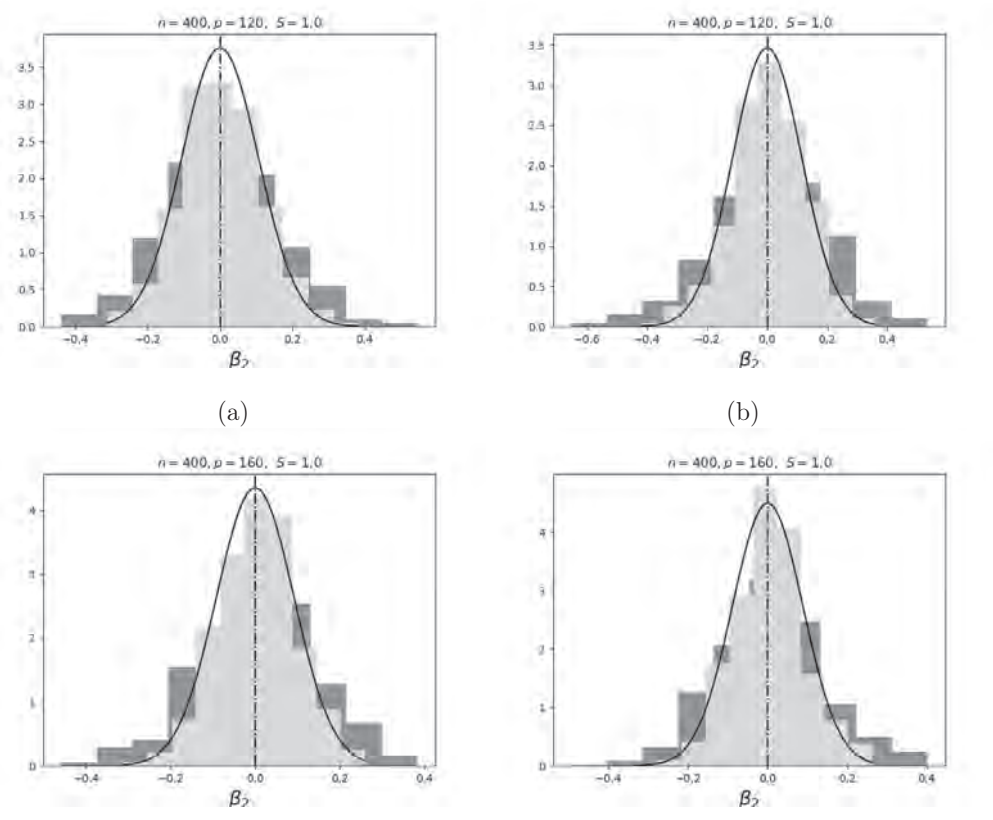


Figure 3.12: Distributions of the values of the non-zero component $\hat{\mathbf{e}}_1 \cdot \boldsymbol{\beta}$ of the association vector, both for the standard ML Cox regression estimator (dark grey) and for the new estimator inferred by our de-biasing algorithm (light grey), for the same data as those used in Figure (3.10,3.11,3.13). Top row: $p = 120$ (so $\zeta = 0.3$), $t_{\max} = 2.0$; bottom row: $p = 160$ (so $\zeta = 0.4$), $t_{\max} = 6.0$. We also show as a vertical dashed line the location of the true value $\hat{\mathbf{e}}_1 \cdot \boldsymbol{\beta}_0 = 1$, and as a solid curve the predicted Gaussian asymptotic distribution of the new estimator, with average 1 and variance $v_\star^2/\kappa_\star^2 p$. We observe that the new estimator indeed removed successfully the overfitting-induced bias of the ML one, while its distribution exhibits finite size effects.

In all cases we see that the proposed algorithm succeeds in correcting for the overfitting bias: in Figure 3.10 the blue lines are distributed around lies around the (red) diagonal, and in Figure 3.11 the modes of the histograms of $S_n, \mathbf{e}_1 \cdot \tilde{\boldsymbol{\beta}}_n$ and $\mathbf{e}_2 \cdot \tilde{\boldsymbol{\beta}}_n$ reproduce the true values $S = 1,\ \mathbf{e}_1 \cdot \boldsymbol{\beta}_0 = 1$ and $\mathbf{e}_2 \cdot \boldsymbol{\beta}_0 = 0$.

In order for our algorithm to converge, it is necessary that the ML estimator actually exists. Hence we expect that the algorithm will converge when $\zeta = p/n$ is sufficiently far from the critical value that marks the censoring equivalent of the phase transition derived in [22]. In a finite size simulation that transition is not sharp, and there will therefore be data instances even below the phase transition line for which the ML estimator does not exist.

The (future) inclusion of a ridge penalization should cure these pathologies, which are due to the use of ML regression, rather than to the algorithm itself. Nevertheless, estimating correctly the hazard rate of the event under study will always be more difficult when only a few primary events are observed in a sample. This implies that any algorithm must inevitably become less reliable as $\langle \Delta \rangle$ decreases toward zero.



(a) (b)

(c) (d)

Figure 3.13: Distributions of the values of the zero component $\hat{\mathbf{e}}_2 \cdot \boldsymbol{\beta}$ of the association vector, both for the standard ML Cox regression estimator (dark grey) and for the new estimator inferred by our de-biasing algorithm (light grey), for the same data as those used in Figure (3.10,3.11,3.12). Top row: $p = 120$ (so $\zeta = 0.3$), $t_{\max} = 2.0$; bottom row: $p = 160$ (so $\zeta = 0.4$), $t_{\max} = 6.0$.We also show as a vertical dashed line the location of the true value $\hat{\mathbf{e}}_2 \cdot \boldsymbol{\beta}_0 = 0$, and as a solid curve the predicted Gaussian asymptotic distribution of the new estimator, here with zero average and variance $v_\star^2/\kappa_\star^2 p$. Here, as expected, already the ML estimator was unbiased, since the theory predicts the overfitting-induced bias to take the form of a multiplicative factor $\kappa_\star$ (which would here just multiply the number zero).

Chapter 4

Proportional asymptotics of piecewise exponential proportional hazards models

This chapter is accepted for publication in the proceeding of the 15th Workshop on Stochastic Models, Statistics and Their Applications (2024). An updated version of this chapter is available on ArXiv `https://arxiv.org/abs/2501.18995`

4.1 Abstract

We study the flexible piece-wise exponential model in a high dimensional setting where the number of covariates p grows proportionally to the number of observations n and under the hypothesis of random uncorrelated Gaussian designs. We prove rigorously that the optimal ridge penalized log-likelihood of the model converges in probability to the saddle point of a surrogate objective function. The technique of proof is the Convex Gaussian Min-Max theorem of Thrampoulidis, Oymak and Hassibi. An important consequence of this result, is that we can study the impact of the ridge regularization on the estimates of the parameter of the model and the prediction error as a function of the ratio p/n. Furthermore these results represents a first step toward rigorously proving the (conjectured) correctedness of several results obtained with the heuristic replica method for the Cox semi-parametric model.

4.2 Introduction

In medicine and healthcare, survival analysis is extensively used to assess the impact of medical interventions, predict patient outcomes, and estimate disease prognosis [21, 20, 29, 6]. Survival data are often analysed in the framework of the Cox model [8], which has several desirable properties in the classical regime where number of covariates (p) is assumed to be small compared to the number of subjects (n) available in a study [1, 38]. However, modern applications p is often comparable or larger than n. In this regime the maximum partial likelihood estimator might not even exist, and when it does is frequently biased and exhibits a large variance [7, 26], making it an unreliable tool for statistical inference and prediction. Such undesirable properties are unavoidable and have been known since the 60' in the statistical community, starting from the works of Kolmogorov and Huber [43, 11].

In the last two decades a vast amount of theory on sparse regression in high dimensional generalized linear models [40], and the Cox model [2, 31], has been established. This line of research builds on the assumption that the underlying "true" model is sparse and guarantees

consistency of estimators obtained by a penalized maximum likelihood approach when: i) the regularizer is cleverly chosen, i.e. lasso and variants of the latter[37], and ii) the regularization strength is appropriately tuned [5]. The theoretical machinery of this line of work, i.e. the theory of empirical processes [41, 42, 39], imposes only minor conditions over the distribution of the covariates and the results are often expressed as bounds on performance metrics, with precise probabilistic guarantees [39, 43, 24, 5, 31].

An alternative "average case" approach to the same subject, starting with the pioneering work [13], which used statistical physics methodology (i.e. the replica method), focuses instead on the study of the average values of performance metrics in the proportional asymptotic regime where the number of covariates $p = p_n = \zeta n$ grows linearly with the number of observations $n \to \infty$, so that the $\lim_{n\to\infty} p_n/n = \zeta \in \mathbb{R}$. This approach has been recently rigorous using different techniques such as Leave One Out method [10, 34], Approximate Message Passing [9] and the Convex Gaussian Min max Theorem (CGMT)[36, 35, 27]. The price to pay for this alternative and "sharper" description is more assumptions on the data generating process. In particular a standard, mathematically convenient (albeit often ideal) assumption is usually undertaken, namely that the covariates follow a standard multivariate normal distribution.

In this manuscript, we study the asymptotic behaviour of the ridge penalized maximum likelihood estimator for the piece-wise exponential proportional hazards model via a rigorous formulation of the typical case approach mentioned above. The piece-wise exponential proportional hazards model is a parametric proportional hazards model where the base hazard rate function is postulated to be piece-wise constant, throughout a set of *user defined* intervals [12, 19]. By applying the CGMT, we prove rigorously that the optimal ridge penalized log-likelihood of the model converges in probability to the saddle point of a surrogate objective function. Furthermore we show that the expected (typical) value of several prediction metrics, and some training metrics, can be computed precisely by computing the location of this saddle point. In principle, the Cox semi-parametric model can be recovered by an appropriate *data dependent* choice of the time intervals, i.e. by assuming the base hazard rate to be constant between un-censored observations [3, 4]. However, the proof presented in this manuscript is not directly applicable since it hinges on the fact the number of parameters describing the base hazard rate is finite as $n \to \infty$ (which is not the case for the Cox model).

The manuscript is structured as follows. In section 4.3 we introduce the model under study and the related notation; the main theoretical results are presented in section 4.4, whilst a brief sketch of the proof can be found in section 4.5 referring to the appendices for the technical details. We illustrate the agreement of theory and simulations in section 4.6 via numerical experiments where we quantify the prediction error by means of the concordance-index and an oracle version of the integrated brier score. Concluding remarks are in section 4.7.

4.3 Setting

We assume that the time at which a subject fails (i.e. experiences the event) is generated as

$$Y|\mathbf{X} \sim f_0(.|\mathbf{X}'\boldsymbol{\beta}_0), \quad \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p) \tag{4.1}$$

We focus on right censoring, being the most common in applications. In this scenario, instead of observing the actual response Y directly, we observe the event time T and the censoring indicator Δ defined as follows

$$T = \min\{Y, C\}, \ \Delta = \mathbf{1}\big[T < C\big] \tag{4.2}$$

where C is the censoring time, i.e. the time at which the subject drops out of the study. We shall further assume that censoring is uninformative, i.e. $Y \perp C|\mathbf{X}$, and that the censoring is at random and not dependent on the subject's covariates $C \perp \mathbf{X}$. Given the data generated according to (4.1) and (4.2), the statistician assumes a proportional hazards model

$$\Delta, T|\mathbf{X} \sim \left(\lambda(T|\boldsymbol{\omega})\mathrm{e}^{\mathbf{X}'\boldsymbol{\beta}}\right)^{\Delta} \mathrm{e}^{-\Lambda(T|\boldsymbol{\omega})\exp(\mathbf{X}'\boldsymbol{\beta})} \times f_C(T)^{1-\Delta} S_C(T)^{\Delta} \ , \tag{4.3}$$

with

$$\lambda(.|\boldsymbol{\omega}) := \frac{\mathrm{d}}{\mathrm{d}t}\Lambda(t|\boldsymbol{\omega}) \ , \tag{4.4}$$

where $C \sim f_C$ and $S_C(x) = \int_x^{+\infty} f_c(y)\mathrm{d}y$ are, respectively, the density and cumulative distribution function of the censoring risk, $\Lambda(.|\boldsymbol{\omega}) : \mathbb{R}^+ \to \mathbb{R}^+$ is the cumulative hazard for the primary risk and $\lambda(.|\boldsymbol{\omega})$ is the corresponding hazard. The piece-wise exponential proportional hazards model postulates that λ is a linear combination of piece-wise functions, i.e. zero-th order spline function, with user specified knots $\tau_1, \ldots, \tau_{\ell+1}$ [12, 19] as follows

$$\lambda(t|\boldsymbol{\omega}) := \sum_{k=1}^{\ell} \psi_k(t)\exp(\omega_k) = \boldsymbol{\psi}(t)'\exp(\boldsymbol{\omega}) \tag{4.5}$$

with

$$\boldsymbol{\psi}(t) = \big(\psi_0(t), \ldots, \psi_{\ell-1}(t)\big)', \quad \psi_k(t) := \boldsymbol{I}[\tau_k < t < \tau_{k+1}], \quad k = 1, \ldots, \ell \ . \tag{4.6}$$

The parametric form of the cumulative hazard can be obtained via integration and reads

$$\Lambda(t|\boldsymbol{\omega}) := \boldsymbol{\Psi}(t)'\exp(\boldsymbol{\omega}) \tag{4.7}$$

where

$$\boldsymbol{\Psi}(t) = \big(\Psi_0(t), \ldots, \Psi_{\ell-1}(t)\big)', \quad \Psi_k(t) := \boldsymbol{I}[t > \tau_k]\min\{t - \tau_k, \tau_{k+1} - \tau_k\} \ . \tag{4.8}$$

We notice that this is a special case of a B-spline parameterization of the baseline hazard function. The model is, in principle, arbitrary flexible in the sense that increasing the number of intervals allows to fit more and more precisely the shape of the "true" (unknown) baseline hazard rate. It has been noticed in literature [16] that the locations of the knots do not generally impact on the quality of the fit, but the number of knots is crucial. In this manuscript we take a penalized splines (P-spline) approach, use several equi-spaced knots and penalize the respective coefficients via a ridge-like regularizer. To keep the setting simple we consider a simple uniform ridge regularization (although smoothness penalties based on finite differences might be easily considered in the present setting), i.e. the model is fitted to the data by minimizing the following objective function

$$\mathfrak{L}_n(\boldsymbol{\omega}, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i'\boldsymbol{\beta}, \boldsymbol{\omega}, T_i, \Delta_i) - \frac{1}{p}\sum_{i=1}^{n} \Delta_i \log\lambda(t_i|\boldsymbol{\omega}) + \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2 + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \tag{4.9}$$

with the function g, defined as

$$g(x, \boldsymbol{\omega}, T, \Delta) = \Lambda(T|\boldsymbol{\omega})\mathrm{e}^x - \Delta x \ . \tag{4.10}$$

The piece-wise exponential model lets us control the "complexity" of the parametrization of the base hazard rate h via the number of intervals $\ell + 1$ (user specified) in which the latter is assumed to be constant. This is very convenient when the sample size n is large (as we shall assume): in this case the number of parameters definingthe survival function is only ℓ numbers, and not n as for Cox model [23, 22]. The ridge penalization is often introduced in practice to improve numerical stability by enforcing strong convexity of $\mathfrak{L}_n(\boldsymbol{\omega}, \boldsymbol{\beta})$ in its arguments and the minimizer always exists unique.

4.4 Technical Results

We need to introduce some definitions from convex analysis. In particular we will need the following.

Definition 1 (Moreau envelope and proximal operator)**.** *Given a convex function $f : \mathbb{R} \to \mathbb{R}$ the Moreau envelope function evaluated at x with parameter α, is defined as*

$$\mathcal{M}_{f(.)}(x, \nu) := \min_{y} \Big\{ \frac{1}{2\nu}(y - x)^2 + f(y) \Big\} \ . \tag{4.11}$$

The point at which the minimum is attained is called the proximal operator of f

$$\text{prox}_{f(.)}(x, \nu) := \arg\min_{y} \Big\{ \frac{1}{2\nu}(y - x)^2 + f(y) \Big\} \ . \tag{4.12}$$

These functions are frequently encountered in convex analysis problems as they posses several desirable properties, e.g. $\mathcal{M}_{f(.)}(., \nu > 0)$ is "smooth" irrespective of the smoothness of f and the set of minimizers of $\mathcal{M}_{f(.)}(., \nu > 0)$, $\forall \nu > 0$ coincides with that of the minimizers of $f(.)$ [30].

The main technical results of the manuscript are established in the following theorems and corollaries. In short the asymptotic value of several quantities of interest derived from the Penalized Maximum Likelihood estimator can be obtained by studying an asymptotically equivalent problem, which is low dimensional and hence quickly and easily solvable (in particular for large values of p), instead of studying directly the original high dimensional problem (4.9).

Theorem 1 (Asymptotically equivalent scalar optimization problem)**.** *Let $\zeta = \lim_{n\to\infty} p(n)/n$ and assume that the data are generated as in (4.1, 4.2). Then*

$$\min_{\boldsymbol{\omega},\boldsymbol{\beta}} \mathfrak{L}_n(\boldsymbol{\omega}, \boldsymbol{\beta}) \xrightarrow[n\to\infty]{P} \min_{\boldsymbol{\omega},w,v} \max_{\phi\geq 0} \inf_{\tau>0} \mathcal{L}(\boldsymbol{\omega}, w, v, \phi, \tau) \tag{4.13}$$

where

$$\begin{aligned} \mathcal{L}(\boldsymbol{\omega}, w, v, \phi, \tau) \ &:= \ \mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0 + vQ, \tau/\phi\big) - \Delta \log \lambda(T|\boldsymbol{\omega}) \Big] + \\ &+ \ \phi\big(\tau/2 - v\sqrt{\zeta}\big) + \frac{1}{2}\eta(v^2 + w^2) + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \end{aligned} \tag{4.14}$$

with $Z_0, Q \sim \mathcal{N}(0, 1)$, $Z_0 \perp Q$.

The convergence in probability of the optimum value of the objective function implies the following.

Theorem 2 (Replica Symmetric equations)**.** *Let $\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\omega}}_n$ be the (unique) minimizer of $\mathfrak{L}_n$ as defined in (4.9), then*

$$\hat{w}_n := \frac{\boldsymbol{\beta}_0'\hat{\boldsymbol{\beta}}_n}{\|\boldsymbol{\beta}_0\|} \xrightarrow[n\to\infty]{P} w_\star \tag{4.15}$$

$$\hat{v}_n := \big\|\mathbf{P}_{\perp\boldsymbol{\beta}_0}\hat{\boldsymbol{\beta}}_n\big\| = \Big\|\Big(\boldsymbol{I} - \frac{\boldsymbol{\beta}_0\boldsymbol{\beta}_0'}{\|\boldsymbol{\beta}_0\|^2}\Big)\hat{\boldsymbol{\beta}}_n\Big\| \xrightarrow[n\to\infty]{P} v_\star \tag{4.16}$$

$$\hat{\boldsymbol{\omega}}_n \xrightarrow[n\to\infty]{P} \boldsymbol{\omega}_\star \ . \tag{4.17}$$

The values $\boldsymbol{\omega}_\star, w_\star, v_\star, \tau_\star, \phi_\star$ identify the saddle point of $\mathcal{L}$ and solve the following set of self consistent equations

$$v^2\zeta = \mathbb{E}_{T,Z_0,Q}\Big[\big\|\hat{\xi} - wZ_0 - vQ\big\|^2\Big] \tag{4.18}$$

$$w(1+\eta\tau/\phi) = \mathbb{E}_{T,Z_0,Q}\Big[Z_0\hat{\xi}\Big] \tag{4.19}$$

$$v(1-\zeta+\eta\tau/\phi) = \mathbb{E}_{T,Z_0,Q}\Big[Q\hat{\xi}\Big] \tag{4.20}$$

$$\tau = v\sqrt{\zeta} \tag{4.21}$$

$$\omega_k = \frac{1}{\eta\rho}\mathbb{E}\Big[\Delta\psi_k(T)\Big] + \tag{4.22}$$

$$-W_0\bigg(\frac{1}{\eta\rho}\mathbb{E}\Big[\mathrm{e}^{\hat{\xi}}\psi_k(T)\Big]\exp\Big\{\frac{1}{\eta\rho}\mathbb{E}\Big[\Delta\Psi_k(T)\Big]\Big\}\bigg),\ k=1,\dots,\ell, \tag{4.23}$$

where

$$\begin{aligned}\hat{\xi}(Z_0,Q,T) &:= \mathrm{prox}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0+vQ,\tau/\phi) = \\ &= wZ_0+vQ+\Delta\tau/\phi - W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau/\phi+wZ_0+vQ}\Big)\ .\end{aligned} \tag{4.24}$$

The theorems (1,2) above are a generalization of previous results [35, 25] to the piece-wise exponential model. We briefly sketch the proof ideas in the next section and we delegate the details to the supplementary material 4.8.2,4.8.3,4.8.4. The assumption that the covariate are uncorrelated is because we want to focus mostly on the part of the proof that depends on the model (which is a novel application). When the covariates are correlated, the proof is complicated by additional terms that depend on the spectrum of the covariates and the regularization function, see for instance [25]. A consequence of the theorem above is the following.

Corollary 1 (Surrogate for out sample linear predictor)**.** *Let $f:\mathbb{R}^{1+l}\to\mathbb{R}$ and $\tilde{\mathbf{X}}$ a newly generated covariate vector*

$$f(\tilde{\mathbf{X}}'\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\omega}}_n) \xrightarrow[n\to\infty]{d} f\big(w_\star Z_0 + v_\star Q, \boldsymbol{\omega}_\star\big) \tag{4.25}$$

Proof. By Slutsky's Lemma [42] we have that for a "fresh" covariate vector $\tilde{\mathbf{X}}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_p)$

$$\tilde{\mathbf{X}}'\hat{\boldsymbol{\beta}}_n = \frac{\tilde{\mathbf{X}}'\boldsymbol{\beta}_0}{\|\boldsymbol{\beta}_0\|}\frac{\boldsymbol{\beta}_0'\hat{\boldsymbol{\beta}}_n}{\|\boldsymbol{\beta}_0\|} + \tilde{\mathbf{X}}'(\boldsymbol{I}_p - \frac{\boldsymbol{\beta}_0\boldsymbol{\beta}_0'}{\|\boldsymbol{\beta}_0\|^2})\hat{\boldsymbol{\beta}}_n \overset{d}{=} \tilde{Z}_0 w_n + \tilde{Q}v_n \xrightarrow[n\to\infty]{d} \tilde{Z}_0 w_\star + \tilde{Q}v_\star, \tag{4.26}$$

with

$$\tilde{Z}_0, \tilde{Q} \sim \mathcal{N}(0,1)\ , \tag{4.27}$$

because of the convergence in probability of w_n, v_n (4.15,4.16). The conclusion follows by the continuous mapping theorem [42] (page 7 theorem 2.3). □

The corollary above allows to precisely compute prediction metrics like the ones that we study in section (4.6).

4.5 Proof sketch

Introducing Lagrange multipliers $\boldsymbol{\phi}$, one can equivalently re-write the minimization problem as a saddle point problem as follows

$$\ell_n(\mathbf{X},\mathbf{T},\eta) = \min_{\boldsymbol{\omega},\boldsymbol{\beta},\boldsymbol{\xi}} \sup_{\boldsymbol{\phi}} \Big\{ \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\xi}_i,\boldsymbol{\omega},\Delta_i,T_i) - \Delta_i \log\lambda(T_i|\boldsymbol{\omega}) + \\ -\boldsymbol{\phi}'\big(\boldsymbol{\xi} - \mathbf{X}\boldsymbol{\beta}\big) + \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2 + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \Big\} \ . \tag{4.28}$$

Notice that $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\mathbf{P}_{\boldsymbol{\beta}_0}\boldsymbol{\beta} + \mathbf{X}\mathbf{P}_{\perp\boldsymbol{\beta}_0}\boldsymbol{\beta} \overset{d}{=} \mathbf{Z}_0\beta_{\|} + \tilde{\mathbf{X}}\boldsymbol{\beta}_{\perp}$, where $\overset{d}{=}$ indicates equality in distribution with $\beta_{\|} := \frac{\boldsymbol{\beta}_0'\boldsymbol{\beta}}{\|\boldsymbol{\beta}_0\|}$, $\boldsymbol{\beta}_{\perp} := \mathbf{P}_{\perp\boldsymbol{\beta}_0}\boldsymbol{\beta} = (\boldsymbol{I} - \boldsymbol{\beta}_0\boldsymbol{\beta}_0'/\|\boldsymbol{\beta}_0\|^2)\boldsymbol{\beta}$, $\mathbf{Z}_0 := \mathbf{X}\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\| \sim \mathcal{N}(\mathbf{0},\boldsymbol{I}_n)$ and $\tilde{\mathbf{X}} := \mathbf{X}\mathbf{P}_{\boldsymbol{\beta}_0}$ $(\tilde{\mathbf{X}})_{i,j} \sim \mathcal{N}(0,1)$ and $\tilde{\mathbf{X}} \perp \mathbf{T}$. Using this fact, we have reduced the original problem into the form

$$\ell_n(\mathbf{X},\mathbf{T},\eta) \overset{d}{=} \min_{\boldsymbol{\omega},\boldsymbol{\beta},\boldsymbol{\xi}} \sup_{\boldsymbol{\phi}} \Big\{ \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\xi}_i,\boldsymbol{\omega},\Delta_i,T_i) - \Delta_i \log\lambda(T_i|\boldsymbol{\omega}) + \\ -\boldsymbol{\phi}'\big(\boldsymbol{\xi} - \mathbf{Z}_0\beta_{\|} - \tilde{\mathbf{X}}\boldsymbol{\beta}_{\perp}\big) + \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2 + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \Big\} \tag{4.29}$$

that can be attacked with the Convex Gaussian Min-Max theorem (CGMT), first introduced in [36] as a generalization of Gordon's Gaussian comparison inequalities [14]. The CGMT is reported without proof below for the reader's convenience. We refer to [36, 35] for a detailed proof.

Theorem 3 (Convex Gaussian Min Max Theorem)**.** *Let $\mathcal{S}_{\mathbf{y}} \subset \mathbb{R}^n$, $\mathcal{S}_{\mathbf{z}} \subset \mathbb{R}^p$ be compact and convex sets, ψ be continuous and convex-concave on $\mathcal{S}_{\mathbf{z}} \times \mathcal{S}_{\mathbf{y}}$, and $\mathbf{X} \in \mathbb{R}^{n\times p}, \mathbf{Q} \in \mathbb{R}^n, \mathbf{G} \in \mathbb{R}^p$ all have entries iid standard normal*

$$\begin{aligned} \Phi(\mathbf{X}) &:= \min_{\mathbf{y}\in\mathcal{S}_{\mathbf{y}}} \max_{\mathbf{z}\in\mathcal{S}_{\mathbf{z}}} \Big\{ \mathbf{y}'\mathbf{X}\mathbf{z} + \psi(\mathbf{z},\mathbf{y}) \Big\} \\ \phi(\mathbf{G},\mathbf{Q}) &:= \min_{\mathbf{y}\in\mathcal{S}_{\mathbf{y}}} \max_{\mathbf{z}\in\mathcal{S}_{\mathbf{z}}} \Big\{ \|\mathbf{y}\|\mathbf{G}'\mathbf{z} + \|\mathbf{z}\|\mathbf{Q}'\mathbf{y} + \psi(\mathbf{z},\mathbf{y}) \Big\} \end{aligned}$$

Then

$$\forall \mu \in \mathbb{R}, t \in \mathbb{R}^+, \ P\Big[\Big|\Phi(\mathbf{X}) - \mu\Big| > t\Big] \leq 2P\Big[\Big|\Phi(\mathbf{G},\mathbf{Q}) - \mu\Big| \geq t\Big] \ . \tag{4.30}$$

Furthermore, let $\mathcal{S} \subset \mathcal{S}_{\mathbf{y}}$ an open subset and $\mathcal{S}^c := \mathcal{S}_{\mathbf{y}} \setminus \mathcal{S}$. Denote $\phi_{\mathcal{S}^c}(\mathbf{G},\mathbf{Q})$ the optimal cost of the surrogate process, when the minimization is now constrained over $\mathcal{S}^c$. If there exist constants $\bar{\phi} < \bar{\phi}_{\mathcal{S}^c}$, such that $\phi(\mathbf{G},\mathbf{Q}) \xrightarrow[n\to\infty]{P} \bar{\phi}$, $\phi_{\mathcal{S}^c}(\mathbf{G},\mathbf{Q}) \xrightarrow[n\to\infty]{P} \bar{\phi}_{\mathcal{S}^c}$, then, denoting with $\hat{y}$ the value of y at the saddle point, we have

$$\lim_{n\to\infty} P\Big[\hat{\mathbf{y}} \in \mathcal{S}\Big] = 1 \ . \tag{4.31}$$

Theorem (3) above tells us that if one is able to prove that ϕ converges in probability to some deterministic value, then the same is true for Φ. The CGMT applies to min-max problems over compact, convex sets. Hence we "artificially" restrict the saddle point problem (4.29) onto compact, convex sets. Intuition suggests that if a saddle point exists and the set is sufficiently large, then there is not going to be any difference between the bounded and

unbounded problem. Hence from now on the min over $\boldsymbol{\beta}, \boldsymbol{\xi}$ and the max over $\boldsymbol{\phi}$ operations are understood over convex, compact sets (so that they actually exist). Following Theorem (3), we consider an auxiliary optimization problem

$$\tilde{\ell}_n(\mathbf{Q},\mathbf{G},\mathbf{T}) = \min_{\boldsymbol{\omega},\boldsymbol{\beta},\boldsymbol{\xi}} \max_{\boldsymbol{\phi}} \Big\{ \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\xi}_i,\boldsymbol{\omega},\Delta_i,T_i) - \Delta_i \log\lambda(T_i|) + \\ +\boldsymbol{\beta}_\perp'\mathbf{G}\|\boldsymbol{\phi}\| - \boldsymbol{\phi}'\big(\boldsymbol{\xi} - \beta_\parallel \mathbf{Z}_0 - \|\boldsymbol{\beta}_\perp\|\mathbf{Q}\big) + \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2 + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \Big\}$$

with $\mathbf{G}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_{(p-1)})$ and $\mathbf{Q}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_n)$, $\mathbf{G}\perp\mathbf{Q}$ and $\mathbf{G},\mathbf{Q}\perp\mathbf{T},\mathbf{Z}_0$. We can optimize over the direction of $\boldsymbol{\phi}$ at fixed length $\|\boldsymbol{\phi}\|=\phi$

$$\tilde{\ell}_n(\mathbf{Q},\mathbf{G},\mathbf{T}) = \min_{\boldsymbol{\omega},\boldsymbol{\beta},\boldsymbol{\xi}} \max_{\phi\geq 0} \Big\{ \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\xi}_i,\boldsymbol{\omega},\Delta_i,T_i) - \Delta_i \log\lambda(T_i|\boldsymbol{\omega}) \\ +\phi\big\|\boldsymbol{\xi} - \beta_\parallel\mathbf{Z}_0 - \|\boldsymbol{\beta}_\perp\|\mathbf{Q}\big\| + \boldsymbol{\beta}_\perp'\mathbf{G}\phi - \Delta_i\log\lambda(T_i|\boldsymbol{\omega}) + \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2 + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \Big\}.$$

The problem above depends on $\boldsymbol{\beta}_\perp$ via $\|\boldsymbol{\beta}_\perp\|$ and $\cos\theta$, with θ the angle between $\boldsymbol{\beta}_\perp$ and $\mathbf{G}$. Hence it is not guaranteed to be convex-concave (as *cos* is not convex on the whole interval $[0,2\pi]$), but it has been shown in [35] (page 22 point 6) that (4.32) can be used in place of (4.32) in the CGMT as $n\to\infty$, i.e. the min-max order can be "swapped" asymptotically. At this points, the minimization over the direction of $\boldsymbol{\beta}_\perp$ at fixed $\|\boldsymbol{\beta}_\perp\|=v$ yields

$$\tilde{\ell}_n(\mathbf{Q},\mathbf{G},\mathbf{T}) = \min_{\boldsymbol{\omega},w,v} \max_{\phi} \min_{\boldsymbol{\xi}} \Big\{ \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\xi}_i,\boldsymbol{\omega},\Delta_i,T_i) - \Delta_i\log\lambda(T_i|\boldsymbol{\omega}) + \\ +\phi\big\|\boldsymbol{\xi} - w\mathbf{Z}_0 - v\mathbf{Q}\big\| - v\phi\|\mathbf{G}\| + \frac{1}{2}\eta(v^2+w^2) + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \Big\} \tag{4.32}$$

where we also defined $w := \boldsymbol{\beta}_\parallel$. Following [35], we use the variational representation $\|\mathbf{y}\| = \inf_{\tau\geq 0}\Big\{\frac{1}{2}\Big(\tau\|\mathbf{y}\|^2 + \frac{1}{\tau}\Big)\Big\}$ for the norm. Then

$$\tilde{\ell}_n(\mathbf{Q},\mathbf{G},\mathbf{T}) = \min_{\boldsymbol{\omega},w,v} \max_{\phi} \inf_{\boldsymbol{\xi},\tau} \Big\{ \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\xi}_i,\boldsymbol{\omega},\Delta_i,T_i) - \Delta_i\log\lambda(T_i|\boldsymbol{\omega}) + \\ +\frac{1}{2\tau}\big\|\boldsymbol{\xi} - w\mathbf{Z}_0 - v\mathbf{Q}\big\|^2 + \phi(\tau/2 - v\|\mathbf{G}\|) + \frac{1}{2}\eta(v^2+w^2) + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \Big\}.$$

Taking the re-scaling $\tau\to\sqrt{n}\tau,\ \phi\to\phi/\sqrt{n}$ we recognize the (averaged) Moreau envelope (as defined in the main previous section)

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta_i,T_i)}\big(wZ_{0,i} + vQ_i, \tau/\phi\big) = \\ \frac{1}{n}\min_{\boldsymbol{\xi}}\Big\{\sum_{i=1}^{n} g(\boldsymbol{\xi}_i,\boldsymbol{\omega},\Delta_i,T_i) + \frac{\phi}{2\tau}\big\|\boldsymbol{\xi} - w\mathbf{Z}_0 - v\mathbf{Q}\big\|^2\Big\}. \tag{4.33}$$

Hence, we are finally left with the saddle point problem

$$\tilde{\ell}_n(\mathbf{Q},\mathbf{G},\mathbf{T}) = \min_{\boldsymbol{\omega},w,v}\max_{\phi}\inf_{\tau}\mathcal{L}_n(\boldsymbol{\omega},w,v,\phi,\tau) \tag{4.34}$$

where

$$\mathcal{L}_n(\boldsymbol{\omega}, w, v, \phi, \tau) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{M}_{g(.,\boldsymbol{\omega},\Delta_i,T_i)}\big(wZ_{0,i} + vQ_i, \tau/\phi\big) - \Delta_i \log\lambda(T_i|\boldsymbol{\omega}) + \\ +\phi\big(\tau/2 - v\|\mathbf{G}\|/\sqrt{n}\big) + \frac{1}{2}\eta(v^2+w^2) + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \ .$$

The weak law of large numbers and a concentration argument for $\|\mathbf{G}\|$ imply, see appendix 4.8.2, that

$$\mathcal{L}_n(\boldsymbol{\omega}, w, v, \phi, \tau) \xrightarrow[n\to\infty]{P} \mathcal{L}(\boldsymbol{\omega}, w, v, \phi, \tau) \tag{4.35}$$

pointwise for $\|\boldsymbol{\omega}\| \leq C_{\boldsymbol{\omega}}, 0 \leq w \leq C_{\boldsymbol{\beta}}, 0 \leq v \leq C_{\boldsymbol{\beta}}, \phi \geq 0$ and $\tau > 0$, where

$$\begin{aligned}\mathcal{L}(\boldsymbol{\omega}, w, v, \phi, \tau) \ &:= \ \mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0 + vQ, \tau/\phi\big) - \Delta\log\lambda(T|\boldsymbol{\omega})\Big] + \\ &+ \ \phi\big(\tau/2 - v\sqrt{\zeta}\big) + \frac{1}{2}\eta(v^2+w^2) + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \ .\end{aligned} \tag{4.36}$$

The pointwise convergence in probability above, together with the fact that the functions $\mathfrak{L}_n, \mathcal{L}$ are concave in ϕ and convex in $\boldsymbol{\omega}, w, v, \xi, \tau$ imply, see appendix 4.8.3, that

$$\min_{\boldsymbol{\omega},w,v}\max_{\phi>0}\inf_{\tau>0} \mathcal{L}_n(\boldsymbol{\omega}, w, v, \phi, \tau) \xrightarrow[n\to\infty]{P} \min_{w,v}\max_{\phi>0}\inf_{\tau>0} \mathcal{L}(\boldsymbol{\omega}, w, v, \phi, \tau) \ . \tag{4.37}$$

By the CGMT, we then have that

$$\min_{\boldsymbol{\beta},\boldsymbol{\omega}}\Big\{\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i'\boldsymbol{\beta}, T_i) + \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2\Big\} \xrightarrow[n\to\infty]{P} \min_{\boldsymbol{\omega},w,v}\max_{\phi>0}\min_{\tau>0} \mathcal{L}(w, v, \phi, \tau) \ . \tag{4.38}$$

Furthermore the pointwise convergence (4.35) and the strict convexity of $\mathcal{L}$ in $w, v, \boldsymbol{\omega}$ implies the convergence in probability of the minimizer (4.15,4.16,4.17) via (4.31) see appendix 4.8.4.

4.6 Numerical experiments

In the following we simulate the model under study and compare various quantities, such as goodness of fit and prediction metrics, against the theory for different values of the regularizer η which controls the amount of ridge shrinking on $\hat{\boldsymbol{\beta}}_n$. The data simulations are carried out as follows. We take the sample size $n = 400$ fixed and vary the number of covariates p via ζ as $p = \zeta n$. The latent survival time are generated from a Log-logistic proportional hazard model

$$Y_i|\mathbf{X}_i \sim -\frac{\mathrm{d}}{\mathrm{d}t}S_0(t|\mathbf{X}_i), \tag{4.39}$$

$$S_0(t|\mathbf{X}_i) = \exp\{-\Lambda_0(.)\exp(\mathbf{X}_i'\boldsymbol{\beta}_0)\}, \ \Lambda_0(t) := \log\Big(1 + t^2/2\Big) \tag{4.40}$$

where $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ and $\boldsymbol{\beta}_0 = \mathbf{e}_1$. With these choices the expected fraction of censored events is 40%, i.e. only 60% of the n subjects experiences the event on average. We also take $\boldsymbol{\beta}_0 \sim \mathbb{S}_{p-1}$, the true associations are drawn at random from the unit sphere in $\mathbb{R}^p$. Notice that our choice of $\boldsymbol{\beta}_0$ does not imply any loss in generality since the distribution of $\mathbf{X}$ and the ridge penalty are rotationally invariant, hence any $\boldsymbol{\beta}_0$ with the same length will yield the same statistics for the population of Y_i.

For each value of ζ (or equivalently p) we simulated 50 datasets and computed the penalized maximum likelihood estimator $\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\omega}}_n$ by numerical minimization of (4.9) along a regularization path for η at fixed $\alpha = 0.01$. We used a variant of the algorithm proposed in [32]. In all the plots that follow the markers are empirical averages, while the errorbars are empirical standard deviations computed over these realizations. We chose to use 11 time points $\tau_{k=1}^{\ell}$ equi-spaced between 0 and 3 (the end of the study) as the knots of the piece-wise parametrization of the hazard rate, this allows for a flexible approximation of the hazard rate. For the solution of the RS equations, we computed numerically the solution via fixed point iteration, to a tolerance of 1.0^{-8}. The expectations in (4.18,4.19,4.20,4.22) are approximated as population averages, with a population size $m = 2 \cdot 10^3$.

Before proceeding to the examine the results of the numerical experiments, we first need to introduce which metrics were employed to score the predictive ability of the model. The following subsection deals with this.

4.6.1 Evaluation metrics for Survival Predictions

In order to evaluate prediction accuracy in survival analysis, it is generally advised to always consider at least two metrics: one for the calibration and another for the discrimination ability of the model. Calibration, quoting from [33], "refers to the agreement between observed outcome and predictions". Discrimination is the ability of the model to separate individuals with different risks scores. The rationale being that a good model should assign shorter survival times to subjects with higher risk score. To evaluate the discriminative ability of the model, we used Harrell's c index [17, 16]

$$HC_n = \frac{\sum_{i=1}^{n} \Delta_i \sum_{j=1}^{n} \Theta(T_j - T_i) \mathbf{1}\Big[f(\mathbf{X}_j) > f(\mathbf{X}_i)\Big]}{\sum_{i=1}^{n} \Delta_i \sum_{j=1}^{n} \Theta(T_j - T_i)} \ . \tag{4.41}$$

This quantity is close to one if the model has perfect discrimination ability. Random guessing would give a c-index of 0.5. Instead of evaluating the calibration of the model directly via the Brier Score, we compute the following

$$IBS_{ideal} = \mathbb{E}_{\mathbf{X}}\left[\int \Big(\hat{S}(t|\mathbf{X}) - S_0(t|\mathbf{X})\Big)^2 \mathrm{d}t \right] . \tag{4.42}$$

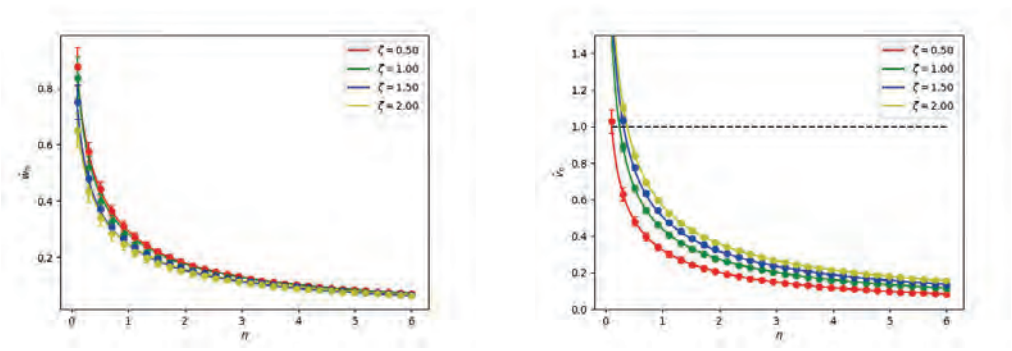
This quantity measures the integrated Mean Squared Error of the estimated survival function with respect to the true one. The "ideal" in underscore is due to the fact that the quantity above is what any estimator of the Integrated Brier Score (IBS) aims ideally to estimate. The ideal IBS (4.42) can only be computed for simulated data because in that case we know the $S_0(.|\mathbf{X})$ used to generate the data. Notice that (4.42) can only be used to score the model relative to another one, hence we prefer to use the ratio

$$R_{IBS} = \frac{\mathbb{E}_{\mathbf{X}}\left[\int \Big(\hat{S}(t|\mathbf{X}) - S_0(t|\mathbf{X})\Big)^2 \mathrm{d}t \right]}{\mathbb{E}_{\mathbf{X}}\left[\int \Big(\hat{S}_{null}(t) - S_0(t|\mathbf{X})\Big)^2 \mathrm{d}t \right]} \tag{4.43}$$

where $\hat{S}_{null}$ is the estimated survival function when no covariates are included in the model. Thanks to corollary (1) we can easily compute the metrics above for the test set via the solution of the RS equations (see theorem 2).

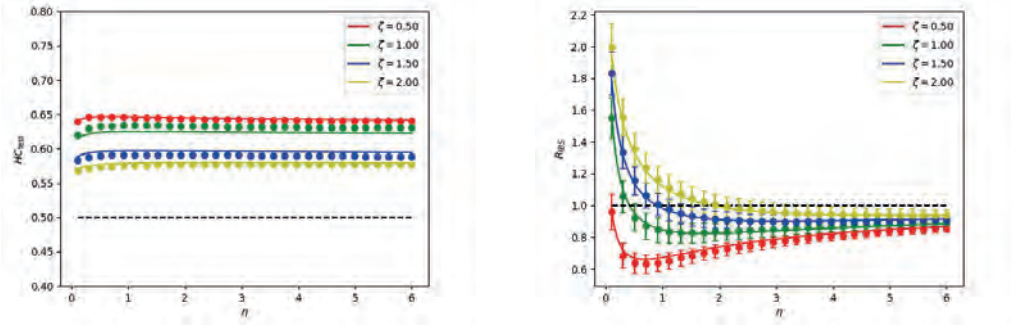
4.6.2 Comparison of theory and simulations

Figures (4.1,4.2) show the comparison between the theoretical values obtained by solving the Replica Symmetric equations (i.e. the non-linear system in corollary 2) as solid lines and the results obtained by the simulations as markers with errorbars. As expected, because of the convergence in probability established in (2), the solid line describe accurately the position of the markers for different values of the ratio ζ.



(a) (b)

Figure 4.1: Simulated data (markers and errorbars) against the theory (solid lines). Figures (4.1a,4.1b) show the value of $\hat{w}_n$ and $\hat{v}_n$ defined in (4.15, 4.16) along a regularization path for $\eta \in (0.1, 6)$ with $\alpha = 0.01$.



(a) (b)

Figure 4.2: Simulated data (markers and errorbars) against the theory (solid lines): (left) the Test c-index; (right) R_{IBS} as defined in (4.43), along a regularization path for $\eta \in (0.1, 6)$ with $\alpha = 0.01$. The errorbars for the figure (4.2a) have been removed to aid visualization.

We notice that the test c-index is virtually independent from the penalization strength as one can see in (4.2a) and its maximal value decreases with ζ. Whilst from figure (4.2b) we deduce that as ζ increases, the minimum of R_{IBS} is attained at a larger value of η, i.e. more regularization is needed just to do slightly better than the null model.

4.7 Conclusion

In conclusion, we have applied the Gaussian Convex Min Max theorem to a flexible parametric model for survival data, i.e. the piece-wise exponential model. With the theoretical guarantees obtained, we investigated the effect of the ridge penalization onto the predictive ability of the model. This represents a first step into the full formalization of a set of previous results obtained for the Cox model [7, 26], in the sense that our final goal will be to generalize the present proof to the Cox semi parametric model. The proof technique used here relies completely on the assumption of Gaussian covariates, since it hinges on the Convex Gaussian Min Max theorem [36]. This approach allows us to rigorously prove heuristic (but, in the end, exact) results obtained via the Replica Method of statistical physics [7], when the parametrization of the base hazard rate is finite dimensional. Further work is required to rigorously establish more recent results for the semi-parametric case [26]. What is interesting is that the heuristic way actually requires only that the linear predictor follows (asymptotically) a Normal law. So we expect that these results holds in more general scenarios that the limited one studied here. It would be ideal to be able to put these several conjectures on a firm theoretical basis and indeed recent investigations seek to do this, at least in some specific settings [15, 18, 28].

Supporting Information

Additional simulations, together with the routines used to compute the estimators, solve the Replica Symmetric equations and plot the figures of the paper are available in GitHub at `https://github.com/EmanueleMassa/SMSA_2024`.

Bibliography

[1] PK Andersen and RD Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[2] J Bradic, J Fan, and J Jiang. Regularization for cox's proportional hazards model with np-dimensionality. *The Annals of Statistics*, 39(6):3092–3120, 2011.

[3] NE Breslow. Discussion on professor cox's paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):202–220, 1972.

[4] NE Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45–57, 1975.

[5] P Bühlmann and S van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg, 2011.

[6] SR Cole and GH Michael. Survival analysis in infectious disease research: describing events in time. *AIDS (London, England)*, 24, 2010.

[7] ACC Coolen, JE Barrett, P Paga, and CJ Perez-Vicente. Replica analysis of overfitting in regression models for time-to-event data. *Journal of Physics A: Mathematical and Theoretical*, 50(37):375001, aug 2017.

[8] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[9] DL Donoho and A Montanari. High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969, 2013.

[10] N El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170:95–175, 2018.

[11] N El Karoui, D Bean, PJ Bickel, C Lim, and B Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[12] M Friedman. Piecewise Exponential Models for Survival Data with Covariates. *The Annals of Statistics*, 10(1):101 – 113, 1982.

[13] E Gardner and B Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, jan 1988.

[14] Y Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50, 1985.

[15] Q Han and Y Shen. Universality of regularized regression estimators in high dimensions. *The Annals of Statistics*, 51(4):1799 – 1823, 2023.

[16] FE Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Graduate Texts in Mathematics. Springer, 2001.

[17] FE Harrell, RM Califf, DB Pryor, KL Lee, and RA Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982.

[18] H Hu and YM Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69:1932–1964, 2020.

[19] JD Kalbfleisch and RL Prentice. Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2):267–278, 08 1973.

[20] JD Kalbfleisch and RL Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley, 2011.

[21] JP Klein and ML Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer New York, 2005.

[22] H Kvamme and Ø Borgan. Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 2022.

[23] H Kvamme, Ø Borgan, and I Scheel. Time-to-event prediction with neural networks and cox regression. *J. Mach. Learn. Res.*, 20:129:1–129:30, 2019.

[24] J Lederer. *Fundamentals of High-Dimensional Statistics: With Exercises and R Labs*. Springer Texts in Statistics. Springer International Publishing, 2021.

[25] B Loureiro, C Gerbelot, H Cui, S Goldt, F Krzakala, M Mézard, and L Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114001, nov 2022.

[26] E Massa, A Mozeika, and ACC Coolen. Replica analysis of overfitting in regression models for time to event data: the impact of censoring. *Journal of Physics A: Mathematical and Theoretical*, 57(12):125003, mar 2024.

[27] L Miolane and A Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313 – 2335, 2021.

[28] A Montanari and B Saeed. Universality of empirical risk minimization. *ArXiv*, abs/2202.08832, 2022.

[29] J Ramjith, C Andolina, T Bousema, and MA Jonker. Flexible time-to-event models for double-interval-censored infectious disease data with clearance of the infection as a competing risk. *Frontiers in Applied Mathematics and Statistics*, 8, 2022.

[30] RT Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1997.

[31] K Shengchun and B Nan. Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Statistica Sinica*, 2014.

[32] NR Simon, JH Friedman, TJ Hastie, and R Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39 5:1–13, 2011.

[33] EW Steyerberg, AJ Vickers, NR Cook, T Gerds, M Gonen, N Obuchowski, MJ Pencina, and MW Kattan. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 2010.

[34] P Sur and EJ Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[35] C Thrampoulidis, E Abbasi, and B Hassibi. Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[36] C Thrampoulidis, S Oymak, and B Hassibi. The gaussian min-max theorem in the presence of convexity, 2015.

[37] R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[38] AA Tsiatis. A large sample study of cox's regression model. *The Annals of Statistics*, 9(1):93 – 108, 1981.

[39] S van de Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[40] SA van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614 – 645, 2008.

[41] A Van der Vaart and J Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York, 2013.

[42] AW Van Der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000.

[43] MJ Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

4.8 Appendix

4.8.1 Properties of the Moreau envelope for the Piece-wise exponential model

In the following we assume that $\|\boldsymbol{\omega}\| \leq C_{bomega}$, $w, v \leq C_{\boldsymbol{\beta}}$ and $\tau > 0$, furthermore $\tau_1 < \tau_2 < \cdots < \tau_{\ell+1} < \infty$.

Proposition 1 (Integrability)**.** *The random function* $\boldsymbol{\omega}, w, v, \tau \to \mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0 + vQ, \tau) - \Delta \log \lambda(T|\boldsymbol{\omega})$ *is absolutely integrable.*

Proof. By definition

$$\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0 + vQ, \tau) \leq \frac{1}{2\tau}\big\|wZ_0 + vQ\big\|^2 + g(0, \boldsymbol{\omega}, \Delta, T) \ . \tag{4.44}$$

hence

$$\min_x g(x, \boldsymbol{\omega}, \Delta, T) \leq \mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0 + vQ, \tau) \leq \tag{4.45}$$

$$\frac{\lambda}{2\tau}\big\|wZ_0 + vQ\big\|^2 + g(0, \boldsymbol{\omega}, \Delta, T) \ . \tag{4.46}$$

This implies that

$$\begin{aligned}
&\Big|\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0 + vQ, \tau)\Big| \leq \\
&\qquad \max\Big\{\frac{\lambda}{2\tau}\big\|wZ_0 + vQ\big\|^2 + \big|g(0, \boldsymbol{\omega}, \Delta, T)\big|, \big|\min_x g(x, \boldsymbol{\omega}, \Delta, T)\big|\Big\} \\
&\qquad \leq \frac{\lambda}{2\tau}\big\|wZ_0 + vQ\big\|^2 + \big|g(0, \boldsymbol{\omega}, \Delta, T)\big| + \big|\min_x g(x, \boldsymbol{\omega}, \Delta, T)\big| \ .
\end{aligned} \tag{4.47}$$

Now we show that the expectation of the right hand side exists finite. For the first term we have $\mathbb{E}\Big[\big\|wZ_0 + vQ\big\|^2\Big] \leq w^2 + v^2$, which is bounded by hypothesis. For the second term, we notice that

$$\Lambda(T|\boldsymbol{\omega}) := \sum_{k=1}^{\ell} \exp(\omega_k)\Psi_k(T) \leq \sum_{k=1}^{\ell} \exp(\omega_k)(\tau_{k+1} - \tau_k) \tag{4.48}$$

hence, via Cauchy Swartz inequality

$$\mathbb{E}\Big[\big|g(0,\boldsymbol{\omega},\Delta,T)\big|\Big] = \mathbb{E}\Big[\Big|\Delta\Lambda(T|\boldsymbol{\omega})\Big|\Big] \leq \sum_{k=1}^{\ell}\exp(\omega_k)(\tau_{k+1}-\tau_k) \leq$$
$$\leq \sqrt{\sum_{k=1}^{\ell}\exp(2\omega_k)}\sqrt{\sum_{k=1}^{\ell}(\tau_{k+1}-\tau_k)^2} \tag{4.49}$$

which is bounded by a finite constant by assumption that $\boldsymbol{\omega}$ is in a compact set and $\tau_{\ell+1}$ is finite. Then

$$\big|\Delta\log\lambda(T|\boldsymbol{\omega})\big| \leq \sum_{k=1}^{\ell}\psi_k(T)\big|\omega_k\big| \tag{4.50}$$

which has a finite expectation value, since $\psi_k(T)$ is an indicator function over the interval $[\tau_k,\tau_{k+1}]$. Finally we have

$$\min_x g(x,\boldsymbol{\omega},\Delta,T) = \Delta\big(1-\log(\Delta)+\log\Lambda(T|\boldsymbol{\omega})\big) \tag{4.51}$$

and we see that

$$\Big|\Delta\big(1-\log\Delta+\log\Lambda(T|\boldsymbol{\omega})\big)\Big| \leq 1+\Big|\log\Lambda(T|\boldsymbol{\omega})\Big| \leq$$
$$1+\Big|\log\sum_{k=1}^{l}\exp(\omega_k)(\tau_{k+1}-\tau_k)\Big| < \infty\ . \tag{4.52}$$

Since $\|\boldsymbol{\omega}\|$ is bounded by a large constant by assumption. □

Proposition 2 (Finite variance)**.** *The random function* $\boldsymbol{\omega},w,v,\tau \to \mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0+vQ,\tau)-\Delta\log\lambda(T|\boldsymbol{\omega})$ *has a finite variance.*

Proof. We compute the second moment, as we have already shown that the expectation exists finite. Using (4.47) we have

$$\Big(\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0+vQ,\tau)-\Delta\log\lambda(T|\boldsymbol{\omega})\Big)^2 \leq \tag{4.53}$$
$$\Big(\frac{\lambda}{2\tau}\big\|wZ_0+vQ\big\|^2+\big|g(0,\boldsymbol{\omega},\Delta,T)\big|+\big|\min_x g(x,\boldsymbol{\omega},\Delta,T)\big|\Big)^2 \tag{4.54}$$

Computing the square we see that:
1) $\mathbb{E}\Big[\big\|wZ_0+vQ\big\|^4\Big] = w^4\mathbb{E}\Big[Z_0^4\Big]+v^4\mathbb{E}\Big[Q^4\Big]$ is finite,
2) $\mathbb{E}\Big[\big|g(0,\boldsymbol{\omega},\Delta,T)\big|^2\Big] \leq \big(\sum_{k=1}^{\ell}\exp(\omega_k)(\tau_{k+1}-\tau_k)\big)^2 \leq$
$\big(\sum_{k=1}^{\ell}\exp(2\omega_k)\big)\big(\sum_{k=1}^{\ell}(\tau_{k+1}-\tau_k)^2\big)$ which is finite,
3) the term $\min_x g(x,\boldsymbol{\omega},\Delta,T)$ is bounded by a deterministic constant (4.52),and hence so it is its square. The cross terms can be similarly bounded via the Cauchy-Schwartz inequality. □

Proposition 3 (Finite variance of the derivative)**.** *The random function* $\boldsymbol{\omega},w,v,\tau \to \dot{g}(.,\boldsymbol{\omega},\Delta,T)(wZ_0+vQ,\tau)$ *has a finite variance.*

Proof. Using (4.48)

$$\Big|\dot{g}(.,\boldsymbol{\omega},\Delta,T)(wZ_0+vQ,\tau)\Big| = \Big|\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{wZ_0+vQ}-\Delta\Big| =$$
$$\leq 1+\mathrm{e}^{wZ_0+vQ}\sqrt{\sum_{k=1}^{\ell}\exp(2\omega_k)}\sqrt{\sum_{k=1}^{\ell}(\tau_{k+1}-\tau_k)^2} \tag{4.55}$$

which has a finite expectation. We then show that the second moment exists finite. Just notice that

$$\Big|\dot{g}(.,\boldsymbol{\omega},\Delta,T)(wZ_0+vQ,\tau)\Big|^2 \leq$$
$$1+2\mathrm{e}^{wZ_0+vQ}\Big(\sum_{k=1}^{l}\exp(2\omega_k)\Big)^{1/2}\Big(\sum_{k=1}^{\ell}(\tau_{k+1}-\tau_k)^2\Big)^{1/2}+$$
$$+\mathrm{e}^{2(wZ_0+vQ)}\sum_{k=1}^{\ell}\exp(2\omega_k)\sum_{k=1}^{\ell}(\tau_{k+1}-\tau_k)^2 \tag{4.56}$$

has a finite expectation □

Proposition 4 (Limits of the Moreau envelope)**.**

$$\lim_{\tau\to\infty}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(x,\tau) = \Delta\big(1-\log(\Delta)+\log\Lambda(T|\boldsymbol{\omega})\big) \tag{4.57}$$
$$\lim_{\tau\to 0}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(x,\tau) = \Lambda(T|\boldsymbol{\omega})\mathrm{e}^x-\Delta x\ . \tag{4.58}$$

Proof. For (4.58), observe that

$$\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(x,\alpha\big) := \min_{\xi}\Big\{\frac{1}{2\alpha}(\xi-x)^2+g(\xi,T)\Big\} =$$
$$\min_{\phi}\Big\{\frac{1}{2}\phi^2+g(x+\sqrt{\alpha}\phi,\boldsymbol{\omega},\Delta,T)\Big\} \tag{4.59}$$

with $\hat{\phi}=\arg\min_{\phi}\Big\{\frac{1}{2}\phi^2+g(x+\sqrt{\alpha}\phi,\boldsymbol{\omega},\Delta,T)\Big\}=-\sqrt{\alpha}\dot{g}(x+\sqrt{\alpha}\hat{\phi},\boldsymbol{\omega},\Delta,T)$. Sending $\alpha\to 0^+$ we get (4.58). For (4.57), notice that

$$\frac{\partial}{\partial\alpha}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(x,\alpha\big) = -\frac{1}{2\alpha^2}\big(\mathrm{prox}_{g(.,\boldsymbol{\omega},\Delta,T)}(x,\alpha)-x\big)^2\leq 0\ . \tag{4.60}$$

Then

$$\lim_{\alpha\to\infty}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(x,\alpha\big) = \inf_{\alpha>0}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(x,\alpha\big) = \tag{4.61}$$
$$\min_{\xi}\inf_{\alpha>0}\Big\{\frac{1}{2\alpha}(\xi-x)^2+g(\xi,\boldsymbol{\omega},\Delta,T)\Big\} = \min_{\xi}\ g(\xi,\boldsymbol{\omega},\Delta,T)\ .$$

□

Proposition 5 (Limits of the Expected Moreau envelope)**.**

$$\lim_{\alpha\to\infty}\mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0+vQ,\alpha\big)\Big] =$$
$$\mathbb{E}_T\Big[\Delta\big(1-\log\Delta+\log\Lambda(T|\boldsymbol{\omega})\big)\Big] \tag{4.62}$$
$$\lim_{\alpha\to 0^+}\mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0+vQ,\alpha\big)\Big] =$$
$$\mathbb{E}_{T,Z_0,Q}\Big[g\big(wZ_0+vQ,T\big)\Big]\ . \tag{4.63}$$

Proof. For the limit of the expectation we use the Dominated Convergence theorem. The Moreau envelope is integrable (proposition 1) and its limits exist (proposition 4). Hence it suffices to show that the limits are themselves integrable. First for (4.62) we see that

$$\mathbb{E}\Big[\Big|\Delta\big(1-\log\Delta+\log\Lambda(T|\boldsymbol{\omega})\big)\Big|\Big] \leq 1+\mathbb{E}\Big[\Big|\log\Lambda(T|\boldsymbol{\omega})\Big|\Big] \leq$$
$$1+\Big|\log\sum_{k=1}^{\ell}\exp(\omega_\nu)(\tau_{k+1}-\tau_k)\Big| < \infty\ . \tag{4.64}$$

For (4.63), notice that

$$\mathbb{E}\Big[\Big|\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{wZ_0+vQ}-\Delta(wZ_0+vQ)\Big|\Big] \leq$$
$$\mathbb{E}\Big[\Big|\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{wZ_0+vQ}\Big|\Big]+\mathbb{E}\Big[\Big|\Delta(wZ_0+vQ)\Big|\Big] \tag{4.65}$$

and

$$\mathbb{E}\Big[\Big|\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{wZ_0+vQ}\Big|\Big] \leq \sum_{k=1}^{\ell}\exp(\omega_\nu)(\tau_{k+1}-\tau_k)\mathrm{e}^{2(w^2+v^2)} < \infty \tag{4.66}$$
$$\mathbb{E}\Big[\Big|\Delta(wZ_0+vQ)\Big|\Big] \leq \mathbb{E}\Big[\Big|wZ_0+vQ\Big|\Big] \leq \sqrt{2/\pi}\sqrt{w^2+v^2} < \infty \tag{4.67}$$

which implies the desiderata. □

Proposition 6 (Derivatives of the Moreau envelope 1)**.**

$$\frac{\partial}{\partial x}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(x,\tau) = \frac{1}{\tau}\big(x-\mathrm{prox}_{g(.,\boldsymbol{\omega},\Delta,T)}(x,\tau)\big) =$$
$$\Delta-\frac{1}{\tau}W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+x}\Big) \tag{4.68}$$
$$\frac{\partial}{\partial\tau}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(x,\tau) = -\frac{1}{2\tau^2}\big(x-\mathrm{prox}_{g(.,\boldsymbol{\omega},\Delta,T)}(x,\tau)\big)^2 =$$
$$-\frac{1}{2}\Big(\Delta-\frac{1}{\tau}W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+x}\Big)\Big)^2\ . \tag{4.69}$$

Proof. The proposition follows from differentiation and definition of proximal mapping. □

Proposition 7 (Derivatives of the expected Moreau envelope 1)**.**

$$\frac{\partial}{\partial w}\mathbb{E}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0+vQ,\tau)\Big] =$$
$$\frac{1}{\tau}\mathbb{E}\Big[Z_0\Big(\Delta\tau-W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+x}\Big)\Big)\Big] \tag{4.70}$$
$$\frac{\partial}{\partial v}\mathbb{E}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0+vQ,\tau)\Big] =$$
$$\frac{1}{\tau}\mathbb{E}\Big[Q\Big(\Delta\tau-W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+x}\Big)\Big)\Big] \tag{4.71}$$
$$\frac{\partial}{\partial\tau}\mathbb{E}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0+vQ,\tau)\Big] =$$
$$-\frac{1}{2\tau^2}\mathbb{E}\Big[\Big(\Delta\tau-W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+x}\Big)\Big)^2\Big]\ . \tag{4.72}$$

Proof. We show that the conditions of the Dominated converge theorem of Lebesgue are satisfied simultaneously for all derivatives. Via proposition 6 we see that

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial w}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\tau)\Big|\Big] \leq \frac{1}{\tau}\mathbb{E}\Big[Z_0^2\Big]^{1/2}\mathbb{E}\Big[(\Delta\tau - W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)^2\Big]^{1/2} \tag{4.73}$$

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial v}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\tau)\Big|\Big] \leq \frac{1}{\tau}\mathbb{E}\Big[Q^2\Big]^{1/2}\mathbb{E}\Big[(\Delta\tau - W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)^2\Big]^{1/2} \tag{4.74}$$

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial \tau}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\tau)\Big|\Big] = \frac{1}{2\tau^2}\mathbb{E}\Big[(\Delta\tau - W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)^2\Big]\ . \tag{4.75}$$

Thus it is sufficient to show that the last term is bounded to show integrability of all the first derivatives.
Since $\tau > 0$ and for $x \geq 0$ it holds $W_0(x) \geq 0$, we have that

$$\Big(\Delta\tau - W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)\Big)^2 \leq \tau^2 + W_0^2\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)\ . \tag{4.76}$$

We now use the bound $W_0(x) \leq \log(x+1) \leq x$ which is valid for $x \geq 0$, obtaining

$$\frac{1}{\tau^2}\mathbb{E}\left[\left(\Delta\tau - W_0\Big(\tau\Lambda(T)\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)\right)^2\right] \leq 1 + 1\mathbb{E}\Big[\Lambda^2(T|\boldsymbol{\omega})\mathrm{e}^{2(\Delta\tau+wZ_0+vQ)}\Big]\ . \tag{4.77}$$

which can be shown to be finite using(4.48), since $\|\boldsymbol{\omega}\|, w, v$ are bounded by a constant, $\tau > 0$ and τ_ℓ is finite by hypothesis. □

Proposition 8 (Partial derivatives of the expected Moreau envelope 2)**.**

$$\frac{\partial}{\partial\omega_k}\mathbb{E}\left[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0+vQ,\tau)\right] = \mathbb{E}\Big[\Psi_k(T)\mathrm{e}^{\mathrm{prox}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0+vQ,\tau)}\exp(\omega_k)\Big] \tag{4.78}$$

$$\frac{\partial}{\partial\omega_k}\mathbb{E}\Big[\Delta\log\lambda(T|\boldsymbol{\omega})\Big] = \mathbb{E}\Big[\Delta\psi_k(T)\Big] \tag{4.79}$$

Proof. We show that the conditions of the dominated convergence theorem are satisfied. First we compute

$$\begin{aligned}\frac{\partial}{\partial\omega_k}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}(wZ_0+vQ,\tau) &= \Psi_k(T)\exp(\omega_k)\frac{W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)}{\tau\Lambda(T|\boldsymbol{\omega})}\\ \frac{\partial}{\partial\omega_k}\Delta\log\lambda(T|\boldsymbol{\omega}) &= \Delta\psi_k(T)\end{aligned} \tag{4.80}$$

where we used that

$$\exp\Big\{\text{prox}_{g(.,T)}(wZ_0+vQ,\tau)\Big\} =$$
$$= \exp\Big\{wZ_0+vQ+\Delta\tau - W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)\Big\} =$$
$$\frac{W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+x}\Big)}{\tau\Lambda(T|\boldsymbol{\omega})}\ .$$

Notice that

$$\mathbb{E}\left[\Psi_k(T)\exp(\omega_k)\frac{W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)}{\tau\Lambda(T|\boldsymbol{\omega})}\right] \leq$$
$$(\tau_{k+1}-\tau_k)\mathbb{E}_{Z_0,Q}\left[\frac{W_0\Big(\tau\Lambda(T|\boldsymbol{\omega})\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big)}{\tau\Lambda(T|\boldsymbol{\omega})}\right]$$
$$\leq (\tau_{k+1}-\tau_k)\mathbb{E}_{Z_0,Q}\Big[\mathrm{e}^{\Delta\tau+wZ_0+vQ}\Big], \tag{4.81}$$

where for the last inequality we used that $W_0(x)\leq\log(x+1)\leq x$, for $x\geq 0$. This implies that the above exists finite. Furthermore

$$\mathbb{E}\Big[\Delta\psi_k(T)\Big] \leq \mathbb{E}\Big[\psi_k(T)\Big] \tag{4.82}$$

which is finite since ψ_k is the indicator function of the interval (τ_k,τ_{k+1}).

□

4.8.2 Pointwise convergence of $\mathcal{L}_n$ to $\mathcal{L}$

Proposition 9. *Let*

$$\mathcal{L}_n(\boldsymbol{\omega},w,v,\phi,\tau) := \frac{1}{n}\sum_{i=1}^{n}\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta_i,T_i)}\big(wZ_{0,i}+vQ_i,\tau/\phi\big) - \Delta_i\log\lambda(T_i|\boldsymbol{\omega}) +$$
$$+\phi\big(\tau/2 - v\|\mathbf{G}\|/\sqrt{n}\big) + \frac{1}{2}\eta(v^2+w^2) + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2 \tag{4.83}$$

and

$$\mathcal{L}(\boldsymbol{\omega},w,v,\phi,\tau) := \mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0+vQ,\tau/\phi\big)\Big] - \mathbb{E}\Big[\Delta\log\lambda(T|\boldsymbol{\omega})\Big] +$$
$$+\phi\big(\tau/2 - v\sqrt{\zeta}\big) + \frac{1}{2}\eta(v^2+w^2) + \frac{1}{2}\alpha\|\boldsymbol{\omega}\|^2\ . \tag{4.84}$$

Then $\mathcal{L}_n(w,v,\boldsymbol{\omega},\phi,\tau)\xrightarrow[n\to\infty]{P}\mathcal{L}(w,v,\boldsymbol{\omega},\phi,\tau)$ *pointwise in* $w,v,\tau,\boldsymbol{\omega},\phi$.

Proof. Since $\|\mathbf{G}\|$ follows a chi distribution with $p-1$ degrees of freedom

$$\mathbb{E}\Big[\|\mathbf{G}\|\Big] = \sqrt{2}\frac{\Gamma(\frac{p}{2})}{\Gamma(\frac{p-1}{2})} = \sqrt{p-2}+o_p(1), \tag{4.85}$$
$$\mathbb{V}\Big[\|\mathbf{G}\|^2\Big] = p-1-\mathbb{E}\Big[\|\mathbf{G}\|\Big]^2 = 1+o_p(1), \tag{4.86}$$

hence we conclude $\|\mathbf{G}\|/\sqrt{n} \xrightarrow[n\to\infty]{P} \sqrt{\zeta}$ (e.g. via Chebishev inequality). Furthermore, by the weak law of large numbers, given proposition (2), we have that

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{M}_{g(.,\boldsymbol{\omega},\Delta_i,T_i)}\big(wZ_{0,i}+vQ_i,\tau/\phi\big) - \Delta_i \log\lambda(T_i|\boldsymbol{\omega}) \xrightarrow[n\to\infty]{P} \tag{4.87}$$

$$\mathbb{E}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0+vQ,\tau/\phi\big) - \Delta\log\lambda(T|\boldsymbol{\omega})\Big]\ . \tag{4.88}$$

□

4.8.3 Asymptotic convergence of the saddle point of the AO

In this section we show that

$$\min_{\boldsymbol{\omega},w,v\geq 0}\ \max_{0\leq\phi\leq\lambda_{\max}}\ \min_{\tau>0}\ \mathcal{L}_n(\boldsymbol{\omega},w,v,\phi,\tau) \xrightarrow[n,p\to\infty]{P} \min_{\boldsymbol{\omega},w,v\geq 0}\ \max_{0\leq\phi\leq\lambda_{\max}}\ \min_{\tau>0}\ \mathcal{L}(\boldsymbol{\omega},w,v,\phi,\tau) \tag{4.89, 4.90}$$

The idea is to use the so-called *convexity lemma*, this guarantees that point-wise convergence of convex functions over compact sets implies uniform convergence over the latter.

Lemma 1 (Convexity lemma)**.** *Let $E\subset\mathbb{R}^d$ be open, convex and F_n a sequence of proper, convex functions and f a deterministic function, both defined on E, such that*

$$F_n(x) \xrightarrow{P} f(x), \forall\ x\in E\ . \tag{4.91}$$

Then F_n converges uniformly to f over all compact subsects of E, i.e.

$$\sup_{x\in A}\Big|F_n(x)-f(x)\Big| \xrightarrow{P} 0\ . \tag{4.92}$$

The proof can be found in [1] (page 1116 theorem 2.1). In particular we will use the following lemma, which is a consequence of the former.

Lemma 2 (Min-convergence – Open Sets)**.** *Consider a sequence of proper, convex stochastic functions $F_n:(0,\infty)\to\mathbb{R}$, and, a deterministic function $f:(0,\infty)\to\mathbb{R}$, such that:*

(a) $F_n(x)\xrightarrow{P} f(x),\ \forall\ x>0,$

(b) $\exists z>0\ :\ f(x)>\inf_{x>0} f(x),\ \forall x\geq z\iff \lim_{x\to\infty} f(x)=+\infty.$

Then

$$\inf_{x>0} F_n(x) \xrightarrow{P} \inf_{x>0} f(x)\ . \tag{4.93}$$

The proof can be found in [35] (lemma A6 page 29). The assumption (b) of lemma 2 above, is known as level-bounded condition. The following lemma, gives an equivalent characterization of a level bounded convex function.

Lemma 3 (Level bounded convex functions)**.** *Let $f:(0,\infty)\to\mathbb{R}$ be convex, then the following statements are equivalent:*

(a) $\exists z>0\ :\ f(x)>\inf_{x>0} f(x),\ \forall x\geq z,$

(b) $\lim_{x\to\infty} f(x)=+\infty.$

The proof can be found in [35] (lemma A7 page 29). We will proceed sequentially from the innermost operation to the outermost.

Minimization over τ

Fix $\boldsymbol{\omega}, w, v, \phi \geq 0$ and let us denote $\mathcal{L}_n^{\boldsymbol{\omega},w,v,\phi}(\tau) := \mathcal{L}_n(\boldsymbol{\omega}, w, v, \phi, \tau)$ and $\mathcal{L}^{\boldsymbol{\omega},w,v,\phi}(\tau) := \mathcal{L}(\boldsymbol{\omega}, w, v, \phi, \tau)$.

Proposition 10.

$$\inf_{\tau \geq 0} \mathcal{L}_n^{\boldsymbol{\omega},w,v,\phi}(\tau) \xrightarrow{P} \inf_{\tau > 0} \mathcal{L}^{\boldsymbol{\omega},w,v,\phi}(\tau) \ . \tag{4.94}$$

Proof. Since $\mathcal{L}_n^{\boldsymbol{\omega},w,v,\phi}(.)$ is convex, and $\mathcal{L}_n^{\boldsymbol{\omega},w,v,\phi}(.) \xrightarrow{P} \mathcal{L}^{\boldsymbol{\omega},w,v,\phi}(.)$ on $(0,\infty)$, also $\mathcal{L}^{\boldsymbol{\omega},w,v,\phi}(.)$ is convex. If $\phi > 0$, then $\lim_{\tau\to\infty} \mathcal{L}^{\boldsymbol{\omega},w,v,\phi}(\tau) = +\infty$, since $\phi\tau \to \infty$ ($\phi > 0$). Hence via lemma (3), we see that the conditions of lemma (2) are satisfied and the conclusion follows. If $\phi = 0$, then

$$\begin{aligned}\inf_{\tau \geq 0} \mathcal{L}_n^{\boldsymbol{\omega},w,v,\phi=0}(\tau) &= \lim_{\alpha\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_{0,i} + vQ_i, \alpha\big) - \Delta_i \log \lambda(T_i|\boldsymbol{\omega}) = \\ &= \frac{1}{n} \sum_{i=1}^{n} \min_{\xi} \ g(\xi, \boldsymbol{\omega}, \Delta_i, T_i) - \Delta_i \log \lambda(T_i|\boldsymbol{\omega}) \ . \end{aligned} \tag{4.95}$$

Under our assumptions also

$$\begin{aligned}\inf_{\tau \geq 0} \mathcal{L}^{\boldsymbol{\omega},w,v,\phi=0}(\tau) &= \lim_{\alpha\to\infty} \mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0 + vQ, \alpha\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] = \\ &= \mathbb{E}_T\Big[\min_{\xi} \ g(\xi, \boldsymbol{\omega}, \Delta, T) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] \end{aligned} \tag{4.96}$$

and via the weak law of large numbers we obtain the claim. □

Maximization over ϕ

Fix $\boldsymbol{\omega}, w, v$, let us denote $\mathcal{L}_n^{\boldsymbol{\omega},w,v}(\phi) := \inf_{\tau>0} \mathcal{L}_n(\boldsymbol{\omega}, w, v, \phi, \tau)$ and $\mathcal{L}^{\boldsymbol{\omega},w,v}(\phi) := \inf_{\tau>0} \mathcal{L}(\boldsymbol{\omega}, w, v, \phi, \tau)$.

Proposition 11.

$$\sup_{\phi \geq 0} \mathcal{L}_n^{\boldsymbol{\omega},w,v}(\phi) \xrightarrow{P} \sup_{\phi \geq 0} \mathcal{L}^{\boldsymbol{\omega},w,v}(\phi) \ . \tag{4.97}$$

Proof. Notice that $\mathcal{L}_n^{\boldsymbol{\omega},w,v}(\phi)$ is concave, as the pointwise minima of concave functions. The same is true for $\mathcal{L}^{\boldsymbol{\omega},w,v}(\phi)$ because $\mathcal{L}_n^{\boldsymbol{\omega},w,v}(.) \xrightarrow{P} \mathcal{L}_n^{\boldsymbol{\omega},w,v}(.)$ point wise.

If $\phi = 0$, then $\mathcal{L}_n^{\boldsymbol{\omega},w,v>0}(0) \xrightarrow{P} \mathcal{L}^{\boldsymbol{\omega},w,v>0}(0)$ for any v, as shown before. Else $\phi > 0$. We want to use again lemma (2): if we show that $\lim_{\phi\to\infty} \mathcal{L}_n^{\boldsymbol{\omega},w,v}(\phi) = -\infty$, then we are done, as the function is concave and level bounded. It holds by definition that

$$\mathcal{L}^{\boldsymbol{\omega},w,v}(\phi) \leq \lim_{\tau\to 0^+} \mathcal{L}(\boldsymbol{\omega}, w, v, \phi, \tau) \tag{4.98}$$

where

$$\begin{aligned}&\lim_{\tau\to 0^+} \mathcal{L}(\boldsymbol{\omega}, w, v, \phi, \tau) = \\ &\quad \lim_{\tau\to 0^+} \mathbb{E}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0 + vQ, \tau/\phi\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] - \phi v\sqrt{\zeta} = \\ &\quad \mathbb{E}\Big[g\big(wZ_0 + vQ, \boldsymbol{\omega}, \Delta, T\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] - \phi v\sqrt{\zeta} \ . \end{aligned} \tag{4.99}$$

If $v > 0$, then $\lim_{\phi\to\infty} \mathcal{L}^{\boldsymbol{\omega},w,v}(\phi) = -\infty$, since the expectation of g exists finite for fixed $\boldsymbol{\omega}, w, v$, and lemma (2) gives

$$\sup_{\phi\geq 0} \mathcal{L}_n^{\boldsymbol{\omega},w,v}(\phi) \xrightarrow{P} \sup_{\phi\geq 0} \mathcal{L}^{\boldsymbol{\omega},w,v}(\phi) \ . \tag{4.100}$$

On the other hand if $v = 0$, it is not straightforward to conclude the desiderata. Observe that

$$\begin{aligned} \lim_{\phi\to\infty} \inf_{\tau>0} \mathcal{L}(\boldsymbol{\omega}, w, 0, \phi, \tau) &\leq \sup_{\phi\geq 0} \inf_{\tau>0} \mathcal{L}(\boldsymbol{\omega}, w, 0, \phi, \tau) \\ &\leq \sup_{\phi\geq 0} \lim_{\tau\to 0^+} \mathcal{L}(\boldsymbol{\omega}, w, 0, \phi, \tau) \end{aligned} \tag{4.101}$$

and

$$\begin{aligned} \lim_{\phi\to\infty} \inf_{\tau>0} \mathcal{L}_n(\boldsymbol{\omega}, w, 0, \phi, \tau) \leq \sup_{\phi\geq 0} \inf_{\tau>0} \mathcal{L}_n(\boldsymbol{\omega}, w, 0, \phi, \tau) \leq \\ \sup_{\phi\geq 0} \lim_{\tau\to 0^+} \mathcal{L}_n(\boldsymbol{\omega}, w, 0, \phi, \tau) \ . \end{aligned} \tag{4.102}$$

We now show that

$$\begin{aligned} \lim_{\phi\to\infty} \inf_{\tau>0} \mathcal{L}_n(\boldsymbol{\omega}, w, 0, \phi, \tau) \xrightarrow{P} \lim_{\phi\to\infty} \inf_{\tau>0} \mathcal{L}(\boldsymbol{\omega}, w, 0, \phi, \tau) \\ \lim_{\tau\to 0^+} \mathcal{L}_n(\boldsymbol{\omega}, w, 0, \phi, \tau) \xrightarrow{P} \lim_{\tau\to 0^+} \mathcal{L}(\boldsymbol{\omega}, w, 0, \phi, \tau) \ . \end{aligned} \tag{4.103}$$

Which implies

$$\sup_{\phi\geq 0} \inf_{\tau>0} \mathcal{L}_n(\boldsymbol{\omega}, w, 0, \phi, \tau) \xrightarrow{P} \sup_{\phi\geq 0} \inf_{\tau>0} \mathcal{L}(\boldsymbol{\omega}, w, 0, \phi, \tau) \tag{4.104}$$

and as a consequence the desired. By the weak law of large numbers, we have that

$$\begin{aligned} \lim_{\tau\to 0^+} \mathcal{L}_n(\boldsymbol{\omega}, w, v = 0, \phi, \tau) &= \frac{1}{n} \sum_{i=1}^{n} g\big(wZ_0, \boldsymbol{\omega}, \Delta, T\big) - \Delta_i \log \lambda(T_i|\boldsymbol{\omega}) \\ &\xrightarrow{P} \mathbb{E}_{\Delta,T,Z_0,Q}\Big[g\big(wZ_0, \boldsymbol{\omega}, \Delta, T\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] = \\ &\lim_{\tau\to 0^+} \mathcal{L}(\boldsymbol{\omega}, w, v = 0, \phi, \tau) \ . \end{aligned} \tag{4.105}$$

The "other side" requires more work. Notice that

$$\begin{aligned} &\lim_{\phi\to\infty} \inf_{\tau>0} \mathcal{L}_n(\boldsymbol{\omega}, w, v = 0, \phi, \tau) = \\ &\lim_{\phi\to\infty} \inf_{\kappa>0} \frac{1}{n} \sum_{i=1}^{n} \mathcal{M}_{g(.,\boldsymbol{\omega},\Delta_i,T_i)}\big(wZ_{0,i}, \kappa\big) - \Delta_i \log \lambda(T_i|\boldsymbol{\omega}) + \phi^2\kappa/2 \end{aligned}$$

The sequence of functions
$f_n^{\boldsymbol{\omega},w,\phi}(\kappa) := \frac{1}{n}\sum_{i=1}^{n} \mathcal{M}_{g(.,\boldsymbol{\omega},\Delta_i,T_i)}\big(wZ_{0,i}, \kappa\big) - \Delta_i \log \lambda(T_i|\boldsymbol{\omega}) + \phi^2\kappa/2$:

- are convex in κ,
- $f_n'(0^+) > 0$, for ϕ, n sufficiently large, since $\mathbb{V}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0, \kappa\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] < \infty$
- $\lim_{\kappa\to\infty} f_n^{\boldsymbol{\omega},w,\phi}(\kappa) = \infty$.

Because of convexity of $f_n^{\boldsymbol{\omega},w,\phi}$, these imply that the infimum is at $\kappa = 0$, for sufficiently large n, ϕ. Hence

$$\lim_{\phi\to\infty} \inf_{\tau>0} \mathcal{L}_n(\boldsymbol{\omega}, w, 0, \phi, \tau) \xrightarrow{P} \mathbb{E}_{T,Z_0,Q}\Big[g\big(wZ_0, \boldsymbol{\omega}, \Delta, T\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] \ . \tag{4.106}$$

Similarly

$$\lim_{\phi\to\infty} \inf_{\tau>0} \mathcal{L}(\boldsymbol{\omega}, w, 0, \phi, \tau) = \\ \lim_{\phi\to\infty} \inf_{\kappa>0} \mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0, \kappa\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] + \phi^2\kappa/2 \ . \tag{4.107}$$

The function
$f^{\boldsymbol{\omega},w,\phi}(\kappa) := \mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,\boldsymbol{\omega},\Delta,T)}\big(wZ_0, \kappa\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] + \phi^2\kappa/2$:

- is convex in κ,
- $f'(0^+) > 0$, for ϕ sufficiently large,
- $\lim_{\kappa\to\infty} f^{\boldsymbol{\omega},w,\phi}(\kappa) = \infty$.

Because of convexity of $f^{\boldsymbol{\omega},w,\phi}$, these imply that the infimum is at $\kappa = 0$. Hence

$$\lim_{\phi\to\infty} \inf_{\kappa>0} \mathcal{L}(\boldsymbol{\omega}, w, 0, \phi, \tau) = \mathbb{E}_{T,Z_0,Q}\Big[g\big(wZ_0, \boldsymbol{\omega}, \Delta, T\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] \tag{4.108}$$

and we have shown

$$\lim_{\phi\to\infty} \inf_{\tau>0} \mathcal{L}_n(\boldsymbol{\omega}, w, 0, \phi, \tau) = \frac{1}{n}\sum_{i=1}^{n} g\big(wZ_0, \boldsymbol{\omega}, \Delta, T\big) \\ \xrightarrow{P} \mathbb{E}_{T,Z_0,Q}\Big[g\big(wZ_0, \boldsymbol{\omega}, \Delta, T\big) - \Delta \log \lambda(T|\boldsymbol{\omega})\Big] = \\ \lim_{\phi\to\infty} \inf_{\kappa>0} \mathcal{L}(\boldsymbol{\omega}, w, 0, \phi, \tau) \ .$$

□

Minimization over (w, v)

Let us denote $\mathcal{L}_n(\boldsymbol{\omega}, w, v) := \sup_{\phi\geq 0} \mathcal{L}_n^{\boldsymbol{\omega},w,v>0}(\phi)$ and $\mathcal{L}(\boldsymbol{\omega}, w, v) := \sup_{\phi\geq 0} \mathcal{L}^{\boldsymbol{\omega},w,v}(\phi)$.

Proposition 12. *The function $\mathcal{L}(\boldsymbol{\omega}, w, v)$ admits a unique minimizer for $\boldsymbol{\omega} \in \mathcal{S}_{\boldsymbol{\omega}}$, $w \in [0, C_{\boldsymbol{\beta}}]$ and $v \in [0, C_{\boldsymbol{\beta}}]$ and*

$$\min_{(\boldsymbol{\omega},w,v)} \mathcal{L}_n(\boldsymbol{\omega}, w, v) \xrightarrow[n\to\infty]{P} \min_{(\boldsymbol{\omega},w,v)} \mathcal{L}(\boldsymbol{\omega}, w, v) \ . \tag{4.109}$$

Proof. The functions $\mathcal{L}_n$ are convex in their arguments, as the pointwise maxima of convex functions. The same is true for $\mathcal{L}$, furthermore $\mathcal{L}_n \xrightarrow{P} \mathcal{L}$ point wise. Let us denote $\mathcal{S}_{\boldsymbol{\omega},w,v} = \mathcal{S}_{\boldsymbol{\omega}} \times [0, C_{\boldsymbol{\beta}}] \times [0, C_{\boldsymbol{\beta}}]$. The function $\mathcal{L}(\boldsymbol{\omega}, w, v)$ is strongly convex in w, v and strictly convex in $\boldsymbol{\omega}$ (it is sufficient to compute the second derivative, which is always non negative on $\mathcal{S}_{\boldsymbol{\omega}}$). In particular $\mathcal{L}(\boldsymbol{\omega}, w, v)$ is strongly convex in its arguments if we insist that $\forall \mu,\ P[t_\mu < T < t_{\mu+1}] \geq \delta > 0$ and hence admits a unique minimizer. If $\boldsymbol{\omega}_\star, w_\star, v_\star$ is in the interior of $\mathcal{S}_{\boldsymbol{\omega},w,v}$, then we can readily conclude via the convexity lemma (1), that

$$\min_{(\boldsymbol{\omega},w,v)} \mathcal{L}_n(\boldsymbol{\omega}, w, v) \xrightarrow[n\to\infty]{P} \min_{(\boldsymbol{\omega},w,v)} \mathcal{L}(\boldsymbol{\omega}, w, v) \ . \tag{4.110}$$

□

4.8.4 Convergence in probability of the minimizer(s)

In order to prove (4.15,4.16,4.17), we want to use Theorem 3 an in particular implication (4.31). To do so, let us define

$$\mathcal{S}_\epsilon := \left\{(\boldsymbol{\beta}, \boldsymbol{\omega}) \in \mathcal{S}_{\boldsymbol{\beta}} \times \mathcal{S}_{\boldsymbol{\omega}} : \ \left|w_n - w_\star\right| < \epsilon, \left|v_n - v_\star\right| < \epsilon, \left|\boldsymbol{\omega} - \boldsymbol{\omega}_\star\right| < \epsilon\right\} \tag{4.111}$$

with $\mathcal{S}_{\boldsymbol{\beta}}, \mathcal{S}_{\boldsymbol{\omega}}$ two compact subset of $\mathbb{R}^p$ and $\mathbb{R}^d$, respectively, and w_n, v_n defined as

$$w_n := \frac{\boldsymbol{\beta}_0' \boldsymbol{\beta}}{\|\boldsymbol{\beta}_0\|}, \quad v_n := \left\|\mathbf{P}_{\perp \boldsymbol{\beta}_0} \boldsymbol{\beta}\right\| = \left\| \left(\boldsymbol{I} - \frac{\boldsymbol{\beta}_0 \boldsymbol{\beta}_0'}{\|\boldsymbol{\beta}_0\|^2} \right) \boldsymbol{\beta} \right\| . \tag{4.112}$$

By the convexity lemma we have uniform convergence $\mathcal{L}_n(\boldsymbol{\omega}, w, v) \xrightarrow[n\to\infty]{P} \mathcal{L}(\boldsymbol{\omega}, w, v)$ for $(\boldsymbol{\omega}, w, v) \in \mathcal{B} \subset \mathcal{S}_{\boldsymbol{\beta}} \times \mathcal{S}_{\boldsymbol{\omega}}$, with $\mathcal{B}$ compact. As a consequence, since $\mathcal{S}_\epsilon^c := \mathcal{S}_{\boldsymbol{\beta}} \times \mathcal{S}_{\boldsymbol{\omega}} \setminus \mathcal{S}_\epsilon$ is compact, we have that

$$\min_{(\boldsymbol{\omega}, w, v) \in \mathcal{S}_\epsilon^c} \mathcal{L}_n(\boldsymbol{\omega}, w, v) \xrightarrow[n\to\infty]{P} \min_{(\boldsymbol{\omega}, w, v) \in \mathcal{S}_\epsilon^c} \mathcal{L}(\boldsymbol{\omega}, w, v) \ . \tag{4.113}$$

The function $\mathcal{L}$ above is strongly convex in its arguments, hence it has a unique minimizer $(w_\star, v_\star, \boldsymbol{\omega}_\star)$ as defined in (4.15,4.16,4.17) respectively. In this case it must hold that

$$\min_{(\boldsymbol{\omega}, w, v) \in \mathcal{S}_\epsilon^c} \mathcal{L}(\boldsymbol{\omega}, w, v) > \min_{(\boldsymbol{\omega}, w, v)} \mathcal{L}(\boldsymbol{\omega}, w, v) \ . \tag{4.114}$$

Via Theorem 3 implication (4.31) this implies that

$$\lim_{n\to\infty} P\Big[(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\omega}}_n) \in \mathcal{S}_\epsilon\Big] = 1 \tag{4.115}$$

which is the desired.

Chapter 5

Observable asymptotics of regularized Cox regression models with standard Gaussian designs: a statistical mechanics approach

An updated version of this chapter is published as: **Massa, E** and Coolen, A C C, *Observable asymptotics of regularized Cox regression models with standard Gaussian designs: a statistical mechanics approach*, Journal of Physics A: Mathematical and Theoretical, 58, 105001, 2025. https://dx.doi.org/10.1088/1751-8121/adb8ad

5.1 Abstract

We study the asymptotic behaviour of Regularized Maximum Partial Likelihood Estimator (RMPLE) in the proportional limit, considering an arbitrary regularizer and assuming that the covariates $\mathbf{X}_i \in \mathbb{R}^p$ follow a multivariate Gaussian law with covariance $\boldsymbol{I}_p/p$ for each $i = 1, \dots, n$. In order to efficiently compute the estimator under investigation, we propose a modified Approximate Message Passing (AMP) algorithm, that we name COX-AMP, and compare its performance with the Coordinate-wise Descent (CD) algorithm, which is taken as reference. By means of the Replica method, we derive a set of six Replica Symmetric equations that we show to correctly describe the average behaviour of the estimators when the sample size and the number of covariates is large and commensurate. These equations cannot be solved in practice, since the data generating process (that we are trying to estimate) is not known. However, the update equations of COX-AMP suggest the construction of a local field that can in turn be used to accurately estimate all the RS order parameters of the theory *without* actually solving the RS equations. We emphasize that this approach can be applied when the estimator is computed via any method and is not restricted to COX-AMP.

5.2 Introduction

When the number of features is a sufficiently large fraction of the number of observations, the estimator obtained by maximization of the Cox partial likelihood is not well defined. Hence the necessity of adding a regularization term in order to "fit" the model, i.e. define a well posed optimization problem. Sparsity inducing regularizations are widely adopted in this

context when the goal of the analysis is to obtain a more "parsimonious" model. By shrinking some of the coefficient exactly to zero, these are capable of performing variable selection. Because of this appealing characteristic, these regularizations, e.g. Lasso, Elastic Net, MCP, SCAD to name a few, have been investigated in the statistical literature for Generalized Linear Models [24, 4], but also for the Cox model [2, 14]. In particular, [2] showed that, with the Lasso and folded convex regularizations, the Regularized Maximum Partial Likelihood Estimator (RMPLE) is equal to a (biased) oracle estimator which knows the underlying true model, with overwhelming probability as the sample size n diverges. Besides some mild assumptions on the data generating process, their results require that the number of true active component s (the number of covariates that are effectively associated with the survival time) does not grow linearly with the number of observations n, i.e. $s = O(n^\alpha)$ for $\alpha \in (0, 1)$.

Here we characterize the behaviour of the RMPLE under the assumption that $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p/p)$ for all subjects $i = 1, \ldots, n$ and in the proportional regime, where the number of observation n, the number of covariates p and, importantly, the number of active covariates s, diverge proportionally, i.e. $p = \zeta n$, $s = \nu p$. This is achieved by means of the replica method from statistical physics [10, 6, 15], which has already been applied in the recent past to study the behaviour of sparse linear regression [12, 23, 18]. The replica method, under the Replica Symmetric (RS) ansatze, leads to six coupled non-linear equations, whose solution grants access to virtually any metric of interest, e.g. the mean squared error for the estimator of the association and of the survival function, but also the fraction of false positive and negatives in support recovery (as originally noted in [22] for logistic regression). We show that the prediction of the theory agree perfectly with the result of simulations.

We propose and test a generalization of the Approximate Message Passing algorithm for the Cox model, which we refer to as COX-AMP, in order to compute the RMPLE. We show by numerical simulations that the resulting algorithm leads to estimators that are very close (in L2 distance) to the estimators obtained via the Coordinate-wise Descent (CD) algorithm [25], which is taken as a reference. The update equations of COX - AMP suggest the construction of a "local field", that we use to estimate the order parameters of the theory from the data, i.e. without the need of actually solving the RS equations. We show via numerical simulations, that these estimators, obtained without the knowledge of the data generating process, perform extremely well, i.e. they are very close to the numerical solution of the RS equations, which conversely requires the knowledge of the data generating process. Always by numerical simulations, we show that the "local field" might be computed directly also from the estimator obtained by the CD algorithm. Our approach is a generalization of what is done in [23] by means of the Expectation Consistent or Adaptive TAP method [19], which, however, cannot be directly applied here.

We highlight that estimating the RS order parameter from the data is preferable to the direct solution of the RS equations, since the knowledge of the data generating process is not required. This paves the road to applications, where the only available quantities are the data and the estimators (computed from the data).

5.3 Setting and notation

Focusing on the case of right censored data, time to event data comprise observations reporting: i) the event indicator Δ_i, equal to 1 if the subject experienced the event and 0 otherwise, ii) the observed time to event T_i if $\Delta_i = 1$, or the censoring time else,and iii) the covariate vector $\mathbf{X}_i = (X_1, \ldots, X_p) \in \mathbb{R}^p$, i.e. the list of characteristics of the subject. The most widely adopted model to deal with time to event data is the Cox semi-parametric

proportional hazards model [13, 11, 7]. If censoring is non-informative, then the likelihood density of the event Δ in the time interval $[t, t+dt)$ given covariates $\mathbf{X}_i$ and under the proportional hazards assumption reads

$$p(\Delta, t|\mathbf{X}_i'\boldsymbol{\beta}) = f(t|\mathbf{X}_i'\boldsymbol{\beta})^{\Delta} S_C(t)^{\Delta} \times S(t|\mathbf{X}_i'\boldsymbol{\beta})^{1-\Delta} f_c(t)^{1-\Delta} \ , \tag{5.1}$$

where we indicated with $'$ the dot product, i.e. for $\boldsymbol{a}, \boldsymbol{v} \in \mathbb{R}^d$, we have $\boldsymbol{a}'\boldsymbol{v} = \sum_{k=1}^{d} a_k v_k$. Above, f_c and $S_C(x) = \int_x^{\infty} f_c(x)\mathrm{d}x$ are , respectively,the density and survival functions for the censoring risk, while

$$S(t|\mathbf{X}_i'\boldsymbol{\beta}) := \exp\{-\Lambda(t)\exp(\mathbf{X}'\boldsymbol{\beta})\}, \tag{5.2}$$

$$f(t|\mathbf{X}_i'\boldsymbol{\beta}) := -\frac{\mathrm{d}}{\mathrm{d}t} S(t|\mathbf{X}_i'\boldsymbol{\beta}) = \lambda(t)\exp\{\mathbf{X}_i'\boldsymbol{\beta} - \Lambda(t)\exp(\mathbf{X}_i'\boldsymbol{\beta})\}, \tag{5.3}$$

are the survival (5.2) and density (5.3) functions for the primary risk (the risk under investigation). The latter' are parametrized by: i) the association parameters $\boldsymbol{\beta} \in \mathbb{R}^p$, ii) the cumulative hazard function Λ . The minus log-likelihood for a data-set of i.i.d censored observations $\{(\Delta_i, T_i, \mathbf{X}_i)\}_{i=1}^n$, reads

$$\mathcal{L}_n\big(\boldsymbol{\beta}, \lambda\big) = \sum_{i=1}^{n} \Big\{ \Lambda(T_i)\mathrm{e}^{\mathbf{X}_i'\boldsymbol{\beta}} - \Delta_i\big(\log\lambda(T_i) + \mathbf{X}_i'\boldsymbol{\beta}\big)\Big\} \tag{5.4}$$

$$= \sum_{i=1}^{n} \Big\{ g\big(\mathbf{X}_i'\boldsymbol{\beta}, \Lambda(T_i), \Delta_i\big) - \Delta_i\big(\log\lambda(T_i)\big)\Big\} \ , \tag{5.5}$$

where we introduced

$$g(x, y, z) := \exp(x)y - zx \ , \tag{5.6}$$

in order to highlight the part of the likelihoof function that depends on the linear predictors $\mathbf{X}_i'\boldsymbol{\beta}$. By first minimizing with respect to λ one obtains the so-called Breslow estimator [3, 9]:

$$\hat{\lambda}_n(t, \boldsymbol{\beta}) = \sum_{k=1}^{n} \frac{\Delta_k \delta(t - t_k)}{\sum_{j=1}^{n} \Theta(t_j - t_k)\mathrm{e}^{\mathbf{X}_j'\boldsymbol{\beta}}} \ , \tag{5.7}$$

where $\Theta(x)$ denotes the Heaviside step-function, equal to one if $x \geq 0$ and zero else, and $\delta(x)$ the Dirac's delta distribution. Substituting in (5.4) and disregarding terms which do not depend on $\boldsymbol{\beta}$ it is possible to obtain the logarithm of the Cox partial likelihood

$$\mathcal{PL}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \Big\{ \log\Big(\frac{1}{n}\sum_{j=1}^{n} \Theta(T_j - T_i)\mathrm{e}^{\mathbf{X}_j'\boldsymbol{\beta}}\Big) - \mathbf{X}_i'\boldsymbol{\beta}\Big\} \ . \tag{5.8}$$

Optimizing the Cox partial likelihood one obtains an estimator $\hat{\boldsymbol{\beta}}_n$ and, consequently, an estimator for the cumulative hazard function of the primary risk via the Nelson-Aalen estimator

$$\hat{\Lambda}_n(t) := \sum_{i=1}^{n} \frac{\Delta_i \Theta(t - T_i)}{\sum_{j=1}^{n} \Theta(T_j - T_i)\mathrm{e}^{\mathbf{X}_j'\hat{\boldsymbol{\beta}}_n}} \tag{5.9}$$

which is obtained via integration of (5.7) with respect to t. In the following, we assume that a convex regularizer r is added to (5.8), i.e. we are interested in the properties of the estimator that minimizes

$$\mathcal{PL}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \Big\{ \log\Big(\frac{1}{n}\sum_{j=1}^{n} \Theta(T_j - T_i)\mathrm{e}^{\mathbf{X}_j'\boldsymbol{\beta}}\Big) - \mathbf{X}_i'\boldsymbol{\beta}\Big\} + r(\boldsymbol{\beta}) \ . \tag{5.10}$$

5.4 Computing the Regularized Maximum Partial Likelihood Estimator (RMPLE)

The RMPLE can be computed in different manners, depending on the regularization r. A widely applicable strategy is Coordinate Wise descent. This algorithm is easy to implement and has the advantage of being applicable also when the regularization is not differentiable, e.g. for the Lasso. We denote

$$\mathrm{NA}(T_i, \mathbf{X}\hat{\boldsymbol{\beta}}_n) := \sum_{l=1}^{n} \frac{\Delta_i \Theta(T_i - T_l)}{\sum_{j=1}^{n} \Theta(T_j - T_l) \mathrm{e}^{\mathbf{X}_j' \hat{\boldsymbol{\beta}}_n}} \ , \tag{5.11}$$

and

$$\mathrm{NA}(\mathbf{T}, \mathbf{X}\hat{\boldsymbol{\beta}}_n) = \Big(\mathrm{NA}(T_1, \mathbf{X}_1'\hat{\boldsymbol{\beta}}_n), \ldots, \mathrm{NA}(T_n, \mathbf{X}_n'\hat{\boldsymbol{\beta}}_n) \Big) \ . \tag{5.12}$$

A variant of the coordinate wise descent algorithm of [25] can be summarized as in Algorithm 1. We give a brief derivation in 5.10.5.

Algorithm 1 Coordinate wise path solution for Cox model

$\boldsymbol{\beta}^0 \leftarrow \mathbf{0}_p$
$\boldsymbol{\Lambda}^0 \leftarrow \mathrm{NA}(\mathbf{T}, \mathbf{0}_n)$
$\alpha \leftarrow \alpha_{\max}$
for α in path **do**
    err $\leftarrow 1$
    while err $\geq$ tol **do**
        $\mathbf{s}(\boldsymbol{\beta}^t, \Lambda^t) \leftarrow \Big\{ \Lambda^t(\mathbf{T}) \mathrm{e}^{\mathbf{X}\boldsymbol{\beta}^t} - \boldsymbol{\Delta} \Big\} \mathbf{X}$
        $\mathbf{M}(\boldsymbol{\beta}^t, \Lambda^t) \leftarrow \mathbf{X}' \Big\{ \Lambda^t(\mathbf{T}) \mathrm{e}^{\mathbf{X}\boldsymbol{\beta}^t} \Big\} \mathbf{X}$
        $\boldsymbol{\varphi}^t \leftarrow \boldsymbol{\beta}^t$
        for $1 \leq k \leq p$ **do**
            $\psi_k \leftarrow \big\{ \mathbf{e}_k' \mathbf{M}(\boldsymbol{\beta}^t, \Lambda^t) \big\{ \boldsymbol{\beta}^t - (\boldsymbol{I} - \mathbf{e}_k \mathbf{e}_k') \boldsymbol{\varphi}^t \big\} - s_k(\boldsymbol{\beta}^t, \Lambda^t) \big\} / \mathbf{e}_k' \mathbf{M}(\boldsymbol{\beta}^t, \Lambda^t) \mathbf{e}_k$
            $1/\hat{\tau}_k \leftarrow \mathbf{e}_k' \mathbf{M}(\boldsymbol{\beta}^t, \Lambda^t) \mathbf{e}_k$
            $\varphi_k^t \leftarrow \mathrm{prox}_{r(.)}\big(\psi_k, \tau_k\big)$
        end for
        $\boldsymbol{\beta}^{t+1} \leftarrow \boldsymbol{\varphi}^t$
        $\Lambda(\mathbf{T})^{t+1} \leftarrow \mathrm{NA}(\mathbf{T}, \mathbf{X}\boldsymbol{\beta}^{t+1})$
        err $\leftarrow \sqrt{\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2 + \|\Lambda^{t+1}(\mathbf{T}) - \Lambda^t(\mathbf{T})\|^2}$
    end while
end for

When the covariates are not correlated and gaussian with mean zero, a variant of the Approximate Message Passing (AMP) algorithm [8] has been proposed to compute the Maximum A Posteriori estimator for Generalized Linear Models [20]. Here we propose the use a modification of the Generalized-AMP, in order to compute the Penalized Partial Likelihood estimator. We call the resulting algorithm COX-AMP, see Algorithm 2, we give a derivation in 5.10.6.

Numerical experiments with the elastic net regularization,

$$\mathrm{r}(\boldsymbol{\beta}) := \alpha \|\boldsymbol{\beta}\|_1 + \eta \frac{1}{2} \|\boldsymbol{\beta}\|_2^2 \ , \quad \alpha = \rho\lambda, \quad \eta = \rho(1-\lambda) \ , \tag{5.13}$$

Algorithm 2 COX-AMP algorithm for MAP in Cox regression

Require: $\hat{\boldsymbol{\beta}}^0 \leftarrow \mathbf{0}_p$, $\zeta \leftarrow p/n$ tol $\leftarrow 1.0e-8$, max epochs $\leftarrow 300$

$\boldsymbol{\tau}^0 = 0,\ \hat{\boldsymbol{\tau}}^0 = 0, \boldsymbol{\xi}^0 = \mathbf{X}\hat{\boldsymbol{\beta}}^0$
flag = True, err $= \infty$, t $= 0$
while err $\geq$ tol and flag **do**
    $t \leftarrow t + 1$
    $\hat{\Lambda}_n^t(\mathbf{T}) \leftarrow \mathrm{NA}(\mathbf{T}, \mathrm{prox}_{g(.,\hat{\Lambda}^{t-1}(\mathbf{T}),\boldsymbol{\Delta})}(\boldsymbol{\xi}^{t-1}, \tau^{t-1}))$
    err $\leftarrow$ err $+ \|\hat{\Lambda}_n^t(\mathbf{T}) - \hat{\Lambda}^{t-1}(\mathbf{T})\|_2^2$
    $\boldsymbol{\xi}^t \leftarrow \mathbf{X}\hat{\boldsymbol{\beta}}^{t-1} + \tau^{t-1}\dot{\mathcal{M}}_{g(.,\hat{\Lambda}^t(\mathbf{T}),\boldsymbol{\Delta})}(\boldsymbol{\xi}^{t-1}, \tau^{t-1})$
    err $\leftarrow$ err $+ \|\boldsymbol{\xi}^t - \boldsymbol{\xi}^{t-1}\|_2^2$
    $\hat{\tau}^t \leftarrow \zeta / \left\langle \ddot{\mathcal{M}}_{g(.,\hat{\Lambda}^t(\mathbf{T}),\boldsymbol{\Delta})}(\boldsymbol{\xi}^t, \tau^{t-1}) \right\rangle$
    err $\leftarrow$ err $+ (\hat{\tau}^t - \hat{\tau}^{t-1})^2$
    $\boldsymbol{\psi}^t \leftarrow \hat{\boldsymbol{\beta}}^{t-1} - \hat{\tau}^t \mathbf{X}' \dot{\mathcal{M}}_{g(.,\hat{\Lambda}^t(\mathbf{T}),\boldsymbol{\Delta})}(\boldsymbol{\xi}^t, \tau^{t-1})$
    $\hat{\boldsymbol{\beta}}^t \leftarrow \mathrm{prox}_{r(.)}(\boldsymbol{\psi}^t, \hat{\tau}^t)$
    err $\leftarrow$ err $+ \|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}^{t-1}\|_2^2$
    $\tau^t \leftarrow \hat{\tau}^t \left\langle \mathrm{prox}'_{r(.)}(\boldsymbol{\psi}^t, \hat{\tau}^t) \right\rangle$
    err $\leftarrow$ err $+ (\tau^t - \tau^{t-1})^2$
    err $\leftarrow \sqrt{\text{err}}$
    if t $\geq$ max epochs **then**
        flag = False
    end if
end while

show, see figures (5.1), that the estimator computed via Cox - AMP is very close (in L2 norm) to the estimator computed via Coordinate-wise Descent. However, we noticed that CD always converge, whilst Cox - AMP might not always converge and generally requires a damped update, especially when the regularization strength is "small" and ζ "large". To be more precise, it is well known that the Maximum Partial Likelihood estimator does not exist (with probability approaching one in the asymptotic limit) past a critical threshold ζ_c which depends on the data generating process. Hence we think that the instability of Cox AMP algorithm at small regularization strength is due to the explosion of the variance of the estimator in the limit of vanishing regularization for ζ sufficiently large.

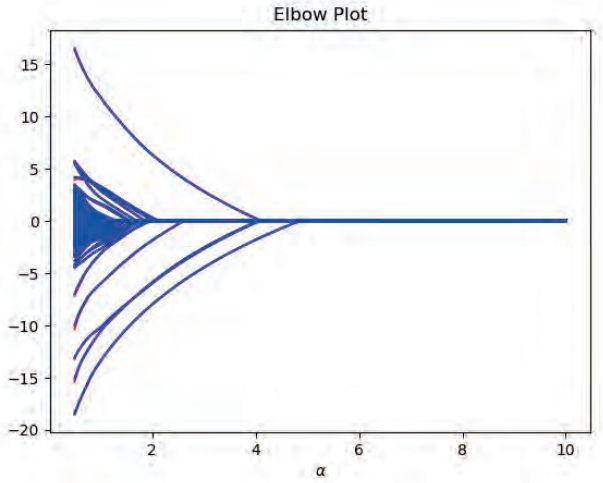


Figure 5.1: **Elbow plot of $\hat{\boldsymbol{\beta}}_n(\alpha)$** computed via Cox-AMP in red and Cox-CD in blue. The data are generated from a Log-logistic proportional hazards model $\Lambda_0(t) = \log(1+t^2/2)$ with uniform censoring between $\tau_1 = 1.0$ and $\tau_2 = 2.0$. The associations $\boldsymbol{\beta}_0$ are sampled as from a Gauss-Bernoulli distribution $\mathcal{P}_{\beta_0}(x) = (1-\nu)\delta(x) + \nu \exp\{-x^2/2\sigma^2\}/\sqrt{2\pi}\sigma$, with sparsity $\nu = 0.01$ and σ adjusted to have $\|\boldsymbol{\beta}_0\|_2^2/p = \theta_0^2 = 1.0$. The L1 ratio $\lambda = 0.95$. The L2 relative distance between $\hat{\boldsymbol{\beta}}_n(\alpha)$ computed via Cox-AMP and Cox-CD is always less than 0.05 in all our simulations.

5.5 Typical behaviour of the RMPLE

In the statistical physics approach to optimization, the optimal value of the objective function (5.10) is equivalent to minus the zero temperature free energy density of a fictitious physical system, i.e.

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\mathcal{P}\mathcal{L}_n(\hat{\boldsymbol{\beta}}_n)\Big] = \lim_{n\to\infty}\lim_{\gamma\to\infty}\frac{1}{n\gamma}\mathbb{E}_{\mathcal{D}}\Big[-\log\int \mathrm{e}^{-\gamma\mathcal{H}_n(\boldsymbol{\beta}|\mathcal{D})}\mathrm{d}\boldsymbol{\beta}\Big], \tag{5.14}$$

with Hamiltonian equal to the minus penalized partial likelihood

$$\mathcal{H}_n(\boldsymbol{\beta}|\mathcal{D}) = \sum_{i=1}^{n}\Delta_i\Big[\log\Big(\frac{1}{n}\sum_{j=1}^{n}\Theta(T_j - T_i)\mathrm{e}^{\mathbf{X}_i'\boldsymbol{\beta}}\Big) - \mathbf{X}_i'\boldsymbol{\beta}\Big] + r(\boldsymbol{\beta}) \tag{5.15}$$

where we indicated with $\mathcal{D}$ the data-set, i.e. the set of tuples $\{\Delta_j, T_j, \mathbf{X}_j\}_{j=1}^n$. The $\boldsymbol{\beta}$s are the degrees of freedom of the system and the data-set $\mathcal{D} = \{\Delta_j, T_j, \mathbf{X}_j\}_{j=1}^n$ plays the role of the quenched disorder.

We compute the right hand side of (5.14) via the replica method under the following assumptions over the data generating process: the event indicator and event time are generated as

$$\Delta_i = \Theta(C_i - Y_i), \quad T_i = \min\{Y_i, C_i\} \tag{5.16}$$

where $C \sim f_C$ is the (latent) random variable censoring time and

$$Y_i|\mathbf{X}_i \sim f_0(.|\mathbf{X}_i'\boldsymbol{\beta}_0), \quad \mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{p}\boldsymbol{I}_p) \tag{5.17}$$

is the (latent) censoring time. This implies that the true (unknown) generative model is

$$f(\Delta, t|\mathbf{X}'\boldsymbol{\beta}_0) = f_0(t|\mathbf{X}'\boldsymbol{\beta}_0)^{\Delta} S_C(t)^{\Delta} \times S_0(t|\mathbf{X}'\boldsymbol{\beta}_0)^{1-\Delta} f_c(t)^{1-\Delta} \ . \tag{5.18}$$

Notice that we do not require that the model is correctly specified, i.e. we do not assume that the data are generated from a proportional hazard model. The scaling of the covariates is a standard assumption in virtually any previous study of high dimensional regression. The full replica derivation is included in 5.10.1 and gives

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\mathcal{P}\mathcal{L}_n(\hat{\boldsymbol{\beta}}_n)\Big] = \underset{w,v,\tau,\hat{w},\hat{v},\hat{\tau}}{\text{extr}} \mathcal{F}(w, v, \tau, \hat{w}, \hat{v}, \hat{\tau}) \ . \tag{5.19}$$

We now introduce some useful definition that will recur throughout the manuscript and in particular to define $\mathcal{F}$. For a convex function $b : \mathbb{R}^d \to \mathbb{R}$, according to [21], we introduce:

- the Moureau envelope $\mathcal{M}_{b(.)}$

$$\mathcal{M}_{b(.)}(\mathbf{x}, \alpha) = \min_{\mathbf{z}\in\mathbb{R}^d}\Big\{\frac{1}{2\alpha}\|\mathbf{z}-\mathbf{x}\|^2 + b(\mathbf{z})\Big\}, \tag{5.20}$$

- the proximal mapping operator $\text{prox}_{b(.)}$

$$\text{prox}_{b(.)}(\mathbf{x}, \alpha) = \arg\min_{\mathbf{z}\in\mathbb{R}^d}\Big\{\frac{1}{2\alpha}\|\mathbf{z}-\mathbf{x}\|^2 + b(\mathbf{z})\Big\} \ . \tag{5.21}$$

Furthermore it is handy to define the limit of the empirical distribution of the entries of $\boldsymbol{\beta}_0$ as

$$\mathcal{P}_{\beta_0}(x) := \lim_{p\to\infty} \frac{1}{p}\sum_{\mu=1}^{p} \delta(x - \mathbf{e}_\mu'\boldsymbol{\beta}_0) \ , \tag{5.22}$$

so that a component of $\boldsymbol{\beta}_0$ can be seen as a draw from the distribution (5.22) above. The zero temperature (disordered averaged) free energy in the Replica Symmetric ansatze reads

$$\begin{aligned}
\mathcal{F}(w, v, \tau, \hat{w}, \hat{v}, \hat{\tau}) &= -\frac{\zeta}{2\hat{\tau}}\Big((w-\hat{w})^2 + v^2 + \hat{v}^2(1-\tau/\hat{\tau})\Big) \\
&+ \zeta\mathbb{E}_{Z,\beta_0}\Big[\mathcal{M}_{\mathrm{r}(.)}\Big(\hat{w}\frac{\beta_0}{\theta_0} + \hat{v}Z, \hat{\tau}\Big)\Big] \\
&+ \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\mathcal{M}_{g(.,\Lambda(T),\Delta)}\Big(wZ_0 + vQ, \tau\Big)\Big] + \text{const}
\end{aligned} \tag{5.23}$$

where $Z_0, Q \sim \mathcal{N}(0,1)$, $Z_0 \perp Q$,

$$\Delta, T|Z_0 \sim f(\Delta, t|\theta_0 Z_0),\ \theta_0 := \|\boldsymbol{\beta}_0\|/\sqrt{p},\ Z_0, Q \sim \mathcal{N}(0,1),\ Z_0 \perp Q \ , \tag{5.24}$$

and the function Λ is defined at w, v, τ as the self consistent solution of

$$\Lambda(t) = \mathbb{E}_{\Delta,T,Z_0,Q}\left[\frac{\Delta\Theta(t-T)}{\mathcal{S}(T)}\right], \tag{5.25}$$

$$\mathcal{S}(t) = \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\Theta(T-t)\mathrm{e}^{wZ_0+vQ+\tau\Delta - W_0\left(\tau\mathrm{e}^{\tau\Delta+wZ_0+vQ}\Lambda(T)\right)}\Big], \tag{5.26}$$

with $W_0(.)$ the (real branch of) Lambert W-function, i.e the solution of $W_0(x)\exp W_0(x) = x,\ x > -1/\mathrm{e}$.

The point of extremum in (5.19) satisfies the replica symmetric (RS) equations

$$
\begin{aligned}
w &= \mathbb{E}_{Z,\beta_0}\Big[\beta_0\varphi\Big]/\theta_0 && (5.27)\\
\hat{v}\frac{\tau}{\hat{\tau}} &= \mathbb{E}_{Z,\beta_0}\Big[Z\varphi\Big] && (5.28)\\
(w^2+v^2) &= \mathbb{E}_{Z,\beta_0}\Big[\varphi^2\Big] && (5.29)\\
\hat{w} &= w - \frac{\hat{\tau}}{\zeta\tau}\Big(w - \mathbb{E}_{\Delta,T,Z_0,Q}\Big[Z_0\xi\Big]\Big) && (5.30)\\
\zeta\tau/\hat{\tau} &= \mathbb{E}_{\Delta,T,Z_0,Q}\Big[Q\xi\Big] && (5.31)\\
\zeta\hat{v}^2 &= \frac{\hat{\tau}^2}{\tau^2}\mathbb{E}_{\Delta,T,Z_0,Q}\Big[(\xi - wZ_0 - vQ)^2\Big] && (5.32)
\end{aligned}
$$

where

$$
\begin{aligned}
&\xi := \mathrm{prox}_{g(.,\Lambda(T),\Delta)}(wZ_0 + vQ, \tau) = && (5.33)\\
&wZ_0 + vQ + \tau\Delta - W_0\Big(\tau \mathrm{e}^{\tau\Delta + wZ_0 + vQ}\Lambda(T)\Big)\\
&\varphi := \mathrm{prox}_{r(.)}(\hat{w}\beta_0/\theta_0 + \hat{v}Z, \hat{\tau})\ . && (5.34)
\end{aligned}
$$

5.5.1 Interpretation

We show in 5.10.4 that

$$
\mathcal{P}(\Delta, t, h) := \lim_{n\to\infty} \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\delta(t - T_i)\delta_{\Delta,\Delta_i}\delta(x - \mathbf{X}_i'\hat{\boldsymbol{\beta}})\Big] \tag{5.35}
$$

$$
\mathcal{P}_{\varphi}(x) := \lim_{n\to\infty} \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{p}\sum_{k=1}^{p}\delta(x - \mathbf{e}_k'\hat{\boldsymbol{\beta}})\Big]\ . \tag{5.36}
$$

admit the following expressions in the Replica Symmetric ansatze

$$
\begin{aligned}
\mathcal{P}_{\xi}(\Delta, t, h) &= \mathbb{E}_{\Delta',T',Z_0',Q'}\Big[\delta(t - T')\delta_{\Delta,\Delta'}\delta\big(h - \xi_\star\big)\Big] && (5.37)\\
\mathcal{P}_{\varphi}(x) &= \mathbb{E}_{\beta_0,Z}\Big[\delta\big(x - \varphi_\star\big)\Big] && (5.38)
\end{aligned}
$$

where $\xi_\star, \varphi_\star$ are as in (5.33,5.34), computed at the fixed point of the RS equations, i.e. at $w_\star, v_\star, \tau_\star, \hat{w}_\star, \hat{v}_\star, \hat{\tau}_\star$ and $\Lambda_\star(.)$ is the function solving (5.25, 5.26) evaluated at the fixed point. Hence in the proportional regime and for $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{p}\boldsymbol{I}_p)$, we have

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\ell_1(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n(T_i), \Delta_i, T_i)\Big] =\\
&\qquad\qquad \sum_{\Delta\in\{1,0\}}\int \mathcal{P}_{\xi}(\Delta, t, h)\ell_1(h, \Lambda_\star(t), \Delta, t)\mathrm{d}t\mathrm{d}h, && (5.39)\\
&\lim_{p\to\infty} \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{p}\sum_{\mu=1}^{p}\ell_2(\mathbf{e}_\mu'\hat{\boldsymbol{\beta}}_n)\Big] = \int \mathcal{P}_{\varphi}(x)\ell_2(x)\ \mathrm{d}x\ , && (5.40)
\end{aligned}
$$

for any "reasonable" functions ℓ_1, ℓ_2. This can be compactly re-written as

$$\boldsymbol{\xi} := \text{prox}_{g(.,\Lambda(T),\Delta)}(w\mathbf{Z}_0 + v\mathbf{Q}, \tau) \quad \underset{n\to\infty}{\overset{d}{\approx}} \quad \mathbf{X}\hat{\boldsymbol{\beta}}_n \tag{5.41}$$

$$\boldsymbol{\varphi} := \text{prox}_{r(.)}(\hat{w}\boldsymbol{\beta}_0/\theta_0 + \hat{v}\mathbf{Z}, \hat{\tau}) \quad \underset{n\to\infty}{\overset{d}{\approx}} \quad \hat{\boldsymbol{\beta}}_n \ , \tag{5.42}$$

where $\underset{n\to\infty}{\overset{d}{\approx}}$ means approximately equal in distribution in the limit and $\mathbf{Z}_0 = (Z_{0,1}, \ldots, Z_{0,n})$, $\mathbf{Q} = (Q_1, \ldots, Q_n)$, $\mathbf{Z} = (Z_1, \ldots, Z_n)$ and the proximal operators act component-wise. In turn (5.40) or (5.42) provides the following interpretation of the values $w_\star, v_\star$

$$\hat{w}_n := \frac{\boldsymbol{\beta}_0'\hat{\boldsymbol{\beta}}_n}{\sqrt{\boldsymbol{\beta}_0'\boldsymbol{\beta}_0}} \underset{n\to\infty}{\approx} w_\star, \quad \hat{v}_n^2 := \hat{\boldsymbol{\beta}}_n'\hat{\boldsymbol{\beta}}_n - w_n^2 \underset{n\to\infty}{\approx} v_\star^2 \ . \tag{5.43}$$

Similarly (5.39) or (5.41) implies that the Replica Symmetric functional order parameter $\Lambda(.)$ is equivalent to the Nelson-Aalen estimator computed over a large (ideally infinite) data-set, i.e.

$$\hat{\Lambda}_n(.) \underset{n\to\infty}{\approx} \Lambda_\star(.) \ . \tag{5.44}$$

5.6 Numerical solution of the RS equations with known data-generating process

In this section we compare the solution of the RS equations (5.25, 5.26, 5.27, 5.28, 5.29, 5.30, 5.31, 5.32) with the finite sample size metrics computed from synthetic data when the data generating process is perfectly known. This is done to benchmark the prediction of the Replica Symmetric theory.

We first detail on the simulation protocol. In all the simulations of this section the true associations are generated as follows

$$\boldsymbol{\beta}_0 = \Big(\theta_0 \ \sqrt{p}\ \boldsymbol{U}_s, \mathbf{0}_{p-s}\Big), \quad \boldsymbol{U}_s \sim \text{Unif}(\mathbb{S}_{s-1}) \ . \tag{5.45}$$

When $p, s \to \infty$ we have that

$$\mathcal{P}_{\beta_0}(x) = \nu\frac{1}{\sqrt{2\pi}\sigma}\text{e}^{-\frac{1}{2}x^2/\sigma^2} + (1-\nu)\delta(x), \quad \nu = s/p \in [0,1], \quad \sigma = \theta_0/\sqrt{\nu} \ . \tag{5.46}$$

where $\mathcal{P}_{\beta_0}$ is the limiting empirical distribution of the entries of $\boldsymbol{\beta}_0$, defined in (5.22). The sparsity of the signal is controlled by ν which is the fraction of "active" covariates that is actually correlated with the outcome. Under this assumption, and with the elastic net regularization (5.13), the replica symmetric equations can be simplified further, see 5.10.3,

to a somewhat simpler form

$$
\begin{aligned}
w &= 2\frac{1}{1+\eta\hat{\tau}}\hat{w}\Phi(\chi_1) && (5.47)\\
\tau &= 2\frac{1}{1+\eta\hat{\tau}}\hat{\tau}\Big\{\nu\Phi(\chi_1)+(1-\nu)\Phi(\chi_0)\Big\} && (5.48)\\
\frac{1}{2}(v^2+w^2) &= \frac{1}{(1+\eta\hat{\tau})^2}\Big\{\nu\Big((1+1/\chi_1^2)\Phi(\chi_1)-\phi(\chi_1)\Big)+\\
&+ (1-\nu)\Big((1+1/\chi_0^2)\Phi(\chi_0)-\phi(\chi_0)\Big)\Big\} && (5.49)\\
\hat{w} &= w-\frac{\hat{\tau}}{\zeta\tau}\Big(w-\mathbb{E}_{\Delta,T,Z_0,Q}\Big[Z_0\xi\Big]\Big) && (5.50)\\
v(1-\zeta\tau/\hat{\tau}) &= \mathbb{E}_{\Delta,T,Z_0,Q}\Big[Q\xi\Big] && (5.51)\\
\zeta\hat{v}^2 &= \frac{\hat{\tau}^2}{\tau^2}\mathbb{E}_{\Delta,T,Z_0,Q}\Big[(\xi-wZ_0-vQ)^2\Big] && (5.52)
\end{aligned}
$$

where ξ is defined in (5.33),

$$
\chi_0 := \frac{\alpha\hat{\tau}}{\hat{v}}, \quad \chi_1 := \frac{\alpha\hat{\tau}}{\sqrt{\hat{v}^2+\hat{w}^2/\nu}}, \tag{5.53}
$$

and we indicated with $\phi(x) := \exp\{-\frac{1}{2}x^2\}/\sqrt{2\pi}$ the density function and with $\Phi(x) := \int_x^{+\infty}\phi(t)\mathrm{d}t$ the complementary distribution function of a standard normal random variable. The numerical solution of these self -consistent equations is obtained via fixed point iteration. The integral are approximated as averages of populations of size $5\cdot 10^3$. The latent survival data are generated from a Log-logistic proportional hazard model

$$
\begin{aligned}
Y|\mathbf{X} &\sim -\frac{\mathrm{d}}{\mathrm{d}t}S_0(t|\mathbf{X}),\\
S_0(t|\mathbf{X}) &= \exp\{-\Lambda_0(t)\exp(\mathbf{X}'\boldsymbol{\beta}_0)\},\ \Lambda_0(t) := \log\left(1+\mathrm{e}_0^{\phi}t^{\rho_0}\right), && (5.54)
\end{aligned}
$$

and the latent censoring time is sampled uniformly in the interval $[\tau_1,\tau_2]$, i.e.

$$
C \sim \mathrm{Unif}[\tau_1,\tau_2]\ . \tag{5.55}
$$

The observations Δ, T are then generated according to

$$
\Delta = \mathbf{1}[Y<C], \quad Y=\min\{Y,C\}\ . \tag{5.56}
$$

To make the setting above concrete, we choose: $\theta_0=1, \phi_0=-\log 2, \rho_0=2$ and $\tau_1=1.0, \tau_2=2.0$, with $\rho=0.75$ and $\nu=0.05$. The quantities computed from simulations are always displayed as dots, indicating the averages over 50 repetitions (optimization with Coordinate-wise Descent), with errorbars, the corresponding finite sample standard deviations.

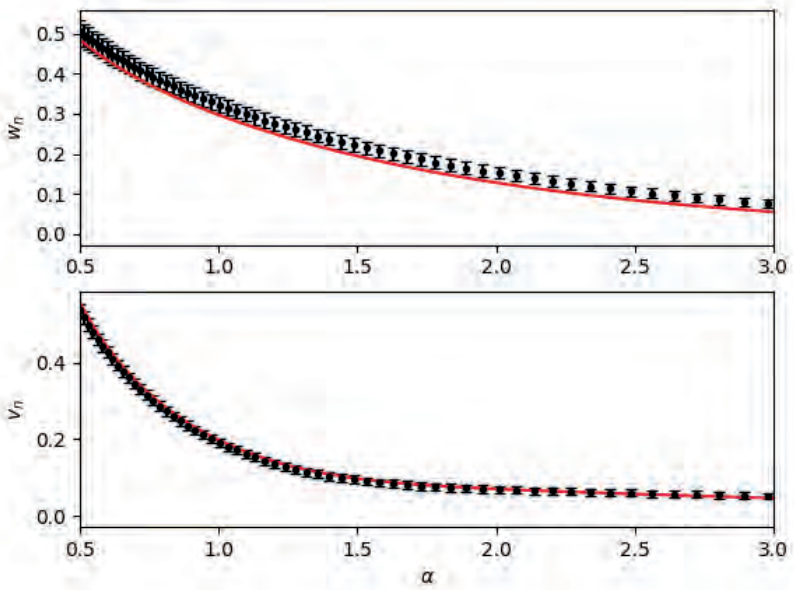


Figure 5.2: **Comparison** between the numerical solution of the RS equations against their finite sample counter part $w_n := (\hat{\boldsymbol{\beta}}_n' \boldsymbol{\beta}_0 / \|\boldsymbol{\beta}_0\|)/\sqrt{p}$, $v_n := \sqrt{\|\hat{\boldsymbol{\beta}}_n^2/p - w_n^2\|}$. Notice that here we use the knowledge of $\boldsymbol{\beta}_0$ to compute w_n, v_n. Here $\zeta = 2.0$ and $p = 2000$, i.e. $n = 1000$.

We have seen that the prediction of the RS theory are indeed correct. However, the RS equations cannot be used in practice, since they require perfect knowledge of the data-generating process. To overcome this limitation we need to relate the order parameters to *practically* observable quantities, i.e. quantities that can be computed without knowing the data generating process, i.e. $\boldsymbol{\beta}_0$ and $\Lambda_0(.)$. We show that this is indeed possible, and in a quite straightforward manner, when the RMPLE is obtained via the Cox - AMP Algorithm 2.

5.7 Inferring the RS order parameters via Cox-AMP without solving the RS equations

The COX-AMP algorithm returns τ_n and $\hat{\tau}_n$, which we interpret as finite sample size realizations of $\tau_\star$ and $\hat{\tau}_\star$. This means that $\hat{v}$ can be estimated as v_n from the data, by approximating

$$\zeta \hat{v}^2 = \hat{\tau}^2 \mathbb{E}_{T,Z_0,Q}\Big[\dot{g}\big(\xi, \Lambda(T), \Delta\big)^2\Big] \ , \tag{5.57}$$

with

$$\hat{v}_n^2 = \big\langle \ddot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n, \mathbf{T})\big\rangle \hat{\tau}_n^2/\zeta \ , \tag{5.58}$$

via (5.41), since ζ is known. Above we have indicated with $\langle\rangle$ the empirical average operator, i.e. $\langle \boldsymbol{a} \rangle = \sum_{k=1}^d a_k/d$ for a vector $\boldsymbol{a} \in \mathbb{R}^d$. The functions $g, \dot{g}, \ddot{g}, \hat{\Lambda}$ are taken to act component wise.

Since the large n behaviour of Cox-AMP is predicted by the RS equations, we see that at the fixed point, the local field

$$\boldsymbol{\psi}_\star := \hat{\boldsymbol{\beta}} - \hat{\tau}_\star \mathbf{X}' \dot{\mathcal{M}}_{g(.,\hat{\Lambda}_n(\mathbf{T}),\boldsymbol{\Delta})}(\boldsymbol{\xi}, \tau_\star) \ , \tag{5.59}$$

has a Gaussian distribution

$$\boldsymbol{\psi}_\star \stackrel{d}{=} \hat{w}_\star \boldsymbol{\beta}_0/\theta_0 + \hat{v}_\star \mathbf{Z}, \quad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p) \ . \tag{5.60}$$

As the local field has an explicit expression in the Cox-AMP algorithm, we deduce, by matching the variance, the following equation

$$\frac{1}{p}\big\|\hat{\boldsymbol{\beta}} - \hat{\tau}_\star \mathbf{X}' \dot{\mathcal{M}}_{g(.,\hat{\Lambda}_n(\mathbf{T}),\boldsymbol{\Delta})}(\mathbf{X}\hat{\boldsymbol{\beta}}_n, \tau_\star)\big\|_2^2 = \hat{w}_n^2 + \ \hat{v}_n^2 \ . \tag{5.61}$$

From which we infer $\hat{w}_n$. We now address the slightly more complicated issue of estimating $w_\star$ and $v_\star$ from the data. It can be seen that, by definition of proximal operator for ξ, we have

$$\begin{aligned}
\zeta \hat{v}^2 \tau^2/\hat{\tau}^2 = \tau^2 \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\dot{g}(\xi,T)^2\Big] = \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\big(\xi - wZ_0 - vQ\big)^2\Big] = \\
= \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\xi^2\Big] - 2w\mathbb{E}_{\Delta,T,Z_0,Q}\Big[Z_0\xi\Big] - 2v\mathbb{E}_{\Delta,T,Z_0,Q}\Big[Q\xi\Big] + w^2 + v^2 = \\
= \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\xi^2\Big] + 2w\zeta\frac{\tau}{\hat{\tau}}(w - \hat{w}) + 2v^2\zeta\frac{\tau}{\hat{\tau}} - w^2 - v^2 = \\
= \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\xi^2\Big] - (w^2 + v^2)(1 - 2\zeta\frac{\tau}{\hat{\tau}}) - 2w\hat{w}\zeta\frac{\tau}{\hat{\tau}} \ .
\end{aligned} \tag{5.62}$$

The above implies

$$\mathbb{E}_{\Delta,T,Z_0,Q}\Big[\xi^2\Big] = \zeta\hat{v}^2\tau^2/\hat{\tau}^2 + (w^2 + v^2)(1 - 2\zeta\frac{\tau}{\hat{\tau}}) + 2w\hat{w}\zeta\frac{\tau}{\hat{\tau}} \ . \tag{5.63}$$

On the other hand, the relationship

$$\mathbb{E}_{\Delta,T,Z_0,Q}\Big[\big(\xi + \tau\dot{g}(\xi, \Lambda(T), \Delta)\big)^2\Big] = w^2 + v^2 \ , \tag{5.64}$$

can be deduced by the definition of proximal operator. Hence

$$w\hat{w}\zeta\frac{\tau}{\hat{\tau}} = \frac{1}{2}\mathbb{E}_{\Delta,T,Z_0,Q}\Big[\xi^2\Big] - \frac{1}{2}\zeta\hat{v}^2\tau^2/\hat{\tau}^2 - \frac{1}{2}(w^2 + v^2)(1 - 2\zeta\tau/\hat{\tau}) \ . \tag{5.65}$$

The finite sample counterpart of the previous equations read

$$\frac{1}{n}\|\mathbf{X}\hat{\boldsymbol{\beta}}_n + \tau_n\dot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n(\mathbf{T}), \boldsymbol{\Delta})\|_2^2 = w_n^2 + v_n^2 \tag{5.66}$$

$$w_n\hat{w}_n\zeta\frac{\tau_n}{\hat{\tau}_n} = \frac{1}{2n}\|\mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2 - \frac{1}{2}\zeta\hat{v}_n^2\tau_n^2/\hat{\tau}_n^2 - \frac{1}{2}(w_n^2 + v_n^2)(1 - 2\zeta\tau_n/\hat{\tau}_n) \ . \tag{5.67}$$

From which one can solve for v_n and w_n. The remarkable performance of the estimators obtained as described above in estimating the replica symmetric order parameters $w_\star, v_\star, \tau_\star, \hat{w}_\star, \hat{v}_\star, \hat{\tau}_\star$ can be appreciated in figure(5.3). There we show the estimators summarized as errorbar plots, computed over 50 realizations, when $\zeta = 2.0$ and $p = 2000$, for the elastic net regularization (5.13) when the L1 ratio is $\lambda = 0.75$.

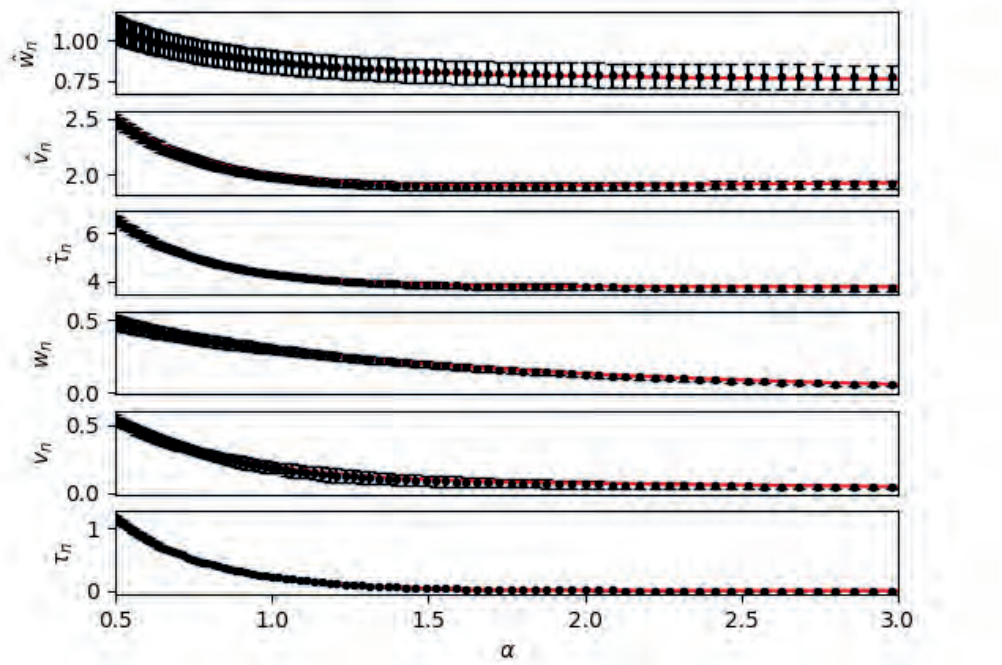


Figure 5.3: **Cox-AMP with elastic net regularization** The estimators of $\hat{w}_\star, \hat{v}_\star, \hat{\tau}_\star, w_\star, v_\star, \tau_\star$ summarized by a black errorbar plot against the true values in red, along a regularization path (different values of α). The data are generated from a Log-logistic proportional hazards model $\Lambda_0(t) = \log(1 + t^2/2)$ with uniform censoring between $\tau_1 = 1.0$ and $\tau_2 = 2.0$. The associations $\boldsymbol{\beta}_0$ are sampled as from a Gauss-Bernoulli distribution $\mathcal{P}_{\beta_0}(x) = (1-\nu)\delta(x) + \nu \exp\{-x^2/2\sigma^2\}/\sqrt{2\pi}\sigma$, with sparsity $\nu = 0.01$ and σ adjusted to have $\|\boldsymbol{\beta}_0\|_2^2/p = \theta_0^2 = 1.0$. The L1 ratio is $\lambda = 0.75$.

We have shown that the RS order parameter can be easily estimated from the estimators $\hat{\boldsymbol{\beta}}_n$ and $\hat{\Lambda}_n$ as computed from the COX-AMP algorithm. The key property of the AMP algorithm is that it estimates also $\hat{\tau}$ and τ (which are not available in other optimization methods), which in turn allows to estimate the local field $\boldsymbol{\psi}_\star$ from the data. It seems natural to wonder if a similar construction is available when $\hat{\boldsymbol{\beta}}_n$ is computed via another method, e.g. Coordinate-wise Descent. We show in the next section how this can be done.

5.8 Inferring the RS order parameters via Cox-CD without solving the RS equations

Within the Coordinate-wise Descent algorithm, we do not infer the value of τ, nor the value of $\hat{\tau}$. Hence we cannot directly observe $\hat{v}_n$ and solve the chain of equations that we have derived in the previous section. However, in some cases it seems possible to infer $\hat{\tau}$ and τ

nevertheless. Noticing that

$$\begin{aligned}\hat{v}\frac{\tau}{\hat{\tau}} &= \mathbb{E}_{Z,\beta_0}\Big[Z\varphi\Big] = \\ &= \hat{v}\mathbb{E}_{Z,\beta_0}\Big[\mathrm{prox}'_{\mathrm{r}(.)}\big(\hat{w}\beta_0/\theta_0 + \hat{v}Z, \hat{\tau}\big)\Big] \ , \end{aligned} \tag{5.68}$$

where we used integration by parts, also known as Stein's lemma, we obtain an alternative form of (5.28)

$$\frac{\tau}{\hat{\tau}} = \mathbb{E}_{Z,\beta_0}\Big[\mathrm{prox}'_{\mathrm{r}(.)}\big(\hat{w}\beta_0/\theta_0 + \hat{v}Z, \hat{\tau}\big)\Big] \ . \tag{5.69}$$

For the elastic net regression for instance, we have

$$\mathrm{prox}_{\mathrm{r}(.)}\big(\hat{w}\beta_0/\theta_0 + \hat{v}Z, \hat{\tau}\big) = \frac{1}{1+\eta\hat{\tau}}\mathrm{st}\big(\hat{w}\beta_0/\theta_0 + \hat{v}Z, \alpha\hat{\tau}\big) \tag{5.70}$$

and

$$\begin{aligned}\mathrm{prox}'_{\mathrm{r}(.)}\big(\hat{w}\beta_0/\theta_0 + \hat{v}Z, \hat{\tau}\big) &= \frac{1}{1+\eta\hat{\tau}}\mathrm{st}'\big(\hat{w}\beta_0/\theta_0 + \hat{v}Z, \alpha\hat{\tau}\big) \\ &= \frac{1}{1+\eta\hat{\tau}}\Theta(|\varphi|) \ , \end{aligned} \tag{5.71}$$

where in the last expression we used the fact that

$$\mathrm{st}'\big(\hat{w}\beta_0/\theta_0 + \hat{v}Z, \alpha\hat{\tau}\big) = \Theta\big(|\hat{w}\beta_0/\theta_0 + \hat{v}Z| - \alpha\hat{\tau}\big) = \Theta(|\varphi|) \ . \tag{5.72}$$

So

$$\frac{\tau}{\hat{\tau}} = \frac{1}{1+\eta\hat{\tau}}\mathbb{E}_{Z,\beta_0}\Big[\Theta(|\varphi|)\Big] \ , \tag{5.73}$$

which might be solved simultaneously with (5.31) for $\hat{\tau}, \tau$. Their finite sample counter-parts read

$$\frac{\tau_n}{\hat{\tau}_n} = \frac{1}{1+\eta\hat{\tau}_n}\|\hat{\boldsymbol{\beta}}_n\|_0 \tag{5.74}$$

$$\zeta\tau_n/\hat{\tau}_n = \Big\langle \frac{\tau_n \ddot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n(\mathbf{T}), \boldsymbol{\Delta})}{1+\tau_n \ddot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n(\mathbf{T}), \boldsymbol{\Delta})} \Big\rangle \ , \tag{5.75}$$

where we indicated with $\|\mathbf{x}\|_0$ the zero "norm" of $\mathbf{x}$, i.e. the number of non-null entries of $\mathbf{x}$. The system above can easily be solved by first solving (5.74) for $\hat{\tau}_n$ as a function of τ_n, i.e.

$$\hat{\tau}_n = \frac{\tau_n}{\|\hat{\boldsymbol{\beta}}_n\|_0 - \eta\tau_n} \tag{5.76}$$

which gives by substitution

$$\zeta\big(\|\hat{\boldsymbol{\beta}}_n\|_0 - \eta\tau_n\big) = \Big\langle \frac{\tau_n \ddot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n(\mathbf{T}), \boldsymbol{\Delta})}{1+\tau_n \ddot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n(\mathbf{T}), \boldsymbol{\Delta})} \Big\rangle \ , \tag{5.77}$$

which can be solved numerically for τ_n, by Newton method for instance. We see in figure(5.4) that the resulting estimators, summarized as black error bar plots, are close to the quantity to be estimated (i.e. the fixed point solution of the RS equations in red). It is indeed difficult

to see differences between the estimators obtained via the CD algorithm (5.4) and the ones obtained via the AMP algorithm (5.3), as it should.

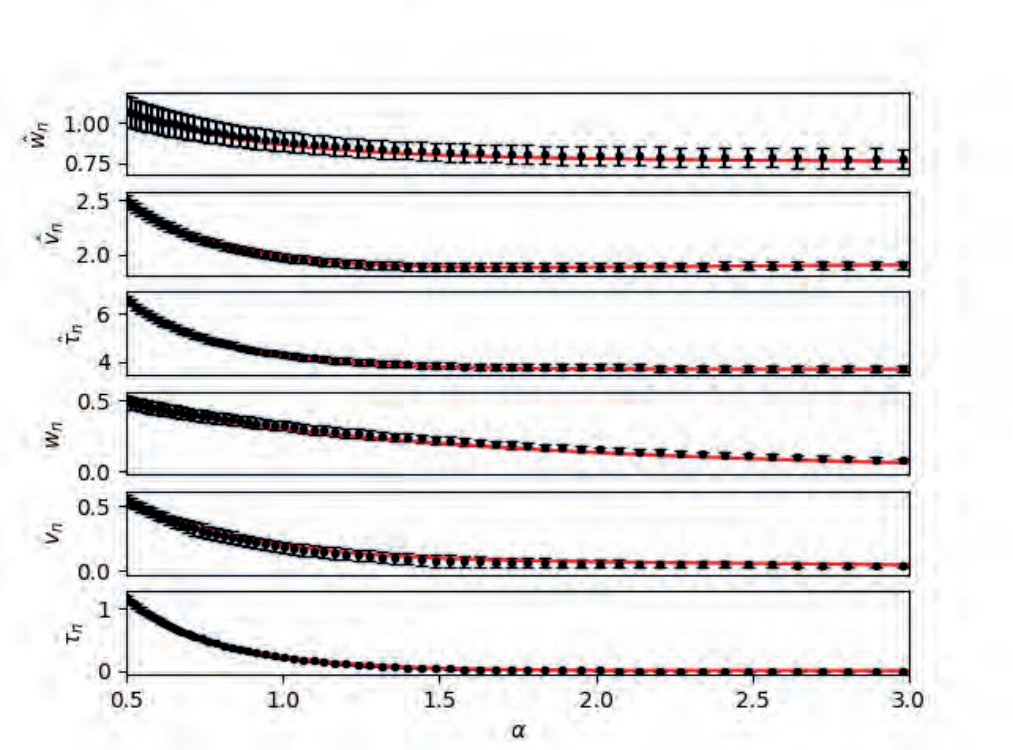


Figure 5.4: **Cox-CD with elastic net regularization** The estimators of $\hat{w}_\star, \hat{v}_\star, \hat{\tau}_\star$ (Left) and $w_\star, v_\star, \tau_\star$ (Right) summarized by a black errorbar plot against the true values in red, along a regularization path (different values of α). The data are generated from a Log-logistic proportional hazards model $\Lambda_0(t) = \log(1 + t^2/2)$ with uniform censoring between $\tau_1 = 1.0$ and $\tau_2 = 2.0$. The associations $\boldsymbol{\beta}_0$ are sampled as from a Gauss-Bernoulli distribution $\mathcal{P}_{\beta_0}(x) = (1-\nu)\delta(x) + \nu \exp\{-x^2/2\sigma^2\}/\sqrt{2\pi}\sigma$, with sparsity $\nu = 0.01$ and σ adjusted to have $\|\boldsymbol{\beta}_0\|_2^2/p = \theta_0^2 = 1.0$. The L1 ratio is $\lambda = 0.75$.

We conclude that the Cox-AMP is not the unique mean to estimate the values of the RS order parameters, i.e. it is not strictly needed. However, the additional equation (5.61), which is necessary to close the estimating equations, is derived via the insight provided by the AMP algorithm and its statistical physics interpretation. This is in agreement with recent results [1], which are however, obtained via an alternative route. From a numerical perspective the non-necessity of using Cox-AMP is reassuring, since we have already remarked that the CD algorithm is more stable.

5.9 Conclusion

We showed via simulations that the Replica Symmetric theory is capable of correctly predicting the behaviour of the Regularized Maximum Partial Likelihood Estimator in the

proportional regime when the number of covariates p, the number of samples n and the number of active covariates s diverge proportionally. When the data generating model is perfectly known, the theory gives the values of several observables of interest, such as the Bias and the Mean Squared Error, as a function of six order parameters $w_\star, v_\star, \tau_\star, \hat{w}_\star, \hat{v}_\star, \hat{\tau}_\star$ and $\theta_0 := \|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2$. However, the data generating process is not available in applications. We show here that one can estimate the RS order parameters *solely* from the data, *without* the knowledge of the data generating process.
We proposed an AMP algorithm tailored to the Cox model in section 5.4 and note that this suggests the construction of a local field

$$\boldsymbol{\psi}_\star := \hat{\boldsymbol{\beta}} - \hat{\tau}_\star \mathbf{X}' \dot{\mathcal{M}}_{g(.,\hat{\Lambda}_n(\mathbf{T}),\boldsymbol{\Delta})}(\boldsymbol{\xi}, \tau_\star) \ , \tag{5.78}$$

which is predicted to have a Gaussian distribution by comparison with the RS equations, i.e.

$$\boldsymbol{\psi}_\star \overset{d}{=} \hat{w}_\star \boldsymbol{\beta}_0/\theta_0 + \hat{v}_\star \mathbf{Z}, \quad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p) \ . \tag{5.79}$$

This identification allows to estimate all the RS order parameters directly from the data, when the RMPLE $\hat{\boldsymbol{\beta}}_n$ is obtained from the COX-AMP algorithm since in that case $\hat{\tau}$ is observed, as explained in sections 5.7. Afterwards, we noticed that, in some cases, it is possible to estimate the local field $\boldsymbol{\psi}$ also from the RMPLE obtained by any other method, e.g. Coordinate Wise Descent, i.e. without the need of explicitly using the AMP algorithm, in section 5.8. The methodology is widely applicable and not only limited to the Cox regression model.

Although we focused on the idealized setting where the covariates are i.i.d. standard Gaussian, estimating the RS order parameters from the data paves the road to future applications with real data. It is clear that a more robust theory would account also for correlations between the covariates. In that case it is not clear how one can estimate the order parameters solely from the data, this will be subject of future investigations.

Data availability statement The python scripts for generating the data, solving the RS equations, fitting the elastic net Cox model and computing the RS order parameters from the data can be found at `https://github.com/EmanueleMassa/Regularized_Cox_model` .

Bibliography

[1] PC Bellec. Observable adjustments in single-index models for regularized m-estimators, 2024.

[2] J Bradic, J Fan, and J Jiang. Regularization for cox's proportional hazards model with np-dimensionality. *The Annals of Statistics*, 39(6):3092–3120, 2011.

[3] NE Breslow. Discussion on professor cox's paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):202–220, 1972.

[4] P Bühlmann and S van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg, 2011.

[5] ACC Coolen, JE Barrett, V Paga, and C J Perez-Vicente. Replica analysis of overfitting in regression models for time-to-event data. *Journal of Physics A: Mathematical and Theoretical*, 50(37):375001, 2017.

[6] ACC Coolen, M Sheikh, A Mozeika, F Aguirre-Lopez, and F Antenucci. Replica analysis of overfitting in generalized linear regression models. *Journal of Physics A: Mathematical and Theoretical*, 53(36):365001, 2020.

[7] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[8] DL Donoho, A Maleki, and A Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[9] M Friedman. Piecewise Exponential Models for Survival Data with Covariates. *The Annals of Statistics*, 10(1):101 – 113, 1982.

[10] E Gardner and B Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271, 1988.

[11] FE Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Graduate Texts in Mathematics. Springer, 2001.

[12] Y Kabashima, T Wadayama, and T Tanaka. A typical reconstruction limit for compressed sensing based on lp-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003, 2009.

[13] JD Kalbfleisch and RL Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Series in Probability and Statistics. Wiley, 2011.

[14] S Kong and B Nan. Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Statistica Sinica*, 24(1):25, 2014.

[15] B Loureiro, C Gerbelot, H Cui, S Goldt, F Krzakala, M Mézard, and L Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114001, 2022.

[16] A Maleki and A Montanari. Analysis of approximate message passing algorithm. In *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–7. IEEE, 2010.

[17] E Massa, A Mozeika, and A C C Coolen. Replica analysis of overfitting in regression models for time to event data: the impact of censoring. *Journal of Physics A: Mathematical and Theoretical*, 57(12):125003, 2024.

[18] K Okajima, X Meng, T Takahashi, and Y Kabashima. Average case analysis of lasso under ultra-sparse conditions. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[19] M Opper, O Winther, and MJ Jordan. Expectation consistent approximate inference. *Journal of Machine Learning Research*, 6(12), 2005.

[20] S Rangan. Generalized approximate message passing for estimation with random linear mixing. *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172, 2010.

[21] RT Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1997.

[22] F Salehi, E Abbasi, and B Hassibi. *The impact of regularization on high-dimensional logistic regression*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[23] T Takahashi and Y Kabashima. A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(7):073405, 2018.

[24] SA van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614 – 645, 2008.

[25] H Zou and T Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.

5.10 Appendix

5.10.1 Replica derivation

The penalized Cox model is defined by the following energy function

$$\mathcal{H}_n(\boldsymbol{\beta}|\mathcal{D}) = \sum_{i=1}^{n} \Delta_i \Big[\log\Big(\frac{1}{n}\sum_{j=1}^{n} \Theta(T_j - T_i) \mathrm{e}^{\mathbf{X}_i'\boldsymbol{\beta}}\Big) - \mathbf{X}_i'\boldsymbol{\beta}\Big] + r(\boldsymbol{\beta}) \tag{5.80}$$

where we indicate with r the elastic net penalization

$$r(\boldsymbol{\beta}) = \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2 + \alpha|\boldsymbol{\beta}| \ . \tag{5.81}$$

Equivalently we can regard the energy as a functional of the empirical distribution

$$P_n(\Delta, t, h|\boldsymbol{\beta}, \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \delta(t - T_i)\delta_{\Delta,\Delta_i}\delta(h - \mathbf{X}_i'\boldsymbol{\beta}) \tag{5.82}$$

where we indicated with $\mathcal{D}$ the data-set, i.e. the set of couples $\{T_j, \mathbf{X}_j\}_{j=1}^n$, which plays here the role of the disorder. Explicitly

$$\mathcal{H}\Big[P_n(.|\boldsymbol{\beta}, \mathcal{D})\Big] = n\mathcal{E}\Big[P_n(.|\boldsymbol{\beta}, \mathcal{D})\Big] + r(\boldsymbol{\beta}), \tag{5.83}$$

$$\mathcal{E}\Big[P_n(.|\boldsymbol{\beta}, \mathcal{D})\Big] =$$

$$\sum_{\Delta=\pm1} \int \Delta\Big(\log S_n(t|\boldsymbol{\beta},\ \mathcal{D}) - h\Big) P_n(\Delta, t, h|\boldsymbol{\beta}, \mathcal{D})\mathrm{d}h\mathrm{d}t, \tag{5.84}$$

$$S_n(t|\boldsymbol{\beta}, \mathcal{D}) = \sum_{\Delta'=\pm1} \int \Theta(t - t')\mathrm{e}^{h'} P_n(\Delta', t,' h'|\boldsymbol{\beta}, \mathcal{D})\mathrm{d}h'\mathrm{d}t' \ . \tag{5.85}$$

It is well known that the information content of inference is quantified by the free energy

$$f(\gamma) := -\lim_{n\to\infty} \frac{1}{n\gamma}\mathbb{E}_{\mathcal{D}}\Big[\log Z_n(\gamma, \mathcal{D})\Big], \quad Z_n(\gamma) := \int \mathrm{e}^{-\gamma\mathcal{H}[P_n(.|\boldsymbol{\beta},\mathcal{D})]}\mathrm{d}\boldsymbol{\beta} \tag{5.86}$$

To compute the average of the logarithm we use the replica trick

$$f(\gamma) = \lim_{n\to\infty}\lim_{r\to 0} f_n^{(r)}(\gamma), \qquad f_n^{(r)}(\gamma) := -\frac{1}{nr}\log \mathbb{E}_{\mathcal{D}}\Big[Z_n^r(\gamma, \mathcal{D})\Big] \ . \tag{5.87}$$

We refer to $f_n^{(r)}(\gamma)$ as the replicated free energy. We now compute this quantity.

The replicated free energy

In order to compute

$$\mathbb{E}_{\mathcal{D}}\Big[Z_n^r(\gamma,\mathcal{D})\Big] \;=\; \int \mathrm{e}^{-\gamma\sum_{\alpha=1}^r \mathcal{H}[P_n(.|\boldsymbol{\beta}_\alpha,\mathcal{D})]}\prod_{\alpha=1}^r \mathrm{d}\boldsymbol{\beta}_\alpha \;, \tag{5.88}$$

We first write $Z_n(\gamma,\mathcal{D})$ in a more convenient form. Let us introduce the functional delta measure

$$\begin{aligned}\delta\Big[\mathcal{P}(.)-P_n(.|\boldsymbol{\beta}_\alpha,\mathcal{D})\Big] &\propto \int \mathrm{e}^{\mathrm{i}n\Big\{\sum_{\Delta=\pm1}\int\hat{\mathcal{P}}(\Delta,t,h)\Big[\mathcal{P}(\Delta,t,h)-P_n(\Delta,t,h|\boldsymbol{\beta},\mathcal{D})\Big]\mathrm{d}t\mathrm{d}h\Big\}}\mathcal{D}\hat{\mathcal{P}}\\ &= \int \mathrm{e}^{\mathrm{i}n\Big\{\sum_{\Delta=\pm1}\int\hat{\mathcal{P}}(\Delta,t,h)\mathcal{P}(\Delta,t,h)\mathrm{d}t\mathrm{d}h\Big\}-\mathrm{i}\sum_{i=1}^n\hat{\mathcal{P}}(\Delta_i,T_i,\mathbf{X}_i'\boldsymbol{\beta}_\alpha)}\mathcal{D}\hat{\mathcal{P}}\end{aligned} \tag{5.89}$$

obtaining

$$\begin{aligned}Z_n(\gamma,\mathcal{D}) \;&=\; \int \mathrm{e}^{n\Big\{\mathrm{i}\sum_{\Delta=\pm1}\int\hat{\mathcal{P}}(\Delta,t,h)\mathcal{P}(\Delta,t,h)\mathrm{d}t\mathrm{d}h-\gamma\mathcal{E}\left[\mathcal{P}(.)\right]\Big\}}\times\\ &\times\;\Big\{\int \mathrm{e}^{-\mathrm{i}\sum_{i=1}^n\hat{\mathcal{P}}(\Delta_i,T_i,\mathbf{X}_i'\boldsymbol{\beta})-\gamma r(\boldsymbol{\beta}/\sqrt{p})}\mathrm{d}\boldsymbol{\beta}\Big\}\;\mathcal{D}\hat{\mathcal{P}}\;\mathcal{D}\mathcal{P}\;.\end{aligned} \tag{5.90}$$

For integer r we then have

$$\begin{aligned}Z_n^r(\gamma,\mathcal{D}) = \int \mathrm{e}^{n\sum_{\alpha=1}^r\Big\{\mathrm{i}\sum_{\Delta=\pm1}\int\hat{\mathcal{P}}_\alpha(\Delta,t,h,\gamma)\mathcal{P}_\alpha(\Delta,t,h,\gamma)\mathrm{d}t\mathrm{d}h-\gamma\mathcal{E}\left[\mathcal{P}_\alpha(.)\right]\Big\}}\times\\ \times\Big\{\int\prod_{\alpha=1}^r \mathrm{e}^{-\mathrm{i}\sum_{i=1}^n\hat{\mathcal{P}}_\alpha(\Delta_i,T_i,\mathbf{X}_i'\boldsymbol{\beta}_\alpha)-\gamma r(\boldsymbol{\beta}_\alpha)}\mathrm{d}\boldsymbol{\beta}_\alpha\Big\}\;\prod_{\alpha=1}^r\mathcal{D}\hat{\mathcal{P}}_\alpha\;\mathcal{D}\mathcal{P}_\alpha\;.\end{aligned} \tag{5.91}$$

Taking the expectation with respect to the data-set

$$\mathbb{E}_{\mathcal{D}}\Big[Z_n^r(\gamma,\mathcal{D})\Big] = \int \mathrm{e}^{n\sum_{\alpha=1}^r\mathcal{A}\left[\hat{\mathcal{P}}_\alpha,\mathcal{P}_\alpha\right]}\mathcal{W}\big[\{\hat{\mathcal{P}}_\alpha\}\big]\prod_{\alpha=1}^r\mathcal{D}\hat{\mathcal{P}}_\alpha\;\mathcal{D}\mathcal{P}_\alpha \tag{5.92}$$

where

$$\mathcal{A}\big[\gamma,\hat{\mathcal{P}}_\alpha,\mathcal{P}_\alpha\big] = \mathrm{i}\sum_{\Delta=\pm1}\int\hat{\mathcal{P}}_\alpha(\Delta,t,h)\mathcal{P}_\alpha(\Delta,t,h)\mathrm{d}t\mathrm{d}h-\gamma\mathcal{E}\big[\mathcal{P}_\alpha(.)\big] \tag{5.93}$$

and

$$\mathcal{W}\big[\gamma,\{\hat{\mathcal{P}}_\alpha\}\big] = \int\Big(\mathbb{E}_{\Delta,T,\mathbf{X}}\Big[\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^r\hat{\mathcal{P}}_\alpha(\Delta,T,\mathbf{X}'\boldsymbol{\beta}_\alpha)}\Big]\Big)^n\mathrm{e}^{-\gamma r(\boldsymbol{\beta}_\alpha)}\prod_{\alpha=1}^r\mathrm{d}\boldsymbol{\beta}_\alpha\;. \tag{5.94}$$

We notice that the expression above depends on $\boldsymbol{\beta}_\alpha$ only via the linear predictors $\mathbf{X}_i'\boldsymbol{\beta}_\alpha$. Since $\mathbf{X}\sim\mathcal{N}(\mathbf{0},\frac{1}{p}\boldsymbol{I}_{p\times p})$, we have that

$$\mathbf{Y}=(Y_0,Y_1,\ldots,Y_r)\sim\mathcal{N}\big(\mathbf{0},\mathbf{C}(\{\boldsymbol{\beta}_\alpha\})\big),\quad Y_\alpha:=\mathbf{X}'\boldsymbol{\beta}_\alpha \tag{5.95}$$

with

$$\mathbf{C}(\{\boldsymbol{\beta}_\alpha\})) = \begin{pmatrix}\theta_0^2 & \mathbf{M}'\\ \mathbf{M} & \mathbf{R}\end{pmatrix} \tag{5.96}$$

where

$$\theta_0^2 = \|\boldsymbol{\beta}_0\|^2/p \tag{5.97}$$
$$\mathbf{M} = (M_\alpha)_{\alpha=1}^r, \quad M_\alpha := \boldsymbol{\beta}_0'\boldsymbol{\beta}_\alpha/p \tag{5.98}$$
$$\mathbf{R} = (R_{\alpha,\rho})_{\alpha,\rho=1}^r, \quad R_{\alpha,\rho} := \boldsymbol{\beta}_\alpha'\boldsymbol{\beta}_\rho/p = R_{\rho,\alpha} \ . \tag{5.99}$$

Then, introducing a matrix delta function, we can write

$$\mathcal{W}\big[\{\hat{\mathcal{P}}_\alpha\}\big] = \int \mathrm{e}^{ip\mathrm{Tr}(\hat{\mathbf{C}}\mathbf{C})+p\phi(\hat{\mathbf{C}})+n\varphi(\gamma,\mathbf{C})}\ \mathrm{d}\hat{\mathbf{C}}\mathrm{d}\mathbf{C} \tag{5.100}$$

with

$$\varphi[\gamma,\mathbf{C},\{\hat{\mathcal{P}}_\alpha(.)\}_{\alpha=1}^r] = \log \mathbb{E}_{\Delta,T,\mathbf{Y}}^C\Big[\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^r \hat{\mathcal{P}}_\alpha(\Delta,T,Y_\alpha)}\Big] \tag{5.101}$$
$$\phi(\gamma,\hat{\mathbf{C}}) = \frac{1}{p}\log\int \mathrm{e}^{-\mathrm{i}p\mathrm{Tr}(\hat{\mathbf{C}}\mathbf{C}(\{\boldsymbol{\beta}_\alpha\}))-\gamma\frac{1}{2}\mathrm{r}(\boldsymbol{\beta}_\alpha)}\prod_{\alpha=1}^r \mathrm{d}\boldsymbol{\beta}_\alpha \ . \tag{5.102}$$

Putting everything together, we have obtained a so-called "saddle point" (functional) integral

$$\mathbb{E}_{\mathcal{D}}\Big[Z_n^r(\gamma,\mathcal{D})\Big] = \int \mathrm{e}^{-n\psi\left[\{\mathcal{P}_\alpha,\hat{\mathcal{P}}_\alpha\}_{\alpha=1}^r,\hat{\mathbf{C}},\mathbf{C}\right]}\prod_{\alpha=1}^r \mathcal{D}\hat{\mathcal{P}}_\alpha\ \mathcal{D}\mathcal{P}_\alpha\mathrm{d}\mathbf{C}\mathrm{d}\hat{\mathbf{C}} \tag{5.103}$$

where

$$\begin{aligned}-\psi\big[\{\mathcal{P}_\alpha,\hat{\mathcal{P}}_\alpha\}_{\alpha=1}^r,\hat{\mathbf{C}},\mathbf{C}\big] &= \sum_{\alpha=1}^r \mathcal{A}\big[\hat{\mathcal{P}}_\alpha,\mathcal{P}_\alpha\big] + \varphi[\gamma,\mathbf{C},\{\hat{\mathcal{P}}_\alpha(.)\}_{\alpha=1}^r] + \\ &+ \mathrm{i}\zeta\mathrm{Tr}(\hat{\mathbf{C}}\mathbf{C}) + \zeta\phi(\gamma,\hat{\mathbf{C}}) \ .\end{aligned} \tag{5.104}$$

Saddle point integration

The idea is now to evaluate the integral via the saddle point method by interchanging the limits $n\to\infty$ and $r\to 0$, as customary in these calculations [5, 6, 15, 17]. We first derive the stationary conditions with respect to the functions $\mathcal{P}_\alpha$ and $\hat{\mathcal{P}}_\alpha$, which are obtained via functional differentiation

$$\mathcal{P}_\alpha(\Delta,t,h,\gamma) = \frac{\mathbb{E}_{\mathbf{Y}}^{\mathbf{C}}\Big[f(\Delta,t|Y_0)\delta(h-Y_\alpha)\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^r \hat{\mathcal{P}}_\alpha(\Delta,t,Y_\alpha,\gamma)}\Big]}{\mathbb{E}_{\Delta,T,\mathbf{Y}}^{\mathbf{C}}\Big[\mathrm{e}^{-\mathrm{i}\sum_{\alpha=1}^r \hat{\mathcal{P}}_\alpha(\Delta,T,Y_\alpha,\gamma)}\Big]} \tag{5.105}$$
$$\mathrm{i}\hat{\mathcal{P}}_\alpha(\Delta,t,h,\gamma) = \gamma\Delta\Big[\log\mathcal{S}_\alpha(t,\gamma) - h\Big] + \gamma\mathrm{e}^h\Lambda_\alpha(t,\gamma) \ . \tag{5.106}$$

where

$$\mathcal{S}_\alpha(t,\gamma) = \sum_{\Delta'=\pm 1}\int \Theta(t-t')\mathrm{e}^{h'}\mathcal{P}_\alpha(\Delta',t,'h')\mathrm{d}h'\mathrm{d}t' \tag{5.107}$$
$$\Lambda_\alpha(t,\gamma) := \sum_{\Delta'=\pm 1}\int \frac{\Delta'\Theta(t-t')\mathcal{P}_\alpha(\Delta',t',h',\gamma)}{\mathcal{S}_\alpha(t',\gamma)}\mathrm{d}t'\mathrm{d}h' \ . \tag{5.108}$$

Equivalently

$$\mathcal{P}_\alpha(\Delta,t,h,\gamma) = \frac{\mathbb{E}^{\mathbf{C}}_{\mathbf{Y}}\Big[f(\Delta,t|Y_0)\delta(h-Y_\alpha)\mathrm{e}^{-\sum_{\alpha=1}^r\gamma\left(\Delta\log\mathcal{S}_\alpha(t,\gamma)+\exp(Y_\alpha)\Lambda_\alpha(t,\gamma)-\Delta Y_\alpha\right)}\Big]}{\mathbb{E}^{\mathbf{C}}_{\Delta,T,\mathbf{Y}}\Big[\mathrm{e}^{-\sum_{\alpha=1}^r\gamma\left(\Delta\log\mathcal{S}_\alpha(T,\gamma)+\exp(Y_\alpha)\Lambda_\alpha(T,\gamma)-\Delta Y_\alpha\right)}\Big]} \,. \tag{5.109}$$

Furthermore at the saddle point we have

$$\begin{aligned}&\varphi[\gamma,\mathbf{C},\{\hat{\mathcal{P}}_\alpha(.)\}_{\alpha=1}^r] = \\ &\varphi(\gamma,\mathbf{C}) = \log\mathbb{E}^{\mathbf{C}}_{\Delta,T,\mathbf{Y}}\Big[\mathrm{e}^{-\sum_{\alpha=1}^r\gamma\left(\Delta\log\mathcal{S}_\alpha(T,\gamma)+\exp(Y_\alpha)\Lambda_\alpha(T,\gamma)-\Delta Y_\alpha\right)}\Big]\end{aligned}$$

and hence

$$\begin{aligned}-\tilde{\psi}(\gamma,\hat{\mathbf{C}},\mathbf{C}) \;&:=\; -\operatorname*{extr}_{\{\mathcal{P}_\alpha,\hat{\mathcal{P}}_\alpha\}}\psi = \mathrm{i}\zeta\mathrm{Tr}(\hat{\mathbf{C}}\mathbf{C}) + \zeta\phi(\gamma,\hat{\mathbf{C}}) + \varphi(\gamma,\mathbf{C}) + \\ &+\; \gamma\sum_{\Delta=\pm1}\int\mathrm{e}^h\sum_{\alpha=1}^r\Lambda_\alpha(t,\gamma)\mathcal{P}_\alpha(\Delta,t,h,\gamma)\mathrm{d}t\mathrm{d}h\end{aligned} \tag{5.110}$$

With a modest amount of foresight we take the following change of variables

$$\mathrm{i}\hat{\mathbf{C}} = \frac{1}{2}\mathbf{D} \tag{5.111}$$

which is expected from previous similar calculations and aids the book-keeping. In principle we could now derive the saddle point equations for the elements of the matrices $\mathbf{C}$ nor $\mathbf{D}$ and then take the limit $r\to 0$. In practice we will assume the replica symmetric ansatze

$$\mathbf{C} = \begin{pmatrix} \theta_0^2 & m & \dots & \dots & m \\ m & \rho & q & \dots & q \\ \vdots & q & \rho & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & q \\ m & q & \dots & q & \rho \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} 0 & \hat{m} & \dots & \dots & \hat{m} \\ \hat{m} & \hat{\rho} & -\hat{q} & \dots & -\hat{q} \\ \vdots & -\hat{q} & \hat{\rho} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\hat{q} \\ \hat{m} & -\hat{q} & \dots & -\hat{q} & \hat{\rho} \end{pmatrix} \tag{5.112}$$

directly from now on. Note that this directly implies that

$$\mathbf{C}^{-1} = \begin{pmatrix} \tilde{\mu}_r & \tilde{m}_r & \dots & \dots & \tilde{m}_r \\ \tilde{m}_r & \tilde{\rho}_r & \tilde{q}_r & \dots & \tilde{q}_r \\ \vdots & \tilde{q}_r & \tilde{\rho}_r & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \tilde{q}_r \\ \tilde{m}_r & \tilde{q}_r & \dots & \tilde{q}v & \tilde{\rho}_r \end{pmatrix} \,. \tag{5.113}$$

After some algebra one obtains that

$$\tilde{\mu}_r = \frac{\rho+q(r-1)}{\theta_0^2(\rho+q(r-1))-rm^2} \tag{5.114}$$

$$\tilde{m}_r = \frac{m}{rm^2-\theta_0^2(\rho+q(r-1))} \tag{5.115}$$

$$\tilde{q}_r = \frac{1}{\rho-q}\,\frac{m^2-q\theta_0^2}{\theta_0^2(\rho+q(r-1))-rm^2} \tag{5.116}$$

$$\tilde{\rho}_r = \frac{1}{\rho-q}\Big(1+\frac{m^2-q\theta_0^2}{\theta_0^2(\rho+q(r-1))-rm^2}\Big) \,. \tag{5.117}$$

We are now in position to compute the "limit" $r\to 0$, i.e. to obtain the replica symmetric symmetric free energy.

Simplification of ϕ within RS ansatz

Let's remind the definition of the potential ϕ

$$\phi(\gamma, \mathbf{D}) = \frac{1}{p} \log \int \mathrm{e}^{-\frac{1}{2} p \mathrm{Tr}(\mathbf{DC}(\{\boldsymbol{\beta}_\alpha\})) - \gamma \mathrm{r}(\boldsymbol{\beta}_\alpha)} \prod_{\alpha=1}^{r} \mathrm{d}\boldsymbol{\beta}_\alpha \ . \tag{5.118}$$

Using the replica symmetric ansatz we get

$$\begin{aligned} \frac{1}{2} p \mathrm{Tr}(\mathbf{DC}(\{\boldsymbol{\beta}_\alpha\})) &= \sum_{\rho=1}^{r} \big(\hat{m} \boldsymbol{\beta}_0' \boldsymbol{\beta}_\rho + \frac{1}{2} \hat{\rho} \boldsymbol{\beta}_\rho' \boldsymbol{\beta}_\rho \big) - \sum_{\rho, \alpha \neq \rho}^{r} \hat{q} \boldsymbol{\beta}_\rho \boldsymbol{\beta}_\alpha \\ &= \sum_{\rho=1}^{r} \big(\hat{m} \boldsymbol{\beta}_0' \boldsymbol{\beta}_\rho + \frac{1}{2} (\hat{\rho} + \hat{q}) \|\boldsymbol{\beta}_\rho\|^2 \big) - \frac{1}{2} \hat{q} \Big\| \sum_{\rho=1}^{r} \boldsymbol{\beta}_\rho \Big\|^2 \end{aligned}$$

and via Gaussian linearization

$$\phi_{RS}^{(r)}(\gamma, \hat{m}, \hat{q}, \hat{\rho}) = \frac{1}{p} \log \mathbb{E}_{\mathbf{Z}} \Bigg[\Bigg(\int \mathrm{e}^{-\frac{1}{2}(\hat{\rho}+\hat{q})\|\mathbf{x}\|^2 - (\hat{m}\boldsymbol{\beta}_0 + \sqrt{\hat{q}}\mathbf{Z})'\mathbf{x} - \gamma \mathrm{r}(\mathbf{x})} \mathrm{d}\mathbf{x} \Bigg)^r \Bigg] \ .$$

Since we are finally interested in taking the limit $r \to 0$ it is convenient to expand the integrand for small r, thus obtaining

$$\phi_{RS}^{(r)}(\gamma, \hat{m}, \hat{q}, \hat{\rho}) = r \frac{1}{p} \mathbb{E}_{\mathbf{Z}} \Bigg[\log \int \mathrm{e}^{-\frac{1}{2}(\hat{\rho}+\hat{q})\|\mathbf{x}\|^2 - (\hat{m}\boldsymbol{\beta}_0 + \sqrt{\hat{q}}\mathbf{Z})'\mathbf{x} - \gamma \mathrm{r}(\mathbf{x})} \mathrm{d}\mathbf{x} \Bigg] + O(r^2) \ .$$

Since the integrand in the expression above factorizes over the components of $\mathbf{x}$ and $\mathbf{Z}$, this can be further re-written in terms of the distribution of the entries of $\boldsymbol{\beta}_0$

$$\phi_{RS}^{(r)}(\gamma, \hat{m}, \hat{q}, \hat{\rho}) = r \mathbb{E}_{\beta_0, Z} \Bigg[\log \int \mathrm{e}^{-\frac{1}{2}(\hat{\rho}+\hat{q})x^2 - (\hat{m}\beta_0 + \sqrt{\hat{q}}Z)x - \gamma \mathrm{r}(x)} \mathrm{d}x \Bigg] + O(r^2) \ .$$

Simplification of φ within RS ansatz

Inserting the replica symmetric ansatze, we obtain

$$f(y_0, y_1, \ldots, y_r) \propto \exp \Big\{ - \Big(\frac{1}{2} \tilde{\mu} y_0^2 + \sum_{\rho=1}^{r} \big(\tilde{m} y_0 y_\rho + \frac{1}{2} (\tilde{\rho} - \tilde{q}) y_\rho^2 \big) + \frac{1}{2} \tilde{q} \big(\sum_{\rho=1}^{r} y_\rho \big)^2 \Big) \Big\} \tag{5.119}$$

and via Gaussian linearization

$$f(y_0, y_1, \ldots, y_r) \propto \mathbb{E}_Q \Big[\exp \Big\{ - \frac{1}{2} \tilde{\mu} y_0^2 - \sum_{\rho=1}^{r} \big[\frac{1}{2} (\tilde{\rho} - \tilde{q}) y_\rho^2 + \big(\tilde{m} y_0 + \mathrm{i} \sqrt{\tilde{q}} Q \big) y_\rho \big] \Big\} \Big] \tag{5.120}$$

with $Q \sim \mathcal{N}(0,1)$. Upon setting $\sqrt{\tilde{\mu}} y_0 = z_0$ we obtain

$$\begin{aligned} &\varphi_{RS}^{(r)}(\gamma, \tilde{\mu}_r, \tilde{m}_r, \tilde{q}_r, \tilde{\rho}_r) \\ &= \log \frac{\mathbb{E}_{\Delta, T, Z_0, Q} \Bigg[\Bigg(\int \mathrm{e}^{-\frac{1}{2}(\tilde{\rho}-\tilde{q})x^2 - \left(\tilde{m}/\sqrt{\tilde{\mu}} Z_0 + \mathrm{i}\sqrt{\tilde{q}} Q \right) x - \gamma g(x, \Lambda^{(r)}(T,\gamma), \Delta)} \frac{\mathrm{d}x}{\sqrt{2\pi}} \Bigg)^r \Bigg]}{\mathbb{E}_{\Delta, T, Z_0, Q} \Bigg[\Bigg(\int \mathrm{e}^{-\frac{1}{2}(\tilde{\rho}-\tilde{q})x^2 - \left(\tilde{m}/\sqrt{\tilde{\mu}} Z_0 + \mathrm{i}\sqrt{\tilde{q}} Q \right) x} \Bigg)^r \Bigg]} \ . \end{aligned}$$

with $Z_0 \sim \mathcal{N}(0,1)$, $Z_0 \perp Q$ and where we have defined

$$g(x, y, z) = \exp(x) y - z x \ . \tag{5.121}$$

Replica symmetric functional equation

Let us introduce, for notational convenience, the random variable $\tilde{\xi}$, with density

$$f^{(r)}_{\tilde{\xi}}(x|Z_0,Q,T) := \frac{\mathrm{e}^{-\left\{\frac{1}{2}(\tilde{\rho}-\tilde{q})x^2+\left(\tilde{m}/\sqrt{\tilde{\mu}}Z_0+\mathrm{i}\sqrt{\tilde{q}}Q\right)x+\gamma g(x,\Lambda^{(r)}(T,\gamma),\Delta)\right\}}}{\mathcal{Z}(\gamma,\Delta,T,Z_0,Q)} \ , \tag{5.122}$$

where the normalization is

$$\mathcal{Z}(\gamma,\Delta,T,Z_0,Q) = \int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\frac{(\tilde{\rho}-\tilde{q})}{\gamma}x^2+\frac{1}{\gamma}\left(\tilde{m}/\sqrt{\tilde{\mu}}Z_0+\mathrm{i}\sqrt{\tilde{q}}Q\right)x+\gamma g(x,\Lambda^{(r)}(T,\gamma),\Delta)\right\}}\mathrm{d}x \ .$$

Using (5.120) we obtain

$$\begin{aligned}\mathcal{P}^{(r)}(\Delta,t,h,\gamma) = &\left(\mathbb{E}_{\Delta,T,Z_0,Q}\Big[\mathcal{Z}^r(\gamma,\Delta,T,Z_0,Q)\Big]\right)^{-1}\times \\ &\times\mathbb{E}_{Z_0,Q}\left[f(\Delta,t|Z_0/\sqrt{\tilde{\mu}})\mathbb{E}_{\tilde{\xi}|Z_0,Q,T}\Big[\delta(h-\tilde{\xi})\Big]\mathcal{Z}^r(\gamma,\Delta,T,Z_0,Q)\right] \ .\end{aligned} \tag{5.123}$$

Let us stop and recall what we have achieved so far. By assuming the RS ansatz we have obtained

$$-\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}_{\mathcal{D}}\Big[Z^r_n(\gamma,\mathcal{D})\Big] = \underset{m,q,\rho,\hat{\mu},\hat{m},\hat{q},\hat{\rho}}{\mathrm{extr}}\tilde{\psi}^{(r)}_{RS}(m,q,\rho,\hat{m},\hat{q},\hat{\rho}) \tag{5.124}$$

with

$$\begin{aligned}-\tilde{\psi}^{(r)}_{RS}(\dots) = &\ \zeta(\mu\hat{\mu}+rm\hat{m}+r(\rho\hat{\rho}-q\hat{q})+r^2q\hat{q})+ \\ &\zeta\phi^{(r)}_{RS}(\gamma,\hat{m},\hat{q},\hat{\rho})+\varphi^{(r)}_{RS}(\gamma,\tilde{\mu}_r,\tilde{m}_r,\tilde{q}_r,\tilde{\rho}_r)+ \\ &\gamma r\mathbb{E}_{\Delta,T,Z_0,Q}\left[\Lambda^{(r)}(t,\gamma)\mathbb{E}_{\tilde{\xi}|Z_0,Q,T}\Big[\exp(\tilde{\xi})\Big]\mathcal{Z}^r(\gamma,\Delta,T,Z_0,Q)\right] \ ,\end{aligned} \tag{5.125}$$

where now

$$\begin{aligned}\mathcal{S}^{(r)}(t',\gamma) &= \mathbb{E}_{\Delta,T,Z_0,Q}\left[\Theta(T-t')\mathbb{E}_{\tilde{\xi}|Z_0,Q,T}\Big[\exp(\tilde{\xi})\Big]\mathcal{Z}^r(\gamma,\Delta,T,Z_0,Q)\right] \\ \Lambda^{(r)}(t,\gamma) &:= \sum_{\Delta'=\pm1}\int\frac{\Delta'\Theta(t-t')}{\mathcal{S}(t',\gamma)}\mathcal{P}^{(r)}(\Delta',t',h',\gamma)\mathrm{d}t'\mathrm{d}h' \ .\end{aligned}$$

The limit $r\to 0$

Taking the limit $r\to 0$, we get the replica symmetric free energy

$$f_{RS}(\gamma) = \underset{\mu,m,q,\rho,\hat{m},\hat{q},\hat{\rho}}{\mathrm{extr}}\tilde{f}_{RS}(\gamma,\mu,m,q,\rho,\hat{m},\hat{q},\hat{\rho}) \tag{5.126}$$

with

$$\tilde{f}_{RS}(\gamma,m,q,\rho,\hat{m},\hat{q},\hat{\rho}) := \lim_{r\to0}\frac{1}{\gamma r}\tilde{\psi}^{(r)}_{RS}(\gamma,m,q,\rho,\hat{m},\hat{q},\hat{\rho}) \ . \tag{5.127}$$

where

$$-\tilde{f}_{RS}(\dots) = \zeta(m\hat{m} + r(\rho\hat{\rho} - q\hat{q})\big) + \zeta\tilde{\phi}_{RS}(\gamma, \hat{m}, \hat{q}, \hat{\rho}) + \tilde{\varphi}_{RS}(\gamma, m, q, \rho) + \mathbb{E}_{\Delta,T,Z_0,Q}\left[\Lambda(t,\gamma)\mathbb{E}_{\tilde{\xi}|Z_0,Q,T}\Big[\exp(\tilde{\xi})\Big]\right], \tag{5.128}$$

and the random variable $\tilde{\xi}$ has the density

$$f_{\tilde{\xi}}(x|Z_0,Q,T) := \frac{\mathrm{e}^{-\frac{1}{\rho-q}\left\{\frac{1}{2}x^2 - \left((m/\theta_0)Z_0 + \sqrt{q-(m/\theta_0)^2}Q\right)x\right\} - \gamma g(x,\Lambda(t,\gamma),\Delta)}}{\int \mathrm{e}^{-\frac{1}{\rho-q}\left\{\frac{1}{2}x^2 - \left((m/\theta_0)Z_0 + \sqrt{q-(m/\theta_0)^2}Q\right)x\right\} - \gamma g(x,\Lambda(t,\gamma),\Delta)}\mathrm{d}x}. \tag{5.129}$$

Above, we introduced the following definitions

$$\begin{aligned}
\tilde{\phi}_{RS}(\gamma,\hat{m},\hat{q},\hat{\rho}) &= \lim_{r\to 0}\frac{1}{\gamma r}\phi^{(r)}_{RS}(\gamma,\hat{m},\hat{q},\hat{\rho}) \\
&= \mathbb{E}_{\beta_0,Z}\left[\log\int \mathrm{e}^{-\frac{1}{2}(\hat{\rho}+\hat{q})x^2 - (\hat{m}\beta_0 + \sqrt{\hat{q}}Z)x - \gamma \mathrm{r}(x)}\mathrm{d}x\right] \\
\tilde{\varphi}_{RS}(\gamma,m,q,\rho) &= \lim_{r\to 0}\frac{1}{\gamma r}\varphi^{(r)}_{RS}(\gamma,\tilde{\mu}_r,\tilde{m}_r,\tilde{q}_r,\tilde{\rho}_r) \\
&= \mathbb{E}_{\Delta,T,Z_0,Q}\left[\log\frac{\int \mathrm{e}^{-\frac{1}{\rho-q}\left\{\frac{1}{2}x^2 - \left((m/\theta_0)Z_0 + \sqrt{q-(m/\theta_0)^2}Q\right)x\right\} - \gamma g(x,\Lambda(t,\gamma),\Delta)}\mathrm{d}x}{\int \mathrm{e}^{-\frac{1}{\rho-q}\left\{\frac{1}{2}x^2 - \left((m/\theta_0)Z_0 + \sqrt{q-(m/\theta_0)^2}Q\right)x\right\}}\mathrm{d}x}\right]
\end{aligned}$$

and we have used that

$$\tilde{\mu} = \lim_{r\to 0}\tilde{\mu}_r = 1/\theta_0^2 \tag{5.130}$$

$$\tilde{m} = \lim_{r\to 0}\tilde{m}_r = -\frac{m}{\theta_0^2(\rho-q)} \tag{5.131}$$

$$\tilde{q} = \lim_{r\to 0}\tilde{q}_r = \frac{1}{\rho-q}\frac{m^2 - q\theta_0^2}{\theta_0^2(\rho-q)} \tag{5.132}$$

$$\tilde{\rho} - \tilde{q} = \lim_{r\to 0}\tilde{\rho}_r - \tilde{q}_r = \frac{1}{\rho-q}. \tag{5.133}$$

Assuming the following scaling

$$\tau = \gamma/(\tilde{\rho} - \tilde{q}) = \gamma(\rho - q) \tag{5.134}$$

we obtain

$$\begin{aligned}
&\tilde{\varphi}_{RS}(\gamma,m,q,\rho) = \\
&= \mathbb{E}_{\Delta,T,Z_0,Q}\left[\frac{1}{\gamma}\log\frac{\int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\left(x-((m/\theta_0)Z_0+\sqrt{q-(m/\theta_0)^2}Q)\right)^2/\tau + g\left(x,\Lambda(T,\gamma),\Delta\right)\right\}}\mathrm{d}x}{\int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\left(x-((m/\theta_0)Z_0+\sqrt{q-(m/\theta_0)^2}Q)\right)^2/\tau\right\}}\mathrm{d}x}\right].
\end{aligned}$$

Similarly, taking the additional rescaling

$$1/\hat{\tau} = (\hat{\rho} + \hat{q})/\gamma, \quad \hat{m} = \gamma\hat{m}, \quad \hat{q} = \gamma^2\hat{q} \tag{5.135}$$

we have

$$\tilde{\phi}_{RS}(\gamma,\hat{m},\hat{q},\hat{\rho}) = \mathbb{E}_{\beta_0,Z}\left[\frac{1}{\gamma}\log\int \mathrm{e}^{-\gamma\left\{\frac{1}{2\hat{\tau}}x^2+(\hat{m}\beta_0+\sqrt{\hat{q}}Z)x+\mathrm{r}(x)\right\}}\mathrm{d}x\right] =$$
$$= \frac{1}{2}\hat{\tau}\left(\hat{m}^2\theta_0^2+\hat{q}\right) + \mathbb{E}_{\beta_0,Z}\left[\frac{1}{\gamma}\log\int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\frac{1}{\hat{\tau}}\left\{\tilde{x}+\hat{\tau}(\hat{m}\beta_0+\sqrt{\hat{q}}Z)\right\}^2+\mathrm{r}(\tilde{x})\right\}}\mathrm{d}\tilde{x}\right] + \text{const}$$

Furthermore, in the limit $r \to 0$, the functional equation reads

$$\mathcal{P}(\Delta,t,h,\gamma) = \mathbb{E}_{Z_0,Q}\left[f(\Delta,t|Z_0/\sqrt{\tilde{\mu}})\mathbb{E}_{\tilde{\xi}|Z_0,Q,T}\left[\delta(h-\tilde{\xi})\right]\right]$$

where

$$\mathcal{S}(t',\gamma) = \mathbb{E}_{\Delta,T,Z_0,Q}\left[\Theta(T-t')\mathbb{E}_{\tilde{\xi}|Z_0,Q,T}\left[\exp(\tilde{\xi})\right]\right]$$
$$\Lambda(t,\gamma) := \mathbb{E}_{\Delta',T'}\left[\frac{\Delta'\Theta(t-T')}{\mathcal{S}(T',\gamma)}\right],$$

with

$$f_{\tilde{\xi}}(x|Z_0,Q,T) := \frac{\mathrm{e}^{-\gamma\left\{\frac{1}{2}\left(x-((m/\theta_0)Z_0+\sqrt{q-(m/\theta_0)^2}Q)\right)^2/\tau+g\left(x,\Lambda(T,\gamma),\Delta\right)\right\}}}{\int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\left(x-((m/\theta_0)Z_0+\sqrt{q-(m/\theta_0)^2}Q)\right)^2/\tau+g\left(x,\Lambda(T,\gamma),\Delta\right)\right\}}\mathrm{d}x}, \tag{5.136}$$

and

$$\int \mathcal{P}(\Delta,t,h,\gamma)\mathrm{d}h = f(\Delta,t|Z_0/\sqrt{\tilde{\mu}}). \tag{5.137}$$

The limit $\gamma \to \infty$

Via Laplace integration we see that

$$-\lim_{\gamma\to\infty}\frac{1}{\gamma}\log\int \mathrm{e}^{-\gamma\left\{\frac{1}{2\alpha}(z-x)^2+b(z)\right\}}\mathrm{d}z = \mathcal{M}_{b(.)}(x,\alpha) \tag{5.138}$$

where $\mathcal{M}_{b(.)}$ is the Moureau envelope of a convex function $b:\mathbb{R}\to\mathbb{R}$, which is defined as

$$\mathcal{M}_{b(.)}(x,\alpha) = \min_z\left\{\frac{1}{2\alpha}(z-x)^2+b(z)\right\}. \tag{5.139}$$

Using the "Laplace-Moureau" identity above (5.138), we obtain

$$\lim_{\gamma\to\infty}\tilde{\phi}_{RS}(\gamma,\hat{m},\hat{q},\hat{\rho}) =$$
$$\frac{1}{2}\hat{\tau}\left(\hat{m}^2\theta_0^2+\hat{q}\right) - \mathbb{E}_{Z,\beta_0}\left[\mathcal{M}_{\mathrm{r}(.)}\left(\hat{m}\beta_0+\sqrt{\hat{q}}Z,\hat{\tau}\right)\right]$$
$$\lim_{\gamma\to\infty}\tilde{\varphi}_{RS}(\gamma,m,q,\rho) =$$
$$-\mathbb{E}_{\Delta,T,Z_0,Q}\left[\mathcal{M}_{\tau g(.,\Lambda(T),\Delta)}\left(m/\theta_0 Z_0+\sqrt{q-(m/\theta_0)^2}Q,\tau\right)\right].$$

In the limit $\gamma \to \infty$, the functional equation (5.136) reads

$$\mathcal{P}(\Delta, t, h) = \mathbb{E}_{Z_0,Z}\Big[f(\Delta, t|Z_0)\delta\Big(h - \xi(T, Z_0, Z)\Big)\Big] \tag{5.140}$$

$$\xi(T, Z_0, Z) := \mathrm{prox}_{g(.,\Lambda(T),\Delta)}\Big((m/\theta_0)Z_0 + \sqrt{q - (m/\theta_0)^2}Z, \tau\Big), \tag{5.141}$$

where we used the Laplace integration method to conclude that for any "well behaved" function $a : \mathbb{R}^d \to \mathbb{R}$

$$a\Big(\mathrm{prox}_{b(.)}(x, \alpha)\Big) = \lim_{\gamma\to\infty} \log \frac{\int \mathrm{e}^{-\gamma\Big\{\frac{1}{2\alpha}(z-x)^2 + b(z)\Big\}} a(z)\mathrm{d}z}{\int \mathrm{e}^{-\gamma\Big\{\frac{1}{2\alpha}(z-x)^2 + b(z)\Big\}} \mathrm{d}z} \tag{5.142}$$

with $\mathrm{prox}_{b(.)}$ the proximal mapping of the convex function b defined as

$$\mathrm{prox}_{b(.)}(x, \alpha) = \arg\min_z \Big\{ \frac{1}{2\alpha}(z - x)^2 + b(z) \Big\} . \tag{5.143}$$

We also have

$$\begin{aligned} \mathcal{S}(t) &= \mathbb{E}_{\Delta,T,Z_0,Z}\Big[\Theta(t - T)\mathrm{e}^{\xi}\Big] \\ \Lambda(t) &:= \mathbb{E}_{\Delta',T'}\left[\frac{\Delta'\Theta(t - T')}{\mathcal{S}(T')}\right] \end{aligned}$$

where ξ is defined in (5.141). Furthermore

$$\begin{aligned} &\lim_{\gamma\to\infty} \mathbb{E}_{\Delta,T,Z_0,Z}\left[\Lambda(t, \gamma)\mathbb{E}_{\xi|Z_0,Q,T}\Big[\exp(\xi)\Big]\right] \\ = \quad &\mathbb{E}_{\Delta,T}\left[\frac{\Delta\mathbb{E}_{\Delta'',T'',Y_0'',Z''}\Big[\Theta(T'' - T)\mathrm{e}^{\xi''}\Big]}{\mathbb{E}_{\Delta',T',Y_0',Z'}\Big[\Theta(T' - T)\mathrm{e}^{\xi'}\Big]}\right] = \mathbb{E}\Big[\Delta\Big] = \text{const.} \end{aligned}$$

Putting all together (for the last time), we have obtained that

$$\begin{aligned} &\lim_{\gamma\to\infty} \tilde{f}_{RS}(\gamma, m, q, \rho, \hat{m}, \hat{q}, \hat{\rho}) = \\ &\qquad \mathcal{F}(m, q, \tau, \hat{m}, \hat{q}, \hat{\tau}) = \lim_{n\to\infty} \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\mathcal{P}\mathcal{L}_n(\hat{\boldsymbol{\beta}}_n)\Big] \end{aligned} \tag{5.144}$$

with

$$\begin{aligned} \mathcal{F}(m, q, \tau, \hat{m}, \hat{q}, \hat{\tau}) = &-\zeta(m\hat{m} + \frac{1}{2}q/\hat{\tau} - \frac{1}{2}\tau\hat{q}) + \frac{1}{2}\hat{\tau}\zeta\Big(\hat{m}^2\theta_0^2 + \hat{q}\Big) + \\ &\zeta\mathbb{E}_{Z,\beta_0}\Big[\mathcal{M}_{\mathrm{r}(.)}\Big(\hat{m}\beta_0 + \sqrt{\hat{q}}Z, \hat{\tau}\Big)\Big] + \\ &\mathbb{E}_{\Delta,T,Z_0,Q}\Big[\mathcal{M}_{g(.,\Lambda(T),\Delta)}\Big(m/\theta_0 Z_0 + \sqrt{q - (m/\theta_0)^2}Q, \tau\Big)\Big] + \text{const} \end{aligned}$$

It is now convenient to take additional change of variables in order to obtain "neater" formulae. Let us define

$$w = m/\theta_0, \quad v = q - m^2/\theta_0^2, \quad \hat{w} = -\hat{\tau}\hat{m}\theta_0, \quad \hat{v} = \hat{\tau}\sqrt{\hat{q}} \tag{5.145}$$

then (abusing the notation yet again)

$$\mathcal{F}(w,v,\tau,\hat{w},\hat{v},\hat{\tau}) = -\frac{\zeta}{2\hat{\tau}}\Big((w-\hat{w})^2+v^2+\hat{v}^2(1-\tau/\hat{\tau})\Big) + \zeta\mathbb{E}_{Z,\beta_0}\Big[\mathcal{M}_{\mathrm{r}(.)}\Big(\hat{w}\frac{\beta_0}{\theta_0}+\hat{v}Z,\hat{\tau}\Big)\Big] + \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\mathcal{M}_{g(.,\Lambda(T),\Delta)}\Big(wZ_0+vQ,\tau\Big)\Big] + \text{const} \tag{5.146}$$

5.10.2 Replica symmetric equations

It is convenient to define

$$\varphi(\beta_0,Z) := \mathrm{prox}_{\mathrm{r}(.)}\Big(\hat{w}\frac{\beta_0}{\theta_0}+\hat{v}Z,\hat{\tau}\Big) \tag{5.147}$$

$$\xi := \mathrm{prox}_{g(.,\Lambda(T),\Delta)}(wZ_0+vQ,\tau) \tag{5.148}$$

then

$$\frac{\partial}{\partial\hat{w}}\mathcal{F} = \frac{\zeta}{\hat{\tau}}w\frac{\zeta}{\hat{\tau}}\mathbb{E}_{Z,\beta_0}\Big[\beta_0\varphi\Big]/\theta_0 \;\rightarrow\; w=\mathbb{E}_{Z,\beta_0}\Big[\beta_0\varphi\Big]/\theta_0 \tag{5.149}$$

$$\frac{\partial}{\partial\hat{v}}\mathcal{F} = \frac{\zeta}{\hat{\tau}}\hat{v}\frac{\tau}{\hat{\tau}}\frac{\zeta}{\hat{\tau}}\mathbb{E}_{Z,\beta_0}\Big[Z\varphi\Big] \;\rightarrow\; \hat{v}\frac{\tau}{\hat{\tau}}=\mathbb{E}_{Z,\beta_0}\Big[Z\varphi\Big] \tag{5.150}$$

$$\begin{aligned}\frac{\partial}{\partial\hat{\tau}}\mathcal{F} &= \frac{\zeta}{2\hat{\tau}^2}\Big((w-\hat{w})^2+v^2+\hat{v^2}(1-2\frac{\tau}{\hat{\tau}})\Big)+\\ &- \frac{\zeta}{2\hat{\tau}^2}\mathbb{E}_{Z,\beta_0}\Big[\big(\varphi-\hat{w}\beta_0-\hat{v}Z\big)^2\Big]=\\ &= \frac{\zeta}{2\hat{\tau}^2}\Big(w^2+v^2\Big)-\frac{\zeta}{2\hat{\tau}^2}\mathbb{E}_{Z,\beta_0}\Big[\varphi^2\Big]\end{aligned} \tag{5.151}$$

$$\frac{\partial}{\partial w}\mathcal{F} = -\zeta(w-\hat{w})/\hat{\tau}+\frac{1}{\tau}\Big(w-\mathbb{E}_{\Delta,T,Z_0,Q}\Big[Z_0\xi\Big]\Big) \tag{5.152}$$

$$\frac{\partial}{\partial v}\mathcal{F} = -\zeta v/\hat{\tau}+\frac{1}{\tau}\Big(v-\mathbb{E}_{\Delta,T,Z_0,Q}\Big[Q\xi\Big]\Big) \tag{5.153}$$

$$\frac{\partial}{\partial\tau}\mathcal{F} = \zeta\frac{\hat{v}^2}{\hat{\tau}^2}-\frac{1}{\tau^2}\mathbb{E}_{\Delta,T,Z_0,Q}\Big[(\xi-wZ_0-vQ)^2\Big]\,. \tag{5.154}$$

Hence

$$w = \mathbb{E}_{Z,\beta_0}\Big[\beta_0\varphi\Big]/\theta_0 \tag{5.155}$$

$$\hat{v}\frac{\tau}{\hat{\tau}} = \mathbb{E}_{Z,\beta_0}\Big[Z\varphi\Big] \tag{5.156}$$

$$(w^2+v^2) = \mathbb{E}_{Z,\beta_0}\Big[\varphi^2\Big] \tag{5.157}$$

$$\hat{w} = w-\frac{\hat{\tau}}{\zeta\tau}\Big(w-\mathbb{E}_{\Delta,T,Z_0,Q}\Big[Z_0\xi\Big]\Big) \tag{5.158}$$

$$v(1-\zeta\tau/\hat{\tau}) = \mathbb{E}_{\Delta,T,Z_0,Q}\Big[Q\xi\Big] \tag{5.159}$$

$$\zeta\hat{v}^2 = \frac{\hat{\tau}^2}{\tau^2}\mathbb{E}_{\Delta,T,Z_0,Q}\Big[(\xi-wZ_0-vQ)^2\Big] \tag{5.160}$$

5.10.3 Explicit computation for the lasso regularization

Here we assume that

$$\mathbf{e}'_\mu \boldsymbol{\beta}_0 \sim \mathcal{N}(0, \theta_0^2/\nu), \quad \nu := s/p \ . \tag{5.161}$$

For the Lasso regularization

$$\begin{aligned} \mathrm{prox}_{\mathrm{r}(.)}(x, \hat{\tau}) &= \mathrm{st}(x, \alpha\hat{\tau}), & (5.162)\\ \mathrm{st}(x, \alpha) &:= \mathrm{relu}(x-\alpha) - \mathrm{relu}(-x-\alpha) & (5.163) \end{aligned}$$

with

$$\mathrm{relu}(x) = x\Theta(x) \ . \tag{5.164}$$

Before proceeding, let us remember some useful fact

$$\begin{aligned} \mathbb{E}_Z\Big[\Theta(\sigma Z + \mu - \alpha)\Big] &= \Phi\big(\frac{\alpha-\mu}{\sigma}\big) & (5.165)\\ \mathbb{E}_Z\Big[Z\Theta(\sigma Z + \mu - \alpha)\Big] &= \frac{\mathrm{e}^{-\frac{1}{2}\left(\frac{\alpha-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}} & (5.166)\\ \mathbb{E}_Z\Big[Z^2\Theta(\sigma Z + \mu - \alpha)\Big] &= \Phi\big(\frac{\alpha-\mu}{\sigma}\big) + \frac{\alpha-\mu}{\sigma}\frac{\mathrm{e}^{-\frac{1}{2}\left(\frac{\alpha-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}} & (5.167) \end{aligned}$$

where

$$\Phi(x) := \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \mathrm{d}x \tag{5.168}$$

The identities above imply

$$\mathbb{E}_Z[Z\mathrm{relu}(\sigma Z + \mu - \alpha)] = \sigma\Phi\big(\frac{\alpha-\mu}{\sigma}\big) \tag{5.169}$$

and hence

$$\mathbb{E}_Z[Z\mathrm{st}(\sigma Z, \alpha)] = \sigma\Phi\big(\frac{\alpha-\mu}{\sigma}\big) + \sigma\Phi\big(\frac{\alpha+\mu}{\sigma}\big) \ . \tag{5.170}$$

This implies that

$$\begin{aligned} \mathbb{E}_{Z,\beta_0}\Big[\beta_0\varphi\Big] &= \hat{w}\mathbb{E}_Z\Big[\Phi\Big(\sqrt{\nu}\frac{\alpha\hat{\tau}-\hat{v}Z}{\hat{w}}\Big) + \Phi\Big(\sqrt{\nu}\frac{\alpha\hat{\tau}+\hat{v}Z}{\hat{w}}\Big)\Big]\\ &= 2\hat{w}\mathbb{E}_Z\Big[\Phi\Big(\sqrt{\nu}\frac{\alpha\hat{\tau}+\hat{v}Z}{\hat{w}}\Big)\Big] \ . & (5.171) \end{aligned}$$

and also

$$\begin{aligned} \mathbb{E}_{Z,\beta_0}\Big[Z\varphi\Big] &= \hat{v}\mathbb{E}_{\beta_0}\Big[\Phi\big(\frac{\alpha\hat{\tau}-\hat{w}\beta_0/\theta_0}{\hat{v}}\big) + \Phi\big(\frac{\alpha\hat{\tau}+\hat{w}\beta_0/\theta_0}{\hat{v}}\big)\Big]\\ &= 2\hat{v}\mathbb{E}_{\beta_0}\Big[\Phi\big(\frac{\alpha\hat{\tau}+\hat{w}\beta_0/\theta_0}{\hat{v}}\big)\Big] \ . & (5.172) \end{aligned}$$

Next we use that

$$\mathbb{E}_X\Big[\Phi(aX+b)\Big] = \Phi\Big(\frac{b}{\sqrt{1+a^2}}\Big) \tag{5.173}$$

and get that

$$\mathbb{E}_Z\Big[\Phi\Big(\sqrt{\nu}\frac{\alpha\hat{\tau}+\hat{v}Z}{\hat{w}}\Big)\Big] = \Phi\Big(\frac{\alpha\hat{\tau}}{\sqrt{\hat{w}^2/\nu+\hat{v}^2}}\Big) \tag{5.174}$$

and

$$\mathbb{E}_{\beta_0}\Big[\Phi\big(\frac{\alpha\hat{\tau}+\hat{w}\beta_0/\theta_0}{\hat{v}}\big)\Big] = \nu\Phi\Big(\frac{\alpha\hat{\tau}}{\sqrt{\hat{w}^2/\nu+\hat{v}^2}}\Big) + (1-\nu)\Phi\Big(\frac{\alpha\hat{\tau}}{\hat{v}}\Big)\ . \tag{5.175}$$

Finally we use that

$$\mathbb{E}_Z[\mathrm{relu}^2(\sigma Z+\mu-\alpha)] = \Phi\big(\frac{\alpha-\mu}{\sigma}\big)\Big(\sigma^2+(\mu-\alpha)^2\Big) - \sigma(\alpha-\mu)\frac{\mathrm{e}^{-\frac{1}{2}\left(\frac{\alpha-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}} \tag{5.176}$$

hence

$$\begin{aligned}\frac{1}{2}\mathbb{E}_{Z,\beta_0}\Big[\varphi^2\Big] &= \nu\Big\{\Phi\big(\frac{\alpha\hat{\tau}}{\sqrt{\hat{v}^2+\hat{w}^2/\nu}}\big)\Big(\hat{v}^2+\hat{w}^2/\nu+\alpha^2\hat{\tau}^2\Big)+\\ &- \sqrt{\hat{v}^2+\hat{w}^2/\nu}\,\alpha\hat{\tau}\frac{\mathrm{e}^{-\frac{1}{2}\left(\frac{\alpha\hat{\tau}}{\sqrt{\hat{v}^2+\hat{w}^2/\nu}}\right)^2}}{\sqrt{2\pi}}\Big\}+\\ &+ (1-\nu)\Big\{\Phi\big(\frac{\alpha\hat{\tau}}{\hat{v}}\big)\Big(\hat{v}^2+\alpha^2\hat{\tau}^2\Big) - \hat{v}\alpha\hat{\tau}\frac{\mathrm{e}^{-\frac{1}{2}\left(\frac{\alpha\hat{\tau}}{\hat{v}}\right)^2}}{\sqrt{2\pi}}\Big\}\ .\end{aligned} \tag{5.177}$$

If we introduce the short-hands

$$\chi_0 := \frac{\alpha\hat{\tau}}{\hat{v}},\quad \chi_1 := \frac{\alpha\hat{\tau}}{\sqrt{\hat{v}^2+\hat{w}^2/\nu}} \tag{5.178}$$

then the first three RS equations take the more compact form

$$w = 2\hat{w}\Phi(\chi_1) \tag{5.179}$$

$$\tau = 2\hat{\tau}\Big\{\nu\Phi(\chi_1)+(1-\nu)\Phi(\chi_0)\Big\} \tag{5.180}$$

$$\begin{aligned}\frac{1}{2}(v^2+w^2) &= \nu\{(1+1/\chi_1^2)\Phi(\chi_1)-G(\chi_1)\}+\\ &+ (1-\nu)\{(1+1/\chi_0^2)\Phi(\chi_0)-G(\chi_0)\}\end{aligned} \tag{5.181}$$

Explicit computation for the elastic net regularization

For the Elastic net regularization

$$\mathrm{prox}_{\mathrm{r}(.)}(x,\hat{\tau}) = \frac{1}{1+\eta\hat{\tau}}\mathrm{st}(x,\alpha\hat{\tau}) \tag{5.182}$$

hence the replica symmetric equations of the main section can be easily recovered.

5.10.4 Distributions

Here we show that the distributions

$$\mathcal{P}_\xi(\Delta,t,h) := \lim_{n\to\infty}\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^n \delta_{\Delta,\Delta_i}\delta(t-T_i)\delta(h-\mathbf{X}_i'\hat{\boldsymbol{\beta}})\Big]\ , \tag{5.183}$$

$$\mathcal{P}_\varphi(x) := \lim_{n\to\infty}\mathbb{E}_{\mathcal{D}}\Big[\frac{1}{p}\sum_{k=1}^p \delta(x-\mathbf{e}_k'\hat{\boldsymbol{\beta}})\Big] \tag{5.184}$$

indeed satisfy

$$\mathcal{P}_\xi(\Delta,t,h) = \mathbb{E}_{\Delta',T',Z_0',Q'}\Big[\delta(t-T')\delta_{\Delta,\Delta'}\delta\big(h-\xi_\star\big)\Big] \tag{5.185}$$

$$\mathcal{P}_\varphi(x) = \mathbb{E}_{\beta_0,Z}\Big[\delta\big(x-\varphi_\star\big)\Big], \tag{5.186}$$

as stated in the main text (5.37, 5.38). For the identity (5.185), observe that $\mathcal{P}_\xi(\Delta,t,h)$ is the functional order parameter of the replica theory (5.82), hence, at the replica symmetric saddle point we have (5.123) and after taking the limit $r\to 0$ and $\gamma\to\infty$ we obtain (5.140), which coincides with (5.185) as

$$\begin{aligned}\mathcal{P}_\xi(\Delta,t,h) &= \mathbb{E}_{Z_0,Z}\Big[f(\Delta,t|Z_0)\delta\Big(h-\xi_\star\Big)\Big] = \\ &= \mathbb{E}_{\Delta',T',Z_0',Z'}\Big[\delta_{\Delta,\Delta'}\delta(t-T')\delta\Big(h-\xi_\star\Big)\Big]\,. \end{aligned}\tag{5.187}$$

The identity (5.186) requires more care. Via Laplace integration we may write

$$\frac{1}{p}\sum_{k=1}^p \delta(x-\mathbf{e}_k'\hat{\boldsymbol{\beta}}) = \lim_{\gamma\to\infty}\frac{1}{p}\sum_{i=1}^p\sum_{k=1}^p \frac{\int \mathrm{e}^{-\gamma\mathcal{H}(\boldsymbol{\beta})}\;\delta(x-\mathbf{e}_\mu'\boldsymbol{\beta})\;\mathrm{d}\boldsymbol{\beta}}{\int \mathrm{e}^{-\gamma\mathcal{H}(\boldsymbol{\beta})}\;\mathrm{d}\boldsymbol{\beta}}\,. \tag{5.188}$$

To compute the expectation over the data-set we use the following alternative replica identity

$$\mathcal{P}_\varphi(x) = \lim_{r\to 0}\lim_{\gamma\to\infty}\lim_{p\to\infty}\frac{1}{r\;p}\sum_{k=1}^p \mathbb{E}_{\mathcal{D}}\Big[\int \mathrm{e}^{-\gamma\sum_{\alpha=1}^r\mathcal{H}(\boldsymbol{\beta}_\alpha|\mathcal{D})}\;\delta(x-\mathbf{e}_k'\boldsymbol{\beta}_1)\;\prod_{\alpha=1}^r \mathrm{d}\boldsymbol{\beta}_\alpha\Big]\,. \tag{5.189}$$

Inserting the functional delta measure as in (5.89) and taking the expectation over $\mathcal{D}$, i.e. the data-set, we obtain

$$\mathcal{P}_\varphi(x) = \int \mathrm{e}^{-n\psi\left[\{\mathcal{P}_\alpha,\hat{\mathcal{P}}_\alpha\}_{\alpha=1}^r,\hat{\mathbf{C}},\mathbf{C}\right]}\mathcal{J}(\gamma,\hat{\mathbf{C}})\prod_{\alpha=1}^r\;\mathcal{D}\hat{\mathcal{P}}_\alpha\;\mathcal{D}\mathcal{P}_\alpha\mathrm{d}\mathbf{C}\mathrm{d}\hat{\mathbf{C}} \tag{5.190}$$

where ψ is defined as in (5.104) and

$$\mathcal{J}(\gamma,\hat{\mathbf{C}}) := \frac{1}{p}\sum_{k=1}^p \frac{\int \mathrm{e}^{-\mathrm{i}p\mathrm{Tr}(\hat{\mathbf{C}}\mathbf{C}(\{\boldsymbol{\beta}_\alpha\}))-\gamma\mathrm{r}(\boldsymbol{\beta}_\alpha)}\delta(x-\mathbf{e}_k'\boldsymbol{\beta}_1)\prod_{\alpha=1}^r\mathrm{d}\boldsymbol{\beta}_\alpha}{\int \mathrm{e}^{-\mathrm{i}p\mathrm{Tr}(\hat{\mathbf{C}}\mathbf{C}(\{\boldsymbol{\beta}_\alpha\}))-\gamma\mathrm{r}(\boldsymbol{\beta}_\alpha)}\prod_{\alpha=1}^r\mathrm{d}\boldsymbol{\beta}_\alpha}\,. \tag{5.191}$$

Now taking $i\hat{\mathbf{C}} = \frac{1}{2}\mathbf{D}$, and replica symmetry, we notice that the expression above can be simplified, at the saddle point, because the regularizer is separable, i.e. $r(\mathbf{x}) = \sum_{l=1}^p r(x_l)$,

$$\begin{aligned}&\mathcal{J}_{RS}^{(r)}(\gamma,\hat{m},\hat{q},\hat{\rho}) = \\ &\frac{1}{p}\sum_{k=1}^p \frac{\int \mathrm{e}^{-\sum_{\rho=1}^r\left(\hat{m}\beta_{0,k}\beta_{\rho,k}+\frac{1}{2}(\hat{\rho}+\hat{q})\beta_{\rho,k}^2\right)+\frac{1}{2}\hat{q}\left(\sum_{\rho=1}^r\boldsymbol{\beta}_{\rho,k}\right)^2-\gamma\mathrm{r}(\beta_{\alpha,k})}\delta(x-\beta_{1,k})\prod_{\alpha=1}^r\mathrm{d}\beta_{\alpha,k}}{\int \mathrm{e}^{-\sum_{\rho=1}^r\left(\hat{m}\beta_{0,k}\beta_{\rho,k}+\frac{1}{2}(\hat{\rho}+\hat{q})\beta_{\rho,k}^2\right)+\frac{1}{2}\hat{q}\left(\sum_{\rho=1}^r\boldsymbol{\beta}_{\rho,k}\right)^2-\gamma\mathrm{r}(\beta_{\alpha,k})}\prod_{\alpha=1}^r\mathrm{d}\beta_{\alpha,k}}\,.\end{aligned}$$

At this point Gaussian linearization gives

$$\begin{aligned}\mathcal{J}_{RS}^{(r)}(\gamma,\hat{m},\hat{q},\hat{\rho}) = \mathbb{E}_{\beta_0}\Bigg[\mathbb{E}_Z\Bigg[\int \mathrm{e}^{-\left(\{\hat{m}\beta_0+\sqrt{\hat{q}}Z\}\beta+\frac{1}{2}(\hat{\rho}+\hat{q})\beta^2\right)-\gamma\mathrm{r}(\beta)}\delta(x-\beta)\mathrm{d}\beta\;\times \\ \times\left(\int \mathrm{e}^{-\left(\{\hat{m}\beta_{0,k}+\sqrt{\hat{q}}Z\}\beta+\frac{1}{2}(\hat{\rho}+\hat{q})\beta^2\right)-\gamma\mathrm{r}(\beta)}\mathrm{d}\beta\right)^{r-1}\Bigg]\times \\ \times\mathbb{E}_Z\Bigg[\left(\int \mathrm{e}^{-\left(\{\hat{m}\beta_0+\sqrt{\hat{q}}Z\}\beta+\frac{1}{2}(\hat{\rho}+\hat{q})\beta^2\right)-\gamma\mathrm{r}(\beta)}\mathrm{d}\beta\right)^r\Bigg]^{-1}\Bigg]\,.\end{aligned}\tag{5.192}$$

Hence, after taking the limit $r \to 0$, we get

$$\mathcal{J}_{RS}(\gamma, \hat{m}, \hat{q}, \hat{\rho}) := \mathbb{E}_{Z,\beta_0}\left[\frac{\int \mathrm{e}^{-\left(\{\hat{m}\beta_0+\sqrt{\hat{q}}Z\}\beta+\frac{1}{2}(\hat{\rho}+\hat{q})\beta^2\right)-\gamma \mathrm{r}(\beta)}\delta(x-\beta)\mathrm{d}\beta}{\int \mathrm{e}^{-\left(\{\hat{m}\beta_0,+\sqrt{\hat{q}}Z\}\beta+\frac{1}{2}(\hat{\rho}+\hat{q})\beta^2\right)-\gamma \mathrm{r}(\beta)}\mathrm{d}\beta}\right] .$$

After taking the rescaling

$$1/\hat{\tau} = (\hat{\rho}+\hat{q})/\gamma, \quad \hat{m} = \gamma\hat{m}, \quad \hat{q} = \gamma^2\hat{q} \tag{5.193}$$

and the limit $\gamma \to \infty$, we get

$$\mathcal{P}_{\varphi}(x) = \mathbb{E}Z, \beta_0\Big[\delta\Big(x - \mathrm{prox}_{\mathrm{r}(.)}\big(\hat{m}_\star\beta_0 + \sqrt{\hat{q}_\star}Z, \hat{\tau}_\star\big)\Big)\Big] . \tag{5.194}$$

The expression above reduces to (5.186) after the change of variables

$$\hat{w} = -\hat{\tau}\hat{m}\theta_0, \quad \hat{v} = \hat{\tau}\sqrt{\hat{q}}, \tag{5.195}$$

and using the definition (5.34).

5.10.5 Coordinate Wise Minimization algorithm for pathwise solution

Consider the function

$$g(\boldsymbol{\beta}, \lambda) := \frac{1}{n}\sum_{i=1}^{n}\Big\{\Lambda(T_i)\mathrm{e}^{\mathbf{X}_i'\boldsymbol{\beta}} - \Delta_i\mathbf{X}_i'\boldsymbol{\beta} - \Delta_i \log\lambda(T_i)\Big\} \tag{5.196}$$

Suppose we want to minimize the objective function

$$f(\boldsymbol{\beta}, \lambda) := g(\boldsymbol{\beta}, \lambda) + \mathrm{r}(\boldsymbol{\beta}) \tag{5.197}$$

with r a separable convex regularization function. In our case we will be interested in

$$\mathrm{r}(\boldsymbol{\beta}) := \alpha|\boldsymbol{\beta}| + \frac{1}{2}\eta\|\boldsymbol{\beta}\|^2 . \tag{5.198}$$

This can be done via coordinate descent. First we minimize over λ, obtaining the Nelson-Aalen estimator

$$\hat{\Lambda}_n(t) = \mathrm{NA}(\{T_j, \mathbf{X}_j'\boldsymbol{\beta}\}) := \sum_{i=1}^{n}\frac{\Delta_i\Theta(t-T_i)}{\sum_{j=1}^{n}\Theta(T_j-T_i)\mathrm{e}^{\mathbf{X}_j'\boldsymbol{\beta}}} . \tag{5.199}$$

Then we minimize over $\boldsymbol{\beta}$ and so on an so forth, until a fixed point, i.e. the updated values for $\boldsymbol{\beta}^{t+1}$ and $\Lambda^{t+1}(T_1), \ldots, \Lambda^{t+1}(T_n)$ are within a user defined tolerance from their previous value. To compute the minimizer in $\boldsymbol{\beta}$ at fixed Λ, we use the algorithm proposed in [25]. The idea is to reduce the problem to an iterative regularized least squared regression. Let us define

$$\ell(\boldsymbol{\beta}, \Lambda) := \frac{1}{n}\sum_{i=1}^{n}\Big\{\Lambda(T_i)\mathrm{e}^{\mathbf{X}_i'\boldsymbol{\beta}} - \Delta_i\mathbf{X}_i'\boldsymbol{\beta}\Big\} . \tag{5.200}$$

A second order expansion in $\boldsymbol{\xi}$ around $\boldsymbol{\phi}$ gives the following approximation for ℓ

$$\tilde{\ell}(\boldsymbol{\beta},\Lambda,\boldsymbol{\phi}) = \ell(\boldsymbol{\phi},\Lambda) + \mathbf{s}(\boldsymbol{\phi},\Lambda)'(\boldsymbol{\beta}-\boldsymbol{\phi}) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\phi})'\boldsymbol{H}(\boldsymbol{\phi},\Lambda)(\boldsymbol{\beta}-\boldsymbol{\phi}) \tag{5.201}$$

where

$$\mathbf{s}(\boldsymbol{\phi},\Lambda) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\Lambda(T_i)\mathrm{e}^{\mathbf{X}_i'\boldsymbol{\phi}} - \Delta_i\Big\}\mathbf{X}_i = \frac{1}{n}(\mathbf{W} - \mathrm{diag}(\boldsymbol{\Delta})\mathbf{X} \tag{5.202}$$

$$\mathbf{M}(\boldsymbol{\phi},\Lambda) = \frac{1}{n}\sum_{i=1}^{n}\Lambda(T_i)\mathrm{e}^{\mathbf{X}_i'\boldsymbol{\phi}}\mathbf{X}_i\mathbf{X}_i' = \frac{1}{n}\mathbf{X}'\mathbf{W}\mathbf{X} \tag{5.203}$$

with $\mathbf{W} := \mathrm{diag}\big(\Lambda(T_i)\exp\{\mathbf{X}_i'\boldsymbol{\phi}\}\big)$. At this points we use a coordinate descent strategy to solve the regularized least square problem

$$\boldsymbol{\beta}^{t+1} = \arg\min_{\boldsymbol{\varphi}}\Big\{f(\boldsymbol{\varphi})\Big\}, \quad f(\boldsymbol{\varphi}) = \tilde{\ell}(\boldsymbol{\varphi},\Lambda^t,\boldsymbol{\beta}^t) + r(\boldsymbol{\varphi}), \tag{5.204}$$

i.e. we minimize along each component keeping the remaining components fixed

$$\begin{aligned}&\frac{\partial}{\partial\varphi_k}f(\boldsymbol{\varphi}) = \\ &s_k(\boldsymbol{\beta}^t,\Lambda^t) + \varphi_k\ \mathbf{e}_k'\mathbf{M}(\boldsymbol{\beta}^t,\Lambda^t)\mathbf{e}_k + \mathbf{e}_k'\mathbf{M}(\boldsymbol{\beta}^t,\Lambda^t)\big\{(\boldsymbol{I}-\mathbf{e}_k\mathbf{e}_k')\boldsymbol{\varphi}-\boldsymbol{\beta}^t\big\} + r'(\varphi_k) = 0\end{aligned} \tag{5.205}$$

which is solved by

$$\begin{aligned}&\varphi_k^{t+1} = \\ &\mathrm{prox}_{r(.)}\Big(\frac{\mathbf{e}_k'\mathbf{M}(\boldsymbol{\beta}^t,\Lambda^t)\big\{\boldsymbol{\beta}^t-(\boldsymbol{I}-\mathbf{e}_k\mathbf{e}_k')\boldsymbol{\varphi}^t\big\} - s_k(\boldsymbol{\beta}^t,\Lambda^t)}{\mathbf{e}_k'\mathbf{M}(\boldsymbol{\beta}^t,\Lambda^t)\mathbf{e}_k}, \frac{1}{\mathbf{e}_k'\mathbf{M}(\boldsymbol{\beta}^t,\Lambda^t)\mathbf{e}_k}\Big)\ .\end{aligned} \tag{5.206}$$

Notice that in the case of the elastic net penalization we get

$$\varphi_k^{t+1} = \frac{1}{1+\eta\tau_k}\mathrm{st}\big(\psi_k,\alpha\tau_k\big) \tag{5.207}$$

where

$$\begin{aligned}&\psi_k := \big\{\mathbf{e}_k'\mathbf{M}(\boldsymbol{\beta}^t,\Lambda^t)\big\{\boldsymbol{\beta}^t-(\boldsymbol{I}-\mathbf{e}_k\mathbf{e}_k')\boldsymbol{\varphi}^t\big\} - s_k(\boldsymbol{\beta}^t,\Lambda^t)\big\}/\mathbf{e}_k'\mathbf{M}(\boldsymbol{\beta}^t,\Lambda^t)\mathbf{e}_k \\ &1/\tau_k := \mathbf{e}_k'\mathbf{M}(\boldsymbol{\beta}^t,\Lambda^t)\mathbf{e}_k\end{aligned}$$

and st is the soft thresholding operator.

5.10.6 Derivation of GAMP via Belief Propagation

Consider the following optimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\Big\{\sum_{i=1}^{n} g\big(\mathbf{X}_i'\boldsymbol{\beta},Y_i\big) + \sum_{\mu=1}^{p} r(\beta_\mu)\Big\} \tag{5.208}$$

where $\mathbf{X}_i \sim \mathcal{N}\big(0,\frac{1}{p}\mathbf{1}\big)$ and $Y_i|\mathbf{X}_i \sim f(.|\mathbf{X}_i'\boldsymbol{\beta}_0)$. Define the Gibbs measure

$$f_\gamma(\boldsymbol{\beta}|\{Y_i,\mathbf{X}_i\}) = \frac{\mathrm{e}^{-\gamma\left\{\sum_{i=1}^{n} g(\mathbf{X}_i'\boldsymbol{\beta},Y_i)+\lambda\sum_{\mu=1}^{p} r(\boldsymbol{\beta}_\mu)\right\}}}{Z_\gamma(\{Y_i,\mathbf{X}_i\})}\ . \tag{5.209}$$

Such a distribution corresponds to a fully connected bypartite factor graph, with p variable nodes and n factor nodes. The (loopy) belief propagation equations read

$$m_{\mu\to i}(\boldsymbol{\beta}_\mu) \propto \mathrm{e}^{-\gamma r(\boldsymbol{\beta}_\mu)} \prod_{j\neq i} m_{j\to\mu}(\beta_\mu) \tag{5.210}$$

$$m_{i\to\mu}(\boldsymbol{\beta}_\mu) \propto \int \mathrm{e}^{-\gamma g(\mathbf{X}_i'\boldsymbol{\beta}, Y_i)} \prod_{\alpha\neq\mu} m_{\alpha\to i}(\beta_\alpha)\mathrm{d}\beta_\alpha \ . \tag{5.211}$$

The tempered Moreau envelope and proximal operator

Let us introduce the tempered Moreau envelope as

$$\mathcal{M}_{f(.)}(x,\alpha,\gamma) := -\frac{1}{\gamma}\log \int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\frac{(z-x)^2}{\alpha}+f(z)\right\}}\mathrm{d}z \ . \tag{5.212}$$

Notice that

$$\dot{\mathcal{M}}_{f(.)}(x,\alpha,\gamma) = \frac{1}{\alpha}\big(x - \mathbb{E}_\gamma[Z]\big) \tag{5.213}$$

$$\ddot{\mathcal{M}}_{f(.)}(x,\alpha,\gamma) = \frac{1}{\alpha} - \frac{\gamma}{\alpha^2}\mathbb{V}_\gamma[Z] \ . \tag{5.214}$$

In analogy with the proximal operator, we define

$$\mathrm{prox}_{f(.)}(x,\alpha,\gamma) := \mathbb{E}_\gamma[Z] = \frac{\int z\mathrm{e}^{-\gamma\left\{\frac{1}{2}\frac{(z-x)^2}{\alpha}+f(z)\right\}}\mathrm{d}z}{\int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\frac{(z-x)^2}{\alpha}+f(z)\right\}}\mathrm{d}z} = x - \alpha\dot{\mathcal{M}}_{f(.)}(x,\alpha,\gamma) \ . \tag{5.215}$$

Based on the previously derived relationships we notice that

$$\mathrm{prox}'_{f(.)}(x,\alpha,\gamma) = \mathbb{V}_\gamma[Z]\gamma/\alpha \ . \tag{5.216}$$

In the limit $\gamma\to\infty$ we have

$$\lim_{\gamma\to\infty} \mathcal{M}_{f(.)}(x,\alpha,\gamma) := \mathcal{M}_{f(.)}(x,\alpha) \tag{5.217}$$

furthermore

$$\lim_{\gamma\to\infty} \dot{\mathcal{M}}_{f(.)}(x,\alpha,\gamma) = \frac{1}{\alpha}\big(x - \mathrm{prox}_{f(.)}(x,\alpha)\big) = \dot{f}\big(\mathrm{prox}_{f(.)}(x,\alpha)\big) \tag{5.218}$$

$$\lim_{\gamma\to\infty} \ddot{\mathcal{M}}_{f(.)}(x,\alpha,\gamma) = \frac{\ddot{f}\big(\mathrm{prox}_{f(.)}(x,\alpha)\big)}{1+\alpha\ddot{f}\big(\mathrm{prox}_{f(.)}(x,\alpha)\big)} \ . \tag{5.219}$$

The last identity is obtained by differentiating $\dot{f}\big(\mathrm{prox}_{f(.)}(x,\alpha)\big)$ with respect to x. Notice that this implies

$$\lim_{\gamma\to\infty} \gamma\mathbb{V}_\gamma[Z] = \alpha\frac{1}{1+\alpha\ddot{f}\big(\mathrm{prox}_{f(.)}(x,\alpha)\big)} \ . \tag{5.220}$$

Approximation of the messages

We start by re-writing the self consistent condition of the messages as

$$\begin{aligned} m_{i\to\mu}(\boldsymbol{\beta}_\mu) &\propto \int \mathrm{e}^{-\gamma g\left(X_{i,\mu}\beta_\mu+\sum_{\alpha\neq\mu}X_{i,\alpha}\beta_\alpha,Y_i\right)}\prod_{\alpha\neq\mu}\mathrm{d}m_{\alpha\to i}(\beta_\alpha) \\ &\propto \int \mathrm{e}^{i\hat{\xi}\left(\xi-X_{i,\mu}\beta_\mu\right)-\sum_{\alpha\neq\mu}\psi_{\beta_\alpha}(i\hat{\xi}X_{i,\alpha})-\gamma g(\xi,Y_i)}\,\mathrm{d}\xi\mathrm{d}\hat{\xi} \end{aligned} \tag{5.221}$$

where

$$\psi_X(q) := \log \mathbb{E}_X\left[\mathrm{e}^{qX}\right] \tag{5.222}$$

is the cumulant generating function. Taking a quadratic approximation of ψ_{β_α} gives

$$\begin{aligned} &m_{i\to\mu}(\boldsymbol{\beta}_\mu) \approx \int \mathrm{e}^{i\hat{\xi}\left(\xi-X_{i,\mu}\beta_\mu-\sum_{\alpha\neq\mu}X_{i,\alpha}\hat{\beta}_{\alpha\to i}\right)-\gamma g(\xi,Y_i)-\frac{1}{2}\frac{1}{\gamma}\left(\sum_{\alpha\neq\mu}\nu_{\alpha\to i}X_{i,\alpha}^2\right)\hat{\xi}^2}\,\mathrm{d}\xi\mathrm{d}\hat{\xi} = \\ &\propto \int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\frac{\left(\xi-X_{i,\mu}\beta_\mu-\sum_{\alpha\neq\mu}X_{i,\alpha}\hat{\beta}_{\alpha\to i}\right)^2}{\sum_{\alpha\neq\mu}\nu_{\alpha\to i}X_{i,\alpha}^2}+g(\xi,Y_i)\right\}}\,\mathrm{d}\xi\ , \end{aligned} \tag{5.223}$$

where $\hat{\beta}_{\alpha\to i},\tau_{\alpha\to i}$ are respectively the mean and the variance (rescaled by the inverse temperature γ) of the belief $m_{\alpha\to i}$. We notice that

$$m_{i\to\mu}(\boldsymbol{\beta}_\mu) \approx \mathrm{e}^{-\gamma\dot{\mathcal{M}}_{f(.)}(X_{i,\mu}(\beta_\mu-\hat{\beta}_{\mu\to i})+\hat{\xi}_i,\tau_i-\nu_{\mu\to i}X_{i,\mu}^2,\gamma)} \tag{5.224}$$

where

$$\xi_i = \sum_\alpha X_{i,\alpha}\hat{\beta}_{\alpha\to i} \tag{5.225}$$

$$\tau_i = \sum_\alpha \nu_{\alpha\to i}X_{i,\alpha}^2\ . \tag{5.226}$$

Assuming $X_{i,\mu}(\beta_\mu-\hat{\beta}_{\mu\to i})$ to be small, we take a quadratic approximation of $\mathcal{M}_{g(.,T_i)}(.,\tau_i-\nu_{\mu\to i}X_{i,\mu}^2,\gamma)$ around ξ_i obtaining

$$\begin{aligned} &\mathcal{M}_{g(.,T_i)}(X_{i,\mu}(\beta_\mu-\hat{\beta}_{\mu\to i})+\xi_i,\tau_i-\nu_{\mu\to i}X_{i,\mu}^2,\gamma)\approx \\ &\quad\mathcal{M}_{g(.,T_i)}(\xi,\tau_i-\nu_{\mu\to i}X_{i,\mu}^2,\gamma)+ \\ &\quad\dot{\phi}_{i\to\mu}X_{i,\mu}(\beta_\mu-\hat{\beta}_{\mu\to i})+\frac{1}{2}\ddot{\phi}_{i\to\mu}X_{i,\mu}^2(\beta_\mu-\hat{\beta}_{\mu\to i})^2 \end{aligned} \tag{5.227}$$

where

$$\dot{\phi}_{i\to\mu} = \dot{\mathcal{M}}_{g(.,T_i)}(\xi_i,\tau_i-\nu_{\mu\to i}X_{i,\mu}^2,\gamma) \tag{5.228}$$

$$\ddot{\phi}_{i\to\mu} = \ddot{\mathcal{M}}_{g(.,T_i)}(\xi_i,\tau_i-\nu_{\mu\to i}X_{i,\mu}^2,\gamma)\ . \tag{5.229}$$

Then

$$m_{i\to\mu}(\boldsymbol{\beta}_\mu) \approx \exp\left\{-\gamma\left(\frac{1}{2}\hat{\tau}_{i\to\mu}\beta_\mu^2-s_{i\to\mu}\beta_\mu\right)\right\}\ . \tag{5.230}$$

where

$$s_{i\to\mu} = -(\dot{\phi}_{i\to\mu}X_{i,\mu}-\ddot{\phi}_{i\to\mu}X_{i,\mu}^2\hat{\beta}_{\mu\to i}) \tag{5.231}$$

$$1/\hat{\tau}_{i\to\mu} = \ddot{\phi}_{i\to\mu}X_{i,\mu}^2\ . \tag{5.232}$$

As a consequence

$$m_{\mu\to i}(\beta_\mu) \approx \exp\Big\{ -\gamma\Big(r(\beta_\mu) + \frac{1}{2}\beta_\mu^2/\hat{\tau}_{\mu\to i} - s_{\mu\to i}\beta_\mu\Big)\Big\} \ . \tag{5.233}$$

where

$$\begin{aligned} s_{\mu\to i} &= \sum_{j\neq i} s_{j\to\mu} & (5.234)\\ 1/\hat{\tau}_{\mu\to i} &= \sum_{j\neq i} \hat{\tau}^{-1}_{j\to\mu} \ . & (5.235) \end{aligned}$$

Finally

$$\begin{aligned} \hat{\beta}_{\mu\to i} &= \hat{\tau}_{\mu\to i} s_{\mu\to i} - \hat{\tau}_{\mu\to i}\dot{\mathcal{M}}_{r(.)}\big(s_{\mu\to i}\hat{\tau}_{\mu\to i}, \hat{\tau}_{\mu\to i}, \gamma\big) & (5.236)\\ \nu_{\mu\to i} &= \hat{\tau}_{\mu\to i}\Big\{1 - \hat{\tau}_{\mu\to i}\ddot{\mathcal{M}}_{r(.)}\big(s_{\mu\to i}\hat{\tau}_{\mu\to i}, \hat{\tau}_{\mu\to i}, \gamma\big)\Big\} \ , & (5.237) \end{aligned}$$

and as a consequence

$$\begin{aligned} \hat{\beta}_{\mu\to i} &= \mathrm{prox}_{r(.)}\big(s_{\mu\to i}\hat{\tau}_{\mu\to i}, \hat{\tau}_{\mu\to i}, \gamma\big) = \mathrm{prox}_{r(.)}\big(\psi_{\mu\to i}, \hat{\tau}_{\mu\to i}, \gamma\big) & (5.238)\\ \nu_{\mu\to i} &= \hat{\tau}_{\mu\to i}\mathrm{prox}'_{r(.)}\big(s_{\mu\to i}\hat{\tau}_{\mu\to i}, \hat{\tau}_{\mu\to i}, \gamma\big) = \hat{\tau}_{\mu\to i}\mathrm{prox}'_{r(.)}\big(\psi_{\mu\to i}, \hat{\tau}_{\mu\to i}, \gamma\big) \ , & (5.239) \end{aligned}$$

where

$$\psi_{\mu\to i} := s_{\mu\to i}\hat{\tau}_{\mu\to i} = \hat{\beta}_{\mu\to i} - \hat{\tau}_{\mu\to i}X_{i,\mu}\dot{\phi}_{i\to\mu} \ . \tag{5.240}$$

TAP approximation

First of all we neglect all the corrections $\nu_{\mu\to i}X_{i,\mu}^2$ so that

$$\begin{aligned} \dot{\phi}_{i\to\mu} &\approx \dot{\phi}_i = \dot{\mathcal{M}}_{g(.,T_i)}(\xi_i, \tau_i, \gamma) & (5.241)\\ \ddot{\phi}_{i\to\mu} &\approx \ddot{\phi}_i = \ddot{\mathcal{M}}_{g(.,T_i)}(\xi_i, \tau_i, \gamma) \ . & (5.242) \end{aligned}$$

Then

$$\begin{aligned} s_{i\to\mu} &\approx -(\dot{\phi}_i X_{i,\mu} - \ddot{\phi}_i X_{i,\mu}^2\hat{\beta}_{\mu\to i}) & (5.243)\\ \hat{\tau}_{i\to\mu} &\approx \ddot{\phi}_i X_{i,\mu}^2 \ . & (5.244) \end{aligned}$$

Notice that

$$\begin{aligned} s_{\mu\to i} &= \sum_j s_{j\to\mu} - s_{i\to\mu} = s_\mu - s_{i\to\mu} & (5.245)\\ 1/\hat{\tau}_{\mu\to i} &= \sum_j 1/(\hat{\tau}_{j\to\mu} - \hat{\tau}_{i\to\mu}) \approx 1/\hat{\tau}_\mu \ , & (5.246) \end{aligned}$$

hence

$$\begin{aligned} \hat{\beta}_{\mu\to i} &\approx \mathrm{prox}_{\mathrm{r}(.)}\big(\psi_\mu, \hat{\tau}_\mu, \gamma\big) - \mathrm{prox}'_{\mathrm{r}(.)}\big(\psi_\mu, \hat{\tau}_\mu, \gamma\big)s_{i\to\mu}\hat{\tau}_\mu = \hat{\beta}_\mu - \nu_\mu s_{i\to\mu} & (5.247)\\ \nu_{\mu\to i} &\approx \nu_\mu = \mathrm{prox}'_{\mathrm{r}(.)}\big(\psi_\mu, \hat{\tau}_\mu, \gamma\big)\hat{\tau}_\mu \ , & (5.248) \end{aligned}$$

where

$$\psi_\mu := s_\mu\hat{\tau}_\mu = \hat{\beta}_\mu - \hat{\tau}_\mu X_{i,\mu}\dot{\phi}_i \tag{5.249}$$

and

$$\hat{\beta}_\mu = \mathrm{prox}_{\mathrm{r}(.)}\big(\psi_\mu, \hat{\tau}_\mu, \gamma\big) \ . \tag{5.250}$$

Finally

$$\begin{aligned} \tau_i &\approx \sum_\alpha \nu_\alpha X_{i,\alpha}^2 \\ \xi_i &\approx \sum_\alpha X_{i,\alpha} \hat{\beta}_\alpha + \sum_\alpha X_{i,\alpha}^2 \nu_\alpha \dot{\phi}_i \end{aligned} \tag{5.251}$$

TAP algorithm for GLMs

The TAP algorithm can be schematized as follows

Algorithm 3 TAP algorithm for Bayes Estimator in GLMs

Require: $\hat{\boldsymbol{\beta}}^0 \leftarrow \mathbf{0}_p$, tol $\leftarrow 1.0e-8$, max epochs $\leftarrow 300$
  $\boldsymbol{\tau}^0 = \mathbf{0}_n,\ \hat{\boldsymbol{\tau}}^0 = \mathbf{0}_p, \boldsymbol{\xi}^0 = \mathbf{X}\hat{\boldsymbol{\beta}}^0$
  flag = True, err $= \infty$, t $= 0$
  while err $\geq$ tol and flag **do**
    $t \leftarrow t+1$
    $\xi_i^t \leftarrow \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{t-1} + \tau_i \dot{\mathcal{M}}_{g(.,T)}(\xi^{t-1}, \tau_i^{t-1}, \gamma),\ \forall i$
    $1/\hat{\tau}_\mu^t \leftarrow \sum_{i=1}^n X_{i,\mu}^2 \ddot{\mathcal{M}}_{g(.,T)}(\xi_i^t, \tau_i^{t-1}, \gamma),\ \forall \mu$
    $\psi_\mu^t \leftarrow \hat{\beta}_\mu^{t-1} - \hat{\tau}_\mu^t \sum_{i=1}^n \dot{\mathcal{M}}_{g(.,T)}(\xi_i^t, \tau_i^{t-1}, \gamma) X_{i,\mu},\ \forall \mu$
    $\hat{\beta}_\mu^t \leftarrow \text{prox}_{r(.)}(\psi_\mu^t, \hat{\tau}_\mu^t, \gamma),\ \forall \mu$
    $\tau_i^t \leftarrow \sum_{\alpha=1}^p \hat{\tau}_\alpha^t\ \text{prox}'_{r(.)}(\psi_\alpha^t, \hat{\tau}_\alpha^t) X_{i,\alpha}^2,\ \forall i$
    err $\leftarrow \sqrt{\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}^{t-1}\|_2^2 + \|\boldsymbol{\tau}^t - \boldsymbol{\tau}^{t-1}\|_2^2 + \|\boldsymbol{\xi}^t - \boldsymbol{\xi}^{t-1}\|_2^2 + \|\hat{\boldsymbol{\tau}}^t - \hat{\boldsymbol{\tau}}^{t-1}\|_2^2}$
    if t $\geq$ max epochs **then**
      flag = False
    end if
  end while

The algorithm above is, in general, difficult to implement, since it requires to compute $\ddot{\mathcal{M}}_{g(.,T)}(.,a,\gamma)$, $\dot{\mathcal{M}}_{g(.,T)}(.,a,\gamma), \text{prox}_{r(.)}(.,a,\gamma)$, $\text{prox}'_{r(.)}(.,a,\gamma)$ at finite $\gamma < \infty$. These quantities are not directly available (an exception is the linear model with ridge regularization) and must be approximated via a Montecarlo scheme, or by numerical integration. In the zero temperature limit, however, the algorithm simplify to

Algorithm 4 TAP algorithm for MAP in GLMs

Require: $\hat{\boldsymbol{\beta}}^0 \leftarrow \mathbf{0}_p$, tol $\leftarrow 1.0e-8$, max epochs $\leftarrow 300$
  $\boldsymbol{\tau}^0 = \mathbf{0}_n,\ \hat{\boldsymbol{\tau}}^0 = \mathbf{0}_p, \boldsymbol{\xi}^0 = \mathbf{X}\hat{\boldsymbol{\beta}}^0$
  flag = True, err $= \infty$, t $= 0$
  while err $\geq$ tol and flag **do**
    $t \leftarrow t+1$
    $\xi_i^t \leftarrow \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{t-1} + \tau_i \dot{\mathcal{M}}_{g(.,T)}(\xi^{t-1}, \tau_i^{t-1}),\ \forall i$
    $1/\hat{\tau}_\mu^t \leftarrow \sum_{i=1}^n X_{i,\mu}^2 \ddot{\mathcal{M}}_{g(.,T)}(\xi_i^t, \tau_i^{t-1}),\ \forall \mu$
    $\psi_\mu^t \leftarrow \hat{\beta}_\mu^{t-1} - \hat{\tau}_\mu^t \sum_{i=1}^n \dot{\mathcal{M}}_{g(.,T)}(\xi_i^t, \tau_i^{t-1}) X_{i,\mu},\ \forall \mu$
    $\hat{\beta}_\mu^t \leftarrow \text{prox}_{r(.)}(\psi_\mu^t, \hat{\tau}_\mu^t),\ \forall \mu$
    $\tau_i^t \leftarrow \sum_{\alpha=1}^p \hat{\tau}_\alpha^t\ \text{prox}'_{r(.)}(\psi_\alpha^t, \hat{\tau}_\alpha^t) X_{i,\alpha}^2,\ \forall i$
    err $\leftarrow \sqrt{\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}^{t-1}\|_2^2 + \|\boldsymbol{\tau}^t - \boldsymbol{\tau}^{t-1}\|_2^2 + \|\boldsymbol{\xi}^t - \boldsymbol{\xi}^{t-1}\|_2^2 + \|\hat{\boldsymbol{\tau}}^t - \hat{\boldsymbol{\tau}}^{t-1}\|_2^2}$
    if t $\geq$ max epochs **then**
      flag = False
    end if
  end while

This can be efficiently implemented. It is advised to damp the AMP iteration to improve numerical stability and we notice empirically that this is the case also here.

Approximate Message Passing for i.i.d. Gaussian covariates

When

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p/p) \ , \tag{5.252}$$

one can further simplify the algorithm and obtain the Approximate Message Passing algorithm originally proposed to solve the Lasso optimization problem [16] and then extended to generalized linear models in [20]. The algorithm can be stated elegantly in vector notation, which provides also a more compact formulation. To this end it is useful to introduce the average operator$\langle\rangle$, which for a vector $\mathbf{x} \in \mathbb{R}^l$ reads

$$\langle \mathbf{x} \rangle = \frac{1}{l} \sum_{k=1}^{l} x_k \ . \tag{5.253}$$

Algorithm 5 AMP algorithm for MAP in GLMs (vector notation)

Require: $\hat{\boldsymbol{\beta}}^0 \leftarrow \mathbf{0}_p$, tol $\leftarrow 1.0e-8$, max epochs $\leftarrow 300$
  $\boldsymbol{\tau}^0 = 0, \ \hat{\boldsymbol{\tau}}^0 = 0, \boldsymbol{\xi}^0 = \mathbf{X}\hat{\boldsymbol{\beta}}^0$
  flag = True, err $= \infty$, t $= 0$
  while err $\geq$ tol and flag **do**
    $t \leftarrow t + 1$
    $\boldsymbol{\xi}^t \leftarrow \mathbf{X}\hat{\boldsymbol{\beta}}^{t-1} + \tau \dot{\mathcal{M}}_{g(.,T)}(\boldsymbol{\xi}^{t-1}, \tau^{t-1})$
    $\hat{\tau}^t \leftarrow \zeta / \left\langle \ddot{\mathcal{M}}_{g(.,T)}(\boldsymbol{\xi}^t, \tau^{t-1}) \right\rangle$
    $\boldsymbol{\psi}^t \leftarrow \hat{\boldsymbol{\beta}}^{t-1} - \hat{\tau}^t \mathbf{X}' \dot{\mathcal{M}}_{g(.,T)}(\boldsymbol{\xi}^t, \tau^{t-1})$
    $\hat{\boldsymbol{\beta}}^t \leftarrow \text{prox}_{r(.)}(\boldsymbol{\psi}^t, \hat{\tau}^t)$
    $\tau^t \leftarrow \hat{\tau}^t \left\langle \text{prox}'_{r(.)}(\boldsymbol{\psi}^t, \hat{\tau}^t) \right\rangle$
    err $\leftarrow \sqrt{\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}^{t-1}\|_2^2 + (\tau^t - \tau^{t-1})^2 + \|\boldsymbol{\xi}^t - \boldsymbol{\xi}^{t-1}\|_2^2 + (\hat{\tau}^t - \hat{\tau}^{t-1})^2}$
    if t $\geq$ max epochs **then**
      flag = False
    end if
  end while

5.10.7 COX-AMP : an AMP algorithm for the Cox model

The optimization of the Cox Partial Likelihood can be regarded as an alternating minimization, as we have already argued in the derivation of the Coordinate-wise Descent algorithm, see 5.10.5. At the fixed point we must have that

$$\dot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n, \hat{\Lambda}_n(\mathbf{T}), \boldsymbol{\Delta}) = \mathbf{0} \tag{5.254}$$

$$\hat{\Lambda}_n(T_i) = \mathrm{NA}(T_i, \mathbf{X}\hat{\boldsymbol{\beta}}_n) \ . \tag{5.255}$$

The idea now is to do a step with Generalized - AMP in $\boldsymbol{\beta}$ at fixed Λ and then update Λ with the Nelson Aalen estimator (5.199) as

$$\hat{\Lambda}_n^t(\mathbf{T}) \leftarrow \mathrm{NA}\big(\mathbf{T}, \mathrm{prox}_{g(.\hat{\Lambda}_n^{t-1}(\mathbf{T}),\boldsymbol{\Delta})}(\boldsymbol{\xi}^t, \tau^{t-1})\big) \tag{5.256}$$

since at the fixed point of the Generalized -AMP algorithm, we must have that

$$\mathrm{prox}_{g(.\hat{\Lambda}_n^t(\mathbf{T}),\boldsymbol{\Delta})}(\boldsymbol{\xi}, \tau) = \mathbf{X}\hat{\boldsymbol{\beta}}_n \ . \tag{5.257}$$

The algorithm is schematized in the main text, see Algorithm 2.

Chapter 6

Practical debiasing with the Covariant Prior in the proportional regime when $p < n$

This chapter is submitted to Journal of Physics : Communications.

6.1 Abstract

We show that the Covariant Prior can be used to effectively de-bias the resulting MAP estimator in the proportional regime, where the number of covariates p grows proportionally to n, the number of samples, with $p < n$. The asymptotics of the resulting MAP estimator can be studied via the statistical physics approach. It is known that the Replica method leads to three equations for three scalar order parameters. Knowledge of the order parameters and of the true signal strength, allows the computation of the asymptotic bias. Here we show that all these quantities might be estimated from the data alone, without actually solving the equations of the theory (which require the knowledge of the data generating process). Once this is done, a de-biased estimator can be computed easily allowing for eventual testing. We emphasize that the proposed methodology does not require to estimate the true covariance matrix, and can be (easily) applied whenever the latter is positive definite.

6.2 Introduction and motivation

When given a data - set of n observations comprising responses $T_1, \dots, T_n$, $T_i \in \mathcal{T} \subset \mathbb{R}$ and covariates $\mathbf{X}_1, \dots, \mathbf{X}_n, \quad \mathbf{X}_i \in \mathbb{R}^p$, and a postulated (or assumed) model that relates the responses to the covariates, the task of the user is to estimate the parameter of the model. We consider here the case where the responses are generated according to an *un-known* model of the form

$$T_i | \mathbf{X}_i \sim f_0(. | \mathbf{X}_i' \boldsymbol{\beta}_0) \tag{6.1}$$

for some distribution $f_0(. | \mathbf{X}_i' \boldsymbol{\beta}_0)$, with unknown parameters $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. The estimator for $\boldsymbol{\beta}_0$ is computed via Bayesian's Maximum A Posteriori, or equivalently the frequentist's Penalized Maximum Likelihood: i.e it is the solution of an optimization problem

$$\hat{\boldsymbol{\beta}}_n := \arg\min_{\beta} \Big\{ \sum_{i=1}^{n} \ell(\mathbf{X}_i' \boldsymbol{\beta}, T_i) + \mathrm{r}(\boldsymbol{\beta}) \Big\} \tag{6.2}$$

where r is a regularization, or penalty function (this is equivalent to minus the logarithm of the prior in Bayesian language), and $\ell(x,t) := -\log f(t|x)$ is the loss associated with the density f of a chosen Generalized Linear Model (GLM), when viewed as a function of the linear predictor $\mathbf{X}_i'\boldsymbol{\beta}$, with $\mathbf{X}_i$ the covariate vector of the i-th subject, and the response T_i. The examples we have in mind are the linear regression

$$\ell(\mathbf{X}_i'\boldsymbol{\beta}, T_i) := \frac{1}{2}(T_i - \mathbf{X}_i'\boldsymbol{\beta})^2, \tag{6.3}$$

the logistic regression model (where instead of a density we have a probability mass function)

$$\ell(\mathbf{X}_i'\boldsymbol{\beta}, T_i) := \log\cosh(\mathbf{X}_i'\boldsymbol{\beta}) - T_i\mathbf{X}_i'\boldsymbol{\beta} + \log 2, \tag{6.4}$$

or the Cox regression model

$$\ell(\mathbf{X}_i'\boldsymbol{\beta}, \lambda, T_i) := \Lambda(T_i)\exp(\mathbf{X}_i'\boldsymbol{\beta}) - \Delta\log\lambda(T_i) + \text{const}\ , \tag{6.5}$$

where $\Lambda(t) := \int_0^t \lambda(s)\mathrm{d}s$ is the integrated cumulative hazard and $\lambda : \mathbb{R}^+ \to \mathbb{R}^+$ is the base hazard rate.

We are interested in the proportional regime where $p = \zeta n$ for $\zeta \in \mathbb{R}^+$ where the consistency of the MAP estimator is, in general, lost. That is, we cannot estimate with asymptotically vanishing error all the components of $\boldsymbol{\beta}_0$, even when f_0 is postulated correctly [1]. Since consistency is a key step in the "classical" ($p \ll n$) asymptotics [4, 5], highly ideal assumptions are needed to set up an asymptotic theory in the proportional regime. These assumptions are mainly adopted for mathematical convenience. A standard one (see [6, 7, 8, 9]) is to consider that the covariates are "sampled" from a correlated Gaussian distribution with a positive definite covariance matrix $\boldsymbol{\Sigma}_0 \succ 0$, mathematically

$$\mathbf{X}_i \sim \mathcal{N}\big(\mathbf{0}, \boldsymbol{\Sigma}_0\big)\ . \tag{6.6}$$

We shall make this assumption in the rest of the manuscript.

In the scaling regime considered in this paper, it is well known that the Maximum Likelihood Estimator (MLE) might not be well defined. The classical example being the Logistic regression model, for which the MLE does not exist already when p is a not too large fraction of n, see [10]. Under the additional assumption of standard gaussian covariates, i.e. $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$ it has been even proven that the Maximum Likelihood does not exists with probability one past a critical value of p/n, depending on the true signal strength $\|\boldsymbol{\beta}_0\|_2$ and the true intercept ϕ_0 of the data generating model (assumed to be correctly specified) [11]. In a previous paper [12] we showed for $p = \zeta n$ with $\zeta < 1$, the simple covariant regularization

$$\mathrm{r}(\boldsymbol{\beta}) = \frac{1}{2}\alpha\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\ , \tag{6.7}$$

might be used in order to guarantee the existence of $\hat{\boldsymbol{\beta}}_n$ past the critical threshold, and, when properly tuned, it has also a debiasing effect (besides the beneficial shrinking effect, which reduces the variance of the estimator). However, in order to properly tune α, one must estimate $\theta_0 := \|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2$. Here we introduce a very simple (easy-to-implement) and fast

[1] If further assumptions on the data generating process are made and a clever choice of the regularization is made, consistency can be restored. This is the case for instance when using the Lasso regularization and assuming that the underlying $\boldsymbol{\beta}_0$ is *very sparse*, i.e. that the number of non-null components of $\boldsymbol{\beta}_0$ does not scale proportionally with the sample size n [1, 2, 3].

procedure to estimate θ_0. This methodology builds upon recent developments [9] and seems to be widely applicable, beyond our principal objective here. Overall, the Replica Symmetric (RS) equations are used as a guiding tool, but the procedure does not require solving any of the resulting RS equations at all. In fact it *involves only observable quantities*, i.e. quantities that can be computed from the data alone, *without any knowledge of the data generating process*, besides what the user is already assuming during the analysis. Once θ_0 is estimated, the construction of unbiased estimates of $\boldsymbol{\beta}_0$ is "but a hop away". In fact we do not even need to compute the value $\alpha_\star(\theta_0)$ which makes the estimator (asymptotically) unbiased, as was proposed originally in [12]. But rather we can directly estimate the bias and correct for it by dividing the resulting estimator by a factor which is determined *solely from the data*. We show via simulations that the resulting estimator $\hat{\boldsymbol{\beta}}_n^{(d)}$ is effectively un-biased and (approximately) normally distributed around $\boldsymbol{\beta}_0$, with a variance that can be easily estimated from the data. An advantage of the covariant prior is that none of the various quantities that are computed from the data requires the knowledge of the true covariance matrix $\boldsymbol{\Sigma}_0$, which is conversely needed for other approaches to de-biasing with other regularizers like Lasso [7, 8, 9]. These methods, however, allow for de-biasing also when $p > n$, whilst the methodology proposed here requires $p < n$ [2].

6.3 Background on the covariant prior / regularization in the proportional regime

The theoretical analysis of the asymptotics of the estimator $\hat{\boldsymbol{\beta}}_n$ obtained as (6.2) has been derived in several papers with different notations [13, 14, 15, 16] and using different methods like the Convex Gaussian Min Max (CGMT) theorem, see [17, 18], or as the State Evolution equations for the Approximate Message Passing algorithm [19, 20, 21]. For a derivation tailored to the peculiar features of the covariant prior via the Cavity Method we refer to [12]. The result is that, asymptotically in n, we have the approximate representation

$$\hat{\boldsymbol{\beta}}_n \approx \frac{w_\star}{\theta_0}\boldsymbol{\beta}_0 + \frac{1}{\sqrt{p}} v_\star \boldsymbol{\Sigma}_0^{-1/2}\mathbf{Z} \ , \quad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p) \tag{6.8}$$

where $\theta_0 := \|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2$, $\zeta := p/n$ and $w_\star, v_\star, \tau_\star$ satisfy a set of self consistent equations which were originally derived using the Replica Method and are hence called the Replica Symmetric (RS) equations. These read

$$w = \frac{\theta_0}{\zeta}\mathbb{E}_{\Delta,T,Z_0,Q}\Big[\tau\dot{g}\big(\xi,T\big)\dot{g}_0(\theta_0 Z_0, T)\Big] \tag{6.9}$$

$$\zeta = \mathbb{E}_{\Delta,T,Z_0,Q}\left[\frac{\tau\ddot{g}\big(\xi,T\big)}{1+\tau\ddot{g}(\xi,T)}\right] \tag{6.10}$$

$$\zeta v^2 = \tau^2\mathbb{E}_{\Delta,T,Z_0,Q}\Big[\dot{g}(\xi,T)^2\Big] \ , \tag{6.11}$$

where

$$g(x,t) := \ell(x,t) + \alpha\frac{1}{2}x^2 \ , \tag{6.12}$$

[2] Otherwise the regularizer is not guaranteed to be strictly convex and so also the whole objective function to be minimized. In this case multiple minima might exists or no minimum might exist at all.

and

$$\begin{aligned}\xi(wZ_0+vQ,T,\Delta) &= \mathrm{prox}_{g(.,T)}(wZ_0+vQ,\tau) &(6.13)\\ &;= \arg\min_z\Big\{\frac{1}{2\tau}(\xi-wZ_0-vQ)^2+g(T,z)\Big\}\\ &= \mathrm{prox}_{\ell(.,T)}\Big(\frac{wZ_0+vQ}{1+\alpha\tau},\frac{\tau}{1+\alpha\tau}\Big)\ . &(6.14)\end{aligned}$$

The RS equations are restated here in a different notation when compared to [12]. The advantage of this equivalent formulation is the explicit dependence on the model, i.e. on g and its derivatives. For completeness, a short derivation is provided in 6.11.1 via the Replica Method. It has been shown [18, 16] that for the class of problems defined by the Generalized Linear Models and under assumptions on the function g (which are satisfied in our case when $\zeta<1$), we have

$$w_n := \frac{\boldsymbol{\beta}_0'\boldsymbol{\Sigma}_0\hat{\boldsymbol{\beta}}_n}{\|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2} \xrightarrow{P} w_\star \tag{6.15}$$

$$v_n^2 := \|\hat{\boldsymbol{\beta}}_n\|_2^2 - \frac{(\boldsymbol{\beta}_0'\boldsymbol{\Sigma}_0\hat{\boldsymbol{\beta}}_n)^2}{\|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2^2} \xrightarrow{P} v_\star^2\ . \tag{6.16}$$

In 6.11.2, we report the main ideas behind the proof of these results for GLMs, and refer to the relevant papers for the technical details.

From (6.8), we deduced that $\hat{\boldsymbol{\beta}}_n$ is geometrically unbiased when using the covariant regularization, i.e. on average it has the same direction of $\boldsymbol{\beta}_0$, but it is shrunk by a factor $w_\star/\theta_0$ (which is effectively less than 1 for sufficiently large α, depending on $\zeta=p/n$). In a recent paper [9] the author has shown that it is possible to compute $w_\star, v_\star, \tau_\star$ *from the data only* in a very simple manner, without the need to solve the RS equations at all. We show here that his results, can also be derived directly from the RS equations. This alternative derivation leads us to circumvent an important limitation of [9], namely the estimation of the signal strength θ_0 in non-linear models. It is clear from (6.8) that the knowledge of θ_0 is essential to de-bias the estimator $\hat{\boldsymbol{\beta}}_n$. In fact the asymptotic bias is a function of the ratio $w_\star/\theta_0$, i.e.

$$\mathrm{Bias}_n(\zeta,\theta_0) = \big\|\mathbb{E}[\hat{\boldsymbol{\beta}}_n]-\boldsymbol{\beta}_0\big\|_2 \to |w_\star/\theta_0-1|\|\boldsymbol{\beta}_0\|_2\ . \tag{6.17}$$

We show that the solution of very simple equation leads to a good estimate of θ_0. Besides the Bias, also the Mean Squared Error can be estimated once an estimate of θ_0 is obtained, and this might be used as an alternative criterion for the selection of α.

6.4 Approximating the cross validation loss without refitting

Because of (6.15, 6.16), we have that for a fresh sample $\tilde{\mathbf{x}}\sim\mathcal{N}(\mathbf{0}_p,\boldsymbol{\Sigma}_0)$,

$$\tilde{\mathbf{x}}'\hat{\boldsymbol{\beta}}_n \overset{d}{=} w_\star Z_0 + v_\star Q\ , \tag{6.18}$$

with $Z_0, Q\sim\mathcal{N}(0,1)$, independent of each other, and of the training data set. The prediction loss is then given by

$$\mathbb{E}_{T,\tilde{\mathbf{x}}}\Big[\ell(\tilde{\mathbf{x}}'\hat{\boldsymbol{\beta}}_n,T)\Big] = \mathbb{E}_{T,Z_0,Q}\Big[\ell(w_\star Z_0+v_\star Q,T)\Big]\ , \tag{6.19}$$

for $T|\tilde{\mathbf{x}}'\boldsymbol{\beta}_0 \sim f_0(.|\tilde{\mathbf{x}}'\boldsymbol{\beta}_0)$. From the replica theory, we know that at the saddle point

$$\xi = w_\star Z_0 + v_\star Q - \tau \dot{g}(\xi, T) \ , \tag{6.20}$$

hence, based on the identification $\xi \overset{d}{\approx} \mathbf{X}'\hat{\boldsymbol{\beta}}_n$, we can approximate the predicted linear predictor as

$$\xi_i^{\text{loo}} := \mathbf{X}_i'\hat{\boldsymbol{\beta}}_n + \tau_n \dot{g}(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_n, T_i) \ . \tag{6.21}$$

This gives an approximation for the prediction loss as

$$\ell_n^{looapprox} = \frac{1}{n}\sum_{i=1}^{n} \ell(\xi_i^{\text{loo}}, T_i) \ , \tag{6.22}$$

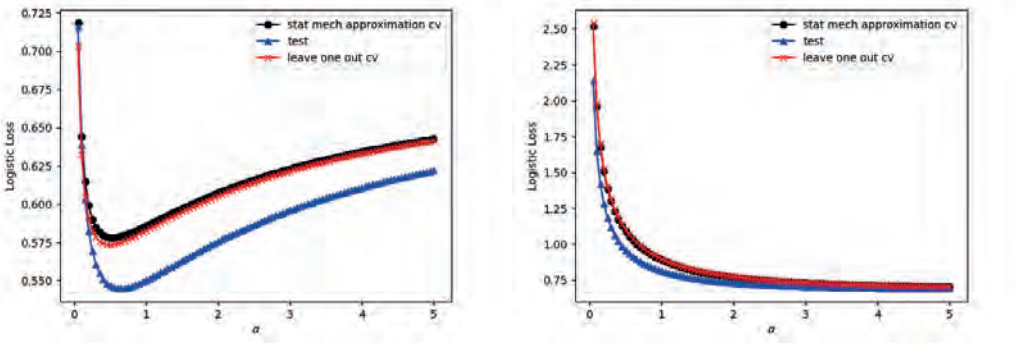
which resembles the leave one out cross validation loss

$$\ell_n^{\text{loocv}} = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_{(i)}, T_i) \ , \tag{6.23}$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the leave one out MAP estimator that solves the optimization problem

$$\hat{\boldsymbol{\beta}}_{(i)} := \arg\min_{\beta} \Big\{ \sum_{j \neq i} -\log f(T_j|\mathbf{X}_j'\boldsymbol{\beta}) + \mathrm{r}(\boldsymbol{\beta}) \Big\} \ . \tag{6.24}$$

It can be seen in Figure (6.1) that the agreement between the approximated leave one out loss (6.22) and its exact counterpart (6.23) is virtually perfect. Indeed the connection between the leave one out linear predictor $\mathbf{X}_i'\hat{\boldsymbol{\beta}}_{(i)}$ and the actual linear predictor $\mathbf{X}_i\hat{\boldsymbol{\beta}}_n$ has been investigated in the literature. In [22] the authors propose a cavity like argument to approximate $\mathbf{X}_i\hat{\boldsymbol{\beta}}_n$ as a function of $\mathbf{X}_i'\hat{\boldsymbol{\beta}}_{(i)}$. The correctness of this approach is then proved in [23], albeit only for linear models. To be more precise, for models that depend only on the linear residuals, i.e $f(T_i|\mathbf{X}_i'\boldsymbol{\beta}) = \tilde{f}(T_i - \mathbf{X}_i\boldsymbol{\beta})$). In [24] a rigorous theory of approximate leave one out is developed along these lines for regularized GLMS. More recently [14] investigated similar questions through the statistical mechanics approach to inference, in particular focusing on the use of Generalize Approximate Message Passing (GAMP) [21].



(a) (b)

Figure 6.1: **Logit regression**. Observed values of leave one out cross validated logistic loss (red line with crosses as markers) against the approximation (6.21) inspired by the RS equations (black line with dot as markers). We also plot the generalization error (blue line with triangle markers) as a reference. Left (a) $\zeta = 0.3$, Right (b) $\zeta = 0.7$. In both cases $\theta_0 = 1.0$ and $n = 200$. In both cases $\theta_0 = 1.0$.

6.5 Estimation of the RS order parameters without solving the RS equations

In order to use the RS equations we need to relate the order parameters to *practically* observable quantities. This idea has been originally proposed in [9], but there a different strategy is used to derive the relationship between the RS order parameters and their observable counterparts.

Based again on the identification $\xi \overset{d}{\approx} \mathbf{X}'\hat{\boldsymbol{\beta}}_n$ we have

$$\mathbb{E}_{\Delta,T,Z_0,Q}\left[\frac{\tau\ddot{g}(\xi,T)}{1+\tau\ddot{g}(\xi,T)}\right] \approx \left\langle \frac{\tau\ddot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n,\mathbf{T})}{1+\tau\ddot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n,\mathbf{T})}\right\rangle \tag{6.25}$$

$$\mathbb{E}_{\Delta,T,Z_0,Q}\left[\dot{g}(\xi,T)^2\right] \approx \tau^2\left\langle \dot{g}\big(\mathbf{X}\hat{\boldsymbol{\beta}}_n,\mathbf{T}\big)^2\right\rangle , \tag{6.26}$$

where we have introduced the empirical average symbol, which for a vector $\mathbf{y} \in \mathbb{R}^l$ is defined as

$$\langle \mathbf{y} \rangle = \frac{1}{l}\sum_{k=1}^{l} y_k, \tag{6.27}$$

and the functions $\dot{g}, \ddot{g}$ act componentwise. Hence we can compute an estimate τ_n of $\tau_\star$ by solving

$$\zeta = \left\langle \frac{\tau_n\ddot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n,\mathbf{T})}{1+\tau_n\ddot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n,\mathbf{T})}\right\rangle , \tag{6.28}$$

since ζ is known. Once this is done, a estimate v_n of $v_\star$ might be obtained as

$$v_n^2 = \frac{1}{\zeta}\tau_n^2\left\langle \dot{g}\big(\mathbf{X}\hat{\boldsymbol{\beta}}_n,\mathbf{T}\big)^2\right\rangle . \tag{6.29}$$

It remains to estimate $w_\star$. In the previous section we have seen that

$$\xi_i^{\text{loo}} := \mathbf{X}_i'\hat{\boldsymbol{\beta}}_n + \tau_n \dot{g}(\mathbf{X}_i'\hat{\boldsymbol{\beta}}_n, T_i) \ , \tag{6.30}$$

is the finite sample (observable) version of the field $w_\star Z_0 + v_\star Q$ appearing in the RS equations. Based on this, we have the identification

$$\frac{1}{n}\|\mathbf{X}'\hat{\boldsymbol{\beta}}_n + \tau_n \dot{g}(\mathbf{X}'\hat{\boldsymbol{\beta}}_n, \mathbf{T})\|_2^2 = w_n^2 + v_n^2 \tag{6.31}$$

from which we get an estimate of w_n as

$$w_n^2 = \frac{1}{n}\|\mathbf{X}'\hat{\boldsymbol{\beta}}_n + \tau_n \dot{g}(\mathbf{X}'\hat{\boldsymbol{\beta}}_n, \mathbf{T})\|_2^2 - v_n^2 \ . \tag{6.32}$$

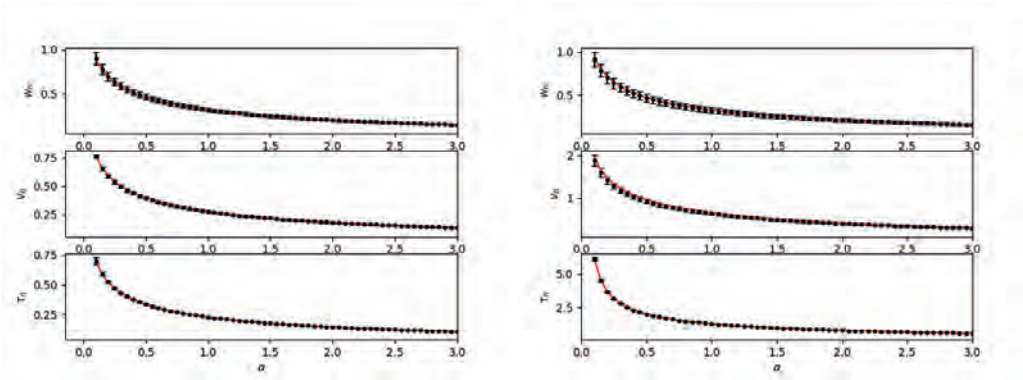
The procedure is easy to implement and leads to very precise estimation of $w_\star, v_\star, \tau_\star$ as one can observe in Figure (6.2). There we display the order parameters w, v, τ as obtained by solving the RS equations. The component of $\boldsymbol{\Sigma}_0$ are fixed as

$$(\boldsymbol{\Sigma}_0)_{j,k}(\epsilon, l) = \delta_{j,k} + \mathbf{1}[|j-k| < l]\epsilon^{|j-k|} \tag{6.33}$$

with $\epsilon = 0.5$ and $l = 7$. This guarantees that our simulations are carried out in a setting that is not close to the uncorrelated one. The data are generated from a Logit model with

$$\boldsymbol{\beta}_0 \sim \nu_0 \text{Unif}(\mathbb{S}_p) \ , \tag{6.34}$$

where $\mathbb{S}_p$ is the unit sphere in $\mathbb{R}^p$, and ν_0 is fixed to have $\theta_0 = \|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2/\sqrt{p}$.



(a) (b)

Figure 6.2: **Logit regression**. The estimated values w_n, v_n, τ_n of the order parameters $w_\star, v_\star, \tau_\star$. The solution of the RS equations are depicted as red lines. The values of the w_n, v_n, τ_n are shown as black error bar plots. Left (a) $\zeta = 0.3$, Right (b) $\zeta = 0.7$. In both cases $\theta_0 = 1.0$ and $n = 500$. The data are generated from a Logit model, with $\boldsymbol{\Sigma}_0$ set according to (6.33) and by sampling $\boldsymbol{\beta}_0$ as in (6.34). *These information is not used to compute the RS order parameters* which are inferred *solely from the data.*

6.6 Our contribution : estimation of the signal strength for arbitrary GLMS

Ideally one would like to have an equation involving only measurable quantities and θ_0. We now show that such an equation can be easily obtained by computing the correlation between the linear predictor $\mathbf{X}_i'\hat{\boldsymbol{\beta}}_n$ and the response T_i. In fact

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{n}\mathbf{T}'\mathbf{X}\hat{\boldsymbol{\beta}}_n\Big] &= \mathbb{E}\Big[T\xi\Big] = \mathbb{E}\Big[T(wZ_0 + vQ - \tau\dot{g}(\xi,T))\Big] = \\ &= w\mathbb{E}\Big[Z_0T\Big] - \tau\mathbb{E}\Big[T\dot{g}(\xi,T)\Big] \end{aligned} \tag{6.35}$$

and hence

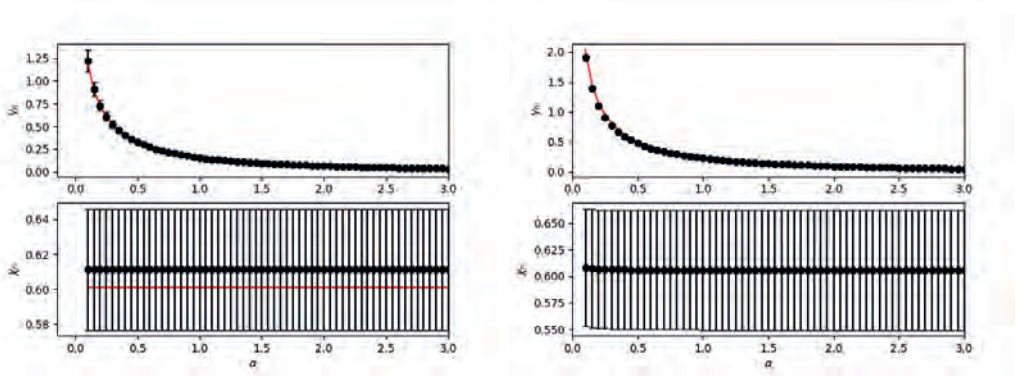
$$\mathbb{E}\Big[T\big\{\xi + \tau\dot{g}(\xi,T)\big\}\Big] = w\mathbb{E}\Big[Z_0T\Big] \tag{6.36}$$

which implies that

$$\begin{aligned}\chi_n &:= \frac{1}{n}\mathbf{T}'\big\{\mathbf{X}\hat{\boldsymbol{\beta}}_n + \tau_n\dot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\approx \mathbb{E}\Big[\frac{1}{n}\mathbf{T}'\big\{\mathbf{X}\hat{\boldsymbol{\beta}}_n + \tau_n\dot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\big\}\Big] = \\ &= \quad w\mathbb{E}\Big[Z_0T\Big] = w\mathbb{E}\Big[Z_0\mathbb{E}[T|Z_0]\Big] = w\mathbb{E}\Big[Z_0\mu(\theta_0Z_0)\Big] = w\theta_0\mathbb{E}\Big[\dot{\mu}(\theta_0Z_0)\Big]\ , \end{aligned} \tag{6.37}$$

where $\mu(\theta_0Z_0) := \mathbb{E}[T|Z_0]$. Notice that this equation is computable for *any* GLM and can be solved both via numerical solution or by means of a grid search. For the Logit regression model we have the simple equation

$$\chi_n/w_n = \mathbb{E}\Big[Z_0\tanh(\theta_0Z_0)\Big] = \theta_0\mathbb{E}\Big[1-\tanh^2(\theta_0Z_0)\Big]\ . \tag{6.38}$$



(a) (b)

Figure 6.3: **Logit regression**. Observed values of γ_n, χ_n against their theoretical values.The solution of the RS equations are depicted as red lines. The values of the $\gamma_n := \|\mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2/n$ and $\chi_n := \mathbf{T}'\big\{\mathbf{X}\hat{\boldsymbol{\beta}}_n + \tau_n\dot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\big\}/n$ are shown as black error bar plots. Left (a) $\zeta = 0.3$, Right (b) $\zeta = 0.7$. In both cases $\theta_0 = 1.0$ and $n = 500$. The data are generated from a Logit model, with $\boldsymbol{\Sigma}_0$ set according to (6.33) and by sampling $\boldsymbol{\beta}_0$ as in (6.34). *This information is not used to compute the estimator* for θ_0 which is computed *solely from the data*.

Other cases can be easily worked out. For the exponential model, for instance, we have

$$\mathbb{E}\Big[Z_0 \log T\Big] = \theta_0 \ , \tag{6.39}$$

since

$$-\log T | Z_0 = \theta_0 Z_0 + G, \quad G \sim \text{Gumbell} \ . \tag{6.40}$$

For the linear model it is easy to verify via direct calculation that

$$\mathbb{E}\Big[Z_0 T\Big] = \theta_0 \ . \tag{6.41}$$

These very simple examples show that θ_0 might be estimated directly from: i)the data and ii) the estimator $\hat{\boldsymbol{\beta}}_n$ (at a fixed α), once a model is postulated. The implicit assumption here is that the model is correctly postulated, which might be violated in practice. If the model is not correctly specified, however, the estimated bias bears no meaning in the first place. Nevertheless, we emphasise that the procedure proposed above might be used to check if there is any (linear) correlation at all between the response and the covariates.

When the model includes nuisance parameters, the situation is slightly more complicated, but the idea remains the same. As done for the method of moments, we just need to match some chosen theoretical statistics with their empirical counterparts computed from the data. For instance, if in linear regression we have an intercept, then this can be estimated via

$$\mathbb{E}\Big[T\Big] = \phi \ , \tag{6.42}$$

as

$$\phi \approx \mathbf{1}'\mathbf{T}/n \ . \tag{6.43}$$

Similarly, the variance of the noise might be estimated via

$$\mathbb{V}\Big[T\Big] = \theta_0^2 + \sigma_0^2 \ . \tag{6.44}$$

As a slightly more involved example, consider the Logit regression model with an intercept

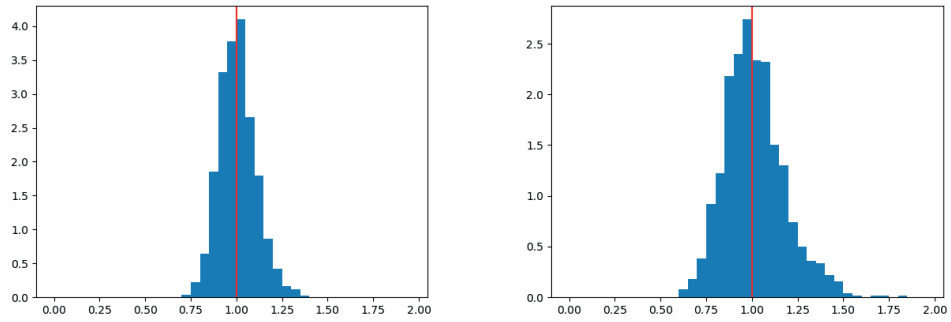
$$T | Z_0 \sim \text{Bernoulli}(p), \quad p := \frac{\mathrm{e}^{\theta_0 Z_0 + \phi_0}}{2\cosh(\theta_0 Z_0 + \phi_0)} \ . \tag{6.45}$$

In this case one imposes the constraints

$$\mathbb{E}\Big[T\Big] = \mathbb{E}\Big[\tanh\big(\theta_0 Z_0 + \phi_0\big)\Big] \approx \langle \mathbf{T} \rangle \tag{6.46}$$

$$\mathbb{E}\Big[Z_0 T\Big] = \mathbb{E}\Big[Z_0 \tanh\big(\theta_0 Z_0 + \phi_0\big)\Big] \approx \frac{1}{n}\mathbf{T}'\big(\mathbf{X}\hat{\boldsymbol{\beta}}_n + \tau_n \dot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n)\big) \ , \tag{6.47}$$

from which ϕ_0 and θ_0 can be easily estimated, for instance via Newton method.



(a) (b)

Figure 6.4: **Logit regression**. The histogram of $m = 1000$ realization of the estimator $\hat{\theta}_n$ obtained by solving the equation (6.37) when $n = 1000$. Left (a) $\zeta = 0.3$, Right (b) $\zeta = 0.7$. In both cases $\theta_0 = 1.0$. The value of the regularization parameter α is selected to minimize the leave one out cross validated minus log-likelihood. We notice that the variance of the estimator increases with the ratio ζ. This is consistent with the larger variance of χ_n. The data are generated from a Logit model, with $\boldsymbol{\Sigma}_0$ set according to (6.33) and by sampling $\boldsymbol{\beta}_0$ as in (6.34).

6.7 Estimation of the component-wise variance for confidence intervals

Via the approximate representation

$$\hat{\boldsymbol{\beta}}_n \approx \frac{w_\star}{\theta_0}\boldsymbol{\beta}_0 + \frac{1}{\sqrt{p}} v_\star \boldsymbol{\Sigma}_0^{-1/2}\mathbf{Z}, \quad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p), \tag{6.48}$$

we have that

$$\mathbf{e}_j'\hat{\boldsymbol{\beta}}_n = \frac{w_\star}{\theta_0}\mathbf{e}_j'\boldsymbol{\beta}_0 + v_\star \mathbf{e}_j'\boldsymbol{\Sigma}_0^{-1/2}\mathbf{Z} \stackrel{d}{=} \frac{w_\star}{\theta_0}\mathbf{e}_j'\boldsymbol{\beta}_0 + \epsilon, \ \epsilon \sim \mathcal{N}\big(0, v_\star^2 \|\boldsymbol{\Sigma}_0^{-1/2}\mathbf{e}_j\|_2^2/p\big) \ .$$

In order to determine an approximate confidence interval, we need to estimate the shrinking factor $w_\star/\theta_0$ in order to center the components of $\hat{\boldsymbol{\beta}}_n$ and the variance to compute the width of the interval. Since we showed in section 6.6 that θ_0 can be estimated, and in section 6.5 that w can be estimated, it remains to estimate $\|\boldsymbol{\Sigma}_0^{-1/2}\mathbf{e}_j\|_2^2$. The estimation of all the components of the covariance matrix is not possible via the empirical covariance matrix, since it is not a consistent estimator (for all its elements) in the proportional limit. An alternative, proposed originally in [9], is to estimate directly $\|\boldsymbol{\Sigma}_0^{-1/2}\mathbf{e}_j\|_2^2 = \mathbf{e}_j\boldsymbol{\Sigma}_0^{-1}\mathbf{e}_j$ as follows .

Consider regressing $\mathbf{X}\mathbf{e}_j$ onto $\mathbf{X}_{(j)} := \mathbf{X}\mathbf{P}_{\perp j} = \mathbf{X}(\boldsymbol{I} - \mathbf{e}_j\mathbf{e}_j')$. It is easy to check (by matching variance and covariance) that $\mathbf{X}\mathbf{e}_j$ admits the conditional representation

$$\mathbf{X}\mathbf{e}_j = \mathbf{X}_{(j)}\gamma_j + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \omega_j^2) \ , \tag{6.49}$$

where

$$\gamma_j = \big(\mathbf{P}_{\perp j}\boldsymbol{\Sigma}_0\mathbf{P}_{\perp j}\big)^{-1}\mathbf{P}_{\perp j}\boldsymbol{\Sigma}_0\mathbf{e}_j \tag{6.50}$$

$$\omega_j^2 = \mathbf{e}_j'\boldsymbol{\Sigma}_0\mathbf{e}_j - \mathbf{e}_j'\boldsymbol{\Sigma}_0\mathbf{P}_{\perp j}\big(\mathbf{P}_{\perp j}\boldsymbol{\Sigma}_0\mathbf{P}_{\perp j}\big)^{-1}\mathbf{P}_{\perp j}\boldsymbol{\Sigma}_0\mathbf{e}_j \ . \tag{6.51}$$

Since

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \tag{6.52}$$
$$\begin{pmatrix} \big(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\big)^{-1} & -\big(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\big)\boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{D}^{-1}\boldsymbol{C}\big(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\big)^{-1} & \boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{C}\big(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}\big)^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix} ,$$

provided $\boldsymbol{D}$ and $\boldsymbol{A}$ are invertible, we have that

$$\mathbf{e}_j'\boldsymbol{\Sigma}_0^{-1}\mathbf{e}_j = 1/\omega_j^2 \ . \tag{6.53}$$

The Least Squares estimator for γ_j reads

$$\hat{\gamma}_j = \big(\mathbf{X}_{(j)}'\mathbf{X}_{(j)}\big)^{-1}\mathbf{X}_{(j)}'\mathbf{X}\mathbf{e}_j = \gamma_j + \big(\mathbf{X}_{(j)}'\mathbf{X}_{(j)}\big)^{-1}\mathbf{X}_{(j)}'\epsilon \ , \tag{6.54}$$

with prediction error given by

$$\mathcal{E}_n \ = \ \frac{1}{n}\|\mathbf{X}\mathbf{e}_j - \mathbf{X}_{(j)}\hat{\gamma}_j\|_2^2 = \frac{1}{n}\|\big(\boldsymbol{I}-\mathbf{X}_{(j)}\big(\mathbf{X}_{(j)}'\mathbf{X}_{(j)}\big)^{-1}\mathbf{X}_{(j)}'\big)\boldsymbol{\epsilon}\|_2^2 \ . \tag{6.55}$$

Noting that the expected prediction error is

$$\begin{aligned} \mathbb{E}\Big[\mathcal{E}_n\Big] \ &= \ \omega_j^2\frac{1}{n}\mathrm{Tr}\big(\boldsymbol{I}-\mathbf{X}_{(j)}\big(\mathbf{X}_{(j)}'\mathbf{X}_{(j)}\big)^{-1}\mathbf{X}_{(j)}'\big) = \\ &= \ \omega_j^2(1-(p-1)/n) \approx \omega_j^2(1-\zeta) \ , \end{aligned} \tag{6.56}$$

we might estimate ω_j^2 by

$$\hat{\omega}_j^2 := \frac{1}{n}\|\mathbf{X}\mathbf{e}_j - \mathbf{X}_{(j)}\hat{\gamma}_j\|_2^2\frac{1}{1-\zeta} \ . \tag{6.57}$$

6.8 Putting it all together : practical application of the RS theory

The results above can be easily used in practice. Once $\hat{\boldsymbol{\beta}}_n$ is obtained for some value of α, we can compute the estimates w_n, v_n, τ_n as detailed in section 6.5 and an estimate θ_n (although we notice in practice that the resulting estimate is, as it should, independent of α) of θ_0. At this point the debiased estimator is obtained as

$$\hat{\boldsymbol{\beta}}^{(d)} = \hat{\boldsymbol{\beta}}_n\theta_n/w_n \ , \tag{6.58}$$

the per-component variance can be estimate as explained in section 6.7. The per-component residuals

$$R_j^{(d)} := \sqrt{p}\hat{\omega}_j\frac{\mathbf{e}_j'\big(\hat{\boldsymbol{\beta}}_n^{(d)}-\boldsymbol{\beta}_0\big)}{v_n} \tag{6.59}$$

should be approximately Gaussian with unit variance and zero mean, i.e.

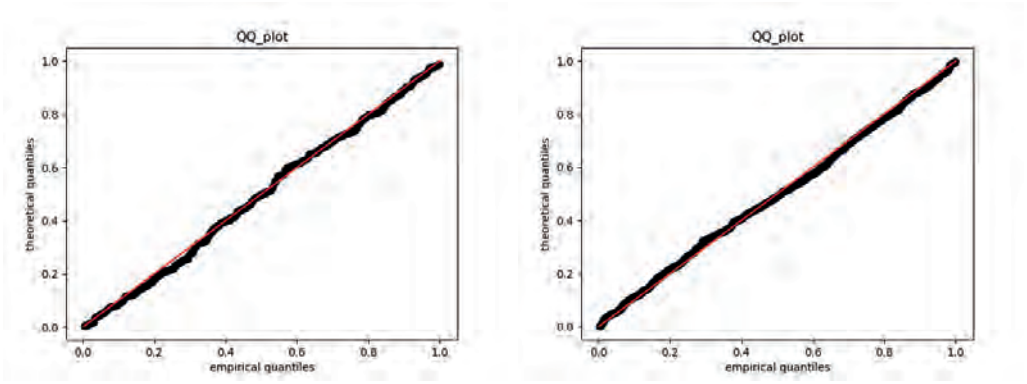
$$R_j^{(d)} \overset{d}{\approx} \mathcal{N}(0,1) \ . \tag{6.60}$$

A QQ-plot can be used to assess the validity of this claim : it requires to compute the empirical quantiles of $R^{(d)}$ and plot them against the theoretical quantiles. If these are

close to each other, i.e if the plot is approximately a line at 45 degrees, then the residuals are approximately gaussian with variance one and mean zero. The quantile function of the Normal distribution is

$$\mathcal{Q}(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2} \mathrm{d}x \ . \tag{6.61}$$

If we indicate with $R^{(d)}_{(1)}$ the lowest value in $R^{(d)}_1, \ldots, R^{(d)}_n$, with $R^{(d)}_{(2)}$ the second lowest value and so on, then the QQ plot is obtained by displaying $\mathcal{Q}(R^{(d)}_{(1)}), \mathcal{Q}(R^{(d)}_{(2)}), \ldots, \mathcal{Q}(R^{(d)}_{(n)})$ against $(1/n, 2/n, \ldots, 1)$. We show in Figure (6.5) the QQ-plot for two realizations, on the left $\zeta = 0.3$ and on th eright $\zeta = 0.7$ of the per-component residuals $R^{(d)}_j$s defined in (6.59). These show convincingly that the per-component residuals $R^{(d)}$ of the debiased estimator (black dots) are effectively distributed according to a standard normal. In order to select a value of α in a data-driven manner we compute a solution path $\hat{\boldsymbol{\beta}}_n(\alpha)$ for a list of values of α and select the value of α that minimizes the leave one out cross validation loss. Instead of computing the actual leave one out loss, we use the approximation explained in section 6.4. Another possible criterion might be to select the value of α that maximizes $w_n(\alpha)/v_n(\alpha)$, i.e. it maximizes the inferred signal and simultaneously minimizes the noise.



(a) (b)

Figure 6.5: **Logit regression**. QQ plots of the residuals of the de-biased estimator in black, i.e. $\hat{R}^{(d)}_j$s defined in (6.59). In red the identity line. The value of the regularization parameter α is selected to minimize the leave one out cross validated minus log-likelihood. Left (a) $\zeta = 0.3$ and right (b) $\zeta = 0.7$, in both cases $\theta_0 = 1.0$ and $n = 1000$. The data are generated from a Logit model, with $\boldsymbol{\Sigma}_0$ set according to (6.33) and by sampling $\boldsymbol{\beta}_0$ as in (6.34).

6.9 Applying the previous ideas to the Cox semiparametric regression model

The approach described above might be applied, in principle, also to models that do not strictly fall under the umbrella of the GLM class, but with density depending solely on the linear predictor. An example is the Cox semiparametric regression model, which assumes

$$f(\Delta_i, T_i | \mathbf{X}_i'\boldsymbol{\beta}, \lambda) = \\ \mathrm{e}^{-\Lambda_0(T_i)\exp(\mathbf{X}_i'\boldsymbol{\beta}) - \Lambda_C(T_i)} \Big(\lambda_0(T_i)\exp(\mathbf{X}_i'\boldsymbol{\beta})\Big)^{\Delta} \lambda_c(T_i)^{1-\Delta_i} \ , \tag{6.62}$$

where Λ_0, Λ_c are the cumulative hazard rate functions of the primary risk and censoring respectively. These satisfy $\Lambda_0(t) = \int_0^\infty \lambda_0(s)\mathrm{d}s$ and $\Lambda_c(t) = \int_0^\infty \lambda_c(s)\mathrm{d}s$, with $\lambda_0, \lambda_c : \mathbb{R}^+ \to \mathbb{R}^+$.

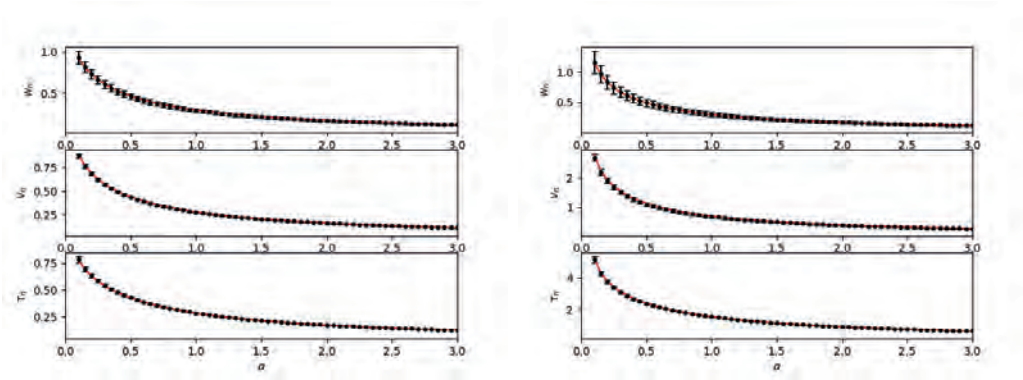
The model is semi-parametric since the shape of the cumulative hazard rate is not constrained to be in any known parametric family of functions, but is inferred non-parametrically via the Nelson Aalen estimator

$$\hat{\Lambda}_n(t) = \sum_{k=1}^{n} \frac{\Delta_k \Theta(t - T_k)}{\sum_{j=1}^{n} \Theta(T_j - T_k) \mathrm{e}^{\mathbf{X}_j' \hat{\boldsymbol{\beta}}_n}} \ . \tag{6.63}$$

Where $\hat{\boldsymbol{\beta}}_n$ is computed by minimizing the sum of minus partial likelihood

$$\mathcal{PL}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i \Big\{ \log\Big(\sum_{j=1}^{n} \Theta(T_j - T_i) \mathrm{e}^{\mathbf{X}_j' \boldsymbol{\beta}} \Big) - \mathbf{X}_i' \boldsymbol{\beta} \Big\} \tag{6.64}$$

and covariant regularization (6.7)



(a) (b)

Figure 6.6: **Cox regression**. The estimated values w_n, v_n, τ_n of the order parameters $w_\star, v_\star, \tau_\star$. The solution of the RS equations are depicted as red lines. The values of the w_n, v_n, τ_n are shown as black error bar plots. The value of the regularization parameter α is selected to maximize the ratio $w_n(\alpha)/v_n(\alpha)$. Left (a) $\zeta = 0.3$, Right (b) $\zeta = 0.7$. In both cases $\theta_0 = 1.0$ and $n = 500$.. *These information is not used to compute the RS order parameters* which are inferred *solely from the data*.

We recently carried out the asymptotic analysis for this model in the proportional regime in [25]. The result is that the estimator $\hat{\boldsymbol{\beta}}_n$ obtained by maximizing (6.64) admits the representation (6.8), with the RS order parameters $w_\star, v_\star, \tau_\star$ solving

$$w = \frac{\theta_0}{\zeta} \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\tau \dot{g}\big(\xi, \Lambda(T), \Delta\big) \dot{g}_0(\theta_0 Z_0, \Lambda_0(T), \Delta) \Big] \tag{6.65}$$

$$\zeta = \mathbb{E}_{\Delta,T,Z_0,Q}\left[\frac{\tau \ddot{g}\big(\xi, \Lambda(T), \Delta\big)}{1 + \tau \ddot{g}(\xi, \Lambda(T), \Delta)} \right] \tag{6.66}$$

$$\zeta v^2 = \tau^2 \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\dot{g}(\xi, \Lambda(T), \Delta)^2 \Big] \ , \tag{6.67}$$

where

$$g(x,\Lambda(T),\Delta) := \ell(x,\Lambda(T),\Delta)+\alpha\frac{1}{2}x^2\ , \tag{6.68}$$

$$\ell(x,\Lambda(T),\Delta) := \Lambda(T)\mathrm{e}^x-\Delta x\ , \tag{6.69}$$

and

$$\begin{aligned}\xi(wZ_0+vQ,T,\Delta) &= \mathrm{prox}_{g(.,\Lambda(T),\Delta)}(wZ_0+vQ,\tau):=\\ &= \arg\min_z\Big\{\frac{1}{2\tau}(\xi-wZ_0-vQ)^2+g(T,z)\Big\}=\\ &= \mathrm{prox}_{\ell(.,\Lambda(T),\Delta)}\Big(\frac{wZ_0+vQ}{1+\alpha\tau},\frac{\tau}{1+\alpha\tau}\Big)=\\ &= \frac{w_\star Z_0+v_\star Q+\tau\Delta}{1+\alpha\tau}-W_0\Big(\frac{\tau\Lambda(T)\mathrm{e}^{\frac{\tau\Delta+w_\star Z_0+v_\star Q}{1+\alpha\tau}}}{1+\alpha\tau}\Big)\ ,\end{aligned} \tag{6.70}$$

with

$$\Lambda(t)=\mathbb{E}_{\Delta,T,Z_0,Q}\left[\frac{\Delta\Theta(t-T)}{\mathbb{E}_{\Delta',T',Z_0',Q'}\Big[\Theta(T'-T)\mathrm{e}^{\xi_\star'}\Big]}\right]\ . \tag{6.71}$$

At the saddle point, the last equations above relates the inferred cumulative hazard rate $\hat{\Lambda}_n$ to the true one Λ_0, through the identification $\hat{\Lambda}_n\approx\Lambda_\star$.

As a consequence we can estimate the order parameters of the theory from the data, by using the same reasoning as in the previous sections, i.e. by solving

$$\zeta = \Big\langle\frac{\tau_n\ddot{g}\big(\mathbf{X}\hat{\boldsymbol{\beta}}_n,\hat{\Lambda}_n(\mathbf{T}),\boldsymbol{\Delta}\big)}{1+\tau_n\ddot{g}\mathbf{X}\hat{\boldsymbol{\beta}}_n,\hat{\Lambda}_n(\mathbf{T}),\boldsymbol{\Delta})}\Big\rangle \tag{6.72}$$

$$\zeta v_n^2 = \tau_n^2\Big\langle\big\{\dot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n,\hat{\Lambda}_n(\mathbf{T}),\boldsymbol{\Delta})\big\}^2\Big\rangle \tag{6.73}$$

$$w_n^2 = \frac{1}{n}\big\|\mathbf{X}\hat{\boldsymbol{\beta}}_n+\tau_n\dot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n,\hat{\Lambda}_n(\mathbf{T}),\boldsymbol{\Delta})\big\|_2^2-v_n^2\ . \tag{6.74}$$

Where the functions inside the empirical average operator are taken to act component-wise.

In [25], we proposed an iterative method that allows to solve the RS equations without the need of knowing the (parameters of the) data generating process. The idea is to infer Λ_c by the Breslow estimator

$$\hat{\Lambda}_c(t)=\sum_{k=1}^n\frac{(1-\Delta_k)\Theta(t-T_k)}{\sum_{j=1}^n\Theta(T_j-T_k)}\ , \tag{6.75}$$

and Λ_0 at fixed θ_0 by solving

$$\tilde{\Lambda}(t,\theta)=\sum_{i=1}^n\frac{\Delta_i\mathbf{1}[T_i<t]}{\sum_{j=1}^n\mathbf{1}[T_j>T_i]\mathbb{E}_{Z|T_j}\Big[\mathrm{e}^{\theta Z}\Big]}\ . \tag{6.76}$$

Then, always at fixed θ_0, the Gaussian latent variables Q, Z_0 used to compute the expectations in the RS equations are drawn independently as follows

$$Q|T=Q \sim \mathcal{N}(0,1) \tag{6.77}$$

$$\begin{aligned}Z_0|T,\Delta &\sim \frac{\big(\lambda_0(t)\exp(\theta_0Z_0)\big)^\Delta\lambda_c(t)^{1-\Delta}\mathrm{e}^{-\Lambda_0(t)\exp(\theta_0Z_0)-\Lambda_c(t)}}{\mathbb{E}_{Z_0}\Big[\big(\lambda_0(t)\exp(\theta_0Z_0)\big)^\Delta\lambda_c(t)^{1-\Delta}\mathrm{e}^{-\Lambda_0(t)\exp(\theta_0Z_0)-\Lambda_c(t)}\Big]}=\\ &= \frac{\mathrm{e}^{\Delta\theta_0Z_0-\Lambda_0(t)\exp(\theta_0Z_0)}}{\mathbb{E}_{Z_0}\Big[\mathrm{e}^{\Delta\theta_0Z_0-\Lambda_0(t)\exp(\theta_0Z_0)}\Big]}\ .\end{aligned} \tag{6.78}$$

Originally we solved for all the order parameters, i.e. w, v, τ (with the approximation $\hat{\Lambda}_n \approx \Lambda$ in the RS equations) while simultaneously enforcing the approximate constraint

$$w^2 + v^2(1-\zeta) \approx \|\mathbf{X}\hat{\boldsymbol{\beta}}_n\|_2^2/n \ , \tag{6.79}$$

since

$$\begin{aligned}
\zeta v^2 &= \tau^2 \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\dot{g}(\xi,T)^2\Big] = \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\big(\xi - wZ_0 - vQ\big)^2\Big] = \\
&= \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\xi^2\Big] - 2w\mathbb{E}_{\Delta,T,Z_0,Q}\Big[Z_0\xi\Big] - 2v\mathbb{E}_{\Delta,T,Z_0,Q}\Big[Q\xi\Big] + w^2 + v^2 = \\
&= \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\xi^2\Big] + 2v^2\zeta - w^2 - v^2 = \mathbb{E}_{\Delta,T,Z_0,Q}\Big[\xi^2\Big] - v^2(1-2\zeta) - w^2 \ ,
\end{aligned} \tag{6.80}$$

implies

$$\mathbb{E}_{\Delta,T,Z_0,Q}\Big[\xi^2\Big] = w^2 + v^2(1-\zeta) \ . \tag{6.81}$$

At convergence we have an estimate θ_n of θ_0 and an estimate $\tilde{\Lambda}$ of Λ_0. However, this procedure is relatively slow, since we need to iterate the RS equations until convergence and at each step we need to infer Λ_0 at the updated value of θ_n. In this manuscript, we realize that solving all the RS equations simultaneously by iterative mapping is not necessary. Since w_n, v_n, τ_n can be inferred from the data only (as detailed in section 6.5), as shown convincingly in Figure (6.6), we can easily compute

$$\xi | Z_0, Q, T = \text{prox}_{g(.,\hat{\Lambda}_n(T),\Delta)}(w_n Z_0 + v_n Q, \tau_n) \tag{6.82}$$

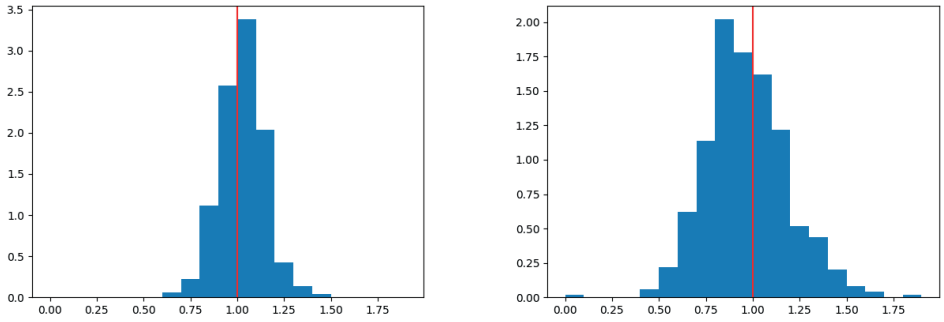
and directly compute all the RS observables (at fixed θ_0). At this point we just need to match the computed value with the observed value in order to infer θ_0. For instance once $\tilde{w}$ is computed from

$$\tilde{w}(\theta,\zeta) = \frac{\theta_0}{\zeta}\sum_{i=1}^{n}\mathbb{E}_{Z_0,Q|T_i,\Delta_i,\theta}\Big[\tau\dot{g}\big(\xi,T_i\big)\dot{g}_0(\theta_0 Z_0, T_i)\Big] \ , \tag{6.83}$$

on a grid of values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_l)$ for θ_0, the best value is selected by matching to w_n which is measured as explained in the section 6.5 from the data, i.e.

$$\hat{\theta}_n = \underset{\theta\in\boldsymbol{\theta}}{\arg\min}\Big\{ \ \big|w_n - \tilde{w}(\zeta,\theta)\big| \ \Big\} \ . \tag{6.84}$$

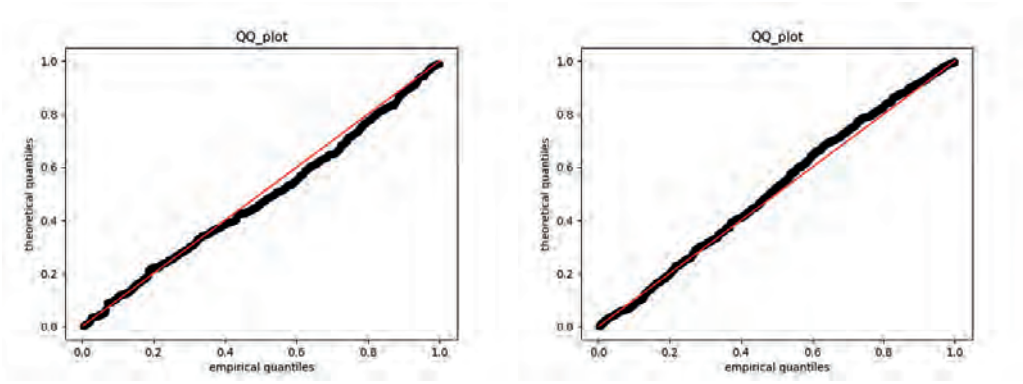
We show the histogram based on $m = 500$ realizations of $\hat{\theta}_n$ for $n = 500$ in Figure (6.7) for $\zeta = 0.3, 0.7$. The data are generated from a Log-logistic proportional hazards model $\Lambda_0(t) = \log(1 + t^2/2)$ with uniform censoring between $\tau_1 = 1.0$ and $\tau_2 = 2$. The $\boldsymbol{\Sigma}_0$ equal to (6.33). We notice that as ζ increases so does the variance of the estimate, similarly as for the estimates for the GLMs in Figure (6.4).



(a) (b)

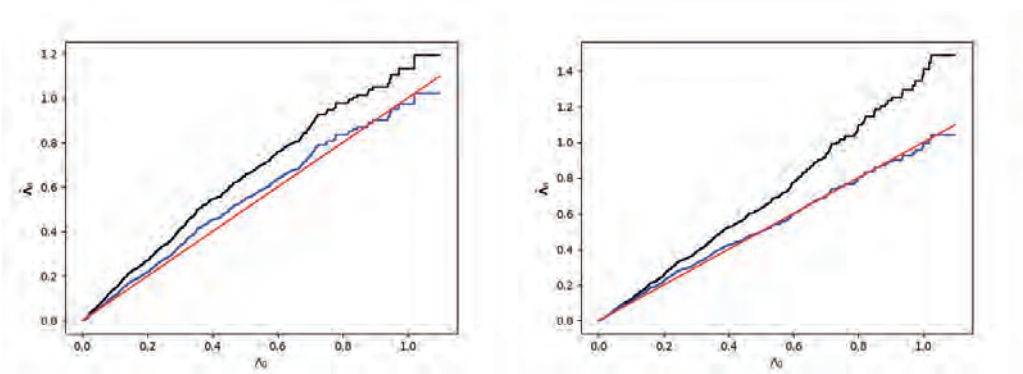
Figure 6.7: **Cox regression**. Histograms of $m = 500$ realization of the estimator $\hat{\theta}_n$ for $n = 500$ defined in (6.84). In red the true value $\theta_0 = 1.0$. Left (a) $\zeta = 0.3$ and right (b) $\zeta = 0.7$. The value of the regularization parameter α is selected to minimize an approximation of the Vervweij - Van Houwelingen cross validation loss (6.87). The data are generated from a Log-logistic proportional hazards model $\Lambda_0(t) = \log(1+t^2/2)$ with uniform censoring between $\tau_1 = 1.0$ and $\tau_2 = 2$. The true covariance matrix $\boldsymbol{\Sigma}_0$ is equal to (6.33) and the true associations $\boldsymbol{\beta}_0$ are sampled as in (6.34).

It is clear that the grid search is a slower (and far less elegant) solution compared to the very simple algorithm proposed for GLMs. On the other hand we empirically observe that using the same strategy as for the GLM does not work for the Cox model. We suspect that this is due to some form of collinearity with the equations used to construct (6.76), but this is a mere speculation. In any case we find that the algorithm proposed returns an approximately de-bias estimator $\hat{\boldsymbol{\beta}}_n^{(d)}$ and is much faster than the previously proposed method. Furthermore the covariant regularization "cancels" the phase transition for the existence of $\hat{\boldsymbol{\beta}}_n$, hence the algorithm is more stable and can be applied for $\zeta < 1$ as one can observe in Figures (6.8, 6.9).



(a) (b)

Figure 6.8: **Cox regression**. QQ plots of two realization of the residuals of the de-biased estimator in black, i.e. $\hat{R}_j^{(d)}$s defined in (6.59). In red the identity line. Left (a) $\zeta = 0.3$ and right (b) $\zeta = 0.7$, in both cases $\theta_0 = 1.0$ and $n = 1000$. The value of the regularization parameter α is selected to minimize an approximation of the Vervweij - Van Houwelingen cross validation loss (6.87). The data are generated from a Log-logistic proportional hazards model $\Lambda_0(t) = \log(1 + t^2/2)$ with uniform censoring between $\tau_1 = 1.0$ and $\tau_2 = 2$. The true covariance matrix $\mathbf{\Sigma}_0$ is equal to (6.33) and the true associations $\boldsymbol{\beta}_0$ are sampled as in (6.34).



(a) (b)

Figure 6.9: **Cox regression**. Scatter plots of two realizations of the inferred cumulative hazard rate for the Breslow estimator $\hat{\Lambda}_n$ in (6.63) as black staircase line, and for the de-biased estimator $\tilde{\Lambda}_n$ as blue staircase line. In red the identity line. Left (a) $\zeta = 0.3$ and right (b) $\zeta = 0.7$, in both cases $\theta_0 = 1.0$ and $n = 1000$. The value of the regularization parameter α is selected to minimize an approximation (6.87) of the Vervweij - Van Houwelingen cross validation loss (6.86). The data are generated from a Log-logistic proportional hazards model $\Lambda_0(t) = \log(1 + t^2/2)$ with uniform censoring between $\tau_1 = 1.0$ and $\tau_2 = 2$. The true covariance matrix $\mathbf{\Sigma}_0$ is equal to (6.33) and the true associations $\boldsymbol{\beta}_0$ are sampled as in (6.34).

In the discussion above, there remains an ambiguity over which value of α should be

chosen to compute the debiased estimator for the Cox model. Leave one out cross validation as explained for GLMs is not applicable in this case. There exists however a generalization specifically devised for the Cox partial likelihood [26]. Defining the leave one out partial likelihood

$$\mathcal{PL}_{(i)}(\boldsymbol{\beta}) := \sum_{k=1\neq i}^{n} \Delta_k \Big\{ \log\Big(\sum_{j=1\neq i}^{n} \Theta(T_j - T_k) \mathrm{e}^{\mathbf{X}_j'\boldsymbol{\beta}} \Big) - \mathbf{X}_k'\boldsymbol{\beta} \Big\} \tag{6.85}$$

and denoting $\hat{\boldsymbol{\beta}}_{(i)}(\alpha)$ the minimizer of $\mathcal{PL}_{(i)}(\boldsymbol{\beta})$ plus the covariant regularization (6.7), the (rescaled) Vervweij - Van Houwelingen cross validation loss can be defined as

$$\mathrm{cvloss}(\alpha) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{PL}_n\big(\hat{\boldsymbol{\beta}}_{(i)}(\alpha)\big) - \mathcal{PL}_{(i)}\big(\hat{\boldsymbol{\beta}}_{(i)}(\alpha)\big) \ . \tag{6.86}$$

This requires re-fitting the Cox model n times (each time for several values of α), as for the leave one out cross validation. For small ζ this operation can be nowadays done quickly by parallelization, however as ζ increases this strategy is bounded to be become slower. For this reason we propose to minimize approximate metric to select α. This is constructed in a similar fashion as to what is done in section 6.4,

$$\mathrm{cvloss}^{\mathrm{approx}}(\alpha) := \frac{1}{n} \sum_{i=1}^{n} \hat{\Lambda}_n(T_i) \mathrm{e}^{\xi_i^{\mathrm{loo}}} - \Delta_i(\xi_i^{\mathrm{loo}} + \log \hat{\lambda}_n(T_i)) \tag{6.87}$$

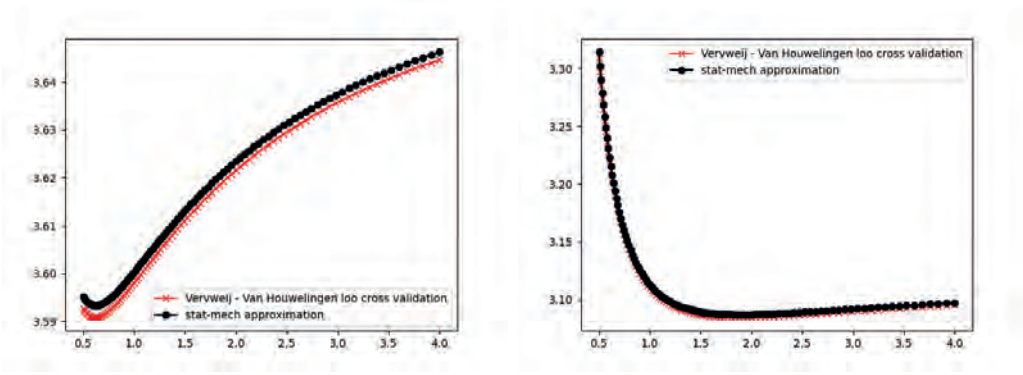
where

$$\xi_i^{\mathrm{loo}} := \mathbf{X}_i'\hat{\boldsymbol{\beta}}_n + \tau_n \hat{\Lambda}_n(T_i) \mathrm{e}^{\mathbf{X}_i'\hat{\boldsymbol{\beta}}_n} \ , \tag{6.88}$$

and

$$\hat{\lambda}_n(T_i) := \hat{\Lambda}_n(T_i) - \hat{\Lambda}_n(T_{i-1}) \, i = 2, \ldots, n, \ \hat{\Lambda}(T_1) = 0 \tag{6.89}$$

is the Breslow estimate of the baseline hazard rate [26]. The main advantage of the approximate cross validation loss (6.87) over its exact counterpart (6.86) is that the former does not require any resampling and can be calculated directly during the construction of the solution path in an easy manner. Numerical experiments show that this approximation is quite satisfactory, see Figure (6.10), at least in the setting studied here.



(a) (b)

Figure 6.10: **Cox regression**. Observed values of the Vervweij - Van Houwelingen cross validation loss (red line with crosses as markers) against the approximation (6.87) inspired by the RS equations (black line with dot as markers). Left (a) $\zeta = 0.3$, Right (b) $\zeta = 0.7$. In both cases $\theta_0 = 1.0$ and $n = 400$. The data are generated from a Log-logistic proportional hazards model $\Lambda_0(t) = \log(1 + t^2/2)$ with uniform censoring between $\tau_1 = 1.0$ and $\tau_2 = 2$. The true associations $\boldsymbol{\beta}_0$ are sampled as in (6.34) and $\boldsymbol{\Sigma}_0$ is set according to (6.33). The constraint $\theta_0 := \|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2 = 1.0$ is enforced by rescaling $\boldsymbol{\beta}_0$.

6.10 Conclusion

Thanks to the insight provided by the RS equations and the recent advancement in [9], the RS order parameters w, v, τ can be estimated as w_n, v_n, τ_n computed *solely* from the data, without the need of solving the RS equations. This is possible for the case of MAP estimation with a regularization term like the covariant prior [12] even for covariates that are correlated with an unknown $\boldsymbol{\Sigma}_0 \succ 0$. Importantly, the use of the covariant prior allows to compute all the necessary unknowns without the knowledge of $\boldsymbol{\Sigma}_0$, which is needed when adopting lasso or ridge [9, 8]. This is an important development for high dimensional statistics in the proportional regime, when $\zeta < 1$, since we no more require knowledge of the parameters of the data generating process. We showed that the simple equation

$$\mathbb{E}\Big[Z_0\mu(\theta_n Z_0)\Big] = \frac{1}{n}\mathbf{T}'\big(\mathbf{X}\hat{\boldsymbol{\beta}}_n + \tau\dot{g}(\mathbf{X}\hat{\boldsymbol{\beta}}_n, \mathbf{T})\big)/w_n \tag{6.90}$$

where $\mu(\theta_0) = \mathbb{E}[T|\theta_0 Z_0]$, can be effectively used to estimate $\theta_0 := \|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2$ from the data, by computing θ_n for Generalized Linear Models. For the Cox model, we propose and test a new, faster, strategy to estimate θ, building on our previous study in [25]. Besides being a simpler strategy , the improvement is in the running time as we do not need to iterate the RS equations to fixed point, but we only need to solve a single scalar equation. This automatically gives also acces to a de-biased Breslow estimator as $\tilde{\Lambda}(., \theta_n)$, see the definition (6.76). Once an estimate θ_n of θ_0 is obtained, we compute the de-biased estimator

$$\hat{\boldsymbol{\beta}}_n^{(d)} := \hat{\boldsymbol{\beta}}_n\theta_n/w_n \ . \tag{6.91}$$

The replica theory gives the representation

$$\hat{\boldsymbol{\beta}}_n \approx \frac{w_\star}{\theta_0}\boldsymbol{\beta}_0 + \frac{1}{\sqrt{p}} v_\star \boldsymbol{\Sigma}_0^{-1/2}\mathbf{Z} \ , \quad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p) \tag{6.92}$$

and because of this, we expect that our debiased estimator should satisfy

$$\hat{\boldsymbol{\beta}}_n^{(d)} - \boldsymbol{\beta}_0 \approx \frac{1}{\sqrt{p}} v_\star \boldsymbol{\Sigma}_0^{-1/2}\mathbf{Z} \ , \tag{6.93}$$

which we showed to be solid, through numerical experiments. In conclusion we have shown that not only for GLMs, but also for the Cox model, the covariant prior is a regularization that allows practical de-biasing, without requiring knowledge of $\boldsymbol{\Sigma}_0$ when $\zeta < 1$. For the regime where $p > n$, we still envisage using (6.90) to complement the estimation of the RS order parameters from the data when different regularizations others than the one studied here are used. The problematic in this case is that the knowledge of $\boldsymbol{\Sigma}_0$ seems to be required to estimate the RS order parameters. This will be subject of future investigations.

Data availability statement The python scripts for generating the data, solving the RS equations, fitting the models and computing the RS order parameters from the data can be found at `https://github.com/EmanueleMassa/Practical_debiasing` .

Bibliography

[1] P Bühlmann and S van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg, 2011.

[2] J Lederer. *Fundamentals of High-Dimensional Statistics: With Exercises and R Labs*. Springer Texts in Statistics. Springer International Publishing, 2021.

[3] PJ Bickel, Y Ritov, and AB Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009.

[4] L Fahrmeir and H Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

[5] SA van de Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[6] T Takahashi and Y Kabashima. A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(7):073405, jul 2018.

[7] A Javanmard and A Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593 – 2622, 2018.

[8] M Celentano, A Montanari, and Y Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194 – 2220, 2023.

[9] PC Bellec. Observable adjustments in single-index models for regularized m-estimators, 2024.

[10] MJ Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(3):310–313, 1981.

[11] EJ Candès and P Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27 – 42, 2020.

[12] E Massa, MA Jonker, and ACC Coolen. Penalization-induced shrinking without rotation in high dimensional glm regression: a cavity analysis. *Journal of Physics A: Mathematical and Theoretical*, 55(48):485002, dec 2022.

[13] ACC Coolen, M Sheikh, A Mozeika, F Aguirre-Lopez, and F Antenucci. Replica analysis of overfitting in generalized linear regression models. *Journal of Physics A: Mathematical and Theoretical*, 53(36):365001, aug 2020.

[14] A Sakata. Prediction errors for penalized regressions based on generalized approximate message passing. *Journal of Physics A: Mathematical and Theoretical*, 56(4):043001, feb 2023.

[15] F Salehi, E Abbasi, and B Hassibi. The impact of regularization on high-dimensional logistic regression. *ArXiv*, abs/1906.03761, 2019.

[16] B Loureiro, C Gerbelot, H Cui, S Goldt, F Krzakala, M Mézard, and L Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model*. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114001, nov 2022.

[17] C Thrampoulidis, S Oymak, and B Hassibi. The gaussian min-max theorem in the presence of convexity, 2015.

[18] C Thrampoulidis, E Abbasi, and B Hassibi. Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

[19] DL Donoho, A Maleki, and A Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[20] A Javanmard and A Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *ArXiv*, abs/1211.5164, 2012.

[21] S Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172, 2011.

[22] N El Karoui, D Bean, PJ Bickel, C Lim, and B Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[23] N El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170:95 – 175, 2017.

[24] KR Rad and A Maleki. A Scalable Estimate of the Out-of-Sample Prediction Error via Approximate Leave-One-Out Cross-Validation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):965–996, 06 2020.

[25] E Massa, A Mozeika, and ACC Coolen. Replica analysis of overfitting in regression models for time to event data: the impact of censoring. *Journal of Physics A: Mathematical and Theoretical*, 57(12):125003, mar 2024.

[26] PJM Verweij and HC Van Houwelingen. Cross-validation in survival analysis. *Statistics in medicine*, 12(24):2305–2314, 1993.

[27] Y Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50, 1985.

6.11 Appendix

6.11.1 Replica Calculation for the Generalized Linear Model

Notice that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n(\mathbf{X},\mathbf{T}) &:= \arg\max_{\boldsymbol{\beta}}\Big\{\sum_{i=1}^{n} g(\mathbf{X}_i\boldsymbol{\beta},T_i)\Big\} = \\ &\stackrel{d}{=} \boldsymbol{\Sigma}_0^{-1/2}\hat{\boldsymbol{\beta}}_n(\tilde{\mathbf{X}},\tilde{\mathbf{T}}) \ ,\end{aligned} \tag{6.94}$$

with

$$\tilde{T}_i|\tilde{\mathbf{X}}_i'\tilde{\boldsymbol{\beta}}_0 = f_0(.|\tilde{\mathbf{X}}_i'\tilde{\boldsymbol{\beta}}_0) \ , \quad \tilde{\mathbf{X}}_i \sim \mathcal{N}(\mathbf{0},\boldsymbol{I}_p/p) \ , \tag{6.95}$$

and $\tilde{\boldsymbol{\beta}}_0 := \boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\sqrt{p}$. So, in order to get rid of $\boldsymbol{\Sigma}_0$ in the derivation, we take the additional change of variables

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}\sqrt{p} \ , \tag{6.96}$$

where the scaling in p is to obtain neater mathematical formulae. We emphasize that the calculation might be carried out without the latter, but at the price of introducing at hoc constants, see [25]. It is well known that the information content of inference is quantified by the free energy

$$f(\gamma) := -\lim_{n\to\infty}\frac{1}{n\gamma}\mathbb{E}_{\mathcal{D}}\Big[\log Z_n(\gamma|\mathcal{D})\Big], \quad Z_n(\gamma|\mathcal{D}) := \int \mathrm{e}^{-\gamma\mathcal{H}_n(\tilde{\boldsymbol{\beta}}|\mathcal{D})}\mathrm{d}\tilde{\boldsymbol{\beta}} \tag{6.97}$$

where the data-set is indicated with $\mathcal{D} := \{(\tilde{T}_1,\tilde{\mathbf{X}}_1),\dots,(\tilde{T}_n,\tilde{\mathbf{X}}_n)\}$, and the Hamiltonian of the problem is defined as

$$\mathcal{H}_n(\tilde{\boldsymbol{\beta}}|\mathcal{D}) = g(\tilde{\mathbf{X}}_i'\tilde{\boldsymbol{\beta}},\tilde{T}_i) \tag{6.98}$$

To compute the average of the logarithm we use the replica trick

$$f(\gamma) = \lim_{n\to\infty}\lim_{r\to 0} f_n^{(r)}(\gamma), \qquad f_n^{(r)}(\gamma) := -\frac{1}{nr}\log\mathbb{E}_{\mathcal{D}}\Big[Z_n^r(\gamma|\mathcal{D})\Big] \ . \tag{6.99}$$

We refer to $f_n^{(r)}(\gamma)$ as the replicated free energy. We now compute this quantity.

The replicated free energy

For integer r we then have

$$Z_n^r(\gamma|\mathcal{D}) = \int \mathrm{e}^{-\gamma \sum_{\alpha=1}^r \mathcal{H}_n(\tilde{\boldsymbol{\beta}}_\alpha|\mathcal{D})} \prod_{\alpha=1}^r \mathrm{d}\tilde{\boldsymbol{\beta}}_\alpha \ . \tag{6.100}$$

Taking the expectation with respect to the data-set

$$\mathbb{E}_{\mathcal{D}}\Big[Z_n^r(\gamma|\mathcal{D})\Big] = \int \Big(\mathbb{E}_{\tilde{\mathbf{X}},T}\Big[\mathrm{e}^{-\gamma \sum_{\alpha=1}^r g(\tilde{\mathbf{X}}'\tilde{\boldsymbol{\beta}}_\alpha,T)}\Big]\Big)^n \prod_{\alpha=1}^r \mathrm{d}\tilde{\boldsymbol{\beta}}_\alpha \ . \tag{6.101}$$

We notice that the expression above depends on $\tilde{\boldsymbol{\beta}}_\alpha$ only via the linear predictors $\tilde{\mathbf{X}}'\tilde{\boldsymbol{\beta}}_\alpha$. Since $\tilde{\mathbf{X}} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p/p)$, we have that

$$\mathbf{Y} = (Y_0, Y_1, \ldots, Y_r) \sim \mathcal{N}\big(\mathbf{0}, \mathbf{C}(\{\tilde{\boldsymbol{\beta}}_\alpha\})\big), \quad Y_\alpha := \tilde{\mathbf{X}}'\tilde{\boldsymbol{\beta}}_\alpha \tag{6.102}$$

with

$$\mathbf{C}(\{\tilde{\boldsymbol{\beta}}_\alpha\})) = \begin{pmatrix} \theta_0^2 & \mathbf{M}' \\ \mathbf{M} & \mathbf{R} \end{pmatrix} \tag{6.103}$$

where

$$\begin{aligned} \theta_0^2 &= \|\tilde{\boldsymbol{\beta}}_0\|^2/p & (6.104)\\ \mathbf{M} &= (M_\alpha)_{\alpha=1}^r, \quad M_\alpha := \tilde{\boldsymbol{\beta}}_0'\tilde{\boldsymbol{\beta}}_\alpha/p & (6.105)\\ \mathbf{R} &= (R_{\alpha,\rho})_{\alpha,\rho=1}^r, \quad R_{\alpha,\rho} := \tilde{\boldsymbol{\beta}}_\alpha'\tilde{\boldsymbol{\beta}}_\rho/p = R_{\rho,\alpha} \ . & (6.106) \end{aligned}$$

It is convenient to introduce

$$\varphi\big(\gamma, \mathbf{C}\big) = \log \mathbb{E}_{T,\mathbf{Y}}^C\Big[\mathrm{e}^{-\gamma \sum_{\alpha=1}^r g(Y_\alpha,T)}\Big] \tag{6.107}$$

$$\phi(\gamma, \hat{\mathbf{C}}) = \frac{1}{p}\log \int \mathrm{e}^{-ip\mathrm{Tr}(\hat{\mathbf{C}}\mathbf{C}(\{\tilde{\boldsymbol{\beta}}_\alpha\}))} \prod_{\alpha=1}^r \mathrm{d}\tilde{\boldsymbol{\beta}}_\alpha \ . \tag{6.108}$$

Putting everything together, we have obtained a so-called "saddle point" integral

$$\mathbb{E}_{\mathcal{D}}\Big[Z_n^r(\gamma, \mathcal{D})\Big] = \int \mathrm{e}^{-n\psi\left(\gamma, \hat{\mathbf{C}}, \mathbf{C}\right)} \mathrm{d}\mathbf{C}\mathrm{d}\hat{\mathbf{C}} \tag{6.109}$$

where

$$-\psi\big(\gamma, \hat{\mathbf{C}}, \mathbf{C}\big) = \varphi\big(\gamma, \mathbf{C}\big) + i\zeta\frac{1}{r}\mathrm{Tr}(\hat{\mathbf{C}}\mathbf{C}) + \zeta\phi(\gamma, \hat{\mathbf{C}}) \ .$$

Replica symmetric ansatze

The idea is now to evaluate the integral via the saddle point method by interchanging the limits $n \to \infty$ and $r \to 0$, as customary in these calculations [13]. With a modest amount of foresight we take the following change of variables

$$i\hat{\mathbf{C}} = \frac{1}{2}\mathbf{D} \tag{6.110}$$

which is expected from previous similar calculations and aids the book-keeping. In principle we could now derive the saddle point equations for the elements of the matrices $\mathbf{C}$ nor $\mathbf{D}$ and then take the limit $r \to 0$. In practice we will assume the replica symmetric ansatze

$$\mathbf{C} = \begin{pmatrix} \theta_0^2 & m & \dots & \dots & m \\ m & \rho & q & \dots & q \\ \vdots & q & \rho & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & q \\ m & q & \dots & q & \rho \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} 0 & \hat{m} & \dots & \dots & \hat{m} \\ \hat{m} & \hat{\rho} & -\hat{q} & \dots & -\hat{q} \\ \vdots & -\hat{q} & \hat{\rho} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\hat{q} \\ \hat{m} & -\hat{q} & \dots & -\hat{q} & \hat{\rho} \end{pmatrix} \tag{6.111}$$

directly from now on. Note that this directly implies that

$$\mathbf{C}^{-1} = \begin{pmatrix} \tilde{\mu}_r & \tilde{m}_r & \dots & \dots & \tilde{m}_r \\ \tilde{m}_r & \tilde{\rho}_r & \tilde{q}_r & \dots & \tilde{q}_r \\ \vdots & \tilde{q}_r & \tilde{\rho}_r & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \tilde{q}_r \\ \tilde{m}_r & \tilde{q}_r & \dots & \tilde{q}v & \tilde{\rho}_r \end{pmatrix} . \tag{6.112}$$

After some algebra one obtains that

$$\tilde{\mu}_r = \frac{\rho + q(r-1)}{\theta_0^2(\rho + q(r-1)) - rm^2} \tag{6.113}$$

$$\tilde{m}_r = \frac{m}{rm^2 - \theta_0^2(\rho + q(r-1))} \tag{6.114}$$

$$\tilde{q}_r = \frac{1}{\rho - q} \frac{m^2 - q\theta_0^2}{\theta_0^2(\rho + q(r-1)) - rm^2} \tag{6.115}$$

$$\tilde{\rho}_r = \frac{1}{\rho - q}\Big(1 + \frac{m^2 - q\theta_0^2}{\theta_0^2(\rho + q(r-1)) - rm^2}\Big) . \tag{6.116}$$

Simplification of ϕ within RS ansatz

Let's remind the definition of the potential ϕ

$$\phi(\gamma, \hat{\mathbf{C}}) = \frac{1}{p} \log \int \mathrm{e}^{-\frac{1}{2} p \mathrm{Tr}(\mathbf{DC}(\{\tilde{\boldsymbol{\beta}}_\alpha\}))} \prod_{\alpha=1}^{r} \mathrm{d}\tilde{\boldsymbol{\beta}}_\alpha . \tag{6.117}$$

Using the Replica Symmetric ansatz we get

$$\begin{aligned} \frac{1}{2} p \mathrm{Tr}(\mathbf{DC}(\{\tilde{\boldsymbol{\beta}}_\alpha\})) &= \sum_{\rho=1}^{r} \big(\hat{m}\tilde{\boldsymbol{\beta}}_0'\tilde{\boldsymbol{\beta}}_\rho + \frac{1}{2}\hat{\rho}\tilde{\boldsymbol{\beta}}_\rho'\tilde{\boldsymbol{\beta}}_\rho\big) - \sum_{\rho,\alpha\neq\rho}^{r} \hat{q}\tilde{\boldsymbol{\beta}}_\rho'\tilde{\boldsymbol{\beta}}_\alpha = \\ &= \sum_{\rho=1}^{r} \big(\hat{m}\tilde{\boldsymbol{\beta}}_0'\tilde{\boldsymbol{\beta}}_\rho + \frac{1}{2}(\hat{\rho} + \hat{q})\|\tilde{\boldsymbol{\beta}}_\rho\|^2\big) - \frac{1}{2}\hat{q}\Big\|\sum_{\rho=1}^{r} \tilde{\boldsymbol{\beta}}_\rho\Big\|^2 \end{aligned}$$

Then via Gaussian linearization, we get

$$\begin{aligned}\phi^{(r)}_{RS}(\gamma,\hat{m},\hat{q},\hat{\rho}) &= \\ &= \frac{1}{p}\log \mathbb{E}_{\mathbf{Z}}\Big[\Big(\int \mathrm{e}^{-\left\{\frac{1}{2}(\hat{\rho}+\hat{q})\|\mathbf{x}\|^2-(\hat{m}\tilde{\boldsymbol{\beta}}_0+\sqrt{\hat{q}/p}\mathbf{Z})'\mathbf{x}\right\}}\mathrm{d}\mathbf{x}\Big)^r\Big] \\ &= \frac{1}{p}\sum_{\mu=1}^{p}\log \mathbb{E}_Z\Big[\Big(\int \mathrm{e}^{-\left\{\frac{1}{2}(\hat{\rho}+\hat{q})y^2-(\hat{m}\mathbf{e}'_\mu\tilde{\boldsymbol{\beta}}_0+\sqrt{\hat{q}/p}Z)y\right\}}\mathrm{d}y\Big)^r\Big]+\text{const} = \\ &= r\frac{\hat{m}^2\theta_0^2+\hat{q}}{\hat{\rho}+\hat{q}}+\frac{r}{2}\log(\hat{\rho}+\hat{q})\ . \end{aligned} \tag{6.118}$$

Simplification of φ within RS ansatz

Inserting the replica symmetric ansatze, we obtain

$$\begin{aligned}&f(y_0,y_1,\dots,y_r)\propto \\ &\quad\exp\Big\{-\Big(\frac{1}{2}\tilde{\mu}y_0^2+\sum_{\rho=1}^{r}\big(\tilde{m}y_0y_\rho+\frac{1}{2}(\tilde{\rho}-\tilde{q})y_\rho^2\big)+\frac{1}{2}\tilde{q}\big(\sum_{\rho=1}^{r}y_\rho\big)^2\Big)\Big\}\end{aligned} \tag{6.119}$$

and via Gaussian linearization

$$\begin{aligned}&f(y_0,y_1,\dots,y_r)\propto \\ &\quad\mathbb{E}_Q\Big[\exp\Big\{-\frac{1}{2}\tilde{\mu}y_0^2-\sum_{\rho=1}^{r}\big[\frac{1}{2}(\tilde{\rho}-\tilde{q})y_\rho^2+\big(\tilde{m}y_0+i\sqrt{\tilde{q}}Q\big)y_\rho\big]\Big\}\Big]\ ,\end{aligned} \tag{6.120}$$

with $Q\sim\mathcal{N}(0,1)$. Upon setting $\sqrt{\tilde{\mu}}y_0=z_0$, we obtain

$$\begin{aligned}&\varphi^{(r)}_{RS}(\gamma,\tilde{\mu}_r,\tilde{m}_r,\tilde{q}_r,\tilde{\rho}_r)= \\ &\qquad\log\frac{\mathbb{E}_{T,Z_0,Q}\Big[\Big(\int \mathrm{e}^{-\frac{1}{2}(\tilde{\rho}-\tilde{q})x^2-\left(\tilde{m}/\sqrt{\tilde{\mu}}Z_0+i\sqrt{\tilde{q}}Q\right)x-\gamma g(x,T)}\frac{\mathrm{d}x}{\sqrt{2\pi}}\Big)^r\Big]}{\mathbb{E}_{T,Z_0,Q}\Big[\Big(\int \mathrm{e}^{-\frac{1}{2}(\tilde{\rho}-\tilde{q})x^2-\left(\tilde{m}/\sqrt{\tilde{\mu}}Z_0+i\sqrt{\tilde{q}}Q\right)x}\Big)^r\Big]}\end{aligned}$$

with $Z_0\sim\mathcal{N}(0,1)$, $Z_0\perp Q$.

Replica Symmetric free energy

Let us stop and recall what we have achieved so far. By assuming the RS ansatz we have obtained

$$\gamma f(\gamma)=-\lim_{n\to\infty}\frac{1}{n}\log\mathbb{E}_{\mathcal{D}}\Big[Z_n^r(\gamma,\mathcal{D})\Big]=\operatorname*{extr}_{m,q,\rho,\hat{\mu},\hat{m},\hat{q},\hat{\rho}}\tilde{\psi}^{(r)}_{RS}(m,q,\rho,\hat{m},\hat{q},\hat{\rho}) \tag{6.121}$$

with

$$\begin{aligned}-\psi^{(r)}_{RS}(\dots)=r\zeta\big(m\hat{m}+(\rho\hat{\rho}-q\hat{q})+rq\hat{q}\big)+ \\ \zeta\phi^{(r)}_{RS}(\gamma,\hat{m},\hat{q},\hat{\rho})+\varphi^{(r)}_{RS}(\gamma,\tilde{\mu}_r,\tilde{m}_r,\tilde{q}_r,\tilde{\rho}_r)\ .\end{aligned} \tag{6.122}$$

The limit $r \to 0$

Taking the limit $r \to 0$, we get the replica symmetric free energy

$$f_{RS}(\gamma) = \underset{m,q,\rho,\hat{m},\hat{q},\hat{\rho}}{\text{extr}} \tilde{f}_{RS}(\gamma, \mu, m, q, \rho, \hat{m}, \hat{q}, \hat{\rho}) \tag{6.123}$$

with

$$\begin{aligned}\tilde{f}_{RS}(\gamma, m, q, \rho, \hat{m}, \hat{q}, \hat{\rho}) := \lim_{r\to 0} \frac{1}{\gamma r} \psi^{(r)}_{RS}(\gamma, m, q, \rho, \hat{m}, \hat{q}, \hat{\rho}) \\ \zeta(m\hat{m} + (\rho\hat{\rho} - q\hat{q})\big) + \zeta\tilde{\phi}_{RS}(\gamma, \hat{m}, \hat{q}, \hat{\rho}) + \tilde{\varphi}_{RS}(\gamma, m, q, \rho) \ .\end{aligned}$$

Above, we introduced the following definitions

$$\begin{aligned}\tilde{\phi}_{RS}(\gamma, \hat{m}, \hat{q}, \hat{\rho}) &= \lim_{r\to 0} \frac{1}{\gamma r} \phi^{(r)}_{RS}(\gamma, \hat{m}, \hat{q}, \hat{\rho}) \\ &= \frac{1}{2\gamma} \frac{\hat{m}^2\theta_0^2 + \hat{q}}{(\hat{\rho} + \hat{q})} + \frac{1}{2\gamma} \log(\hat{\rho} + \hat{q}) \\ \tilde{\varphi}_{RS}(\gamma, m, q, \rho) &= \lim_{r\to 0} \frac{1}{\gamma r} \varphi^{(r)}_{RS}(\gamma, \tilde{\mu}_r, \tilde{m}_r, \tilde{q}_r, \tilde{\rho}_r) = \\ &= \mathbb{E}_{T,Z_0,Q}\left[\log \frac{\int \mathrm{e}^{-\frac{1}{\rho-q}\left\{\frac{1}{2}x^2 - \left((m/\theta_0)Z_0 + \sqrt{q-(m/\theta_0)^2 Q}\right)x\right\} - \gamma g(x,T)} \mathrm{d}x}{\int \mathrm{e}^{-\frac{1}{\rho-q}\left\{\frac{1}{2}x^2 - \left((m/\theta_0)Z_0 + \sqrt{q-(m/\theta_0)^2 Q}\right)x\right\}} \mathrm{d}x} \right]\end{aligned}$$

and we have used that

$$\tilde{\mu} = \lim_{r\to 0} \tilde{\mu}_r = 1/\theta_0^2 \tag{6.124}$$

$$\tilde{m} = \lim_{r\to 0} \tilde{m}_r = -\frac{m}{\theta_0^2(\rho - q)} \tag{6.125}$$

$$\tilde{q} = \lim_{r\to 0} \tilde{q}_r = \frac{1}{\rho - q} \frac{m^2 - q\theta_0^2}{\theta_0^2(\rho - q)} \tag{6.126}$$

$$\tilde{\rho} - \tilde{q} = \lim_{r\to 0} \tilde{\rho}_r - \tilde{q}_r = \frac{1}{\rho - q} \ . \tag{6.127}$$

Assuming the following scaling

$$\tau = \gamma/(\tilde{\rho} - \tilde{q}) = \gamma(\rho - q) \tag{6.128}$$

we obtain

$$\begin{aligned}&\tilde{\varphi}_{RS}(\gamma, m, q, \rho) = \\ &\mathbb{E}_{T,Z_0,Q}\left[\frac{1}{\gamma} \log \frac{\int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\left(x - ((m/\theta_0)Z_0 + \sqrt{q-(m/\theta_0)^2 Q})\right)^2/\tau + g\left(x,T\right)\right\}} \mathrm{d}x}{\int \mathrm{e}^{-\gamma\left\{\frac{1}{2}\left(x - ((m/\theta_0)Z_0 + \sqrt{q-(m/\theta_0)^2 Q})\right)^2/\tau\right\}} \mathrm{d}x} \right] .\end{aligned}$$

Similarly, taking the additional rescaling

$$1/\hat{\tau} = (\hat{\rho} + \hat{q})/\gamma, \quad \hat{m} = \gamma\hat{m}, \quad \hat{q} = \gamma^2\hat{q} \ , \tag{6.129}$$

we get

$$\tilde{\phi}_{RS}(\gamma, \hat{m}, \hat{q}, \hat{\tau}) = \frac{1}{2\hat{\tau}}\left(\hat{m}^2\theta_0^2 + \hat{q}\right) + \frac{1}{2\gamma} \log(\gamma\hat{\tau}) \ . \tag{6.130}$$

The limit $\gamma \to \infty$

Taking the limit $\gamma \to \infty$ of the disorder averaged free energy, we obtain the (disorder averaged) internal energy per particle as

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{n}\mathcal{H}_n(\tilde{\boldsymbol{\beta}}|\mathcal{D})\Big] = \operatorname*{extr}_{m,q,\tau,\hat{m},\hat{q},\hat{\tau}} \mathcal{F}(m,q,\tau,\hat{m},\hat{q},\hat{\tau})$$
$$\mathcal{F}(m,q,\tau,\hat{m},\hat{q},\hat{\tau}) = \lim_{\gamma\to\infty} \tilde{f}_{RS}(\gamma,m,q,\tau,\hat{m},\hat{q},\hat{\tau}) \ . \tag{6.131}$$

Via Laplace integration we see that

$$-\lim_{\gamma\to\infty}\frac{1}{\gamma}\log\int \mathrm{e}^{-\gamma\left\{\frac{1}{2\alpha}\|\mathbf{z}-\mathbf{x}\|^2+b(\mathbf{z})\right\}}\mathrm{d}\mathbf{z} = \mathcal{M}_{b(.)}(\mathbf{x},\alpha) \tag{6.132}$$

where $\mathcal{M}_{b(.)}$ is the Moureau envelope of a convex function $b:\mathbb{R}\to\mathbb{R}$, which is defined as

$$\mathcal{M}_{b(.)}(\mathbf{x},\alpha) = \min_{\mathbf{z}}\Big\{\frac{1}{2\alpha}\|\mathbf{z}-\mathbf{x}\|^2 + b(\mathbf{z})\Big\} \ . \tag{6.133}$$

Using the "Laplace-Moureau" identity above (6.132), we obtain

$$\lim_{\gamma\to\infty}\tilde{\varphi}_{RS}(\gamma,m,q,\rho) = -\frac{1}{\tau}\mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,T)}\Big(m/\theta_0 Z_0 + \sqrt{q-(m/\theta_0)^2}Q,\tau\Big)\Big] \ .$$

Furthermore

$$\lim_{\gamma\to\infty}\tilde{\phi}_{RS}(\gamma,\hat{m},\hat{q},\hat{\rho}) = \frac{1}{2\hat{\tau}}\Big(\hat{m}^2\theta_0^2+\hat{q}\Big)$$

We have obtained the following

$$\mathcal{F}(m,q,\tau,\hat{m},\hat{q},\hat{\tau}) = -\zeta(m\hat{m}+\frac{1}{2}\hat{\tau}q-\frac{1}{2}\tau\hat{q}) - \frac{1}{2\hat{\tau}}\zeta\Big(\hat{m}^2\theta_0^2+\hat{q}\Big) +$$
$$+\mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,T)}\Big(m/\theta_0 Z_0+\sqrt{q-(m/\theta_0)^2}Q,\tau\Big)\Big] + \text{const}$$

It is now easy to see that we can easily "extremize" over the "hat" variables. Taking derivatives we get

$$\frac{\partial}{\partial\hat{m}}\mathcal{F} = \zeta\Big(m+\hat{m}\theta_0^2/\hat{\tau}\Big) = 0 \quad\to\quad \hat{m} = -m\hat{\tau}/\theta_0^2 \tag{6.134}$$
$$\frac{\partial}{\partial\hat{q}}\mathcal{F} = \frac{\zeta}{2}\Big(\tau-1/\hat{\tau}\Big) = 0 \quad\to\quad \hat{\tau} = 1/\tau \tag{6.135}$$
$$\frac{\partial}{\partial\hat{\tau}}\mathcal{F} = \frac{\zeta}{2}\Big(q-(\hat{m}^2\theta_0^2-\hat{q})/\hat{\tau}^2\Big) = 0 \quad\to\quad \hat{q} = \big(q-m^2/\theta_0^2\big)/\tau^2 \ . \tag{6.136}$$

Substituting these identities into $\mathcal{F}$ gives another function which, with abuse of notation, we keep referring to as $\mathcal{F}$

$$\mathcal{F}(m,q,\tau) = \mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,T)}\big(m/\theta_0 Z_0+\sqrt{q-(m/\theta_0)^2}Q,\tau\big)\Big] +$$
$$-\frac{\zeta}{2\tau}\big(q-m^2/\theta_0^2\big) + \text{const} \tag{6.137}$$

At this point it is convenient to take the (invertible) change of variables

$$w = m/\theta_0, \quad v = q-m^2/\theta_0^2 \tag{6.138}$$

so that, again with abuse of notation, we obtain

$$\mathcal{F}(w,v,\tau) = \mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,T)}\big(wZ_0+vQ,\tau\big)\Big] - \frac{\zeta}{2\tau}v^2 + \text{const} \ . \tag{6.139}$$

Replica Symmetric equations

In view of the next steps we remember some known fact about the Moureau envelop that will come in handy to compute the derivatives

$$\frac{\partial}{\partial x}\mathcal{M}_{g(.,T)}(x,\tau) = \frac{1}{\tau}\big(x - \text{prox}_{\tau g(.,T)}(x)\big) = \dot{g}(\text{prox}_{\tau g(.,T)}(x), T) \tag{6.140}$$

$$\frac{\partial}{\partial \tau}\mathcal{M}_{g(.,T)}(x,\tau) = -\frac{1}{2\tau^2}\big(\text{prox}_{\tau g(.,T)}(x) - x\big)^2 = -\frac{1}{2\tau^2}\big\{\dot{g}(\text{prox}_{\tau g(.,T)}(x), T)\big\}^2 \ . \tag{6.141}$$

Taking derivatives with respect to v, we get

$$\begin{aligned}\frac{\partial}{\partial v}\mathcal{F} &= -\zeta v/\tau + \mathbb{E}_{T,Z_0,Q}\Big[Q\Big(wZ_0 + vQ - \text{prox}_{g(.,T)}(wZ_0 + vQ, \tau)\Big)\Big]/\tau \\ &= (1-\zeta)v/\tau - \mathbb{E}_{T,Z_0,Q}\Big[\text{prox}'_{g(.,T)}(wZ_0 + vQ, \tau)\Big]v/\tau = 0\end{aligned} \tag{6.142}$$

so

$$1 - \zeta = \mathbb{E}_{T,Z_0,Q}\Big[\text{prox}'_{\tau g(.,T)}(wZ_0 + vQ)\Big] \ . \tag{6.143}$$

Taking derivatives with respect to w, we get

$$\frac{\partial}{\partial w}\mathcal{F} = \theta_0 \mathbb{E}_{T,Z_0,Q}\Big[Z_0\Big(wZ_0 + vQ - \text{prox}_{g(.,T)}(wZ_0 + vQ, \tau)\Big)\Big]/\tau = 0 \tag{6.144}$$

which implies

$$w = \mathbb{E}_{T,Z_0,Q}\Big[Z_0 \text{prox}_{g(.,T)}(wZ_0 + vQ, \tau)\Big] \ . \tag{6.145}$$

Finally derivation with respect to τ returns

$$\frac{\partial}{\partial \tau}\mathcal{F} = \frac{1}{2}\zeta v^2/\tau^2 - \mathbb{E}_{T,Z_0,Q}\Big[\Big(\text{prox}_{\tau g(.,T)}(k\theta_0 Z_0 + vQ) - wZ_0 + vQ\Big)^2\Big] = 0 \tag{6.146}$$

which gives

$$\zeta v^2 = \mathbb{E}_{T,Z_0,Q}\Big[\Big(\text{prox}_{\tau g(.,T)}(k\theta_0 Z_0 + vQ) - wZ_0 + vQ\Big)^2\Big] \ . \tag{6.147}$$

To sum up, introducing the shorthand

$$\xi = \text{prox}_{\tau g(.,T)}\big(wZ_0 + vQ\big) \tag{6.148}$$

which satisfies by definition

$$\xi = wZ_0 + vQ - \tau \dot{g}(\xi, T) \ , \tag{6.149}$$

we have obtained the following RS equations

$$\zeta v^2 = \mathbb{E}_{T,Z_0,Q}\Big[\Big(\dot{g}(\xi,T)\Big)^2\Big] \tag{6.150}$$

$$1 - \zeta = \mathbb{E}_{T,Z_0,Q}\Big[\frac{1}{1 + \tau \ddot{g}(\xi,T)}\Big] \tag{6.151}$$

$$w = \mathbb{E}_{T,Z_0,Q}\Big[Z_0 \xi\Big] \ . \tag{6.152}$$

The last two equations can be simplified further to obtain the equations that appear in the main text

$$\zeta v^2 = \mathbb{E}_{T,Z_0,Q}\Big[\Big(\dot{g}(\xi,T)\Big)^2\Big] \tag{6.153}$$

$$\zeta = \mathbb{E}_{T,Z_0,Q}\Big[\frac{\tau\ddot{g}(\xi,T)}{1+\tau\ddot{g}(\xi,T)}\Big] \tag{6.154}$$

$$\zeta w = \theta_0\mathbb{E}_{T,Z_0,Q}\Big[\dot{g}(\xi,T)\dot{g}_0(\theta_0 Z_0,T)\Big] \ . \tag{6.155}$$

Notice that the Replica Method attributes the following interpretation to the scalars $w_\star, v_\star$ that solve the above equations:

$$w_\star \approx \frac{1}{\sqrt{p}}\frac{\tilde{\boldsymbol{\beta}}_0'\hat{\boldsymbol{\beta}}_n(\tilde{\mathbf{X}},\tilde{\mathbf{T}})}{\|\tilde{\boldsymbol{\beta}}_0\|_2} \stackrel{d}{=} \frac{1}{\sqrt{p}}\frac{\boldsymbol{\beta}_0'\boldsymbol{\Sigma}_0\hat{\boldsymbol{\beta}}_n(\mathbf{X},\mathbf{T})}{\|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2} \tag{6.156}$$

$$v_\star^2 \approx \frac{1}{p}\Big(\|\hat{\boldsymbol{\beta}}_n(\tilde{\mathbf{X}},\tilde{\mathbf{T}})\|^2 - \frac{\{\tilde{\boldsymbol{\beta}}_0'\hat{\boldsymbol{\beta}}_n(\tilde{\mathbf{X}},\tilde{\mathbf{T}})\}^2}{\|\tilde{\boldsymbol{\beta}}_0\|_2^2}\Big) \tag{6.157}$$

$$\stackrel{d}{=} \|\hat{\boldsymbol{\beta}}_n(\mathbf{X},\mathbf{T})\|_2^2 - \frac{(\boldsymbol{\beta}_0'\boldsymbol{\Sigma}_0\hat{\boldsymbol{\beta}}_n(\mathbf{X},\mathbf{T}))^2}{\|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2^2} \ . \tag{6.158}$$

This can be made fully rigorous in for GLMs, via the Convex Gaussian Min Max Theorem [17, 18, 16]. We briefly explain how in the next section, referring to the relevant articles for the technical details.

6.11.2 Proof sketch of the replica results for GLMs via the CGMT

Here we consider uncorrelated covariates for convenience, i.e. when $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{I}_p/p)$, the result can be mapped back to the correlated case as done for the Replica Calculation. Introducing Lagrange multipliers $\boldsymbol{\phi}$, one can equivalently re-write the minimization problem as a saddle point problem as follows

$$\Psi_n(\mathbf{X},\mathbf{T}) = \min_{\boldsymbol{\beta},\boldsymbol{\xi}}\sup_{\boldsymbol{\phi}}\Big\{\big\langle g(\boldsymbol{\xi},\mathbf{T})\big\rangle - \boldsymbol{\phi}'\big(\boldsymbol{\xi}-\mathbf{X}\boldsymbol{\beta}\big)\Big\} \ , \tag{6.159}$$

where

$$\big\langle g(\boldsymbol{\xi},\mathbf{T})\big\rangle := \frac{1}{n}\sum_{i=1}^n g(\xi_i,T_i) \ , \tag{6.160}$$

for notational convenience. Notice that $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\mathbf{P}_{\boldsymbol{\beta}_0}\boldsymbol{\beta} + \mathbf{X}\mathbf{P}_{\perp\boldsymbol{\beta}_0}\boldsymbol{\beta} \stackrel{d}{=} \mathbf{Z}_0\beta_\parallel + \tilde{\mathbf{X}}\boldsymbol{\beta}_\perp$, where $\stackrel{d}{=}$ indicates equality in distribution with $\beta_\parallel := \frac{\boldsymbol{\beta}_0'\boldsymbol{\beta}}{\|\boldsymbol{\beta}_0\|}$, $\boldsymbol{\beta}_\perp := \mathbf{P}_{\perp\boldsymbol{\beta}_0}\boldsymbol{\beta} = (\boldsymbol{I} - \boldsymbol{\beta}_0\boldsymbol{\beta}_0'/\|\boldsymbol{\beta}_0\|^2)\boldsymbol{\beta}$, $\mathbf{Z}_0 := \mathbf{X}\boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\| \sim \mathcal{N}(\mathbf{0},\boldsymbol{I}_n)$ and $\mathbf{X}_\perp := \mathbf{X}\mathbf{P}_{\boldsymbol{\beta}_0}$ $(\mathbf{X}_\perp)_{i,j} \sim \mathcal{N}(0,1)$ and $\mathbf{X}_\perp \perp \mathbf{T}$. Using this fact, we have reduced the original problem into the form

$$\Psi_n(\mathbf{X},\mathbf{T}) \stackrel{d}{=} \min_{\boldsymbol{\beta},\boldsymbol{\xi}}\sup_{\boldsymbol{\phi}}\Big\{\big\langle g(\boldsymbol{\xi},\mathbf{T})\big\rangle - \boldsymbol{\phi}'\big(\boldsymbol{\xi}-\mathbf{Z}_0\beta_\parallel - \mathbf{X}_\perp\boldsymbol{\beta}_\perp\big)\Big\} \tag{6.161}$$

that can be attacked with the Convex Gaussian Min-Max theorem (CGMT), first introduced in [17] as a generalization of Gordon's Gaussian comparison inequalities [27]. The CGMT is reported without proof below for the reader's convenience. We refer to [17, 18] for a detailed proof.

Theorem 4 (Convex Gaussian Min Max Theorem)**.** *Let $\mathcal{S}_{\mathbf{y}} \subset \mathbb{R}^n$, $\mathcal{S}_{\mathbf{z}} \subset \mathbb{R}^p$ be compact and convex sets, ψ be continuous and convex-concave on $\mathcal{S}_{\mathbf{z}} \times \mathcal{S}_{\mathbf{y}}$, and $\mathbf{X} \in \mathbb{R}^{n\times p}, \mathbf{Q} \in \mathbb{R}^n, \mathbf{G} \in \mathbb{R}^p$ all have entries iid standard normal*

$$\begin{aligned}\Phi(\mathbf{X}) &:= \min_{\mathbf{y}\in\mathcal{S}_{\mathbf{y}}} \max_{\mathbf{z}\in\mathcal{S}_{\mathbf{z}}} \Big\{ \mathbf{y}'\mathbf{X}\mathbf{z} + \psi(\mathbf{z},\mathbf{y}) \Big\} \\ \phi(\mathbf{G},\mathbf{Q}) &:= \min_{\mathbf{y}\in\mathcal{S}_{\mathbf{y}}} \max_{\mathbf{z}\in\mathcal{S}_{\mathbf{z}}} \Big\{ \|\mathbf{y}\|\mathbf{G}'\mathbf{z} + \|\mathbf{z}\|\mathbf{Q}'\mathbf{y} + \psi(\mathbf{z},\mathbf{y}) \Big\}\end{aligned}$$

Then

$$\forall \mu \in \mathbb{R}, t \in \mathbb{R}^+,\ P\Big[\Big|\Phi(\mathbf{X}) - \mu\Big| > t\Big] \leq 2P\Big[\Big|\Phi(\mathbf{G},\mathbf{Q}) - \mu\Big| \geq t\Big]\ . \tag{6.162}$$

Furthermore, let $\mathcal{S} \subset \mathcal{S}_{\mathbf{y}}$ an open subset and $\mathcal{S}^c := \mathcal{S}_{\mathbf{y}} \setminus \mathcal{S}$. Denote $\phi_{\mathcal{S}^c}(\mathbf{G},\mathbf{Q})$ the optimal cost of the surrogate process, when the minimization is now constrained over $\mathcal{S}^c$. If there exist constants $\bar{\phi} < \bar{\phi}_{\mathcal{S}^c}$, such that $\phi(\mathbf{G},\mathbf{Q}) \xrightarrow{P} \bar{\phi}$, $\phi_{\mathcal{S}^c}(\mathbf{G},\mathbf{Q}) \xrightarrow{P} \bar{\phi}_{\mathcal{S}^c}$, then, denoting with $\hat{y}$ the value of y at the saddle point, we have

$$\lim_{n\to\infty} P\Big[\hat{\mathbf{y}} \in \mathcal{S}\Big] = 1\ . \tag{6.163}$$

Theorem (4) above tells us that if one is able to prove that ϕ converges in probability to some deterministic value, then the same is true for Φ. The CGMT applies to min-max problems over compact, convex sets. Hence we "artificially" restrict the saddle point problem (6.161) onto compact, convex sets. Intuition suggests that if a saddle point exists and the set is sufficiently large, then there is not going to be any difference between the bounded and unbounded problem. Hence from now on the min over $\boldsymbol{\beta}, \boldsymbol{\xi}$ and the max over $\boldsymbol{\phi}$ operations are understood over convex, compact sets (so that they actually exist). Following Theorem (4), we consider an auxiliary optimization problem

$$\begin{aligned}&\tilde{\Psi}_n(\mathbf{Q},\mathbf{G},\mathbf{T}) = \\ &\min_{\boldsymbol{\beta},\boldsymbol{\xi}} \max_{\boldsymbol{\phi}} \Big\{ \big\langle g(\boldsymbol{\xi},\mathbf{T}) \big\rangle - \boldsymbol{\phi}'\big(\boldsymbol{\xi} - \beta_{\|}\mathbf{Z}_0 - \|\boldsymbol{\beta}_{\perp}\|\mathbf{Q}\big) + \boldsymbol{\beta}_{\perp}'\mathbf{G}\|\boldsymbol{\phi}\| \Big\}\end{aligned} \tag{6.164}$$

with $\mathbf{G} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{(p-1)})$ and $\mathbf{Q} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$, $\mathbf{G} \perp \mathbf{Q}$ and $\mathbf{G}, \mathbf{Q} \perp \mathbf{T}, \mathbf{Z}_0$. We can optimize over the direction of $\boldsymbol{\phi}$ at fixed length $\|\boldsymbol{\phi}\| = \phi$

$$\begin{aligned}&\tilde{\Psi}_n(\mathbf{Q},\mathbf{G},\mathbf{T}) = \\ &\min_{\boldsymbol{\beta},\boldsymbol{\xi}} \max_{\phi\geq 0} \Big\{ \big\langle g(\boldsymbol{\xi},\mathbf{T}) \big\rangle + \phi\big\|\boldsymbol{\xi} - \beta_{\|}\mathbf{Z}_0 - \|\boldsymbol{\beta}_{\perp}\|\mathbf{Q}\big\| + \boldsymbol{\beta}_{\perp}'\mathbf{G}\phi \Big\}\ .\end{aligned} \tag{6.165}$$

The problem above depends on $\boldsymbol{\beta}_{\perp}$ via $\|\boldsymbol{\beta}_{\perp}\|$ and $\cos\theta$, with θ the angle between $\boldsymbol{\beta}_{\perp}$ and $\mathbf{G}$. Hence it is not guaranteed to be convex-concave (as *cos* is not convex on the whole interval $[0, 2\pi]$), but it has been shown in [18] (page 22 point 6) that (6.165) can be used in place of (6.164) in the CGMT as $n \to \infty$, i.e. the min-max order can be "swapped" asymptotically. At this points, the minimization over the direction of $\boldsymbol{\beta}_{\perp}$ at fixed $\|\boldsymbol{\beta}_{\perp}\| = v$ yields

$$\tilde{\Psi}_n(\mathbf{Q},\mathbf{G},\mathbf{T}) = \tag{6.166}$$

$$\min_{w,v} \max_{\phi} \min_{\boldsymbol{\xi}} \Big\{ \big\langle g(\boldsymbol{\xi},\mathbf{T}) \big\rangle + \phi\big\|\boldsymbol{\xi} - w\mathbf{Z}_0 - v\mathbf{Q}\big\| - v\phi\|\mathbf{G}\| \Big\} \tag{6.167}$$

where we also defined $w := \boldsymbol{\beta}_{\|}$. Following [18], we use the variational representation $\|\mathbf{y}\| = \inf_{\alpha\geq 0}\Big\{\frac{1}{2}\Big(\alpha\|\mathbf{y}\|^2 + \frac{1}{\alpha}\Big)\Big\}$ for the norm. Then

$$\tilde{\Psi}_n(\mathbf{Q},\mathbf{G},\mathbf{T}) \ = \ \min_{w,v} \max_{\phi} \inf_{\boldsymbol{\xi},\alpha} \Big\{ \big\langle g(\boldsymbol{\xi},\mathbf{T}) + \frac{1}{2\alpha}\big\|\boldsymbol{\xi} - w\mathbf{Z}_0 - v\mathbf{Q}\big\|^2 \Big\}\ . \tag{6.168}$$

Taking the re-scaling $\alpha \to \sqrt{n}\alpha,\ \phi \to \phi/\sqrt{n}$ we recognize the (averaged) Moreau envelope (as defined in the main previous section)

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{M}_{g(.,T_i)}\big(wZ_{0,i} + vQ_i, \alpha/\phi\big) = \\ \frac{1}{n}\min_{\boldsymbol{\xi}}\Big\{ \sum_{i=1}^{n} g(\boldsymbol{\xi}_i, T_i) + \frac{\phi}{2\alpha}\sum_{i=1}^{n}\big(\xi - wZ_{0,i} - vQ_i\big)^2 \Big\} \ . \tag{6.169}$$

Hence, we are finally left with the saddle point problem

$$\tilde{\Psi}_n(\mathbf{Q}, \mathbf{G}, \mathbf{T}) = \min_{w,v}\max_{\phi}\inf_{\alpha} \mathcal{L}_n(w, v, \phi, \alpha) \tag{6.170}$$

where

$$\mathcal{L}_n(w, v, \phi, \alpha) = \\ \frac{1}{n}\sum_{i=1}^{n} \mathcal{M}_{g(.,T_i)}\big(wZ_{0,i} + vQ_i, \alpha/\phi\big) + \phi\big(\alpha/2 - v\|\mathbf{G}\|/\sqrt{n}\big) \ . \tag{6.171}$$

Since $\|\mathbf{G}\|$ follows a chi distribution with $p-1$ degrees of freedom

$$\mathbb{E}\Big[\|\mathbf{G}\|\Big] = \sqrt{2}\frac{\Gamma(\frac{p}{2})}{\Gamma(\frac{p-1}{2})} = \sqrt{p-2} + o_p(1), \tag{6.172}$$

$$\mathbb{V}\Big[\|\mathbf{G}\|^2\Big] = p - 1 - \mathbb{E}\Big[\|\mathbf{G}\|\Big]^2 = 1 + o_p(1), \tag{6.173}$$

hence we conclude $\|\mathbf{G}\|/\sqrt{n} \xrightarrow{P} \sqrt{\zeta}$ (e.g. via Chebishev inequality). Furthermore, provided

$$\mathbb{V}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,T)}\big(wZ_0 + vQ, \alpha/\phi\big)\Big] < \infty \ , \tag{6.174}$$

we have that

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{M}_{g(.,T_i)}\big(wZ_{0,i} + vQ_i, \alpha/\phi\big) \xrightarrow{P} \tag{6.175}$$

$$\mathbb{E}\Big[\mathcal{M}_{g(.,T)}\big(wZ_0 + vQ, \alpha/\phi\big)\Big] \ , \tag{6.176}$$

by the weak law of large numbers . Hence

$$\mathcal{L}_n(w, v, \phi, \alpha) \xrightarrow{P} \mathcal{L}(w, v, \phi, \alpha) \tag{6.177}$$

pointwise for $0 \le w, v \le C_{\boldsymbol{\beta}}, \phi \ge 0$ and $\alpha > 0$, where

$$\mathcal{L}(w, v, \phi, \alpha) := \mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,T)}\big(wZ_0 + vQ, \alpha/\phi\big)\Big] + \phi\big(\alpha/2 - v\sqrt{\zeta}\big) \ . \tag{6.178}$$

The pointwise convergence in probability above, together with the fact that the functions $\mathcal{L}_n$ is concave in ϕ and convex in w, v, ξ, α imply, see [18, 16], that

$$\min_{w,v}\max_{\phi>0}\inf_{\alpha>0} \mathcal{L}_n(w, v, \phi, \alpha) \xrightarrow{P} \min_{w,v}\max_{\phi>0}\inf_{\alpha>0} \mathcal{L}(w, v, \phi, \alpha) \ . \tag{6.179}$$

By the CGMT, we then have that

$$\min_{\boldsymbol{\beta}}\Big\{\big\langle g(\mathbf{X}\boldsymbol{\beta},\mathbf{T})\big\rangle\Big\} \xrightarrow{P} \min_{w,v}\max_{\phi>0}\min_{\alpha>0}\mathcal{L}(w,v,\phi,\alpha)\ . \tag{6.180}$$

The convergence above can be translated into the statements

$$\hat{w}_n := \frac{\boldsymbol{\beta}_0'\hat{\boldsymbol{\beta}}_n}{\|\boldsymbol{\beta}_0\|} \xrightarrow{P} w_\star \tag{6.181}$$

$$\hat{v}_n := \big\|\mathbf{P}_{\perp\boldsymbol{\beta}_0}\hat{\boldsymbol{\beta}}_n\big\| = \Big\|\Big(\boldsymbol{I}-\frac{\boldsymbol{\beta}_0\boldsymbol{\beta}_0'}{\|\boldsymbol{\beta}_0\|^2}\Big)\hat{\boldsymbol{\beta}}_n\Big\| \xrightarrow{P} v_\star\ . \tag{6.182}$$

via Corollary 6.163.

In the next section we show the validity of the RS equations and the match with the optimality conditions of (6.178).

6.11.3 Replica symmetric equations

The optimality conditions can be obtained by differentiation of $\mathcal{L}$.

Proposition 13 (Derivatives of the Expected Moreau envelope)**.** *If*

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial\alpha}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi)\Big|\Big] < \infty \tag{6.183}$$

then, defining

$$\xi := \mathrm{prox}_{g(.,T)}\big(wZ_0+vQ,\alpha/\phi\big)\ , \tag{6.184}$$

we have

$$\begin{aligned}
\frac{\partial}{\partial w}\mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi)\Big] &= \mathbb{E}_{T,Z_0,Q}\Big[\dot{g}(\xi,T)Z_0\Big]\\
\frac{\partial}{\partial v}\mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi)\Big] &= \mathbb{E}_{T,Z_0,Q}\Big[\dot{g}(\xi,T)Q\Big]\\
\frac{\partial}{\partial \alpha}\mathbb{E}_{T,Z_0,Q}\Big[\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi)\Big] &= -\frac{1}{2\phi}\mathbb{E}_{T,Z_0,Q}\Big[\big\{\dot{g}(\xi,T)\big\}^2\Big]\ .
\end{aligned}$$

Proof. Since

$$\frac{\partial}{\partial x}\mathcal{M}_{g(.,T)}(x,\tau) = \frac{1}{\tau}\big(x-\mathrm{prox}_{g(.,T)}(x,\tau)\big) = \dot{g}\big(\mathrm{prox}_{g(.,T)}(x,\tau),T\big) \tag{6.185}$$

$$\begin{aligned}
\frac{\partial}{\partial \tau}\mathcal{M}_{g(.,T)}(x,\tau) &= -\frac{1}{2\tau^2}\big(x-\mathrm{prox}_{g(.,T)}(x,\tau)\big)^2 =\\
&-\frac{1}{2}\Big\{\dot{g}\big(\mathrm{prox}_{g(.,T)}(x,\tau),T\big)\Big\}^2\ ,
\end{aligned} \tag{6.186}$$

we have

$$\frac{\partial}{\partial w}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi) = \dot{g}(\xi,T)Z_0 \tag{6.187}$$

$$\frac{\partial}{\partial v}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi) = \dot{g}(\xi,T)Q \tag{6.188}$$

$$\frac{\partial}{\partial \alpha}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi) = -\frac{1}{2\phi}\big\{\dot{g}(\xi,T)\big\}^2\ . \tag{6.189}$$

Notice that

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial w}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi)\Big|\Big] \leq \mathbb{E}\Big[\big\{\dot{g}(\xi,T)\big\}^2\Big] = \\ = 2\phi\mathbb{E}\Big[\Big|\frac{\partial}{\partial \alpha}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi)\Big|\Big] \ . \tag{6.190}$$

Hence, if

$$\mathbb{E}\Big[\Big|\frac{\partial}{\partial \alpha}\mathcal{M}_{g(.,T)}(wZ_0+vQ,\alpha/\phi)\Big|\Big] < \infty \tag{6.191}$$

all the first derivatives are integrable and via the Dominated Convergence theorem we have shown that expectation and derivatives can be interchanged. □

Taking derivatives of (6.178), and using integration by parts to simplify the expressions, we get

$$\phi^2 = \mathbb{E}_{T,Z_0,Q}\Big[\big\{\dot{g}(\xi,T)\big\}^2\Big] \tag{6.192}$$

$$w\zeta = \frac{\alpha}{\phi}\theta_0\mathbb{E}_{T,Z_0,Q}\Big[\dot{g}(\xi,T)\dot{g}_0(\theta_0,T)\Big] \tag{6.193}$$

$$\phi\sqrt{\zeta} = v\mathbb{E}_{T,Z_0,Q}\Big[\frac{\ddot{g}(\xi,T)}{1+\frac{\alpha}{\phi}\ddot{g}(\xi,T)}\Big] \tag{6.194}$$

$$\alpha = v\sqrt{\zeta} \ . \tag{6.195}$$

Taking $\alpha/\phi = \tau$, we obtain the RS equations

$$v^2\zeta = \tau^2\mathbb{E}_{T,Z_0,Q}\Big[\big\{\dot{g}(\xi,T)\big\}^2\Big] \tag{6.196}$$

$$w\zeta = \tau\theta_0\mathbb{E}_{T,Z_0,Q}\Big[\dot{g}(\xi,T)\dot{g}_0(\theta_0,T)\Big] \tag{6.197}$$

$$\zeta = \mathbb{E}_{T,Z_0,Q}\Big[\frac{\tau\ddot{g}(\xi,T)}{1+\tau\ddot{g}(\xi,T)}\Big] \tag{6.198}$$

$$\tau\phi = v\sqrt{\zeta} \ . \tag{6.199}$$

Chapter 7

Conclusion and outlook

The main aim of this thesis was to extend the theoretical framework of previous studies on the subject of Survival Analysis with high dimensional data in the proportional regime, where estimators are known to be inconsistent, and eventually generate new methodology to improve inference and/ or prediction.

Asymptotic theory for the Cox model through the lens of statistical physics The first achievement of this thesis is the extension of the asymptotic theory for the Cox regression model in the proportional regime [7, 27, 8], alias "theory of overfitting", by : i) allowing for censored observations, ii) avoiding the use of the variational approximation to solve the Replica Symmetric equations of the theory, iii) allowing for an arbitrary regularization (not only ridge as done in [27, 8]) and iv) showing that, at least in some cases, all the parameters of the theory can be measured, and the theory can be used in practice (this was not the case in [7, 27, 8] which assumed the parameters of the data generating process to be known). These points have been achieved in different stages of my studies. In chapter 3, in collaboration with the co-authors, we introduced a slightly different analytical approach that inspired an efficient strategy to solve the Replica Symmetric equations that arise from the asymptotic theory of Cox regression. Solving these equations requires the knowledge of the data generating process, which is not available in applications with real data (where we actually want to estimate the data generating model). Thus, we proposed, and numerically tested, a self-consistent algorithm that approximates the data generating process in order to solve the RS equations starting from the data. We concluded that the proposed estimator is effectively unbiased. This approach has the drawback of being very slow. An improved version of this idea can be found in chapter 6. There we showed that, when $\zeta < 1$ and the covariant regularization (see chapter 2) is used, all the RS order parameters of the theory can be estimated from the data directly, i.e. without the need of solving the RS equations of the theory. We stress that previous papers already suggested [30, 3] strategies to estimate the RS order parameters solely from the data. Our original contribution here is to propose a much simpler numerical strategy that can be used to estimate the signal strength directly from the data, by solving a single scalar equation. Chapter 5 dealt with the asymptotic theory of regularized Cox regression, with an arbitrary separable regularization. In order to carry out the replica derivation neatly, we require the covariates to be uncorrelated. This is a highly ideal, but standard assumption in several asymptotic studies in the proportional regime [26, 29, 10]. Under these hypotheses, we introduce a variant of the Approximate Message Passing algorithm adapted to the Cox regression model, that we named COX-AMP, to compute efficiently the regularized maximum partial likelihood estimator. We

notice that the update equations of COX-AMP suggest the construction of a quantity that can be used in turn to estimate all the RS order parameters directly from the data, without the need of solving the (six) RS equations of the theory (which require the knowledge of the data generating process). The knowledge of the RS order parameters, can then be used to estimate the signal strength, or estimate the generalization error.

These findings, however, leave a number of open questions. A first one is "can we rigorously justify the results obtained via the replica method for the Cox regression model?" In this direction, I studied in chapter 4 the asymptotics of the Piece-Wise Exponential model (with ridge regularization to guarantee strong convexity), which can be regarded as a flexible parametric alternative to the Cox model. Technically, the main achievement of this chapter is to show that the Convex Gaussian Min-max Theorem of [31] is applicable to this model, when the number of knots used in the parametrization of the base hazard is finite as $n \to \infty$. As a consequence, I prove that the solution of the Replica Symmetric equations correctly describe the limit in probability of the overlaps of the Replica Symmetric theory. An interesting problem for the future is to extend the present proof to incorporate also the semi-parametric Cox model, to which the present proof does not apply.
Another point of practical relevance is to obtain a generalization of the results of chapter 5 when the covariates are correlated. For the Linear Model (LM), this has been studied rigorously in [6, 17]. For Generalized Linear Models, this has been recently investigated [3]. However, when the covariates are correlated, all these results require the knowledge of the population precision matrix (the inverse of the population covariance matrix). This might be, in principle, problematic in the proportional regime, as we already pointed out that consistent estimation of the population precision matrix is impossible without imposing sparsity constraints on the latter. Hence, another interesting open question is "can we estimate the order parameters of the Replica Symmetric theory without estimating the (population) precision matrix when $\zeta \geq 1$"?

Covariant prior as a practical de-biasing procedure The second achievement of this thesis is having shown that the covariant prior is an easy-to-implement regularization that can be used to obtain well-defined (and potentially unbiased) estimators for all $\zeta < 1$ in the proportional regime where $p = \zeta n$. This is interesting when the Maximum Likelihood estimator or the Maximum Partial Likelihood estimator does not exist past a critical value $\zeta_c < 1$, e.g. for the Logit and the Cox regression models.
In chapter 2 we showed that the resulting estimator is, on average, aligned with the true vector of parameters ($\boldsymbol{\beta}_0$). This is not true, in general, for shrinkage priors as e.g. ridge or lasso. Thus, by appropriately tuning the regularization parameter of the covariant prior, one can obtain an unbiased MAP estimator. However, such value of the regularization strength depends on : i) the true signal strength $\theta_0 := \|\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\beta}_0\|_2$ (with $\boldsymbol{\Sigma}_0^{1/2}$ the population covariance matrix), and ii) other nuisance parameters (when present), e.g. the true intercept in Logit regression for instance. These values are clearly unknown and must be estimated. In chapter 6, we showed that *all* the Replica Symmetric order parameters of the theory can be easily estimated solely from the data, i.e. for GLMs θ_0 and the nuisance parameters can be estimated by solving a set of simple moment matching conditions. For the Cox regression model, a slightly more involved procedure must be adopted since the model is semi-parametric, however this "boils down" to solve a single moment matching condition. The knowledge of the RS order parameters and of θ_0 allows one to de-bias the Maximum A Posteriori estimator for *any* value of the regularization strength and of the covariance matrix. This simplifies the procedure proposed in chapter 3 since we can just choose the value of the

regularization strength that minimizes, for instance, an estimate of the generalization error, as normally done in cross validation. Fortunately, once the RS order parameters are estimated, one can readily obtain an estimate of the leave one out cross validated generalization error and thus select a value of the regularization strength in an efficient and data driven manner. Although the de-biased estimator $\hat{\boldsymbol{\beta}}_n$ remains inconsistent (for the whole vector $\boldsymbol{\beta}_0$), it is needed for testing the components of $\boldsymbol{\beta}_0$, i.e. establish if a component is non-zero with a certain confidence level. Whilst other de-biasing methods exist, see [6, 17], and allow for $\zeta \geq 1$, these require the estimation of the *whole* population precision matrix (the inverse of the population covariance matrix). There exist strategies to achieve this under sparsity constraints, see [17]. However, this is generally impossible in the proportional regime without prior assumptions, as the covariance matrix is an inconsistent estimator in the proportional regime. In contrast, the de-biased estimator obtained with the covariant prior only requires the estimation of the diagonal of the precision matrix, which can be easily achieved as we explained. It would be interesting to see how the power of the approximate test procedure obtained by this representation compares with the power of the test of the de-biased lasso procedure [6, 17, 16]. A naive direct comparison might be, however, misleading, due to the fact that when using the covariant prior we are not implicitly assuming that the underlying structure of the signal is sparse, as done when using a Lasso regularization. Hence, one should compare the power over all hypotheses, and not only the sparse ones.

Practical applications of the theory It seems now natural to ask "what can be done in practice" with the new knowledge developed in this thesis. The first natural application is de-biasing. In the proportional regime and when $p < n$, we now have the methodology (derived from the Replica Symmetric equations) to estimate both the associations $\boldsymbol{\beta}_0$ and the cumulative risk function Λ_0 in an unbiased manner. However, in order to draw correct conclusions from the data, it is vital to understand under which conditions the theory (derived for Gaussian covariates) holds irrespective of the actual distribution of the covariates. This has been studied in the literature [21], but under conditions which do not include the Cox regression model. A future study might aim at establishing if similar results hold for a more general class of models, including the semiparametric model studied in this thesis. De-biasing is a very particular application which is useful for inference and testing purposes, but not necessarily for prediction ones. Indeed, it is well known that a biased estimator may finally "beat" an unbiased one, in terms of prediction error. Probably the most famous examples of this are the James Stein [28] and Ridge [15] estimators. This leads to the second practical consequence of this thesis. Namely, we are now able to estimate the leave one out cross validation error in a computationally fast and accurate manner, even for the Cox model as shown in chapter 6. Leave one out cross validation aims at estimating the prediction error and is generally used to select the "best" value of one or more regularization parameters (also called hyperparameters), e.g. the strength of the ridge or lasso regularization. However, it requires to fit the model n times, which is computationally slow in the proportional regime, already for not too large values of ζ. For this reason, fast approximations of the leave one out method are needed in the modern high dimensional regime. Several alternatives have been proposed, see for instance [33, 23, 1]. As argued in chapter 6, the asymptotic theory in the proportional regime suggests a very fast approximation of the leave one out cross validation loss that can be easily implemented in practice. It is an interesting point left to future studies to establish if a generalization of this procedure might be applied when an arbitrary regularization is adopted and if this procedure is competitive (in terms of execution time) with recently proposed alternatives.

Future perspectives Besides establishing the rigorous validity of the conjectures in this thesis, other questions remain open for future investigations. Recently, the asymptotics of Generalized Linear Models with non-convex regularization [34, 11, 12] have been established heuristically in [25, 24]. A generalization for the Cox regression studied in this thesis seems to be a natural sequel for the research present here. The use of a non-convex regularization implies that the Replica Symmetric ansatz might not be valid for all the values of the regularization strength, since the objective might not have a single minimum depending on the value of the latter [2]. This would entail generalizing the current theory to include the family of Replica Symmetric Breaking ansatzes [19]. When the optimization problem related to model fitting is non-convex, the dynamics of the algorithm used to perform optimization might also be of interest. There exists statistical physics methodology to address this kind of problems, namely Dynamical Mean Field Theory. This has been successfully applied to investigate different algorithms in toy models of neural networks [20, 13, 5]. It would be interesting to see how this method might be applied to the Cox model studied in this thesis. Another interesting direction is to extend the results of this thesis to allow for covariates sampled from a mixture of Gaussians, as done in several recent studies for Generalized Linear Models [18, 22]. Since any distribution can be approximated arbitrary well by a mixture of (sufficiently many) Gaussians [14], one expects that these results should be more "universal". Indeed, results in this direction have been proven [9, 22].

Although more complex models like the one discussed above might be, in principle, considered and worked out, there is an important issue that arises when one wants to apply the theory in practice: we need to make the theory observable. Let me explain what I mean by observable. In the statistical physics of disordered systems the term "theory" generally refers to a set of self-consistent equations (which, however, must be solved numerically in many cases) linking some control parameters to a set of so-called observables. In this thesis, for instance, we have seen that the RS equations predict the value of the order parameters $w, v, \tau, \hat{w}, \hat{v}, \hat{\tau}$ as a function of : 1) the population covariance matrix $\boldsymbol{\Sigma}_0$, 2) the distribution of (the entries of) $\boldsymbol{\beta}_0$, 3) the remaining true parameters of the model, e.g. the true cumulative hazard Λ_0 in the case of the Cox model or, more compactly said, the true conditional distribution of the response T, given the covariates $\mathbf{X}$. The point is that within the Replica theory these "observables" (the order parameters) are actually not observable : they are defined in terms of quantities that are not known in practice. To give an example, $w = \hat{\boldsymbol{\beta}}_n' \boldsymbol{\Sigma}_0 \boldsymbol{\beta}_0 / \|\boldsymbol{\Sigma}_0 \boldsymbol{\beta}_0\|_2$. In applications, both $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\beta}_0$ are not known ($\boldsymbol{\beta}_0$ is precisely what we want to estimate!). In statistics, conversely, a "theory" must be closed in terms of measurable quantities only. This is why the consistency of $\hat{\boldsymbol{\beta}}_n$ is so important : if one shows that $\hat{\boldsymbol{\beta}}_n$ is (asymptotically) consistent for $\boldsymbol{\beta}_0$ than one might correctly approximate $\boldsymbol{\beta}_0$ with $\hat{\boldsymbol{\beta}}_n$. This is used, for instance, to compute the asymptotic variance of the Maximum Likelihood estimator in the "classical" regime $p \ll n$ [4, 32]. The observability issue is irrelevant when the data (response and/or covariates) are generated by the theoretician, since the unknowns are actually known in this case. The order parameters can be measured, and one can verify a posteriori that if the covariance matrix is such, the $\boldsymbol{\beta}_0$ is such and the remaining parameters are such, then the theory correctly describes the output of the regression. However, in practice one does not know these quantities, and one might question the use of the theory then, since it seems to be in-applicable. Indeed, one of the main challenges that we tried to overcome in this thesis has been to circumvent this problem. Thanks to the insight generated in [3] and chapter 6, we think that this problem is by now solved when $p < n$ and the data are extracted from a correlated Gaussian distribution. On the other hand, it is unclear if all the order parameter are observable when $p > n$ and/or

the covariates are pulled from a Gaussian mixture. This opens up new interesting questions for the future.

Bibliography

[1] A Auddy, H Zou, K Rahnamarad, and A Maleki. Approximate leave-one-out cross validation for regression with ℓ_1 regularizers. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2377–2385. PMLR, 02–04 May 2024.

[2] J Barbier, D Panchenko, and M Sáenz. Strong replica symmetry for high-dimensional disordered log-concave gibbs measures. *Information and Inference: A Journal of the IMA*, 11(3):1079–1108, 12 2021.

[3] PC Bellec. Observable adjustments in single-index models for regularized m-estimators, 2024.

[4] F Bijma, MA Jonker, and AW van der Vaart. *An Introduction to Mathematical Statistics*. Amsterdam University Press, 2017.

[5] T Bonnaire, D Ghio, K Krishnamurthy, F Mignacco, A Yamamura, and G Biroli. High-dimensional non-convex landscapes and gradient descent dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104004, oct 2024.

[6] M Celentano, A Montanari, and Y Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194 – 2220, 2023.

[7] ACC Coolen, JE Barrett, P Paga, and CJ Perez-Vicente. Replica analysis of overfitting in regression models for time-to-event data. *Journal of Physics A: Mathematical and Theoretical*, 50(37):375001, aug 2017.

[8] ACC Coolen, M Sheikh, A Mozeika, F Aguirre-Lopez, and F Antenucci. Replica analysis of overfitting in generalized linear regression models. *Journal of Physics A: Mathematical and Theoretical*, 53(36):365001, aug 2020.

[9] Y Dandi, L Stephan, F Krzakala, B Loureiro, and L Zdeborová. Universality laws for gaussian mixtures in generalized linear models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 54754–54768. Curran Associates, Inc., 2023.

[10] D Donoho and A Montanari. High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969, 2013.

[11] J Fan and R Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[12] LE Frank and JH Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[13] C Gerbelot, E Troiani, F Mignacco, and L Krzakala, Fand Zdeborová. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024.

[14] I Goodfellow, Y Bengio, and A Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[15] AE Hoerl and RW Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[16] A Javanmard and A Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014.

[17] A Javanmard and A Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593 – 2622, 2018.

[18] B Loureiro, G Sicuro, C Gerbelot, F Pacco, Aand Krzakala, and L Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10144–10157. Curran Associates, Inc., 2021.

[19] M Mezard, G Parisi, and M Virasoro. *Spin Glass Theory and Beyond*. WORLD SCIENTIFIC, 1986.

[20] F Mignacco, F Krzakala, P Urbani, and L Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9540–9550. Curran Associates, Inc., 2020.

[21] A Montanari and BN Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

[22] L Pesce, F Krzakala, B Loureiro, and L Stephan. Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27680–27708. PMLR, 23–29 Jul 2023.

[23] K R Rad and A Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):965–996, 06 2020.

[24] A Sakata. Prediction errors for penalized regressions based on generalized approximate message passing. 56(4):043001, feb 2023.

[25] A Sakata and Y Xu. Approximate message passing for nonconvex sparse regularization with stability and asymptotic analysis. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(3):033404, mar 2018.

[26] F Salehi, E Abbasi, and B Hassibi. The impact of regularization on high-dimensional logistic regression. *ArXiv*, abs/1906.03761, 2019.

[27] M Sheikh and ACC Coolen. Analysis of overfitting in the regularized cox model. *Journal of Physics A: Mathematical and Theoretical*, 52(38):384002, aug 2019.

[28] C Stein. Inadmissibility of the usual estimator for the mean of a multivariate distribution: Berkeley. volume 1. edited by: Neyman j, 1956.

[29] P Sur and EJ Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

[30] T Takahashi and Y Kabashima. A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(7):073405, 2018.

[31] C Thrampoulidis, S Oymak, and B Hassibi. The gaussian min-max theorem in the presence of convexity, 2015.

[32] AW van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000.

[33] S Wang, W Zhou, H Lu, A Maleki, and V Mirrokni. Approximate leave-one-out for fast parameter tuning in high dimensions. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5228–5237. PMLR, 10–15 Jul 2018.

[34] CH Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010.

Chapter 8

Closing chapter

8.1 Samenvatting

Tegenwoordig zijn veel datasets hoogdimensionaal, dat wil zeggen dat het aantal geregistreerde kenmerken vergelijkbaar is met, of groter is dan, het aantal waarnemingen. Om een asymptotische theorie te ontwikkelen, moeten aannamen worden gemaakt. Een belangrijke is de schaling van het aantal kenmerken (p) in een statistisch regressiemodel ten opzichte van het aantal waarnemingen (n). Het proportionele regime, waarbij p evenredig is met n, is extreem uitdagend omdat de consistentie van de gehele Maximale A Posteriori schatters verloren gaat. Dit betekent dat de gelijktijdige schatting van alle modelparameters niet mogelijk is, zelfs niet met een divergerend aantal waarnemingen en met een goed afgestelde regularisatie. Inconsistentie bemoeilijkt de studie van het asymptotisch gedrag van MAP-schatters door middel van traditionele wiskundige benaderingen, zoals die gebruikt worden in het "classical" regime, waar p constant blojft ten opzichte van n, of modernere benaderingen, gebruikt in de studie van geregulariseerde schatters in het ultrahoogdimensionale regime, waar p exponentieel mag groeien met n.

In dit proefschrift hebben we de statistische natuurkundige benadering van optimalisatie toegepast om een verbeterde asymptotische theorie te verkrijgen voor de (geregulariseerde) Maximum Partial Likelihood schatter in het proportionele regime, door middel van de Replica methode. Hoofdstuk 2 behandelt de Maximum Partial Likelihood Estimator (MPLE), die wordt verkregen door optimalisatie van de Cox Partiële Likelihood. Het is bekend dat de MPLE mogelijk niet bestaat voor voldoende grote p, afhankelijk van n. De theorie wordt afgeleid (en gevalideerd door numerieke simulaties) onder de aanname dat de Maximum Partial Likelihood Estimator bestaat. In hoofdstuk 4 breiden we de theorie uit naar het gereguleerde geval. De geregulariseerde MPLE (RMPLE), verkregen door optimalisatie van de som van de (min)logaritme van de partiële waarschijnlijkheid van Cox en een convexe regularisatie term, bestaat gegarandeerd voor elke p. We richten ons in het bijzonder op sparsity inducerende regularisaties, zoals de Lasso en het Elastic Net (een geregulariseerde versie van de Lasso).

De Replica methode vindt zijn oorsprong in de statistische fysica van wanordelijke systemen. Tot op heden is deze methode echter theuristisch: ze bestaat uit een reeks algebraïsche stappen die tot een eindresultaat leiden. Hoewel de fysische intuïtie achter de stappen in de Replica methode duidelijk is, zijn deze niet altijd gerechtvaardigd vanuit een streng wiskundig oogpunt. Daarom heb ik ook een deel van mijn promotieonderzoek gewijd aan het begrijpen van de bewijzen achter sommige resultaten die verkregen zijn met de Replica methode. In hoofdstuk 3 kon ik bewijzen dat de Replica methode tot het juiste resultaat

leidt voor een flexibele parametrische versie van het proportionele hazardmodel, genaamd Piece-Wise Exponential (PWE). De bewijsmethode is de Convex Gaussian Min Max stelling, die in het recente verleden naar voren is gekomen als een wiskundig rigoureuze methode om hoogdimensionale convexe optimalisatieproblemen aan te pakken. Aangezien het Cox-model een limiet geval is van het PWE-model (waarvoor het huidige bewijs echter niet geldt), hebben we het vertrouwen in de resultaten van Replica verbeterd.

Op verschillende punten in het proefschrift heen hebben we algoritmen voorgesteld (en getest) om de volgende drie belangrijke problemen aan te pakken: i) snelle optimalisatie van de Regularized Cox Partial Likelihood, ii) schattingen van onbekenden grootheden die nodig zijn om de theorie praktisch toepasbaar te maken en iii) de-biasing.

Wat punt i) betreft: de statistische natuurkundige benadering van optimalisatie maakt het mogelijk om efficiënte optimalisatiealgoritmen af te leiden. Het meest prominente voorbeeld is de klasse van Approximate Message Passing-algoritmen. Uitgaande van de Generalized - AMP, die wordt gebruikt om geregulariseerde Generalized Linear Models (GLM's) te fitten, hebben we in hoofdstuk 5 het COX-AMP algoritme voorgesteld. Dit is een AMP-algoritme dat specifiek is afgestemd op de bijzondere kenmerken van het Cox semi-parametrische model.

Over punt ii): de asymptotische theorie verkregen door de Replica methode hangt af van verschillende onbekende parameters van het datagenererende proces. In simulaties zijn deze grootheden bekend en kan de juistheid van de theorie numeriek gecontroleerd worden. Dit is wat in eerdere studies is gedaan. Om de theorie echter toepasbaar te maken in de praktijk met echte data, moeten deze onbekenden geschat worden. Hiertoe introduceerden we in hoofdstuk 3 een zelfconsistent algoritme, dat enigzins lijkt op verwachtingsmaximalisatie, en dat gericht is op het schatten van de signaalsterkte (L2 norm van de ware associatievector) en de cumulatieve hazard rate. Daarna, in hoofdstuk 5, stelden we vast dat men eigenlijk alle ordeparameters kan schatten zonder de RS-vergelijkingen te moeten oplossen, maar eerder door ze als leidraad te gebruiken. Tenslotte laten we, gebruikmakend van de ideeën uit hoofdstuken 5 en 3, in hoofdstuk 6 zien dat ook de signaalsterkte geschat kan worden op een veel eenvoudigere manier waarbij geen ingewikkelde zelfconsistente set vergelijkingen opgelost hoeft te worden, maar eerder de wortel van een (impliciete) functie gevonden kan worden.

Wat betreft iii): we hebben in hoofdstuk 1 laten zien dat een eenvoudige veralgemening van de nokregularisatie kan worden gebruikt om onbevooroordeelde, zij het inconsistente, schattingen van de parameters van een Generalized Linear Model (GLM) te verkrijgen. We noemden de prior die overeenkomt met een dergelijke regularisatie "covariant", omdat de resulterende schatter gemiddeld dezelfde richting heeft als de Maximum Likelihood en niet geometrisch bevooroordeeld is zoals in het geval van een nulgemiddelde Gaussische prior (Nokregularisatie). Als de signaalsterkte eenmaal is geschat zoals uitgelegd in hoofdstuk 6, kan de covariante prior effectief worden gebruikt om een debiaschatter (voor GLM's) te verkrijgen door de regularisatiesterkte goed af te stellen. In hoofdstuk 6 hebben we echter opgemerkt dat deze niet zo eenvoudige strategie volledig omzeild kan worden en dat men eenvoudig de debiasingfactoren kan berekenen en een asymptotisch onvertekende schatter kan verkrijgen. Deze alternatieve, en veel eenvoudigere, procedure bouwt voort op de ideeën ontwikkeld in hoofdstuk 5.

Een meer gedetailleerd overzicht van de bijdrage van elk hoofdstuk is te vinden in hoofdstuk 7. Daar bespreken we ook de beperkingen van de resultaten in dit proefschrift en kijken we naar mogelijke toekomstige onderzoeksrichtingen.

8.2 Summary

Nowadays, many data-sets are "high dimensional", that is the number of recorded features is comparable to, or larger than, the number of observations. In order to develop an asymptotic theory, assumptions must be undertaken. A key one is the scaling of number of features (p) included in a statistical regression model with respect to the number of observations (n). The proportional regime, where p is proportional to n, is extremely challenging because consistency of the whole Maximum A Posteriori estimators is lost. This means that the simultaneous estimation of all the model's parameters is not possible, even with a diverging number of observations and with a properly tuned regularization. Inconsistency hinders the study of the asymptotical behaviour of MAP estimators by means of traditional mathematical approaches, like those used in the "classical" regime, where p is small with respect to n, or more modern ones, used in the study of regularized estimators in the ultra high dimensional regime, where p is allowed to grow exponentially with n.

In this thesis, we adopted the statistical physics approach to optimization in order to obtain an improved asymptotic theory for the (Regularized) Maximum Partial Likelihood estimator in the proportional regime, by means of the Replica method. Chapter 3 deals with the Maximum Partial Likelihood Estimator (MPLE), which is obtained by optimization of the Cox Partial Likelihood. In chapter 5, we extend the theory to the regularized case. The Regularized MPLE (RMPLE), obtained by optimization of the objective given by the sum of the (minus) logarithm of the Cox partial likelihood and a convex regularization, is guaranteed to exists for any p. We focus in particular on sparsity inducing regularizations, such as the Lasso and the Elastic Net (a regularized version of the Lasso).

Whilst the physical intuition behind the steps undertaken in the Replica method is clear, these are not always justified from a rigorous mathematical point of view. For this reason, I also dedicated part of my PhD studies to understand the proofs behind some of the results obtained by the Replica method. In chapter (4), I was able to prove that the Replica method leads to the correct result for a flexible parametric version of the proportional hazards model, called Piece-Wise Exponential (PWE). The method of proof is the Convex Gaussian Min Max theorem, which is emerged in the recent past as a mathematically rigorous method to tackle high dimensional convex optimization problems. Since the Cox model is a limiting case of the PWE model (for which , however, the present proof does not apply), we have improved "confidence" in the Replica results.

Throughout the thesis, we have proposed (and tested) algorithms to address three key issues : i) fast optimization of the Regularized Cox Partial Likelihood, ii) estimations of unknowns needed to make the theory practically applicable and iii) de-biasing.

Concerning point i) : the statistical physics approach to optimization allows to derive efficient optimization algorithms. The most prominent example being the class of Approximate Message Passing algorithms. Starting from the Generalized - AMP, which is used to fit regularized Generalized Linear Models (GLMs), we proposed in 5 the COX-AMP algorithm. This is an AMP algorithm specifically tailored to the peculiar features of the Cox semi-parametric model.

About point ii) : the asymptotic theory obtained by the Replica method depends on several unknown parameters of the data generating process. In simulations, these quantities are known, and the correctness of the theory can be numerically checked. This is what has been done in previous studies. However, in order to make the theory applicable in practice with real data, these unknowns must be estimated. To this aim, we introduced in chapter 3 a self consistent algorithm, loosely resembling the expectation maximization one, which

aims at estimating the signal strength (L2 norm of the true association vector) and the cumulative hazard rate. Afterwards, in chapter 5, we noticed that one can actually estimate all the order parameters without needing to solve the RS equations, but rather using them as a guiding tool. Finally, using ideas of chapter 5 and 3, we show in chapter 6 that also the signal strength can be estimated in a much easier manner which does not require solving a complicated self consistent set of equations, but rather finding the root of an (implicit) function.

For what concerns iii) : we showed in chapter 2 that a simple generalization of the ridge regularization, can be used to obtain unbiased, albeit inconsistent, estimates of the parameters of a Generalized Linear Model (GLM). We referred to the prior corresponding to such regularization as "covariant", since the resulting estimator has, on average, the same direction as the Maximum Likelihood one, and is not geometrically biased as in the case of a zero mean Gaussian prior (Ridge regularization). Once the signal strength is estimated as explained in chapter 6, one can effectively use the covariant prior to obtain a debias estimator(for GLMs), by appropriately tuning the regularization strength. However, we noticed in chapter 6, that this not-so-straightforward strategy can be completely by-passed, and one can easily compute the debiasing factors and obtain an asymptotically unbiased estimator. This alternative, and much easier, procedure build on the ideas developed in chapter 5.

A more detailed overview of the contribution of each chapter can be found in the chapter 7. There we also discuss the limitations of the results in this thesis, and look at possible future directions of investigation.

8.3 Research Data Management

Ethics

In this thesis, we used simulated data-sets, in order to correctly test the theoretical results under the hypothetical scenario where the assumptions of the theory are satisfied. In each chapter, we have explained how the simulations were carried out and provided extensive details about the numerical methods used to solve the non-linear Replica Symmetric equations, fit the regression models and generate the data sets.

Findable Accessible

In chapter 2, we provided a detailed explanation of the methodology used to carry out the simulations. We performed the relevant numerical solution, data generation and model fitting in the C language. The C files are available upon request (contact massaemanuele.95@gmail.com, ton.coolen@ru.nl).
In chapter 3, we provided detailed explanation of the methodology used to carry out the simulations. The implementation of the methodology that is there proposed and tested is in the Python language. The Python scripts used in this chapter are available upon request (contact massaemanuele.95@gmail.com, ton.coolen@ru.nl).
In chapter 4, we provided the code used to run the simulations in Github at `https://github.com/EmanueleMassa/SMSA_2024`.
In chapter 5, we provided the code used to run the simulations in Github at `https://github.com/EmanueleMassa/Regularized_Cox_model`.
In chapter 5, we provided the code used to run the simulations in Github at `https://github.com/EmanueleMassa/Practical_debiasing`.

8.4 About the author

Emanuele Massa was born on 19 September 1995 in Genova, Italy. He graduated in 2014 from Liceo Scientifico Carlo Barletti, Ovada (AL), Italy.

He then enrolled in the program Physical Engineering offered at Politecnico di Torino, starting September 2014. He successfully graduated with 109/110 in March 2018, with the thesis "Resistivita' elettrica dei metalli: teoria di Bloch-Boltzmann e verifica su sperimentale su campione Pb-In", under the supervision of Prof. R. Gonnelli.

In September 2018, he was admitted, on a competitive basis, in the International track of the Master "Physics of Complex System". This study program, focusing on Statistical Physics and interdisciplinary applications, is hosted by Politecnico di Torino, International Center for Theoretical Physics (ICTP) and Istituto Superiore di Studi Avanzati (SISSA) in Italy, and Sorbonne Université, Université Paris Diderot and Université Paris Saclay. During the next two years he would have his first encounter with probability and statistical concepts, through the course "Probability and Information theory", taught by Prof. M. Marsili at ICTP. Finally he graduated 110 cum laude/ 110 in October 2020, with the thesis "A coding approach to statistical inference" under supervision of Prof. M.Marsili in ICTP, Trieste.

During the months spent working on his Master thesis he also applied for a position as PhD Candidate working on High Dimensional Survival Analysis, based on the insight provided by statistical physics under the supervision of Prof. A.C.C. Coolen, at Radboud Universiteit, Nijmegen. After being awarded the position, he moved to Nijmegen. Emanuele would go then on spending the next four and half years studying various topics in physics of dysordered system, high dimensional statistics and survival analysis. While studying and performing research, he also was teaching assistant in the course Real World Complex System modelling, given by his supervisor Ton Coolen at Radboud.

With the help of Ton, Emanuele worked a couple of months at Saddle point Europe, after finishing his contract with Radboud Universiteit. There he kept on working on practical implementation of the Replica theory. He extended the fast cross validation strategies initally proposed in this thesis to include the case of correlated covariates, by adopting methodology derived from random matrix theory. Furthermore, he adopted these methodologies to incorporate Ordinal Class regression models.

Afterwards, with the help of Marianne and Kit, he joined the IQ Health department of Radboud Universitair Medisch Centrum. There he focused on extending the Bayesian Federated Inference in order to handle missing covariates.

8.5 Publication List

Journal Articles

1. **Massa, E** and Jonker, MA and Coolen, ACC, *Penalization-induced shrinking without rotation in high dimensional GLM regression: a cavity analysis*, Journal of Physics A: Mathematical and Theoretical, 55, 485002, 2022.

2. **Massa, E** and Mozeika, A and Coolen, ACC, *Replica analysis of overfitting in regression models for time to event data: the impact of censoring*, Journal of Physics A: Mathematical and Theoretical, 57, 125003, 2024.

3. **Massa, E** and Coolen, ACC, *Observable asymptotics of regularized Cox regression models with standard Gaussian designs: a statistical mechanics approach*, Journal of Physics A: Mathematical and Theoretical, 58, 105001, 2025.

Conference Papers

1. **Massa, E**, *Proportional asymptotics of piecewise exponential proportional hazards models*, 15th Workshop on Stochastic Models, Statistics and Their Applications, TU Delft, Delft, The Netherlands, 2024. (Accepted, in press)

Preprints

1. Pazira, H and **Massa, E** and Weijers, JAM and Coolen, ACC and Jonker, MA, *Bayesian Federated Inference for Survival Models*, arXiv, 2024. (submitted to Journal of Applied Statistics).

2. **Massa, E** and Coolen, ACC, *Practical debiasing with the Covariant Prior in the proportional regime when* $p < n$, arXiv, 2025 (submitted to Journal of Physics Communications).

Conference Poster

1. **Massa, E** and Jonker, MA and Roes, KCB and Coolen, ACC ,*Correction of overfitting bias in regression models*, Youth in High Dimensions, ICTP, Trieste, Italy, 2023.

8.6 Acknowledgments

I could have never finished this project if it was not for my personal stubbornness and endurance, but also for all the help I received during these years. With this I want to thank everyone, mentioned or not, that have been part of this experience so far.

Eerst, wil ik graag mijn promotoren A.C.C. Coolen, K.C.B. Roes en co-promotor M.A. Jonker, bedanken voor de begeleiding door deze vijf jaren. Ik ben dankbaar dat zij mij de kans hebben gegeven om onderzoek in dit interessant en snel ontwikkelend vakgebied op de grens van statistiche fysika en hoogdimensionale statistieken te doen. Met name, hartelijk dank aan Ton en Marianne, voor hun geduldigheid door de jaren. Dank dat jullie altijd naar mij geluisterd hebben, over het werk en ook andere dingen. Dank voor jullie technische en emotionele ondersteuning. Dank voor in mij te geloven, ook toen ik niet veel in mijzelf geloofde. Dank voor alle hulp tijdens dit laatste jaar.

Thanks to my office colleagues Peyman, Aaron, Mauricio, Shakeeb, Theodore, and Eduardo, because they made my day less serious at times. I regret that we did not get to know each other more during these years, I wish you all the best in your endeavors, but more general in life.

A big thank to all the nice people who I had the pleasure to work with at Saddle Point Science Europe and Saddle Point Science. A big thank to all the colleagues of IQ Health at Radboud Universitair Medisch Centrum. I hope we will remain in contact.

How could I have survived all these years without Volleyball and Beach Volleyball? Thanks to the Waal international group, it has been always super fun to play with people from all over the world ! Thanks to all the friends and people that I played with at Beach Fabriek. Edyta, Cheng and Sergio : our Tuesday evenings matches have always been fought to the last point, although in three years me and Edyta managed to win only once. Amir, Ruud and Andrea, I got to know you better during the last year, and it is always so much fun to play together. Bedankt aan de mannnen van de woensdag-trainingen. Altijd leuk om samen te spelen. Un grazie anche agli amici Beachers dei dintorni Ovadesi : Gianmaria, Daniele e Valentina. Ogni volta che scendo non vedo l'ora di giocare con voi.

Grazie a tutti gli amici Ovadesi, perché le nostre feste, vacanze ed avventure assieme sono un ricordo indelebile nella mia memoria, e mi fanno sorridere ogni volta che ci penso. Fede e Lollo, il trio é da un po di anni "in wireless", ma quando siamo assieme le risate e le birre sono nella stessa quantitá e qualitá di sempre. Giacomo Bi, sei sempre stato un vero amico, nei momenti migliori, e soprattutto in quelli peggiori. Andrea, Giaggi, Cocco e Subre, siamo separati da un po di chilometri da ormai un po' d'anni, ma quando torno a "Uó" é sempre festa come prima.

Een hartelijke grote dank aan de "schoonfamilie" : Sindy, Johan, Vivian, Daphne, Jeroen en Camiel. Hartelijk bedankt dat jullie mij met open armen hebben ontvangen. Dank voor jullie geduldigheid met mijn Nederlands en voor mij (bijna) te leren Rikken. Dankje voor alle leuke momenten die wij samen meegemaakt hebben, en gaan ervaren.

Un grazie e un grande abbraccio alla famiglia. Grazie ai cugini Lorenzo, Leonardo, Greta, a Zio Paolo, e alle zie Silvia e Donatella. Ultima ma non meno importante, cosa faremmo tutti senza la capo famiglia Nonna Marisa che ha sempre una buona parola per ogniuno di noi.

Federico, grazie semplicemente per esserci, ed essere mio fratello. Non ci sentiamo molto, ma le nostre lunghe chiamate sono sempre importanti per me. Un grazie anche a Laura, che é parte intengrante della famiglia Massa ormai.

Mamma Patrizia e papá Luca, grazie per il vostro infinito amore ed infinita pazienza. Se

mai saró un genitore bravo un quarto di quello che siete con me, mi considererò soddisfatto.

Laatste, maar zeker belangrijkste, de grootste dank aan mijn partner Manon. Het was een onvoorstelbaar en geweldig toeval om zo'n mooie en dapper vrouw te ontmoeten. Lief, jij weet heus wel hoe slecht ik me soms heb gevoeld in deze jaren. Dank je dat je altijd mijn steun bent.

8.7 Donders Graduate School

For a successful research Institute, it is vital to train the next generation of scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School in 2009. The mission of the Donders Graduate School is to guide our graduates to become skilled academics who are equipped for a wide range of professions. To achieve this, we do our utmost to ensure that our PhD candidates receive support and supervision of the highest quality.

Since 2009, the Donders Graduate School has grown into a vibrant community of highly talented national and international PhD candidates, with over 500 PhD candidates enrolled. Their backgrounds cover a wide range of disciplines, from physics to psychology, medicine to psycholinguistics, and biology to artificial intelligence. Similarly, their interdisciplinary research covers genetic, molecular, and cellular processes at one end and computational, system-level neuroscience with cognitive and behavioural analysis at the other end. We ask all PhD candidates within the Donders Graduate School to publish their PhD thesis in de Donders Thesis Series. This series currently includes over 700 PhD theses from our PhD graduates and thereby provides a comprehensive overview of the diverse types of research performed at the Donders Institute. A complete overview of the Donders Thesis Series can be found on our website: `https://www.ru.nl/donders/donders-series`

The Donders Graduate School tracks the careers of our PhD graduates carefully. In general, the PhD graduates end up at high-quality positions in different sectors, for a complete overview see `https://www.ru.nl/donders/destination-our-former-phd`. A large proportion of our PhD alumni continue in academia ($> 50\%$). Most of them first work as a postdoc before growing into more senior research positions. They work at top institutes worldwide, such as University of Oxford, University of Cambridge, Stanford University, Princeton University, UCL London, MPI Leipzig, Karolinska Institute, UC Berkeley, EPFL Lausanne, and many others. In addition, a large group of PhD graduates continue in clinical positions, sometimes combining it with academic research. Clinical positions can be divided into medical doctors, for instance, in genetics, geriatrics, psychiatry, or neurology, and in psychologists, for instance as healthcare psychologist, clinical neuropsychologist, or clinical psychologist. Furthermore, there are PhD graduates who continue to work as researchers outside academia, for instance at non-profit or government organizations, or in pharmaceutical companies. There are also PhD graduates who work in education, such as teachers in high school, or as lecturers in higher education. Others continue in a wide range of positions, such as policy advisors, project managers, consultants, data scientists, web- or software developers, business owners, regulatory affairs specialists, engineers, managers, or IT architects. As such, the career paths of Donders PhD graduates span a broad range of sectors and professions, but the common factor is that they almost all have become successful professionals.

For more information on the Donders Graduate School, as well as past and upcoming defences please visit: `http://www.ru.nl/donders/graduate-school/phd/`

$$\xrightarrow[\to\infty]{P} v_*^2 \cdot \mathbb{E}_{\mathcal{D}}\left[\min_{\beta \in \mathbb{R}^p} \mathcal{L}_n(\beta|\mathcal{D}) \right] = \lim_{r\to\infty} \lim_{r\to}$$

$$\mathcal{P}(t,\xi) := \lim_{n\to\infty} \mathbb{E}_{,z}\left[\frac{1}{n} \sum_{i=1}^{n} \right. = \lim_{r\to 0} \lim_{n\to}$$

$$\left. \delta(t - T_i)\delta(\xi - x_i'\hat{\beta}) \right] \quad \text{prox}_{f(.)}(x,\alpha) =$$

$$\hat{w}_n := \frac{\beta_0' \Sigma_0 \hat{\beta}_n}{\|\Sigma_0^{1/2}\beta_0\|_2} \xrightarrow[n\to\infty]{P} w_* \left[\left| \min_{\omega,\beta} \mathcal{L}_n(\omega,\beta) - \min_{\omega,w,} \right. \right.$$

$$\mathcal{P}\mathcal{L}_n(\beta) = \sum_{i=1}^{n} \Delta_i \left\{ \log \right.$$

$$(\omega,\beta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \Lambda(t|\omega) e^{x_i'\beta} \right.$$