IOP Publishing Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **50** (2017) 375001 (38pp) https://doi.org/10.1088/1751-8121/aa812f

Replica analysis of overfitting in regression models for time-to-event data

A C C Coolen[1,2], J E Barrett[3], P Paga[1,2] and C J Perez-Vicente[4]

[1] Department of Mathematics, King's College London, The Strand, London WC2R 2LS, United Kingdom
[2] Institute for Mathematical and Molecular Biomedicine, King's College London, Hodgkin Building, Guy's Campus, London SE1 1UL, United Kingdom
[3] Department of Primary Care and Public Health Sciences, King's College London, Addison House, Guy's Campus, London SE1 1UL, United Kingdom
[4] Departament de Física Fonamental, Universitat de Barcelona, 08028 Barcelona, Spain

E-mail: ton.coolen@kcl.ac.uk, james.barrett@kcl.ac.uk, pierre.paga@kcl.ac.uk and conrad@ffn.ub.es

Received 9 May 2017, revised 17 July 2017
Accepted for publication 20 July 2017
Published 17 August 2017

Abstract

Overfitting, which happens when the number of parameters in a model is too large compared to the number of data points available for determining these parameters, is a serious and growing problem in survival analysis. While modern medicine presents us with data of unprecedented dimensionality, these data cannot yet be used effectively for clinical outcome prediction. Standard error measures in maximum likelihood regression, such as p-values and z-scores, are blind to overfitting, and even for Cox's proportional hazards model (the main tool of medical statisticians), one finds in literature only rules of thumb on the number of samples required to avoid overfitting. In this paper we present a mathematical theory of overfitting in regression models for time-to-event data, which aims to increase our quantitative understanding of the problem and provide practical tools with which to correct regression outcomes for the impact of overfitting. It is based on the replica method, a statistical mechanical technique for the analysis of heterogeneous many-variable systems that has been used successfully for several decades in physics, biology, and computer science, but not yet in medical statistics. We develop the theory initially for arbitrary regression models for time-to-event data, and verify its predictions in detail for the popular Cox model.

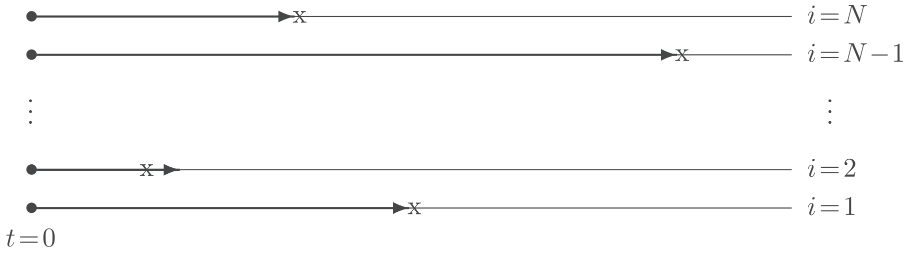
Keywords: replica method, survival analysis, regression

(Some figures may appear in colour only in the online journal)

1. Introduction

In the simplest possible scenario, survival analysis is concerned with data of the following form. We consider a cohort of N individuals, each of whom are at risk of a specified irreversible event, such as the onset of a given disease or death. For each individual i in this cohort we are given p specific measurements $\boldsymbol{z}_i = (z_{i1}, \dots, z_{ip})$ (the covariates) which were taken at a baseline time $t = 0$, as well as the time $t_i > 0$ at which for individual i we either observed the irreversible event, or we ceased our observation without having observed the event yet (the latter case is called 'censoring'). More complex scenarios could involve e.g. having multiple distinct risk types, such as distinct causes of death, or interval censoring, where rather than t_i itself, one is given an interval that contains t_i. The theory developed in this paper can be generalised without serious difficulty to include such extensions, but in the interest of transparency we will focus for now strictly on the simplest case.

$$\begin{array}{ll} \boldsymbol{z}_i \in \mathbb{R}^p: & p\ \textit{covariates of individual}\ i,\ \textit{measured at}\ t=0 \\ t_i > 0: & \textit{event time of individual}\ i\ (\textit{death, onset of disease}, \dots) \end{array}$$



The aim of survival analysis is regression, i.e. to use our data for detecting and quantifying probabilistic patterns (if any) that relate an individual's failure time t to their covariates $\boldsymbol{z}$. Such patterns may allow us to predict individual patients' clinical outcomes, distinguish between high-risk and low-risk patients, reveal general disease mechanisms, or design new data-driven therapeutic interventions (by changing the values of modifiable covariates). For general reviews of the considerable survival analysis literature we refer to textbooks such as [1–4][5]. Being able to use the extracted patterns to predict clinical outcomes for *unseen* patients is the only reliable test of whether our regression results represent true knowledge. Accurate prediction requires that we use as much of the available covariate information as possible, so our focus must be on multivariate regression methods.

Most multivariate survival analysis methods are based on postulating a suitable and plausible parametrisation of the covariate-conditioned event time distribution, whose parameters are estimated from the data via either the maximum likelihood protocol (ML), or (following Bayesian reasoning) via maximum *a posteriori* probability (MAP). The most popular parametrisation is undoubtedly the proportional hazards model of Cox [5], which uses ML inference, and assumes the event time distribution to be of the so-called proportional hazards form $p(t|\boldsymbol{z}) = -\frac{\mathrm{d}}{\mathrm{d}t}\exp[-\exp(\boldsymbol{\beta}\cdot\boldsymbol{z})\Lambda(t)]$. MAP versions of [5] are the so-called 'penalised Cox' or 'ridge' regression models (with Gaussian parameter priors), see e.g. [6, 7]. More complex parametrisation proposals, such as frailty or random effects models [8–11] or latent class models [12], still tend to have proportional hazards type formulae as their building blocks. In all such models the number of parameters is always larger than or equal to the number p of covariates.

[5] Non-medical applications of survival analysis include e.g. the study of the time to component failure in manufacturing, or of the duration of unemployment in economics.

Hence, to avoid overfitting they can be used safely only when $N \gg p$. This limitation was harmless in the 1970s and 1980s, when many of the currently used models were devised, and where one would typically have datasets with $p \sim 10^2$ at most. For the data of post-genome medicine, however, where we regularly have $p \sim 10^{4-6}$, it poses a serious problem which has for instance prevented us from using genomic covariates in rigorous multivariate regression protocols, forcing us instead to work with 'gene signatures'.

Overfitting in survival analysis models [14, 15] can be visualized effectively by combining regression with cross-validation. For the Cox model, for instance, one can use the inferred association parameters $\boldsymbol{\beta}$ of [5] in combination with Breslow's [16] estimator for the base hazard rate (which is the canonical estimator for [5]), to predict whether an event will have happened by a given cutoff time, and compare the fraction of correct predictions in the training set (the data used for regression) to those in a validation set (the unseen data). When drawn as functions of the number of covariates used, the resulting curves typically exhibit the standard fingerprints of overfitting [17, 18]; see figure 1. Simulations with synthetic data [19] showed that the optimal number of covariates in Cox regression (see arrows in figure 1) tends to be roughly proportional to the number of samples N. Given this observed phenomenology, it seems vital before doing multivariate regression to have a tool for estimating the minimum number of samples or events needed to avoid the overfitting regime. To our knowledge, there is no theory in the literature yet for predicting this number, not even for the Cox model [5]. One finds only rules of thumb—e.g. the number of failure events must exceed 10 times the number of independent covariates—and empirical bootstrapping protocols, often based on relatively small scale simulation data [19–21]. This situation is not satisfactory.

To increase our intuition for the problem, we first explore via simple simulation studies the relation between inferred and true parameters in Cox's model [5]. The parameters of [5] are the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ of regression coefficients (where p is the number of covariates), and the base hazard rate $\lambda(t) = \mathrm{d}\Lambda(t)/\mathrm{d}t$. We generated association parameters and covariates randomly from zero-average Gaussian distributions, and corresponding synthetic survival data using Cox's model without censoring (so all N samples correspond to failure events), for different base hazard rates. To understand the nature of the overfitting-induced regression errors we plotted the p pairs $(\beta_\mu, \hat{\beta}_\mu)$ as points in the plane, where β_μ and $\hat{\beta}_\mu$ are the true and inferred association parameters of covariate μ, respectively, calculated via the recipes of [5]. This resulted in scatterplots as shown in figure 2. Simulations were done for different values of the ratio p/N, with multiple independent runs such that the number of points in each panel is identical. The true association parameters were drawn independently from a zero-average Gaussian distribution with $\langle \beta_\mu^2 \rangle = 0.25$ for all μ. Perfect regression would imply finding all points to lie on the diagonal. Rather than a widening of the variance (as with finite sample size regression errors) overfitting-induced errors are somewhat surprisingly seen to manifest themselves mainly as a reproducible tilt of the data cloud, which increases with p/N, and implies a consistent over-estimation of associations: both positive and negative β_μ will always be reported as more extreme than their true values. These observed errors in association parameters appear to be independent of the form of the true base hazard rate. Similarly, we show in figure 3 the inferred integrated base hazard rates $\hat{\Lambda}(t)$ versus time (solid lines), together with the true values (dashed), which again shows consistent and reproducible overfitting errors. A quantitative theory of overfitting that can predict both the observed tilt and width of the data clouds of figure 2 and the deformed inferred hazard rates of figure 3 would enable us to *correct* the inferred parameters of the Cox model for overfitting, and thereby enable reliable regression up to hitherto forbidden ratios of p/N.

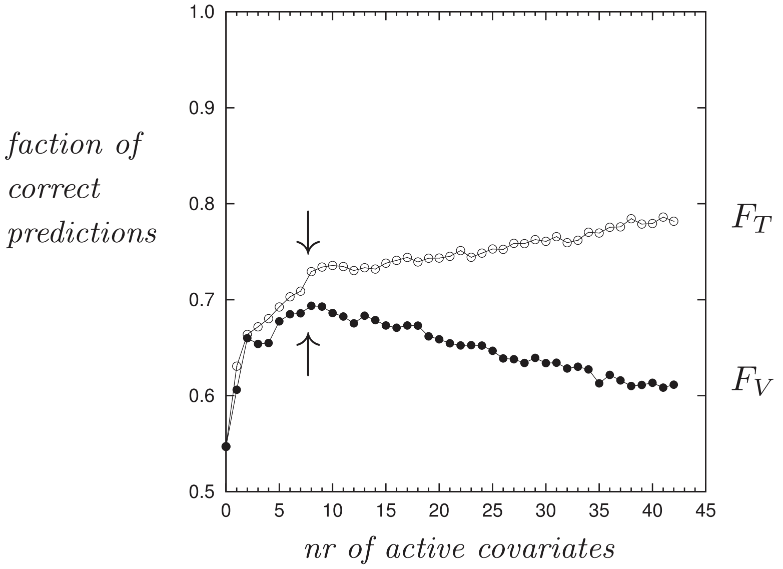


Figure 1. Illustration of overfitting in Cox-type regression. A breast cancer data set [13] containing $N = 309$ samples (129 with recorded events, 180 censored), with clinical and immunological covariates, and disease relapse chosen as event time, was randomly divided into training and validation sets (of roughly equal sizes). L2-regularised Cox regression was used to infer regression coefficients and base hazard rates from the training set (via Breslow's formula [16]), upon which the model was used to predict survival at time $t = 8$ years, for the samples in the training set and for those in the validation set. The fractions of correct predictions are F_T and F_V, respectively. This was repeated multiple times, initially with all covariates, and following repeated iterative removal of the last relevant covariate after each regression. The resulting curves exhibit the standard fingerprints of overfitting: initially the validation performance improves as the number p of retained covariates increases, up to a critical point (here around $p = 6$, see arrows), followed by deterioration as p increases further.

There are mathematical obstacles to the development of a theory of overfitting in survival analysis, which probably explain why it has so far remained an open problem. First, unlike discriminant analysis, it is not immediately clear which error measure to study when outcomes to be predicted are event times. Second, in most survival analysis models (including Cox regression) the estimated parameters are to be solved from coupled transcendental equations, and cannot therefore be written in explicit form. Third, in the overfitting regime one will by definition find even for large N that the inferred parameters depend on the realisation of the data set, while at the more macroscopic level of prediction accuracy there is no such dependence. It is thus not *a priori* clear which quantities to focus on in analytical studies of the regression process, and at which stage in the calculation (if any) averages over possible realisations of the data set may be performed safely.

Our present approach to the problem consists of distinct stages, each removing a specific obstacle, and this is reflected in the structure of our paper. We adapt to time-to-event regression the strategy proposed and executed several decades ago for binary classifiers in the groundbreaking paper by Gardner [22]. We first translate the problem of modelling overfitting into the calculation of a specific information-theoretic generating function, from which we can extract the information we need. Next we use Laplace's argument to eliminate the

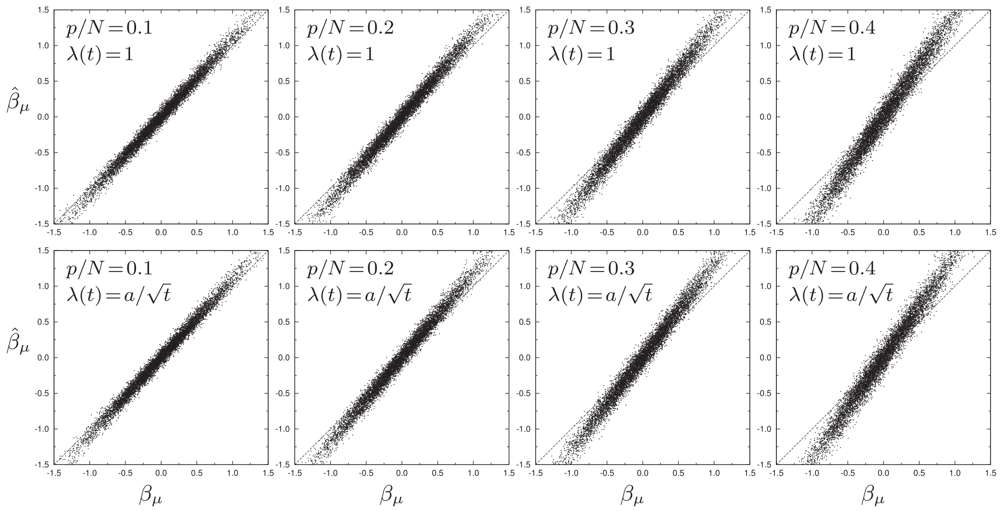


Figure 2. Inferred association parameters (vertical axis) versus true association parameters (horizontal axis) for synthetic survival data generated according to the Cox model, and subsequently analysed with the Cox model. Covariates and true association parameters were drawn randomly from zero-average Gaussian distributions. In all cases $N = 400$, $\langle \beta_\mu^2 \rangle = 0.25$ for all μ, and experiments were repeated such that the total number of points in each panel is identical. Top row: time-independent base hazard rate $\lambda(t) = 1$. Bottom row: time-dependent base hazard rate $\lambda(t) = a/\sqrt{t}$ (dashed), with $a > 0$ chosen such that the average event time is $\langle t \rangle = 1$. The errors in the association parameters induced by overfitting are more dangerous than finite sample size errors, since they mainly take the form of a consistent bias and therefore cannot be 'averaged out'. Moreover, they appear to be independent of the true base hazard rate.

maximisation over model parameters that comes with all ML methods, which is equivalent to writing the ground state energy of a statistical mechanical system as the zero temperature limit of the free energy. The third stage is devoted to making the resulting calculation of the generating function feasible, using the so-called replica method. This method has an impressive track record of several decades in the analysis of complex heterogeneous many-variable systems in physics [23–27], computer science [22, 28], biology [29–31], and economics [32, 33], and enables us to carry out analytically the average of the generating function over all possible realisations of the data set. Finally we exploit steepest descent integration for $N \to \infty$, leading to the identification of the 'natural' macroscopic order parameters of the problem, for which we derive closed equations within the replica symmetric (RS) ansatz. Some technical arguments are placed in appendices, to improve the flow of the paper. We develop our methods initially for generic time-to-event regression models, and then specialise to the Cox model. The final RS equations obtained for the Cox model involve a small number of scalar order parameters, from which we can compute the link between true and inferred regression parameters, and the inferred base hazard rate. The functional saddle point equation for the base hazard rate is rather nontrivial; while we can calculate the asymptotic form of its solution analytically, we limit ourselves mostly to a variational approximation, which already turns out to be quite accurate. We close with a discussion of our results, their implications and applications, and avenues for future work.

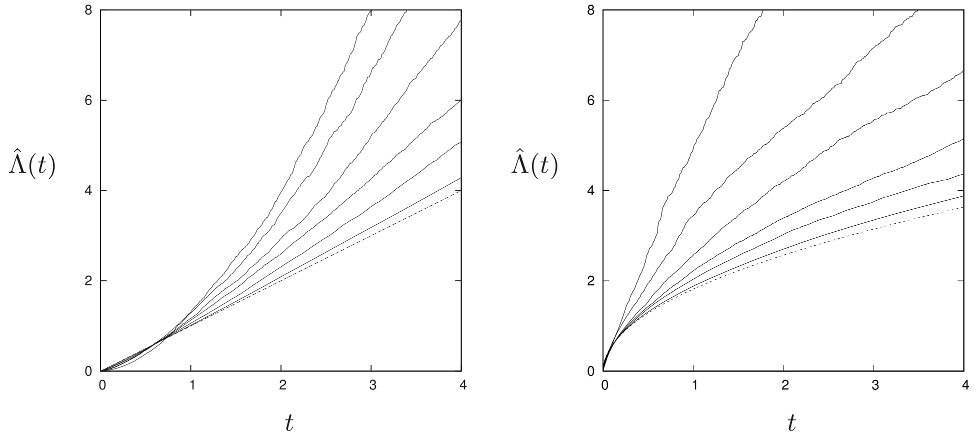


Figure 3. Inferred integrated base hazard rates $\hat{\Lambda}(t) = \int_0^t \mathrm{d}t'\ \hat{\lambda}(t')$ (solid curves, averaged over multiple experiments) for synthetic survival data, generated and subsequently analysed with the Cox model. Covariates and true association parameters were drawn randomly from zero-average Gaussian distributions. In all cases $N = 400$, $\langle \beta_\mu^2 \rangle = 0.25$ for all μ, and $p/N \in \{0.05, 0.15, 0.25, 0.35, 0.45, 0.55\}$ (lower to upper solid curves). Left: data generated with $\lambda(t) = 1$ (dashed). Right: data generated with $\lambda(t) = a/\sqrt{t}$ (dashed), with $a > 0$ chosen such that the average event time is $\langle t \rangle = 1$. The errors induced by overfitting again take the form of a consistent bias: for very short time the base hazard rate is always under-estimated, whereas for large times it is always over-estimated.

2. Overfitting in maximum likelihood models for survival analysis

2.1. Definitions

We assume we have simple time-to-event data $\mathcal{D}$ of the standard type, consisting of N independently drawn samples $i = 1 \ldots N$, with just one active risk and no censoring. Each sample consists of a covariate vector $z_i \in \mathbb{R}^p$, drawn independently from a distribution $P(z)$, and an associated time to event $t_i \in [0, \infty)$, drawn from $P(t|z, \boldsymbol{\theta}^\star)$:

$$\mathcal{D} = \{(z_1, t_1), \ldots, (z_N, t_N)\}. \tag{1}$$

Here $P(t|z, \boldsymbol{\theta}^\star)$ describes a parametrised time-generating model, with q unknown real-valued parameters collected in a vector $\boldsymbol{\theta}^\star \in \mathbb{R}^q$ that we seek to estimate from the data $\mathcal{D}$. We are not interested in estimating $P(z)$, so we take the covariate vectors $\{z_1, \ldots, z_N\}$ as given. The data probability for each parameter choice $\boldsymbol{\theta}$ is

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} P(t_i|z_i, \boldsymbol{\theta}). \tag{2}$$

We next define the empirical distribution of covariates and event times, given the observed data:

$$\hat{P}(t, z|\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \delta(t - t_i)\delta(z - z_i). \tag{3}$$

This allows us to write

$$\begin{aligned}\frac{1}{N}\log P(\mathcal{D}|\boldsymbol{\theta}) &= \int \mathrm{d}t\mathrm{d}\boldsymbol{z}\, \hat{P}(t,\boldsymbol{z}|\mathcal{D})\log P(t|\boldsymbol{z},\boldsymbol{\theta})\\ &= \int \mathrm{d}t\mathrm{d}\boldsymbol{z}\, \hat{P}(t,\boldsymbol{z}|\mathcal{D})\log\Big(\frac{P(t|\boldsymbol{z},\boldsymbol{\theta})}{\hat{P}(t|\boldsymbol{z},\mathcal{D})}\Big)\\ &+ \int \mathrm{d}t\mathrm{d}\boldsymbol{z}\, \hat{P}(t,\boldsymbol{z}|\mathcal{D})\log \hat{P}(t|\boldsymbol{z},\mathcal{D})\\ &= -H(t|\boldsymbol{z},\mathcal{D}) - D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}})\end{aligned} \tag{4}$$

with the conditional differential Shannon entropy of the event time distribution, and the Kullback–Leibler distance [34] between the empirical distribution $\hat{P}(t|\boldsymbol{z},\mathcal{D})$ and the parametrised form $P(t|\boldsymbol{z},\boldsymbol{\theta})$:

$$H(t|\boldsymbol{z},\mathcal{D}) = -\int \mathrm{d}\boldsymbol{z}\, \hat{P}(\boldsymbol{z}|\mathcal{D})\int \mathrm{d}t\, \hat{P}(t|\boldsymbol{z},\mathcal{D})\log\hat{P}(t|\boldsymbol{z},\mathcal{D}) \tag{5}$$

$$D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}}) = \int \mathrm{d}\boldsymbol{z}\, \hat{P}(\boldsymbol{z}|\mathcal{D})\int \mathrm{d}t\, \hat{P}(t|\boldsymbol{z},\mathcal{D})\log\Big(\frac{\hat{P}(t|\boldsymbol{z},\mathcal{D})}{P(t|\boldsymbol{z},\boldsymbol{\theta})}\Big). \tag{6}$$

The parameters $\boldsymbol{\theta}$ estimated via the ML recipe are those that maximise $P(\mathcal{D}|\boldsymbol{\theta})$. According to (4) they minimise the Kullback–Leibler distance $D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}})$ between the empirical covariate-conditioned event time distribution and the parametrised event time distribution with parameter values $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_{\mathrm{ML}} = \mathrm{argmin}_{\boldsymbol{\theta}}\, D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}}). \tag{7}$$

If $N\to\infty$ for fixed p and q, the law of large numbers guarantees that $\lim_{N\to\infty}\hat{P}(t|\boldsymbol{z},\mathcal{D}) = P(t|\boldsymbol{z},\boldsymbol{\theta}^{\star})$ (in a distributional sense), and hence ML regression will indeed estimate the parameters $\boldsymbol{\theta}$ asymptotically correctly, provided the chosen paramerisation is unambiguous:

$$\lim_{N\to\infty}\boldsymbol{\theta}_{\mathrm{ML}} = \mathrm{argmin}_{\boldsymbol{\theta}}\, D(P_{\boldsymbol{\theta}^{\star}}||P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}^{\star}. \tag{8}$$

In this paper, however, we focus on the regime of large datasets with high-dimensional covariate and parameter vectors where overfitting occurs, namely $p,q=\mathcal{O}(N)$ and $N\to\infty$. Here $\hat{P}(t|\boldsymbol{z},\mathcal{D})$ no longer converges to $P(t|\boldsymbol{z},\boldsymbol{\theta}^{\star})$ for $N\to\infty$ in any mathematical sense, the identity (8) is therefore violated, and minimising $D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}})$ as per the ML prescription is no longer appropriate. This is the information-theoretic description of the overfitting phenomenon in survival analysis.

2.2. An information-theoretic measure of under- and overfitting

Maximum likelihood regression algorithms report those parameters $\boldsymbol{\theta}$ for which $P(t,\boldsymbol{z}|\boldsymbol{\theta})$ is as similar as possible to the *empirical* distribution $\hat{P}(t|\boldsymbol{z},\mathcal{D})$, as opposed to the true distribution $P(t|\boldsymbol{z},\boldsymbol{\theta}^{\star})$ from which the data $\mathcal{D}$ were generated. The optimal outcome of regression is for the inferred parameters to be identical to the true ones, i.e. to find $\mathrm{argmin}_{\boldsymbol{\theta}}\, D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}^{\star}$. We therefore define

$$\begin{aligned}E(\boldsymbol{\theta}^{\star},\mathcal{D}) &= \min_{\boldsymbol{\theta}} D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}}) - D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}^{\star}})\\ &= \min_{\boldsymbol{\theta}}\Big\{\frac{1}{N}\sum_{i=1}^{N}\log\Big[\frac{P(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^{\star})}{P(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})}\Big]\Big\}.\end{aligned} \tag{9}$$

This allows us to interpret the value of $E(\boldsymbol{\theta}^\star, \mathcal{D})$ as a measure of ML regression performance:

$$E(\boldsymbol{\theta}^\star, \mathcal{D}) > 0: \text{ underfitting} \tag{10}$$

$$E(\boldsymbol{\theta}^\star, \mathcal{D}) = 0: \text{ optimal parameter estimation} \tag{11}$$

$$E(\boldsymbol{\theta}^\star, \mathcal{D}) < 0: \text{ overfitting.} \tag{12}$$

Optimal regression algorithms would reduce $D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}})$ until $D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}}) = D(\hat{P}_{\mathcal{D}}||P_{\boldsymbol{\theta}^\star})$ and then stop. Maximum likelihood regression will not do this; if it can reduce the Kullback–Leibler distance further it will do so, and thereby cause overfitting. For $N \to \infty$ we expect $E(\boldsymbol{\theta}^\star, \mathcal{D})$ to depend on the data $\mathcal{D}$ only via $P(\boldsymbol{z})$ and $\boldsymbol{\theta}^\star$, this is the fundamental assumption behind any regression. It allows us to focus on the average of $E(\boldsymbol{\theta}^\star, \mathcal{D})$ over all realisations of the data, given $P(\boldsymbol{z})$ and $\boldsymbol{\theta}^\star$:

$$E(\boldsymbol{\theta}^\star) = \Big\langle \min_{\boldsymbol{\theta}} \Big\{ \frac{1}{N} \sum_{i=1}^{N} \log \Big[\frac{P(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}^\star)}{P(t_i|\boldsymbol{z}_i, \boldsymbol{\theta})} \Big] \Big\} \Big\rangle_{\mathcal{D}} \tag{13}$$

in which

$$\begin{aligned}\langle F(t_1, \ldots, t_N; \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) \rangle_{\mathcal{D}} = \int \prod_{i=1}^{N} \Big[\mathrm{d}t_i \mathrm{d}\boldsymbol{z}_i \, P(\boldsymbol{z}_i) P(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}^\star) \Big] \\ \times F(t_1, \ldots, t_N; \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N).\end{aligned} \tag{14}$$

Evaluating $E(\boldsymbol{\theta}^\star)$ analytically for $N \to \infty$ is the focus of this paper. Clearly, if the relevant minimum over $\boldsymbol{\theta}$ corresponds to the true value $\boldsymbol{\theta}^\star$ for all $\mathcal{D}$, then $E(\boldsymbol{\theta}^\star) = 0$.

2.3. Analytical evaluation of the average over data sets

Working out (13) analytically for large N requires first that we deal with the minimisation over $\boldsymbol{\theta}$. This can be done by converting the problem into the calculation of the ground state energy for a statistical mechanical system with degrees of freedom $\boldsymbol{\theta} \in \mathbb{R}^q$ and Hamiltonian[6] $H(\boldsymbol{\theta}) = NE(\boldsymbol{\theta})$:

$$E(\boldsymbol{\theta}^\star) = \lim_{\gamma \to \infty} E_\gamma(\boldsymbol{\theta}^\star) \tag{15}$$

$$\begin{aligned} E_\gamma(\boldsymbol{\theta}^\star) &= -\frac{1}{N} \frac{\partial}{\partial \gamma} \Big\langle \log \int \mathrm{d}\boldsymbol{\theta} \, \mathrm{e}^{-\gamma \sum_{i=1}^{N} \log \left[\frac{P(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}^\star)}{P(t_i|\boldsymbol{z}_i, \boldsymbol{\theta})} \right]} \Big\rangle_{\mathcal{D}} \\ &= -\frac{1}{N} \frac{\partial}{\partial \gamma} \Big\langle \log \int \mathrm{d}\boldsymbol{\theta} \prod_{i=1}^{N} \Big[\frac{P(t_i|\boldsymbol{z}_i, \boldsymbol{\theta})}{P(t_i|\boldsymbol{z}_i, \boldsymbol{\theta}^\star)} \Big]^\gamma \Big\rangle_{\mathcal{D}}. \end{aligned} \tag{16}$$

For finite γ, the quantity $E_\gamma(\boldsymbol{\theta}^\star)$ can be interpreted as the average result of a *stochastic* minimisation, based on carrying out gradient descent on the function $-\log P(\mathcal{D}|\boldsymbol{\theta})$, supplemented by a Gaussian white noise with variance proportional to γ^{-1}.

The remaining obstacle is the logarithm in (16), which prevents the average over all data sets $\mathcal{D}$ from factorising over the samples. This we handle using the so-called replica method, which is based on the identity $\langle \log Z \rangle = \lim_{n \to 0} n^{-1} \log \langle Z^n \rangle$, and to our knowledge has not yet been applied in survival analysis. In the replica method the average $\langle Z^n \rangle$ is carried out for

[6] The rescaling with N of the Hamiltonian is done in anticipation of subsequent limits.

integer n, and the limit $n \to 0$ is taken at the end of the calculation via analytical continuation. Application to (16) leads us after some simple manipulations to a new expression in which the average over data sets *does* factorise over samples:

$$\begin{aligned}
E_\gamma(\boldsymbol{\theta}^\star) &= -\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\Big\langle\Big\{\int \mathrm{d}\boldsymbol{\theta}\prod_{i=1}^N\Big[\frac{P(t_i|\boldsymbol{z}_i,\boldsymbol{\theta})}{P(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star)}\Big]^\gamma\Big\}^n\Big\rangle_{\mathcal{D}} \\
&= -\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int \mathrm{d}\boldsymbol{\theta}^1\ldots\mathrm{d}\boldsymbol{\theta}^n\Big\langle\prod_{i=1}^N\prod_{\alpha=1}^n\Big[\frac{P(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\alpha)}{P(t_i|\boldsymbol{z}_i,\boldsymbol{\theta}^\star)}\Big]^\gamma\Big\rangle_{\mathcal{D}} \\
&= -\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int \mathrm{d}\boldsymbol{\theta}^1\ldots\mathrm{d}\boldsymbol{\theta}^n\Big\{\int \mathrm{d}\boldsymbol{z}\mathrm{d}t\,P(\boldsymbol{z})P(t|\boldsymbol{z},\boldsymbol{\theta}^\star) \\
&\quad\times\prod_{\alpha=1}^n\Big[\frac{P(t|\boldsymbol{z},\boldsymbol{\theta}^\alpha)}{P(t|\boldsymbol{z},\boldsymbol{\theta}^\star)}\Big]^\gamma\Big\}^N. \qquad (17)
\end{aligned}$$

The average over data sets has now been done, and we are left with a completely general explicit expression for $E(\boldsymbol{\theta}^\star)$ in terms of the covariate statistics $P(\boldsymbol{z})$ and the assumed parametrised data generating model $P(t|\boldsymbol{z},\boldsymbol{\theta})$. We will now work out and study this expression for Cox's proportional hazards model [5] with statistically independent zero-average Gaussian covariates.

2.4. Application to Cox regression

In Cox's method [5] the model parameters are a base hazard rate $\lambda(t)\geqslant 0$ (with $t\geqslant 0$) and a vector $\boldsymbol{\beta}\in\mathbb{R}^p$ of regression coefficients. The assumed event time statistics are then of the following form:

$$P(t|\boldsymbol{z},\boldsymbol{\beta},\lambda) = \lambda(t)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}/\sqrt{p}-\exp(\boldsymbol{\beta}\cdot\boldsymbol{z}/\sqrt{p})\Lambda(t)}, \quad \Lambda(t)=\int_0^t \mathrm{d}s\,\lambda(s). \qquad (18)$$

The factors $\sqrt{p}$ only induce an irrelevant scaling factor that will make it easier to take the limit $p\to\infty$. In fact, for large p it is inevitable that the typical association parameter in the Cox model will scale as $\mathcal{O}(p^{-\frac{1}{2}})$, since otherwise one would not find finite nonzero event times.

For simplicity we assume that the covariates are distributed according to $P(\boldsymbol{z})=(2\pi)^{-p/2}\exp(-\frac{1}{2}\boldsymbol{z}^2)$. This restriction of our analysis to uncorrelated covariates is no limitation, since for the Cox model one can always obtain, via a simple mapping, the regression results for data with correlated covariates from those obtained for uncorrelated covariates. This is demonstrated in appendix A.

For the Cox model our general result (17) takes the following form, involving ordinary integration over n-fold replicated vectors $\boldsymbol{\beta}^\alpha$ and functional integration over n-fold replicated base hazard rates λ^α:

$$\begin{aligned}
E_\gamma(\boldsymbol{\beta}^\star,\lambda^\star) &= -\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int\{\mathrm{d}\lambda_1\ldots\mathrm{d}\lambda_n\}\int\mathrm{d}\boldsymbol{\beta}^1\ldots\mathrm{d}\boldsymbol{\beta}^n \\
&\quad\times\Big\{\int\mathrm{d}\boldsymbol{z}\mathrm{d}t\,P(\boldsymbol{z})P(t|\boldsymbol{z},\boldsymbol{\beta}^\star,\lambda^\star)\prod_{\alpha=1}^n\Big[\frac{P(t|\boldsymbol{z},\boldsymbol{\beta}^\alpha,\lambda_\alpha)}{P(t|\boldsymbol{z},\boldsymbol{\beta}^\star,\lambda^\star)}\Big]^\gamma\Big\}^N. \qquad (19)
\end{aligned}$$

To enable efficient further analysis we define the short-hands

$$p(t|\xi,\lambda)=\lambda(t)\mathrm{e}^{\xi-\exp(\xi)\int_0^t\mathrm{d}s\,\lambda(s)} \qquad (20)$$

$$p(\boldsymbol{y}|\boldsymbol{\beta}^0,\ldots,\boldsymbol{\beta}^n)=\int \mathrm{d}\boldsymbol{z}\,P(\boldsymbol{z})\prod_{\alpha=0}^{n}\delta\Big[y_\alpha-\frac{\boldsymbol{\beta}^\alpha\cdot\boldsymbol{z}}{\sqrt{p}}\Big] \tag{21}$$

and the $n+1$-dimensional vector $\boldsymbol{y}=(y_0,\ldots,y_p)$. In addition we rename $(\boldsymbol{\beta}^\star,\lambda^\star)=(\boldsymbol{\beta}^0,\lambda^0)$, so that

$$\begin{aligned}E_\gamma(\boldsymbol{\beta}^0,\lambda_0)=&-\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int\{\mathrm{d}\lambda_1\ldots\mathrm{d}\lambda_n\}\int\mathrm{d}\boldsymbol{\beta}^1\ldots\mathrm{d}\boldsymbol{\beta}^n\\&\times\Big\{\int\mathrm{d}\boldsymbol{y}\,p(\boldsymbol{y}|\boldsymbol{\beta}^0,\ldots,\boldsymbol{\beta}^n)\int\mathrm{d}t\,p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^\gamma\Big\}^N.\end{aligned}\tag{22}$$

All $\{y_\alpha\}$ are linear combinations of Gaussian random variables, so also $p(\boldsymbol{y}|\boldsymbol{\beta}^0,\ldots,\boldsymbol{\beta}^n)$ will be Gaussian (even for most non-Gaussian covariates this would still hold for large p due to the central limit theorem), giving

$$p(\boldsymbol{y}|\boldsymbol{\beta}^0,\ldots,\boldsymbol{\beta}^n)=\frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}[\{\boldsymbol{\beta}\}]\boldsymbol{y}}}{\sqrt{(2\pi)^{n+1}\mathrm{Det}\boldsymbol{C}[\{\boldsymbol{\beta}\}]}}\tag{23}$$

in which the entries of the $(n+1)\times(n+1)$ covariance matrix $\boldsymbol{C}[\{\boldsymbol{\beta}\}]$ are

$$C_{\alpha\rho}[\{\boldsymbol{\beta}\}]=\frac{1}{p}\int\mathrm{d}\boldsymbol{z}\,P(\boldsymbol{z})(\boldsymbol{\beta}^\alpha\cdot\boldsymbol{z})(\boldsymbol{\beta}^\rho\cdot\boldsymbol{z})=\frac{1}{p}\boldsymbol{\beta}^\alpha\cdot\boldsymbol{\beta}^\rho.\tag{24}$$

We introduce integrals over δ-distributions to transport variables to more convenient places, by substituting for each pair (α,ρ):

$$1=\int\mathrm{d}C_{\alpha\rho}\,\delta\Big[C_{\alpha\rho}-C_{\alpha\rho}[\{\boldsymbol{\beta}\}]\Big]=\int\frac{\mathrm{d}C_{\alpha\rho}\mathrm{d}\hat{C}_{\alpha\rho}}{2\pi/p}\,\mathrm{e}^{\mathrm{i}p\hat{C}_{\alpha\rho}\big[C_{\alpha\rho}-C_{\alpha\rho}[\{\boldsymbol{\beta}\}]\big]}.\tag{25}$$

We then obtain, after some simple manipulations,

$$\begin{aligned}E_\gamma(\boldsymbol{\beta}^0,\lambda_0)=&-\frac{\partial}{\partial\gamma}\lim_{n\to 0}\frac{1}{Nn}\log\int\{\mathrm{d}\lambda_1\ldots\mathrm{d}\lambda_n\}\int\frac{\mathrm{d}\boldsymbol{C}\mathrm{d}\hat{\boldsymbol{C}}\,\mathrm{e}^{\mathrm{i}p\sum_{\alpha\rho=0}^{n}\hat{C}_{\alpha\rho}C_{\alpha\rho}}}{(2\pi/p)^{(n+1)^2}}\\&\times\Big\{\int\frac{\mathrm{d}\boldsymbol{y}\,\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}}{\sqrt{(2\pi)^{n+1}\mathrm{Det}\boldsymbol{C}}}\int\mathrm{d}t\,p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^\gamma\Big\}^N\\&\times\int\mathrm{d}\boldsymbol{\beta}^1\ldots\mathrm{d}\boldsymbol{\beta}^n\,\mathrm{e}^{-\mathrm{i}\sum_{\alpha\rho=0}^{n}\hat{C}_{\alpha\rho}\boldsymbol{\beta}^\alpha\cdot\boldsymbol{\beta}^\rho}.\end{aligned}\tag{26}$$

For finite N, expressions such as (26) are of course not easy to use, but as with all statistical theories we will be able to progress upon assuming N to be large[7]. We therefore focus on the asymptotic behaviour of (26) for $N\to\infty$, but with a fixed ratio p/N, and will confirm *a posteriori* the extent to which the resulting theory describes what is observed for large but finite sample sizes.

[7] Note that the standard use of Cox regression away from the overfitting regime, including its formulae for confidence intervals and for p-values (which require Gaussian approximations that build on large N expansions around the most probable parameter values, and assume that uncertainty in base hazard rates can be neglected), is similarly valid only when N is sufficiently large.

3. Asymptotic analysis of overfitting in the Cox model

3.1. Conversion to a saddle-point problem

Following extensive experience with the replica method in other disciplines, with similar definitions, we assume that the two limits $N \to \infty$ and $n \to 0$ commute. The invariance of the right-hand side of (26) under all permutations of the sample indices $i \in \{1, \ldots, N\}$ implies that $E(\boldsymbol{\beta}^0, \lambda_0)$ can depend on the true association parameters $\boldsymbol{\beta}^0$ only via the distribution $P(\beta_0) = p^{-1} \sum_{\mu=1}^{p} \delta[\beta_0 - \beta_\mu^0]$. With a modest amount of foresight we define $S^2 = p^{-1} \sum_{\mu=1}^{p} (\beta_\mu^0)^2$, and obtain

$$
\begin{aligned}
E_\gamma(P, \lambda_0) = &-\frac{\partial}{\partial\gamma} \lim_{n\to 0} \frac{1}{Nn} \log \int \{\mathrm{d}\lambda_1 \ldots \mathrm{d}\lambda_n\} \int \frac{\mathrm{d}\boldsymbol{C}\mathrm{d}\hat{\boldsymbol{C}}\, \mathrm{e}^{\mathrm{i}p\left(\sum_{\alpha\rho=0}^{n} \hat{C}_{\alpha\rho} C_{\alpha\rho} - \hat{C}_{00} S^2\right)}}{(2\pi/p)^{(n+1)^2}} \\
&\times \left\{ \int \frac{\mathrm{d}\boldsymbol{y}\, \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}}{\sqrt{(2\pi)^{n+1} \mathrm{Det}\boldsymbol{C}}} \int \mathrm{d}t\, p(t|y_0, \lambda_0) \prod_{\alpha=1}^{n} \left[\frac{p(t|y_\alpha, \lambda_\alpha)}{p(t|y_0, \lambda_0)}\right]^\gamma \right\}^N \\
&\times\ \mathrm{e}^{p \int \mathrm{d}\beta_0\ P(\beta_0) \log \int \mathrm{d}\beta_1 \ldots \mathrm{d}\beta_n\ \mathrm{e}^{-2\mathrm{i}\beta_0 \sum_{\rho=1}^{n} \hat{C}_{0\rho}\beta_\rho - \mathrm{i} \sum_{\alpha\rho=1}^{n} \hat{C}_{\alpha\rho}\beta_\alpha\beta_\rho}}.
\end{aligned} \tag{27}
$$

Writing the ratio of covariates over samples as $p/N = \zeta$, to be kept fixed in the limit $N \to \infty$, we may take the limit $N \to \infty$ and obtain an integral that can be evaluated using steepest descent:

$$
\begin{aligned}
\lim_{N\to\infty} E_\gamma(P, \lambda_0) = &-\frac{\partial}{\partial\gamma} \lim_{n\to 0} \lim_{N\to\infty} \frac{1}{Nn} \log \int \{\mathrm{d}\lambda_1 \ldots \mathrm{d}\lambda_n\} \\
&\times \mathrm{e}^{-\frac{1}{2}N \log[(2\pi)^{n+1} \mathrm{Det}\boldsymbol{C}]} \int \mathrm{d}\boldsymbol{C}\mathrm{d}\hat{\boldsymbol{C}}\ \mathrm{e}^{\mathrm{i}\zeta N(\sum_{\alpha\rho=0}^{n} \hat{C}_{\alpha\rho}C_{\alpha\rho} - \hat{C}_{00}S^2)} \\
&\times \mathrm{e}^{N \log \int \mathrm{d}\boldsymbol{y}\ \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}} \int \mathrm{d}t\ p(t|y_0,\lambda_0) \prod_{\alpha=1}^{n} \left[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\right]^\gamma} \\
&\times\ \mathrm{e}^{\zeta N \int \mathrm{d}\beta_0\ P(\beta_0) \log \int \mathrm{d}\beta_1 \ldots \mathrm{d}\beta_n\ \mathrm{e}^{-2\mathrm{i}\beta_0 \sum_{\rho=1}^{n} \hat{C}_{0\rho}\beta_\rho - \mathrm{i} \sum_{\alpha\rho=1}^{n} \hat{C}_{\alpha\rho}\beta_\alpha\beta_\rho}} \\
= &\frac{\partial}{\partial\gamma} \lim_{n\to 0} \frac{1}{n} \mathrm{extr}_{\boldsymbol{C},\hat{\boldsymbol{C}},\lambda_1,\ldots,\lambda_n} \Psi[\boldsymbol{C}, \hat{\boldsymbol{C}}; \lambda_1, \ldots, \lambda_n]
\end{aligned} \tag{28}
$$

in which the function to be extremized is

$$
\begin{aligned}
\Psi[\ldots] = &-\mathrm{i}\zeta\Big[\sum_{\alpha\rho=0}^{n} \hat{C}_{\alpha\rho}C_{\alpha\rho} - \hat{C}_{00}S^2\Big] + \frac{1}{2}(n+1)\log(2\pi) + \frac{1}{2}\log \mathrm{Det}\boldsymbol{C} \\
&- \zeta \int \mathrm{d}\beta_0\, P(\beta_0) \log \int \mathrm{d}\beta_1 \ldots \mathrm{d}\beta_n\ \mathrm{e}^{-2\mathrm{i}\beta_0 \sum_{\rho=1}^{n} \hat{C}_{0\rho}\beta_\rho - \mathrm{i} \sum_{\alpha\rho=1}^{n} \hat{C}_{\alpha\rho}\beta_\alpha\beta_\rho} \\
&- \log \int \mathrm{d}\boldsymbol{y}\ \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}} \int \mathrm{d}t\, p(t|y_0, \lambda_0) \prod_{\alpha=1}^{n} \left[\frac{p(t|y_\alpha, \lambda_\alpha)}{p(t|y_0, \lambda_0)}\right]^\gamma.
\end{aligned} \tag{29}
$$

Differentiation with respect to $\hat{C}_{00}$ immediately gives $C_{00} = S^2$. Moreover, for various integrals to be well-defined, the relevant saddle-point must (after contour deformation in the complex plane) be of a form where

$$\alpha,\rho=1\ldots n:\ \ \hat{C}_{\alpha\rho}=-\frac{1}{2}\mathrm{i}D_{\alpha\rho},\ \ \hat{C}_{0\rho}=-\frac{1}{2}\mathrm{i}d_\rho \tag{30}$$

with $D_{\alpha\rho},d_\rho\in\mathbb{R}$, and where the $n\times n$ matrix $\boldsymbol{D}=\{D_{\alpha\rho}\}$ is positive definite. Thus at the relevant saddle-point we will have

$$\begin{aligned}\Psi[\ldots]=&-\frac{1}{2}\zeta\sum_{\alpha\rho=1}^{n}D_{\alpha\rho}C_{\alpha\rho}-\zeta\sum_{\rho=1}^{n}d_\rho C_{0\rho}+\frac{1}{2}(n+1)\log(2\pi)+\frac{1}{2}\log\mathrm{Det}\boldsymbol{C}\\&-\log\int\mathrm{d}\boldsymbol{y}\ \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}\int\mathrm{d}t\,p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^{\gamma}\\&-\ \zeta\int\mathrm{d}\beta_0\,P(\beta_0)\log\int\mathrm{d}\beta_1\ldots\mathrm{d}\beta_n\ \mathrm{e}^{-\beta_0\sum_{\rho=1}^{n}d_\rho\beta_\rho-\frac{1}{2}\sum_{\alpha\rho=1}^{n}D_{\alpha\rho}\beta_\alpha\beta_\rho}\\=&-\frac{1}{2}\zeta\sum_{\alpha\rho=1}^{n}D_{\alpha\rho}C_{\alpha\rho}-\zeta\sum_{\rho=1}^{n}d_\rho C_{0\rho}-\frac{1}{2}\zeta S^2\sum_{\alpha\rho=1}^{n}d_\alpha(\boldsymbol{D}^{-1})_{\alpha\rho}d_\rho\\&+\frac{1}{2}(n+1)\log(2\pi)+\frac{1}{2}\log\mathrm{Det}\boldsymbol{C}\\&-\log\int\mathrm{d}\boldsymbol{y}\ \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}\int\mathrm{d}t\,p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^{\gamma}\\&-\ \zeta\log\int\mathrm{d}\beta_1\ldots\mathrm{d}\beta_n\ \mathrm{e}^{-\frac{1}{2}\sum_{\alpha\rho=1}^{n}D_{\alpha\rho}\beta_\alpha\beta_\rho}.\end{aligned}\tag{31}$$

Variation with respect to the n components $\{d_\alpha\}$ gives $d_\alpha=-S^{-2}\sum_\rho D_{\alpha\rho}C_{0\rho}$, so

$$\begin{aligned}\Psi[\ldots]=&-\frac{1}{2}\zeta\sum_{\alpha\rho=1}^{n}D_{\alpha\rho}\Big[C_{\alpha\rho}-\frac{C_{0\alpha}C_{0\rho}}{S^2}\Big]+\frac{1}{2}(n+1)\log(2\pi)+\frac{1}{2}\log\mathrm{Det}\boldsymbol{C}\\&-\log\int\mathrm{d}\boldsymbol{y}\ \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}\int\mathrm{d}t\,p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^{\gamma}\\&-\ \zeta\log\int\mathrm{d}\beta_1\ldots\mathrm{d}\beta_n\ \mathrm{e}^{-\frac{1}{2}\sum_{\alpha\rho=1}^{n}D_{\alpha\rho}\beta_\alpha\beta_\rho}.\end{aligned}\tag{32}$$

This intermediate result confirms that $\lim_{N\to\infty}E_\gamma(P,\lambda_0)$ indeed depends on the distribution $P(\beta_0)$ only via $S^2=\int\mathrm{d}\beta_0\ P(\beta_0)\beta_0^2$, hence we may henceforth write the former quantity as $E_\gamma(S,\lambda_0)$. Variation with respect to $\boldsymbol{D}$ finally gives $(\boldsymbol{D}^{-1})_{\alpha\rho}=C_{\alpha\rho}-C_{0\alpha}C_{0\rho}/S^2$. Hence we arrive at the following expression, in which the short-hand $\boldsymbol{C}'$ denotes the $n\times n$ matrix with entries $C'_{\alpha\rho}=C_{\alpha\rho}-C_{0\alpha}C_{0\rho}/S^2$ (for $\alpha,\rho=1\ldots n$):

$$E_\gamma(S,\lambda_0)=\frac{\partial}{\partial\gamma}\lim_{n\to0}\frac{1}{n}\mathrm{extr}\,\boldsymbol{C};\lambda_1,\ldots,\lambda_n\Psi[\boldsymbol{C};\lambda_1,\ldots,\lambda_n].\tag{33}$$

$$\begin{aligned}\Psi[\boldsymbol{C};\lambda_1,\ldots,\lambda_n]=&\frac{1}{2}\log\mathrm{Det}\boldsymbol{C}-\frac{1}{2}\zeta\log\mathrm{Det}\boldsymbol{C}'\\&-\log\int\frac{\mathrm{d}\boldsymbol{y}}{\sqrt{2\pi}}\ \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}\int\mathrm{d}t\,p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda_\alpha)}{p(t|y_0,\lambda_0)}\Big]^{\gamma}.\end{aligned}\tag{34}$$

The extremisation over $\boldsymbol{C}$ is to be done subject to $C_{00}=S^2$, and we have removed from $\Psi[\dots]$ those terms that will vanish after taking $n\to 0$ and differentiating with respect to γ.

3.2. Replica symmetric extrema

The replica symmetry ansatz (RS) can be translated into the statement that the solution space of the regression algorithm is ergodic [18, 25, 28], i.e. the typical set of equivalent minima in regression parameter space is connected. Replica symmetric saddle-points of (34) are of the following form:

$$\forall\alpha,\rho=1\dots n:\ \ \lambda_\alpha(t)=\lambda(t),\ \ C_{00}=S^2,\ \ C_{0\alpha}=c_0, \tag{35}$$

$$C_{\alpha\rho}=C\delta_{\alpha\rho}+c(1-\delta_{\alpha\rho}). \tag{36}$$

In appendix B we derive the equations corresponding to the RS ansatz for the stochastic generalization of the Cox model. With the short-hand $\mathrm{D}y=(2\pi)^{-1/2}\mathrm{e}^{-\frac{1}{2}y^2}\mathrm{d}y$, and upon removing terms that vanish upon differentiation by γ, we can summarise these equations in the limit of large data sets, by the following compact expression:

$$\begin{aligned}E_\gamma(S,\lambda_0)=&\int\mathrm{D}y_0\int\mathrm{d}t\,p(t|Sy_0,\lambda_0)\,\Big\{\log p(t|Sy_0,\lambda_0)\\&-\int\mathrm{D}z\left[\frac{\int\mathrm{D}y\,p^\gamma(t|uy+wy_0+vz,\lambda)\log p(t|uy+wy_0+vz,\lambda)}{\int\mathrm{D}y\,p^\gamma(t|uy+wy_0+vz,\lambda)}\right]\Big\}\end{aligned} \tag{37}$$

in which the order parameters $\{u,v,w;\lambda\}$, which are related to the RS order parameters $\{C,c_0,c\}$ via

$$c_0=Sw,\quad c=v^2+w^2,\quad C=u^2+v^2+w^2, \tag{38}$$

are to be evaluated at the saddle point of

$$\begin{aligned}\Psi_{\mathrm{RS}}(u,v,w;\lambda)=&\,\zeta\Big(\frac{v^2}{2u^2}+\log u\Big)\\&+\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\,p(t|Sy_0,\lambda_0)\log\int\mathrm{D}y\,p^\gamma(t|uy+wy_0+vz,\lambda).\end{aligned} \tag{39}$$

3.3. Physical interpretation of order parameters

The physical meaning of the order parameters in the replica symmetric matrix $\boldsymbol{C}$ is found in the usual manner for replica calculations [25], by direct application of our manipulations to the calculation of observables. We will write averages over the stochastic maximization of the data log-likelihood at finite γ, for a fixed training set $\mathcal{D}$, as $\langle\dots\rangle$, and averages over all data sets (as before) as $\langle\dots\rangle_{\mathcal{D}}$. Since the relevant quantities in the theory are found asymptotically to depend on the true association vector $\boldsymbol{\beta}^\star$ only via $S^2=p^{-1}\sum_{\mu=1}^p(\beta_\mu^\star)^2$, there is no need for explicit averages over $\boldsymbol{\beta}^\star$. This results upon application to the Cox model in the following identifications, in the limit $n\to 0$:

$$c_0=\lim_{p\to\infty}\frac{1}{p}\boldsymbol{\beta}^\star\cdot\langle\langle\boldsymbol{\beta}\rangle\rangle_{\mathcal{D}},\quad c=\lim_{p\to\infty}\frac{1}{p}\langle\langle\boldsymbol{\beta}\rangle^2\rangle_{\mathcal{D}},\quad C=\lim_{p\to\infty}\frac{1}{p}\langle\langle\boldsymbol{\beta}^2\rangle\rangle_{\mathcal{D}}. \tag{40}$$

In terms of the transformed order parameters (u,v,w) this becomes

$$u^2 = \lim_{p\to\infty} \frac{1}{p}\langle\langle\boldsymbol{\beta}^2\rangle - \langle\boldsymbol{\beta}\rangle^2\rangle_{\mathcal{D}} \tag{41}$$

$$v^2 = \lim_{p\to\infty} \frac{1}{p}\Big[\langle\langle\boldsymbol{\beta}\rangle^2\rangle_{\mathcal{D}} - \Big(\frac{\boldsymbol{\beta}^\star\cdot\langle\langle\boldsymbol{\beta}\rangle\rangle_{\mathcal{D}}}{|\boldsymbol{\beta}^\star|}\Big)^2\Big] \tag{42}$$

$$w = \lim_{p\to\infty} \frac{1}{\sqrt{p}}\frac{\boldsymbol{\beta}^\star\cdot\langle\langle\boldsymbol{\beta}\rangle\rangle_{\mathcal{D}}}{|\boldsymbol{\beta}^\star|}. \tag{43}$$

Here $\boldsymbol{\beta}$ is the outcome of maximum likelihood regression for data set $\mathcal{D}$ generated with true association parameters $\boldsymbol{\beta}^\star$. Fully random parameter guessing would give $c_0 = c = 0$ and $C > 0$. Perfect regression would imply $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ for all $\mathcal{D}$ and all $\boldsymbol{\beta}^\star$, and hence correspond to $c_0 = c = C = S^2$, giving $u = v = 0$ and $w = S$. It is reassuring to observe that for $\zeta = 0$, expression (37) indeed reproduces $E_\gamma(S, \lambda_0) = 0$ if in the right-hand side we substitute the values $u = v = 0$ and $w = S$.

From (40) follow useful inequalities that must hold at the relevant saddle-point in the limit $n \to 0$, which are consistent with our claim that $u, v, w \geqslant 0$:

$$C \geqslant 0, \quad c \geqslant 0, \quad c_0 \geqslant 0, \quad C \geqslant c, \quad c \geqslant c_0^2/S^2. \tag{44}$$

The first four inequalities are easy to derive. The fifth follows from:

$$\begin{aligned} c = \lim_{p\to\infty}\frac{1}{p}\langle\langle\boldsymbol{\beta}\rangle^2\rangle_{\mathcal{D}} &\geqslant \lim_{p\to\infty}\frac{1}{p}\Big\langle\Big(\frac{\boldsymbol{\beta}^\star}{|\boldsymbol{\beta}^\star|}\cdot\langle\boldsymbol{\beta}\rangle\Big)^2\Big\rangle_{\mathcal{D}} \\ &= \frac{1}{p}\Big(\frac{p}{|\boldsymbol{\beta}^\star|}c_0\Big)^2 = c_0^2/S^2. \end{aligned} \tag{45}$$

If, as suggested by the $\gamma \to \infty$ simulation results shown in appendix A, $\langle\boldsymbol{\beta}\rangle \approx \kappa\boldsymbol{\beta}^\star + \boldsymbol{\xi}$ for some $\kappa > 0$, with a zero-average random vector $\boldsymbol{\xi}$ that reflects data set variability, such that $\langle\boldsymbol{\xi}\rangle_{\mathcal{D}} = \mathbf{0}$ and with amplitude $\lim_{p\to\infty} p^{-1}\sum_{\mu=1}^{p}\langle\xi_\mu^2\rangle_{\mathcal{D}} = \sigma^2$, then we would find the RS saddle point obeying $c_0 = \kappa S^2$ and $c = \kappa^2 S^2 + \sigma^2$. Hence we would find $v = \sigma$ and $\kappa = w/S$, and we would expect $\lim_{\gamma\to\infty} u = 0$ for $\zeta < 1$. Note that the above relations are true given our definition of the event time distribution as $P(t|\mathbf{z}, \boldsymbol{\beta}, \lambda) = -\frac{\mathrm{d}}{\mathrm{d}t}\exp[-\exp(\boldsymbol{\beta}\cdot\mathbf{z}/\sqrt{p})\Lambda(t)]$. If we were to define this distribution instead without the rescaling factor $\sqrt{p}$ as $P(t|\mathbf{z}, \boldsymbol{\beta}, \lambda) = -\frac{\mathrm{d}}{\mathrm{d}t}\exp[-\exp(\boldsymbol{\beta}\cdot\mathbf{z})\Lambda(t)]$ (which is the convention of [5]), then the connection between regression of the form $\langle\boldsymbol{\beta}\rangle \approx \kappa\boldsymbol{\beta}^\star + \boldsymbol{\xi}$ and our order parameters would be:

$$\kappa = w/S, \quad \sigma = v/\sqrt{p}. \tag{46}$$

We conclude that from our RS equations we can extract the dependence on the covariates/samples ratio $\zeta = p/N$ of the two main quantitative characteristics of the data clouds in figure 2: their angle κ and their width σ.

Finally, let us turn to the interpretation of equation (37). We observe that this equation can be written as

$$E_\gamma(S, \lambda_0) = \int \mathrm{d}t\mathrm{d}x\mathrm{d}x'\ \mathcal{P}_\gamma(x, x', t)\log\Big[\frac{p(t|x, \lambda_0)}{p(t|x', \lambda)}\Big] \tag{47}$$

$$\begin{aligned} \mathcal{P}_\gamma(x, x', t) = &\int \mathrm{D}z\mathrm{D}y_0\ \delta[x - Sy_0]\,p(t|Sy_0, \lambda_0) \\ &\times\Big[\frac{\int \mathrm{D}y\ p^\gamma(t|uy + wy_0 + vz, \lambda)\ \delta[x' - uy - wy_0 - vz]}{\int \mathrm{D}y\ p^\gamma(t|uy + wy_0 + vz, \lambda)}\Big]. \end{aligned} \tag{48}$$

If we compare expression (47) with the definition of $E_\gamma(S,\lambda_0)$, which for the Cox model is

$$E_\gamma(S,\lambda_0)=\lim_{N\to\infty}\Big\langle\Big\langle\frac{1}{N}\sum_{i=1}^{N}\log\Big[\frac{p(t_i|\boldsymbol{\beta}^\star\cdot\boldsymbol{z}_i/\sqrt{p},\lambda_0)}{p(t_i|\boldsymbol{\beta}\cdot\boldsymbol{z}_i/\sqrt{p},\lambda)}\Big]\Big\rangle\Big\rangle_{\mathcal{D}} \tag{49}$$

we can infer that

$$\mathcal{P}_\gamma(x,x',t)=\lim_{N\to\infty}\Big\langle\Big\langle\frac{1}{N}\sum_{i=1}^{N}\delta[t-t_i]\;\delta\Big[x-\frac{\boldsymbol{\beta}^\star\cdot\boldsymbol{z}_i}{\sqrt{p}}\Big]\,\delta\Big[x'-\frac{\boldsymbol{\beta}\cdot\boldsymbol{z}_i}{\sqrt{p}}\Big]\Big\rangle\Big\rangle_{\mathcal{D}}. \tag{50}$$

As a consistency test one can confirm that, as an alternative to retracing the replica derivation, the expressions (40) can also be derived explicitly from (48) and (50).

3.4. Derivation of RS saddle point equations

The equations from which to solve the replica symmetric order parameters (u,v,w,λ) are obtained by extremization of (39). Using $\partial\log p(t|\xi)/\partial\xi=1-\mathrm{e}^{\xi}\Lambda(t)$, the three scalar equations are found to be

$$\frac{\zeta}{\gamma u}\Big(\frac{v^2}{u^2}-1\Big)=\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\;p(t|Sy_0,\lambda_0)\times\frac{\int\mathrm{D}y\;y\,p^\gamma(t|uy+wy_0+vz,\lambda)\Big[1-\mathrm{e}^{uy+wy_0+vz}\Lambda(t)\Big]}{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)} \tag{51}$$

$$\zeta\frac{v}{\gamma u^2}=\int\mathrm{D}z\mathrm{D}y_0\;z\int\mathrm{d}t\;p(t|Sy_0,\lambda_0)\Lambda(t)\frac{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)\mathrm{e}^{uy+wy_0+vz}}{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)} \tag{52}$$

$$0=\int\mathrm{D}z\mathrm{D}y_0\;y_0\int\mathrm{d}t\;p(t|Sy_0,\lambda_0)\Lambda(t)\frac{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)\mathrm{e}^{uy+wy_0+vz}}{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)}. \tag{53}$$

Upon integrating by parts over y, we can also write equation (51) as

$$\frac{\zeta}{\gamma u^2}\Big(\frac{v^2}{\gamma u^2}-\frac{1}{\gamma}\Big)=\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\;p(t|Sy_0,\lambda_0)\times\frac{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)\Big[[1-\mathrm{e}^{uy+wy_0+vz}\Lambda(t)]^2-\gamma^{-1}\mathrm{e}^{uy+wy_0+vz}\Lambda(t)\Big]}{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)}. \tag{54}$$

To work out the functional order parameter equation $\delta\Psi_{\mathrm{RS}}(u,v,w;\lambda)/\delta\lambda(s)=0$ we use $\delta\log p(t|\xi)/\delta\lambda(s)=\delta(t-s)/\lambda(s)-\mathrm{e}^{\xi}\theta(t-s)$, and the abbreviation $p(t)=\int\mathrm{D}y_0\;p(t|Sy_0,\lambda_0)$. This gives

$$\begin{aligned}0&=\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\;p(t|Sy_0,\lambda_0)\frac{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)\Big[\frac{\delta(t-s)}{\lambda(s)}-\mathrm{e}^{uy+wy_0+vz}\theta(t-s)\Big]}{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)}\\&=\frac{p(s)}{\lambda(s)}-\int\mathrm{D}z\mathrm{D}y_0\int_s^\infty\mathrm{d}t\;p(t|Sy_0,\lambda_0)\frac{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)\mathrm{e}^{uy+wy_0+vz}}{\int\mathrm{D}y\;p^\gamma(t|uy+wy_0+vz,\lambda)}.\end{aligned} \tag{55}$$

This latter equation can also be written in terms of the distribution (48), giving a form that reduces to Breslow's [16] estimator when we subsequently use the interpretation identity (50):

$$\lambda(t) = \frac{\int \mathrm{d}x\mathrm{d}x'\ \mathcal{P}_\gamma(x,x',t)}{\int_t^\infty \mathrm{d}t' \int \mathrm{d}x\mathrm{d}x'\ \mathcal{P}_\gamma(x,x',t)\mathrm{e}^{x'}}. \qquad (56)$$

The remaining integrations over y in our equations are for finite γ quite nontrivial. They can be expressed in terms of the Laplace transform of the lognormal distribution [36], or mapped onto the core integral in the Random Energy Model [37], both of which could in the past be evaluated analytically only in specific parameter limits.

4. Analysis of the RS equations for the Cox model

4.1. RS equations in the limit $\gamma \to \infty$

The original Cox model [5] corresponds to the limit $\gamma \to \infty$ of our equations. It turns out that the correct scaling with γ of u for $\gamma \to \infty$ is $u = \tilde{u}/\sqrt{\gamma}$; this is suggested by equation (54) and confirms our expectation that follows from the physical meaning of u. Upon substituting $u = \tilde{u}/\sqrt{\gamma}$ as an ansatz into our equations, assuming the other order parameters to have finite $\gamma \to \infty$ limits, allows us to simplify the trio (52)–(54) and the functional equation (55) to

$$\frac{\zeta v}{\tilde{u}^2} = \int \mathrm{D}z\mathrm{D}y_0\ z \int \mathrm{d}t\ p(t|Sy_0,\lambda_0)\Lambda(t)A_1(wy_0+vz,t) \qquad (57)$$

$$0 = \int \mathrm{D}z\mathrm{D}y_0\ y_0 \int \mathrm{d}t\ p(t|Sy_0,\lambda_0)\Lambda(t)A_1(wy_0+vz,t) \qquad (58)$$

$$\frac{\zeta v^2}{\tilde{u}^4} = 1 + \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\ p(t|Sy_0,\lambda_0)\Big[\Lambda^2(t)A_2(y_0,z,t) \\ - 2\Lambda(t)A_1(wy_0+vz,t)\Big] \qquad (59)$$

$$\frac{p(t)}{\lambda(t)} = \int \mathrm{D}z\mathrm{D}y_0 \int_t^\infty \mathrm{d}t'\ p(t'|Sy_0,\lambda_0)A_1(wy_0+vz,t'). \qquad (60)$$

The remaining complexities of the limit are concentrated in

$$\begin{aligned} A_r(\eta,t) &= \lim_{\gamma\to\infty} \frac{\int \mathrm{D}y\ p^\gamma(t|uy+\eta,\lambda)\mathrm{e}^{r(uy+\eta)}}{\int \mathrm{D}y\ p^\gamma(t|uy+\eta,\lambda)} \\ &= \lim_{\gamma\to\infty} \frac{\int \mathrm{d}y\ \mathrm{e}^{-\frac{1}{2}y^2+\gamma\left[uy+\eta-\mathrm{e}^{uy+\eta}\Lambda(t)\right]+r(uy+\eta)}}{\int \mathrm{d}y\ \mathrm{e}^{-\frac{1}{2}y^2+\gamma\left[uy+\eta-\mathrm{e}^{uy+\eta}\Lambda(t)\right]}} \\ &= \lim_{\gamma\to\infty} \frac{\int \mathrm{d}q\ \mathrm{e}^{\gamma\left[-\frac{1}{2}q^2+\tilde{u}q+\eta-\mathrm{e}^{\tilde{u}q+\eta}\Lambda(t)\right]+r(\tilde{u}q+wy_0+vz)}}{\int \mathrm{d}q\ \mathrm{e}^{\gamma\left[-\frac{1}{2}q^2+\tilde{u}q+\eta-\mathrm{e}^{\tilde{u}q+\eta}\Lambda(t)\right]}} \\ &= \left[\mathrm{e}^{\varphi(wy_0+vz,t)\tilde{u}+wy_0+vz}\right]^r \end{aligned} \qquad (61)$$

with

$$\varphi(\eta,t)=\mathrm{argmax}_q\Big\{-\frac{1}{2}q^2+\tilde{u}q+\eta-\mathrm{e}^{\tilde{u}q+\eta}\Lambda(t)\Big\}. \tag{62}$$

After differentiation and rewriting the resulting equation, we find that $\varphi(\eta,t)$ can be written in explicit form in terms of the Lambert W-function [35] as:

$$\varphi(\eta,t)=\tilde{u}-\tilde{u}^{-1}W\Big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+\eta}\Lambda(t)\Big) \tag{63}$$

Hence

$$A_r(\eta,t)=\mathrm{e}^{r\left[\tilde{u}^2+\eta-W\left(\tilde{u}^2\exp(\tilde{u}^2+\eta)\Lambda(t)\right)\right]}. \tag{64}$$

Using the identity $\mathrm{e}^{-W(z)}=W(z)/z$, which follows directly from the definition of the Lambert W-function, we can simplify the above result to

$$A_r(\eta,t)=\Big(\frac{W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+\eta}\Lambda(t)\big)}{\tilde{u}^2\Lambda(t)}\Big)^r. \tag{65}$$

Substitution into our $\gamma\to\infty$ order parameter equations finally gives:

$$\zeta v^2=\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\,p(t|Sy_0,\lambda_0)\Big[\tilde{u}^2-W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big)\Big]^2 \tag{66}$$

$$\zeta v=\int\mathrm{D}z\mathrm{D}y_0\;z\int\mathrm{d}t\,p(t|Sy_0,\lambda_0)W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big) \tag{67}$$

$$0=\int\mathrm{D}z\mathrm{D}y_0\;y_0\int\mathrm{d}t\,p(t|Sy_0,\lambda_0)W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big) \tag{68}$$

$$\frac{p(t)}{\lambda(t)}=\int\mathrm{D}z\mathrm{D}y_0\int_t^\infty\mathrm{d}t'\;p(t'|Sy_0,\lambda_0)\frac{W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t')\big)}{\tilde{u}^2\Lambda(t')}. \tag{69}$$

We observe that the choice $v=0$ always solves (67), but that for $\zeta>0$ it is ruled out by (66). Upon doing integration by parts over z, using $\mathrm{d}W(z)/\mathrm{d}z=W(z)/z[1+W(z)]$ and dismissing the solution $v=0$, we can simplify equation (67) further to

$$\zeta=\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\,p(t|Sy_0,\lambda_0)\frac{W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big)}{1+W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big)}. \tag{70}$$

To compute the corresponding value of the overfitting measure $E(S,\lambda_0)=\lim_{\gamma\to\infty}E_\gamma(S,\lambda_0)$, we substitute $u=\tilde{u}/\sqrt{\gamma}$ into (37) and take the limit $\gamma\to\infty$. This gives, using the short-hands (63) and $p(t)=\int\mathrm{D}y_0\;p(t|Sy_0,\lambda_0)$ and the identity $\exp[-W(z)]=W(z)/z$:

$$
\begin{aligned}
E(S,\lambda_0) &= \int \mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0,\lambda_0)\, \{\log p(t|Sy_0,\lambda_0) - \log\lambda(t) \\
&\quad - \lim_{\gamma\to\infty} \int \mathrm{D}z \frac{\int \mathrm{d}y\; \mathrm{e}^{\gamma[\tilde{u}y+wy_0+vz-\mathrm{e}^{\tilde{u}y+wy_0+vz}\Lambda(t)-\frac{1}{2}y^2]}\Big[\tilde{u}y+wy_0+vz-\mathrm{e}^{\tilde{u}y+wy_0+vz}\Lambda(t)\Big]}{\int \mathrm{d}y\; \mathrm{e}^{\gamma[\tilde{u}y+wy_0+vz-\mathrm{e}^{\tilde{u}y+wy_0+vz}\Lambda(t)-\frac{1}{2}y^2]}} \Bigg\} \\
&= \int \mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0,\lambda_0)\, \big\{\log[\lambda_0(t)/\lambda(t)] - \mathrm{e}^{Sy_0}\Lambda_0(t) \\
&\quad - \int \mathrm{D}z \Big[\tilde{u}\varphi(wy_0+vz,t) - \mathrm{e}^{\tilde{u}\varphi(wy_0+vz,t)+wy_0+vz}\Lambda(t)\Big]\Big\} \\
&= \int \mathrm{d}t\, p(t)\log\Big[\frac{\lambda_0(t)}{\lambda(t)}\Big] - \int \mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0,\lambda_0)\mathrm{e}^{Sy_0}\Lambda_0(t) - \tilde{u}^2 \\
&\quad +\ (1+\frac{1}{\tilde{u}^2}) \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0,\lambda_0) W\Big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\Big).
\end{aligned}
\tag{71}
$$

The second integral can be worked out explicitly:

$$
\begin{aligned}
&\int \mathrm{D}y_0 \int_0^\infty \mathrm{d}t\, p(t|Sy_0,\lambda_0)\mathrm{e}^{Sy_0}\Lambda_0(t) \\
&\qquad = -\int \mathrm{D}y_0 \int_0^\infty \mathrm{d}t\; \mathrm{e}^{Sy_0}\Lambda_0(t)\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{-\exp(Sy_0)\Lambda_0(t)} \\
&\qquad = \int_0^\infty \mathrm{d}x\, x\mathrm{e}^{-x} \;=\; 1.
\end{aligned}
\tag{72}
$$

Therefore

$$
\begin{aligned}
E(S,\lambda_0) &= \int \mathrm{d}t\, p(t)\log\Big[\frac{\lambda_0(t)}{\lambda(t)}\Big] \\
&\quad - (1+\tilde{u}^2)\Big[1 - \frac{1}{\tilde{u}^2}\int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0,\lambda_0) W\Big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\Big)\Big]
\end{aligned}
\tag{73}
$$

In appendix C we study the behaviour of the above equations in the two limits $\zeta \to 0$ and $\zeta \to 1$. For $\zeta \to 0$ we recover the correct solution corresponding to perfect (overfitting-free) regression, as required. For $\zeta \to 1$ we find a phase transition, characterised by divergence of the order parameters $\{\tilde{u}, v, w\}$.

4.2. Numerical and asymptotic solution of RS equations

Solving the coupled order parameter equations (66) and (68)–(70) analytically seems for now too ambitious; solving them numerically is nontrivial, and requires some preparation. To cast the equation for w into a form similar to the others, we need to do partial integration over y_0:

$$
\begin{aligned}
0 &= w\int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0,\lambda_0)\frac{W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big)}{1+W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big)} \\
&\quad +\ S\int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0,\lambda_0) W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big)\Big[1-\mathrm{e}^{Sy_0}\Lambda_0(t)\Big].
\end{aligned}
\tag{74}
$$

We also rewrite the functional equation in a form that involves $\Lambda(t)$ only:

$$\Lambda(t)=\int_0^t \mathrm{d}t'\,p(t')\Big\{\int \mathrm{D}z\mathrm{D}y_0\int_{t'}^{\infty}\mathrm{d}t''\,p(t''|Sy_0,\lambda_0)\frac{W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t'')\big)}{\tilde{u}^2\Lambda(t'')}\Big\}^{-1}. \tag{75}$$

Numerical integration over $t>0$ can be transformed into integration over the survival function $s(t,y_0)=\exp[-\mathrm{e}^{Sy_0}\Lambda_0(t)]\in[0,1]$, using $p(t|Sy_0,\lambda_0)\mathrm{d}t=-\mathrm{d}s$ and $t(s,y_0)=\Lambda_0^{\mathrm{inv}}(\mathrm{e}^{-Sy_0}\log(1/s))$. We also define the short-hand $L(t)=\tilde{u}^2\mathrm{e}^{\tilde{u}^2}\Lambda(t)$. These definitions transform our RS equations to:

$$\zeta v^2=\int \mathrm{D}y_0\mathrm{D}z\int_0^1\mathrm{d}s\,\Big[\tilde{u}^2-W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big)\Big]^2 \tag{76}$$

$$\zeta=\int \mathrm{D}y_0\mathrm{D}z\int_0^1\mathrm{d}s\,\left\{\frac{W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big)}{1+W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big)}\right\} \tag{77}$$

$$\frac{\zeta w}{S}=-\int \mathrm{D}y_0\mathrm{D}z\int_0^1\mathrm{d}s\,\big[1+\log(s)\big]W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big) \tag{78}$$

$$L(t)=\tilde{u}^2\int_0^t\mathrm{d}t'\,p(t')\times\left\{\int \mathrm{D}y_0\mathrm{D}z\int_0^1\mathrm{d}s'\,\frac{\theta[t(s',y_0)-t']}{L(t(s',y_0))}\,W\Big(\mathrm{e}^{wy_0+vz}L(t(s',y_0))\Big)\right\}^{-1}. \tag{79}$$

We next study the functional equation (79) in more detail. We first rewrite it by differentiation with respect to time, and some simple rearrangements, into the more suitable form

$$\tilde{u}^2\frac{p(t)}{\frac{\mathrm{d}}{\mathrm{d}t}L(t)}=\int \mathrm{D}y_0\mathrm{D}z\int_0^1\mathrm{d}s\,\frac{\theta[t(s,y_0)-t]}{L(t(s,y_0))}\,W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big) \tag{80}$$

or, upon further differentiation:

$$-\tilde{u}^2L(t)\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{p(t)}{\frac{\mathrm{d}}{\mathrm{d}t}L(t)}\Big)=\int \mathrm{D}y_0\mathrm{D}z\,W\Big(\mathrm{e}^{wy_0+vz}L(t)\Big)\int_0^1\mathrm{d}s\,\delta[t(s,y_0)-t]. \tag{81}$$

Using $\int_0^1\mathrm{d}s\,\delta[t(s,y_0)-t]=p(t|Sy_0)$, and upon multiplying both sides by $\frac{\mathrm{d}}{\mathrm{d}t}L(t)/p(t)$, this becomes

$$\tilde{u}^2\frac{\mathrm{d}}{\mathrm{d}t}\log\Big(\frac{\mathrm{d}L(t)/\mathrm{d}t}{p(t)}\Big)=\frac{\mathrm{d}\log L(t)}{\mathrm{d}t}\int \mathrm{D}y_0\,\frac{p(t|Sy_0)}{p(t)}\int \mathrm{D}z\,W\Big(\mathrm{e}^{wy_0+vz}L(t)\Big). \tag{82}$$

We write $L(t)$ in the form $L(t)=\Phi(\Lambda_0(t))$, which is always possible since both $L(t)$ and $\Lambda_0(t)$ are monotonic functions of time, and we write $p(t)=\lambda_0(t)g(\Lambda_0(t))$ with

$$g(x)=\int \mathrm{D}y_0\,\mathrm{e}^{Sy_0-x\exp(Sy_0)}. \tag{83}$$

Substitution of these conventions, and working out the various time derivatives, then leads to the following equation from which to solve $\Phi(x)$:

$$\frac{\tilde{u}^2 g(x)}{\mathrm{d}\log\Phi(x)/\mathrm{d}x}\,\frac{\mathrm{d}}{\mathrm{d}x}\log\Big(\frac{\mathrm{d}\Phi(x)/\mathrm{d}x}{g(x)}\Big) = \int \mathrm{D}y_0\;\mathrm{e}^{Sy_0-x\exp(Sy_0)} \times \int \mathrm{D}z\;W\Big(\mathrm{e}^{wy_0+vz}\Phi(x)\Big). \tag{84}$$

We now proceed to calculate the solution $\Phi(x)$ of the above equation, which gives us the form of the inferred integrated base hazard rates $\Lambda(t)$ as shown in figure 3, for large times, i.e. in the regime where $x\to\infty$ and $\Phi(x)\to\infty$. Here we can use use the asymptotic form of the Lambert W-function [35]: $W(z)=\log z-\log(\log z)+\mathcal{O}(\log(\log z)/\log z)$ (for $z\to\infty$), to obtain

$$\frac{\tilde{u}^2 g(x)}{\mathrm{d}\log\Phi(x)/\mathrm{d}x}\,\frac{\mathrm{d}}{\mathrm{d}x}\log\Big(\frac{\mathrm{d}\Phi(x)/\mathrm{d}x}{g(x)}\Big) = g(x)\log\Big(\frac{\Phi(x)}{\log\Phi(x)}\Big)+w\int \mathrm{D}y_0\;y_0\mathrm{e}^{Sy_0-x\exp(Sy_0)} + \int \mathrm{D}y_0\;\mathrm{e}^{Sy_0-x\exp(Sy_0)}\mathcal{O}\Big(\frac{y_0}{\log\Phi(x)},\frac{\log\log\Phi(x)}{\log\Phi(x)}\Big). \tag{85}$$

We can do the remaining integral over y_0 via integration by parts, giving

$$\int \mathrm{D}y_0\;y_0\mathrm{e}^{Sy_0-x\exp(Sy_0)} = S[g(x)+x\frac{\mathrm{d}}{\mathrm{d}x}g(x)]. \tag{86}$$

Hence

$$\frac{\tilde{u}^2\Phi}{\mathrm{d}\Phi/\mathrm{d}x}\,\frac{\mathrm{d}}{\mathrm{d}x}\Big[\log\Big(\frac{\mathrm{d}\Phi}{\mathrm{d}x}\Big)-\log g\Big] = \log\Big(\frac{\Phi}{\log\Phi}\Big)+wS\Big(1+x\frac{\mathrm{d}}{\mathrm{d}x}\log g\Big) + \mathcal{O}\Big(\frac{x\,\mathrm{d}\log g/\mathrm{d}x}{\log\Phi},\frac{\log\log\Phi}{\log\Phi}\Big). \tag{87}$$

To proceed we need the leading orders of $g(x)$. These are derived in appendix D:

$$\log g(x) = -\frac{1}{2S^2}(\log x)^2+\frac{1}{S^2}\log x.\log(\log x)+\mathcal{O}(\log x)\quad (x\to\infty). \tag{88}$$

Our asymptotic equation for $\Phi(x)$ thereby becomes

$$\frac{\tilde{u}^2\Phi}{\mathrm{d}\Phi/\mathrm{d}x}\Big[\frac{\mathrm{d}}{\mathrm{d}x}\log\Big(\frac{\mathrm{d}\Phi}{\mathrm{d}x}\Big)+\frac{\log x}{xS^2}-\frac{\log\log x}{xS^2}+\mathcal{O}(\frac{1}{x})\Big] = \log\Big(\frac{\Phi}{\log\Phi}\Big) + \frac{w}{S}\Big(\log\log x-\log x\Big)+\;\mathcal{O}\Big(1,\frac{\log x}{\log\Phi},\frac{\log\log\Phi}{\log\Phi},\frac{\Phi}{x\mathrm{d}\Phi/\mathrm{d}x}\Big). \tag{89}$$

Inspection of this equation shows that the leading orders of the solution are

$$\Phi(x) = \rho\log x+(1-\rho)\log\log x+o(\log\log x) \tag{90}$$

$$\rho = \frac{w}{2S}\Big(1+\sqrt{1+4\tilde{u}^2/w^2}\Big) \tag{91}$$

or

$$t\gg 1:\;\;\log\Lambda(t) = \rho\log\Lambda_0(t)+(1-\rho)\log(\log\Lambda_0(t))+\ldots. \tag{92}$$

This remarkably simple expression, linking the true and the inferred integrated base hazard rates $\Lambda(t)$ and $\Lambda_0(t)$, predicts that the relation between the two should approach a straight line when shown in a log-log plot. It is not only confirmed by simulations for large times (for which it was derived from our theory) but is in fact found to be quite accurate for all times. This is shown in figure 4, and forms the basis of our variational approximations below.

4.3. Variational approximation

The main complexity of the RS theory is in solving the functional order parameter equation (82). This is the motivation for investigating variational approximations for $\Lambda(t)$. Since our equations were obtained by solving an extremization problem, variational approaches are in the present context both natural and conceptually straightforward. The simulation data in figure 4 suggest writing the functional order parameter in the form $\Lambda(t) = k\Lambda_0^{\rho}(t)$. To compute the new scalar order parameters k and ρ we substitute this expression for $\Lambda(t)$ into the quantity (39) to be extremized. As before we then put $u = \tilde{u}/\sqrt{\gamma}$ and take the limit $\gamma \to \infty$, and find that we need to extremize the following quantity over $(\tilde{u}, v, w, k, \rho)$:

$$\begin{aligned}
\Psi(\tilde{u}, v, w, k, \rho) = {} & \frac{\zeta v^2}{2\tilde{u}^2} + \log k + \log \rho + \int \mathrm{d}t\, p(t) \log \left[\lambda_0(t)\Lambda_0^{\rho-1}(t)\right] \\
& + \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0) \\
& \times \lim_{\gamma\to\infty} \frac{1}{\gamma} \log \int \mathrm{d}y\, \mathrm{e}^{\gamma[\tilde{u}y+wy_0+vz-k\mathrm{e}^{\tilde{u}y+wy_0+vz}\Lambda_0^{\rho}(t)-\frac{1}{2}y^2]} \\
= {} & \frac{\zeta v^2}{2\tilde{u}^2} + \log k + \log \rho + \int \mathrm{d}t\, p(t) \log \left[\lambda_0(t)\Lambda_0^{\rho-1}(t)\right] \\
& + \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0) \\
& \times \max_y \left[\tilde{u}y + wy_0 + vz - k\mathrm{e}^{\tilde{u}y+wy_0+vz}\Lambda_0^{\rho}(t) - \frac{1}{2}y^2\right] \\
= {} & \frac{\zeta v^2}{2\tilde{u}^2} + \log k + \log \rho + \int \mathrm{d}t\, p(t) \log \left[\lambda_0(t)\Lambda_0^{\rho-1}(t)\right] \\
& + \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0) \Big[\tilde{u}\varphi(wy_0 + vz, t) + wy_0 + vz \\
& - k\mathrm{e}^{\tilde{u}\varphi(wy_0+vz,t)+wy_0+vz}\Lambda_0^{\rho}(t) - \frac{1}{2}\varphi^2(wy_0 + vz, t)\Big]
\end{aligned} \tag{93}$$

in which

$$\varphi(\eta, t) = \tilde{u} - \frac{1}{\tilde{u}} W\left(k\tilde{u}^2\mathrm{e}^{\tilde{u}^2+\eta}\Lambda_0^{\rho}(t)\right). \tag{94}$$

It is now easy to derive our order parameter equations, since all contributions to partial derivatives that involve $\varphi(wy_0 + vz, t)$ vanish, by virtue of $\varphi(wy_0 + vz, t)$ maximising the factor between the square brackets. Extremizing (93) over $(\tilde{u}, v, w)$ recovers our earlier equations (76)–(78), with $L(t) = k\tilde{u}^2\mathrm{e}^{\tilde{u}^2}\Lambda_0^{\rho}(t)$, as expected. Extremizing (93) over the new order parameters k and ρ gives:

$$\frac{1}{k} = \int \mathrm{D}y_0\mathrm{D}z \int \mathrm{d}t\, p(t|Sy_0, \lambda_0)\Lambda_0^{\rho}(t)\mathrm{e}^{\tilde{u}^2+wy_0+vz-W\left(k\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda_0^{\rho}(t)\right)} \tag{95}$$

$$\frac{1}{\rho} = k\int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0,\lambda_0)\Lambda_0^\rho(t)\mathrm{e}^{\tilde{u}^2+wy_0+vz-W\left(k\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda_0^\rho(t)\right)}\log\Lambda_0(t)$$
$$-\int \mathrm{d}t\, p(t)\log\Lambda_0(t). \tag{96}$$

Using $W(z)\exp[W(z)] = z$ and our definition of $L(t)$, these two equations can be rewritten as

$$\tilde{u}^2 = \int \mathrm{D}y_0\mathrm{D}z\int_0^1 \mathrm{d}s\; W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big) \tag{97}$$

$$\frac{\tilde{u}^2}{\rho} = \int \mathrm{D}z\mathrm{D}y_0\int_0^1 \mathrm{d}s\; W\big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\big)\Big[\log\log(1/s) - Sy_0\Big]$$
$$-\,\tilde{u}^2\int \mathrm{d}t\, p(t)\log\Lambda_0(t). \tag{98}$$

In the second equation we rewrite the term with the explicit factor y_0, using

$$\int \mathrm{D}z\mathrm{D}y_0\; y_0\int_0^1 \mathrm{d}s\; W\big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\big)$$
$$= \int \mathrm{D}z\mathrm{D}y_0\int_0^1 \mathrm{d}s\; \frac{\partial}{\partial y_0}W\big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\big)$$
$$= \int \mathrm{D}z\mathrm{D}y_0\int_0^1 \mathrm{d}s\; \frac{W\big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\big)}{1+W\big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\big)}\frac{\partial}{\partial y_0}\log\left(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\right)$$
$$= (w-\rho S)\int \mathrm{D}z\mathrm{D}y_0\int_0^1 \mathrm{d}s\; \frac{W\big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\big)}{1+W\big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\big)}. \tag{99}$$

We thus arrive at five relatively simple closed equations from which to solve $(\tilde{u},v,w,k,\rho)$ in our variational approximation. Upon substituting the definition $t(s,y_0) = \Lambda_0^{\mathrm{inv}}(\mathrm{e}^{-Sy_0}\log(1/s))$ we can simplify the argument of Lambert's W-function, which appears in all equations, further to

$$W\Big(\mathrm{e}^{wy_0+vz}L(t(s,y_0))\Big) = W\Big(k\tilde{u}^2\mathrm{e}^{\tilde{u}^2+(w-\rho S)y_0+vz}\log^\rho(1/s)\Big). \tag{100}$$

This enables us to combine the two Gaussian integrals appearing in each order parameter equation by a single zero-average Gaussian integral, with width

$$\sigma(v,w) = \sqrt{(w-\rho S)^2+v^2}. \tag{101}$$

We finally transform the variational order parameter k to $q = k\tilde{u}^2\mathrm{e}^{\tilde{u}^2}$, and evaluate $\int \mathrm{d}t\; p(t)\log\Lambda_0(t) = \int_0^\infty \mathrm{d}x\;\mathrm{e}^{-x}\log x = -C_{\mathrm{E}}$ [38], which involves Euler's constant $C_{\mathrm{E}} = 0.577\,215\,664\,9015\ldots$. We then obtain

$$\zeta v^2 = \int \mathrm{D}x\int_0^1 \mathrm{d}s\;\Big[\tilde{u}^2 - W\Big(q\mathrm{e}^{x\sigma(v,w,\rho)}\log^\rho(1/s)\Big)\Big]^2 \tag{102}$$

$$\zeta = \int \mathrm{D}x\int_0^1 \mathrm{d}s\; \frac{W\Big(q\mathrm{e}^{x\sigma(v,w,\rho)}\log^\rho(1/s)\Big)}{1+W\Big(q\mathrm{e}^{x\sigma(v,w,\rho)}\log^\rho(1/s)\Big)} \tag{103}$$

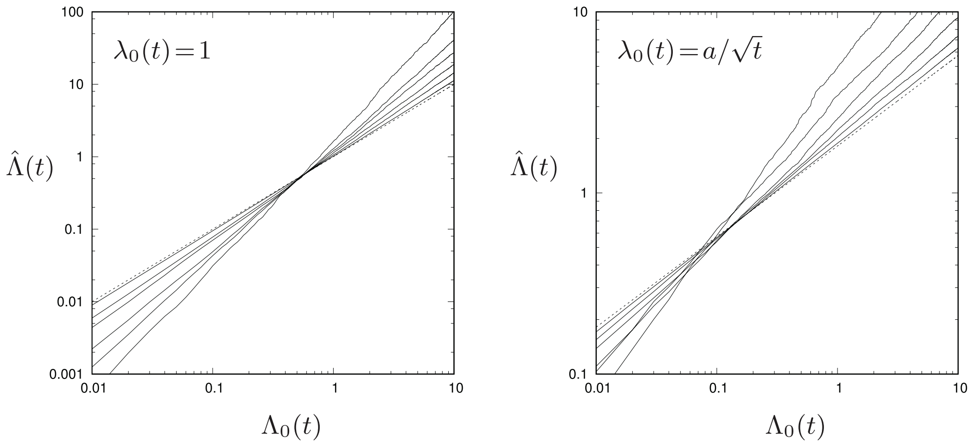


Figure 4. Here we show the simulation data of figure 3 alternatively by drawing the inferred integrated base hazard rates $\hat{\Lambda}(t)$ versus the true values $\Lambda_0(t)$ in log-log plots. We observe that the curves for different values of $\zeta = p/N$ thereby become linear, with high accuracy, for both time-independent (left panel) and time-dependent base hazard rates (right panel). This suggests that $\hat{\Lambda}(t) \approx k\Lambda_0^{\rho}(t)$, with time-independent parameters k and ρ that depend on ζ. The power ρ and the prefactor k both increase with ζ.

$$\frac{\zeta w}{S} = -\int \mathrm{D}x \int_0^1 \mathrm{d}s \, \left[1+\log(s)\right] \, W\Big(q\mathrm{e}^{x\sigma(v,w,\rho)} \log^{\rho}(1/s)\Big) \tag{104}$$

$$\tilde{u}^2 = \int \mathrm{D}x \int_0^1 \mathrm{d}s \, W\Big(q\mathrm{e}^{x\sigma(v,w,\rho)} \log^{\rho}(1/s)\Big) \tag{105}$$

$$\begin{aligned}\frac{\tilde{u}^2}{\rho} = & \int \mathrm{D}x \int_0^1 \mathrm{d}s \, W\Big(q\mathrm{e}^{x\sigma(v,w,\rho)} \log^{\rho}(1/s)\Big) \log\log(1/s) \\ & - S(w-\rho S)\zeta + \tilde{u}^2 C_{\mathrm{E}}.\end{aligned} \tag{106}$$

In the same way we can work out the value of $E(S, \lambda_0)$ for the variational solution, and find:

$$\begin{aligned}E(S,\lambda_0) &= \int \mathrm{d}t \, p(t) \log\left[\frac{\lambda_0(t)}{\lambda(t)}\right] \;=\; -\int \mathrm{d}t \, p(t) \log\left[k\rho\Lambda_0^{\rho-1}(t)\right] \\ &= -\log k - \log\rho - (\rho-1)\int \mathrm{d}t \, p(t) \log\Lambda_0(t) \\ &= -\log k - \log\rho - (\rho-1)\int_0^{\infty} \mathrm{d}x \, \mathrm{e}^{-x} \log x \\ &= -\log k - \log\rho + (\rho-1)C_{\mathrm{E}}.\end{aligned} \tag{107}$$

For $q \to 0$ we may replace $W(q\mathrm{e}^{\sigma x}\log^{\rho}(1/s)) \approx q\mathrm{e}^{\sigma x}\log^{\rho}(1/s)$ and use the integral $\int_0^1 \mathrm{d}s \; \log(1/s)\log\log(1/s) = 1 - C_{\mathrm{E}}$, to recover after some simple expansions the correct $\zeta \to 0$ solution: $\lim_{\zeta\to 0} v = \lim_{\zeta\to 0} \tilde{u} = 0$, $\lim_{\zeta\to 0} w = S$, $\lim_{\zeta\to 0} \rho = \lim_{\zeta\to 0} k = 1$, and $\lim_{\zeta\to 0} E(S,\lambda_0) = 0$.

We observe that our above closed variational equations (102)–(106) are completely independent of the true base hazard rate $\lambda_0(t)$. Hence they predict that the key quantities required for overfitting correction in the Cox model (the slope of the data cloud, and the deformation parameters of the base hazard rate) are independent of the true shape of the base hazard rate.

The easiest protocol for solving our equations numerically is to regard q as an independent parameter, and compute $(\zeta, v, w, \tilde{u}, \rho)$ for each q by iterative mapping. Upon doing so (see figure 5), one finds that the solution always exhibits $\rho = w/S$, within numerical accuracy limitations. We have not yet been able to confirm this analytically, as that would require proving that the solution of our equation obeys

$$\int \mathrm{D}x \int_0^1 \mathrm{d}s\, W\Big(q\mathrm{e}^{xv} \log^{\rho}(1/s)\Big)\Big[\log\log(1/s) + C_{\mathrm{E}} - \frac{1}{\rho}\Big] = 0 \tag{108}$$

but it is for small ζ in agreement with (91) (as it should be). If $\rho = w/S$ is indeed generally true for the solution of our variational equations, it implies that ρ is identical to the slope of the data clouds in figure 2, and that the values of (v, ρ, q) (hence also of the slope and the width of the data clouds in figure 2) are not only independent of $\lambda_0(t)$ but also independent of S. It would also allow us to obtain a more compact closed theory in terms of just three scalar order parameters, as we will show now. Upon making directly the variational ansatz $\Lambda(t) = k\Lambda_0^{\rho}(t)$ with $w = \rho S$, we need to extremize

$$\begin{aligned}\Psi(\tilde{u}, v, k, \rho) = {} & \frac{\zeta v^2}{2\tilde{u}^2} + \log k + \log\rho + \int \mathrm{d}t\, p(t) \log\Big[\lambda_0(t)\Lambda_0^{\rho-1}(t)\Big] \\ & + \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0)\Big[\tilde{u}\varphi(\rho S y_0 + vz, t) + \rho S y_0 + vz \\ & - k\mathrm{e}^{\tilde{u}\varphi(\rho S y_0 + vz, t) + \rho S y_0 + vz}\Lambda_0^{\rho}(t) - \frac{1}{2}\varphi^2(\rho S y_0 + vz, t)\Big]\end{aligned} \tag{109}$$

in which again $\varphi(\eta, t) = \tilde{u} - \tilde{u}^{-1}W(k\tilde{u}^2\mathrm{e}^{\tilde{u}^2+\eta}\Lambda_0^{\rho}(t))$. Following similar manipulations as used for the first variational analysis, and with the previous short-hand $q = k\tilde{u}^2\mathrm{e}^{\tilde{u}^2}$, we find upon extremization of $\Psi(\tilde{u}, v, k, \rho)$ and after elimination of $\tilde{u}$ the following three closed equations for (v, k, ρ):

$$\zeta v^2 = \int \mathrm{D}x \int_0^1 \mathrm{d}s\, \Big[\tilde{u}^2 - W\Big(q\mathrm{e}^{vx}\log^{\rho}(1/s)\Big)\Big]^2 \tag{110}$$

$$\zeta = \int \mathrm{D}x \int_0^1 \mathrm{d}s\, \frac{W\Big(q\mathrm{e}^{vx}\log^{\rho}(1/s)\Big)}{1 + W\Big(q\mathrm{e}^{vx}\log^{\rho}(1/s)\Big)} \tag{111}$$

$$\begin{aligned}\zeta\rho = {} & -\frac{1}{S^2}\int \mathrm{D}x \int_0^1 \mathrm{d}s\, W\Big(q\mathrm{e}^{vx}\log^{\rho}(1/s)\Big)\log\log(1/s) \\ & - \int \mathrm{D}x \int_0^1 \mathrm{d}s\, \Big[1 + \log(s) + (C_{\mathrm{E}} - \frac{1}{\rho})/S^2\Big] W\Big(q\mathrm{e}^{vx}\log^{\rho}(1/s)\Big).\end{aligned} \tag{112}$$

Upon solving the trio (110)–(112), the values of $\tilde{u}$, w and k then follow via

$$\tilde{u}^2 = \int \mathrm{D}x \int_0^1 \mathrm{d}s\, W\Big(q\mathrm{e}^{vx}\log^{\rho}(1/s)\Big), \quad k = \frac{q}{\tilde{u}^2}\mathrm{e}^{-\tilde{u}^2}, \quad w = \rho S. \tag{113}$$

Finally we note that all our equations in this section can also be written in a form that involves only integrations over the interval [0, 1], using the general identity

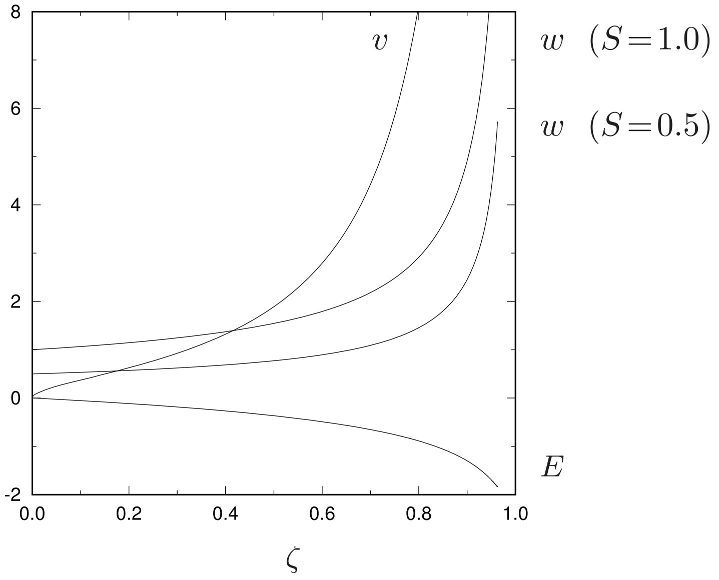


Figure 5. Result of solving numerically the variational equations (110)–(112). The values of v, k, $\rho = w/S$ and E are independent of the strength S of the true associations and independent of the true base hazard rate $\lambda_0(t)$. For $\zeta = 0$ we recover the overfitting-free state $w = S$ and $v = E = 0$. At $\zeta = 1$ a phase transition occurs, marked by divergence of v and w.

$$\int \mathrm{D}x\, f(x) = \int_0^1 \mathrm{d}s\, \frac{f\big(\sqrt{2\log(1/s)}\big) + f\big(-\sqrt{2\log(1/s)}\big)}{2\sqrt{\pi \log(1/s)}}. \tag{114}$$

It is instructive at this stage to test the predictions of the above simple variational equations (110)–(112) against numerical simulations of Cox regression on synthetic data. According to (41)–(43), we must expect to find in our simulations that $v = \lim_{r,N\to\infty} v(r,N)$ and $w = \lim_{r,N\to\infty} w(r,N)$, where

$$v(r,N) = \frac{1}{\zeta N}\Big[\sum_{\mu=1}^{\zeta N}\langle \hat{\beta}_\mu^2\rangle_{\mathcal{D}} - \frac{1}{|\boldsymbol{\beta}^\star|^2}\Big(\sum_{\mu=1}^{\zeta N}\beta_\mu^\star\langle\hat{\beta}_\mu\rangle_{\mathcal{D}}\Big)^2\Big] \tag{115}$$

$$w(r,N) = \frac{1}{\zeta N}\sum_{\mu=1}^{\zeta N}\frac{\beta_\mu^\star\cdot\langle\hat{\beta}_\mu\rangle_{\mathcal{D}}}{|\boldsymbol{\beta}^\star|}. \tag{116}$$

Here $\{\hat{\beta}_\mu\}$ denotes the inferred values of the (rescaled) regression parameters, and the averages $\langle\ldots\rangle_{\mathcal{D}}$ are over r randomly generated data sets. The results of measuring $v(r,N)$ and $w(r,N)$ in numerical simulations are shown in figure 6 together with the variational predictions. In spite of the modest values in our simulations of $N = 200$ and the finite number of training sets over which inferred parameters are averaged in evaluating (115) and (116) (which one expects to generate excess variability), the agreement between the variational predictions and the simulations is seen to be surprisingly good.

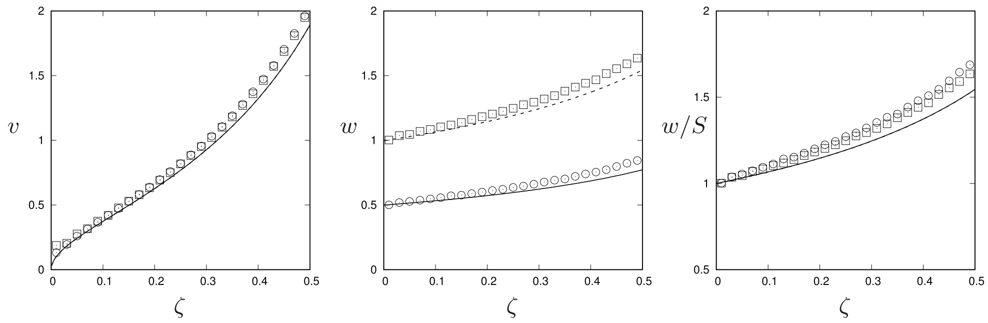


Figure 6. Test of the predictions of the variational equations (110)–(112) against numerical simulations of Cox regression, with $N = 200$, $\lambda_0(t) = 1$, and either $S = 0.5$ (circles) or $S = 1.0$ (squares). Left: order parameter v (solid line) versus $v(r, N)$, see equation (115). Middle: order parameter w (solid line: $S = 0.5$; dashed: $S = 1.0$) versus $w(r, N)$, see equation (116). Right: the corresponding values of w/S. In all cases $r = 10^4$. The simulations confirm the predictions of the theory that both v and w/S are independent of S.

5. Tests and applications

We will now test the variational RS theory (110)–(112) further against numerical simulations, focusing on the the dependence on the ratio ζ of the main characteristics of the regression parameter data clouds of figure 2 (i.e. their slope κ and their width σ), and of the integrated base hazard rates as shown e.g. in figure 3. We know (46) that the theory predicts $\kappa = \rho$ and $\sigma = v/\sqrt{p}$ (for the standard scaling convention of the Cox model [5], i.e. for $p(t|z) = -\frac{\mathrm{d}}{\mathrm{d}t}\exp[-\exp(\boldsymbol{\beta}\cdot z)\Lambda(y)]$), and these predictions are plotted in figure 7 as solid lines, together with the values obtained in regression simulations of the Cox model on synthetic data (markers), for $N = 200$ and $N = 400$, and for two distinct choices for the true base hazard rate $\lambda_0(t)$. Modulo finite size effects, which increase as we approach the phase transition point $\zeta = 1$, there is again good agreement between theory and simulations. The data confirm also the prediction of the variational theory that both κ and σ are independent of the true base hazard rate $\lambda_0(t)$.

In figure 8 we compare the inferred integrated base hazard rates $\hat{\Lambda}(t)$, obtained for synthetic data with $N = 400$, with the predictions of the variational RS theory (110)–(112), for two choices of the base hazard rate. The agreement is satisfactory for times of the order of the typical event times in the data. For larger times (where the theory has to extrapolate to times where available data are at best sparse) one observes increasing deviations, with the variational theory underestimating the impact of overfitting; this is indeed consistent with (92), since the variational approximation captures only the first (leading) term of the exact expansion (92). We can in principle obtain more accurate integrated base hazard rate predictions within the current framework, but this requires that we either solve (numerically) the full RS equations (76)–(79), or develop a more refined variational ansatz for the function $L(t)$.

We found in our simulations that as the ratio $\zeta = p/N$ increases, higher numerical precision is required in solving Cox's equations. For values $N \sim 10^2$–10^3 and $\zeta > 0.4$, using conventional C-code compiled with gcc at double floating point precision (data type 'double') will occasional lead to degeneracies in the equations that cause the association parameters $\hat{\boldsymbol{\beta}}$ to be ill-defined. Upon switching to quadruple floating point precision (data type 'long double') these degeneracies disappeared.

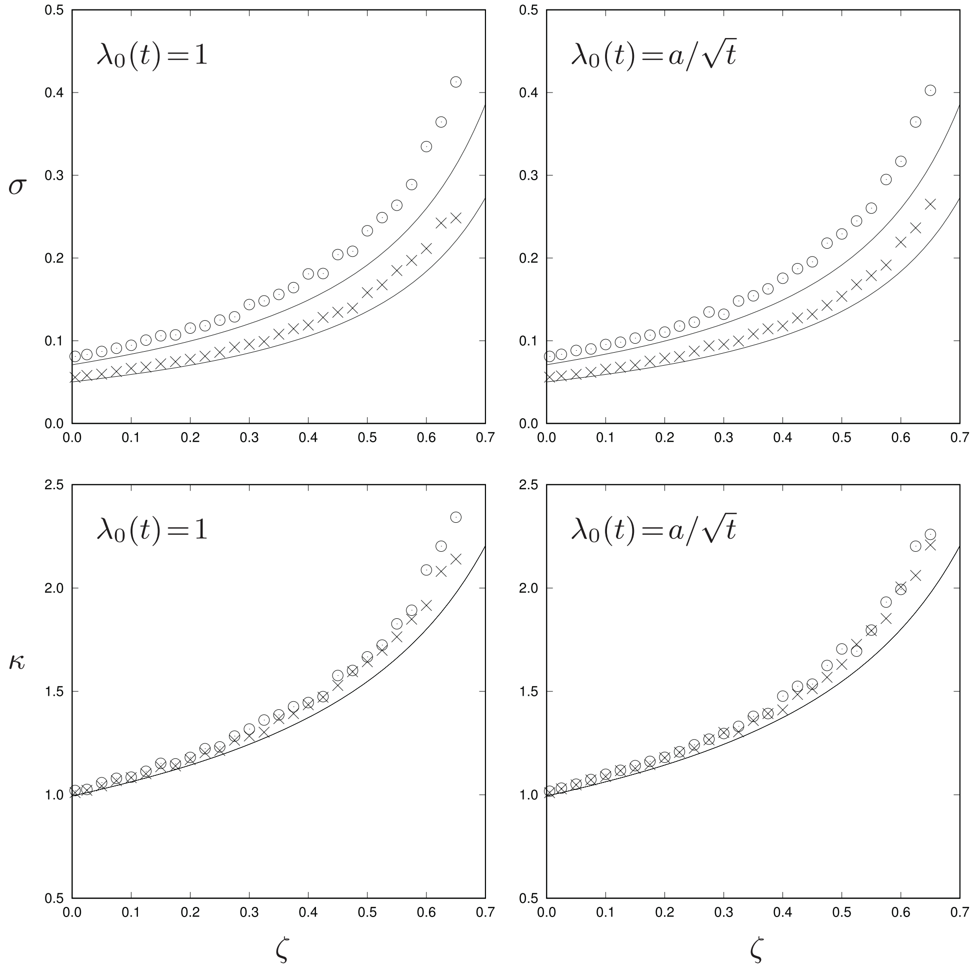


Figure 7. We show the slopes κ and the widths σ of the association parameter data clouds of figure 2, computed from regression simulations carried out on synthetic survival data via least squares fitting, for $N = 200$ (circles) and $N = 400$ (crosses). In all cases $S = 0.5$. Solid lines: predictions of the variational theory, viz. $\sigma = v/\sqrt{p}$ and $\kappa = \rho$ (both of which are independent of $\lambda_0(t)$ and of S). Top row: widths σ, for constant (left) and time-dependent (right) base hazard rates, with $a = \exp(S^2)/\sqrt{2}$ defined such that $\int \mathrm{d}t\ p(t)t = 1$. Bottom row: slopes κ, for constant (left) and time-dependent (right) base hazard rates. Each marker is an average over r independent simulation experiments, such that the product pr is the same for all markers.

The present RS theory has so far been tested only for 'normal' regimes for the parameter S, which represents the typical width of the sum $\sum_\mu \beta^\star_\mu z_\mu/\sqrt{p}$, and hence the typical scale of the covariate-conditioned hazard rates. It turns out that upon carrying out Cox regression for synthetic survival data with large values of ζ and very large values of S, we observe ergodicity breaking: upon plotting true versus inferred association parameters, as in figure 3, for different simulation experiments with the same parameters N and p, we now find multiple data clouds with distinct slopes, as opposed to a single data cloud with unique reproducible characteristics. This suggest that the relevant saddle points in the replica calculation will no

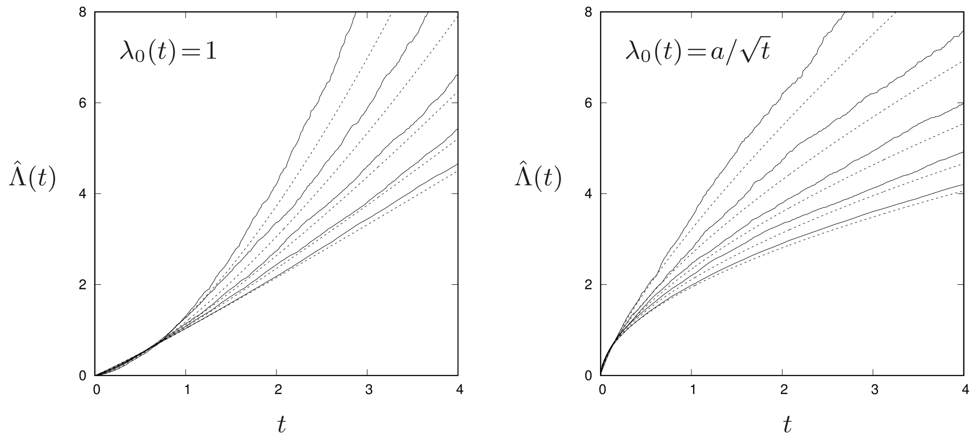


Figure 8. Inferred integrated base hazard rates $\hat{\Lambda}(t)$ (solid curves, averaged over multiple experiments) for synthetic survival data, shown together with the predictions of the variational RS theory (dashed curves) for $\zeta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ (lower to upper curves). In all simulations $N = 400$, $S = 0.5$, and a is defined such that $\int \mathrm{d}t\ p(t)t = 1$.

longer be replica-symmetric. This phenomenology, of which examples are shown in figure 9, can be studied in a natural way within the replica formalism, but it requires so-called RSB (replica symmetry breaking) ansätze for the overlap matrix $\boldsymbol{C}$. One anticipates that for sufficiently large values of ζ there may be a critical value of $S/\sqrt{p}$ that marks an RSB transition, i.e. the onset of non-ergodicity; the preliminary data in figure 9 suggest that this critical value may also depend on the shape of the true base hazard rate. Computing these critical values of S from the replica formalism, in terms of the parameters ζ, S and $\lambda(t)$, will be the subject of a future study.

6. Discussion

The Cox model has been by far the most popular and effective statistical tool for the analysis of time-to-event data in medicine, since its publication nearly half a century ago. However, the demands on statistical methods in 21st century medicine are changing. We can now take measurements on individual patients of unprecedented dimensionality p, such as gene expressions and high-resolution imaging data, but the typical number of samples N in our medical data bases has not grown in proportion. As a result, the condition for maximum likelihood (ML) multivariate regression methods (including the model of Cox) to be applicable, being $p/N \ll 1$ in order to avoid overfitting, is nowadays very often not met. Apart from a few early (and modest) simulation experiments, there appear not to have been any published studies aimed at modelling mathematically the mechanism of overfitting in Cox regression, which is a prerequisite for the development of methods to deal with the overfitting problem. When the dimensionality of the data, relative to the number of available samples, is too high to justify using the multivariate Cox model, medical statisticians and epidemiologists are presently left having to resort to poor alternatives for proper regression: they can either limit *a priori* the number of covariates used in regression (and thereby limit outcome prediction potential), or switch to univariate analysis (which is undesirable since we know that univariate estimates of association parameters correlate poorly with their multivariate counterparts), or work with

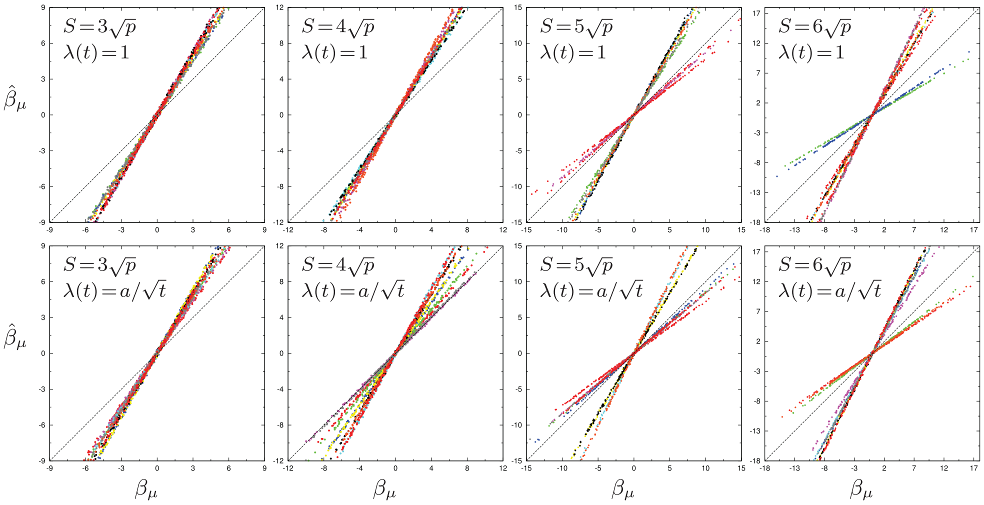


Figure 9. Examples of non-ergodicity in Cox regression, for large values of ζ and S, signalled by the breaking up of the single linear data cloud found for small S into multiple linear clouds, each with distinct slopes (that depend on the realisation of the data set). As in figure 2, we show true versus inferred association coefficients. In all cases $N=500$, $\zeta=0.4$ and $S/\sqrt{p}\in\{3,4,5,6\}$, and all plots show data from 10 independent simulations (where each simulation is given a different colour). Top row: $\lambda_0(t)=1$; bottom row: $\lambda(t)=a/\sqrt{t}$, with a such that $\int \mathrm{d}t\ p(t)t=1$.

so-called 'risk signatures' (which tend to involve ad-hoc definitions, and ad-hoc recipes for interpretation). Thus, expensive and potentially informative high-dimensional clinical data remain under-utilised.

Our regression simulations with synthetic survival data show clearly that the mechanism of overfitting in Cox regression is surprisingly reproducible and consistent: it always leads to a clear bias, which reports association parameter values that are more extreme than their true values, underestimates base hazard rates for short times, and over-estimates base hazard rates for large times. This consistency suggests that it must in principle be possible to model overfitting mathematically, and that (if such modelling is successful) one should be able to correct the outcomes of Cox regression systematically for the impact of overfitting. This, in turn, would allow us to do multivariate regression reliably for significantly larger ratios of the number of covariates over the number of samples, and obtain more accurate and reproducible predictions of clinical outcomes.

In this paper we have presented such a theory, which is built on the mathematical methods of statistical mechanics and inspired by Gardner's famous analysis of binary classifiers [22]. It assumes that N is large, but with p/N finite, and it combines three ideas: (i) the formulation of an information-theoretic measure of overfitting in time-to-event regression, (ii) translating the calculation of this quantity into computing the ground state of a statistical mechanical system, and (iii) dealing with the heterogeneity in the problem (here: the realisation of the data set) with the replica method. Our modeling approach is generic. It is developed initially for arbitrary parametrised time-to-event regression models, but we devote most of our paper to the Cox model, in recognition of its importance and dominance in the medical statistics field. We show that by combining the above three ideas, it is possible to derive explicit macroscopic equations, exact in the asymptotic limit, with which to characterise the regression process for finite values of the ratio p/N. In this paper we assume that the regression process is ergodic, and make the so-called replica symmetric (RS) ansatz for the solution of our equations; this assumption is supported by numerical simulations, provided the true association parameters are not too large.

For the Cox model, the order parameters of the RS theory contain all the relevant information required to quantify the impact of overfitting, but since one of them is a function (the inferred integrated base hazard rate), we introduced a suitable variational approximation, which resulted in a much simpler three-parameter theory. The simplified theory makes various qualitative predictions that are confirmed by regression simulations with synthetic data: that the 'inflation' of inferred association parameters is independent of the amplitude of the true association parameters and of the true base hazard rate, that there is a phase transition when $p/N \to 1$, that the base hazard rate is underestimated for short times and over-estimated for large times, and that the relation between inferred and true integrated base hazard rate is for large times of the form $\log \hat{\Lambda}(t) \sim \rho \log \Lambda_0(t)$, with a parameter ρ that increases with the ratio $\zeta = p/N$. The quantitative agreement between our variational theory and regression simulations with synthetic data is generally very good, modulo finite size fluctuations, including the predicted overfitting-induced bias in association parameters. The only exception is the integrated base rate at large times, where available data are sparse, and where the variational ansatz (which incorporates only the leading order time dependence) under-estimates the impact of overfitting. Upon increasing the values of ζ and S, we observe new phenomenology, such as ergodicity breaking in the regression process (which requires order parameters with broken replica symmetry, or RSB). The calculation of the RSB transition line will be the subject of a subsequent paper.

The present study represents only a first step. It demonstrates that it is possible to model overfitting in Cox regression mathematically, using the replica formalism. We envisage many direct extensions, such as increasing the precision of our predictions by constructing full non-variational solutions to our RS order parameter equations (analytically or numerically), the incorporation of censoring, and the addition of MAP-type regulariser terms. More technical potential follow-up studies could investigate RSB phenomena, including the calculation of the ergodicity breaking transition line, or the impact of having covariate distributions for which the sums $\sum_\mu \beta_\mu z_\mu$ no longer have Gaussian statistics. Casting the net somewhat wider, and given our more general initial formulation of the theory, we expect that there will be other survival analysis models for which a similar overfitting analysis can be done.

Last but certainly not least, we would now like to explore the potential of our methodology to provide practical tools with which to correct multivariate Cox regression analyses of real time-to-event data in medicine for the impact of overfitting. Such tools could be used retrospectively, to determine objectively which past results in the medical literature that were obtained with the Cox method can be trusted, and which perhaps cannot. They should hopefully also lead to more accurate clinical outcome predictions in the future, by allowing medical statisticians to include more covariates in multivariate regression by default, without overfitting danger, and enable the construction of sample size tables for multivariate regression that allow overfitting effects to be taken into account in the design of clinical trials. The results presented in this paper suggest that in the near future, with proper overfitting corrections, reliable multivariate regression for time-to-event data at ratios of up to $p/N \approx 0.5$ or higher will be quite feasible.

Acknowledgments

We would like to thank Bryan Lutchmanen for contributing to the regression simulation studies, and Anita Grigoriadis for the data used to produce figure 1. We are also grateful for support from Saddle Point Science, the Engineering and Physical Sciences Research Council (EPSRC), and the Medical Research Council (MRC) of the United Kingdom.

Appendix A. Covariate correlations in Cox regression

In the absence of censoring, the equations from which to compute the inferred base hazard rate $\hat{\lambda}(t)$ and the inferred association parameters $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ in Cox regression are the following [5]:

$$\hat{\lambda}(t) = \frac{\sum_{i=1}^N \delta(t-t_i)}{\sum_{i=1}^N \theta(t_i-t)\mathrm{e}^{\hat{\boldsymbol{\beta}}\cdot\boldsymbol{z}_i}} \tag{A.1}$$

$$\hat{\boldsymbol{\beta}} = \mathrm{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^N \Big\{\boldsymbol{\beta}\cdot\boldsymbol{z}_i - \log\Big[\sum_{j=1}^N \theta(t_j-t_i)\mathrm{e}^{\boldsymbol{\beta}\cdot\boldsymbol{z}_j}\Big]\Big\}. \tag{A.2}$$

Let us define the average values and correlations of the covariates as $\langle \boldsymbol{z}\rangle = \bar{\boldsymbol{z}}$ and $\langle (z_\mu-\bar{z}_\mu)(z_\nu-\bar{z}_\nu)\rangle = A_{\mu\nu}$, with $\langle f(\boldsymbol{z})\rangle = N^{-1}\sum_{i=1}^N f(\boldsymbol{z}_i)$. We can then simply write the original $\{\boldsymbol{z}_i\}$ in terms of zero-average and uncorrelated covariate vectors $\{\tilde{\boldsymbol{z}}_i\}$, by writing $\boldsymbol{z}_i = \bar{\boldsymbol{z}} + \boldsymbol{A}^{\frac{1}{2}}\tilde{\boldsymbol{z}}_i$. The equation for the regression parameters thereby becomes

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathrm{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^N \Big\{\boldsymbol{\beta}\cdot\bar{\boldsymbol{z}} + \boldsymbol{\beta}\cdot\boldsymbol{A}^{\frac{1}{2}}\tilde{\boldsymbol{z}}_i - \log\Big[\sum_{j=1}^N \theta(t_j-t_i)\mathrm{e}^{\boldsymbol{\beta}\cdot\bar{\boldsymbol{z}}+\boldsymbol{\beta}\cdot\boldsymbol{A}^{\frac{1}{2}}\tilde{\boldsymbol{z}}_j}\Big]\Big\} \\ &= \mathrm{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^N \Big\{(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\beta})\cdot\tilde{\boldsymbol{z}}_i - \log\Big[\sum_{j=1}^N \theta(t_j-t_i)\mathrm{e}^{(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{\beta})\cdot\tilde{\boldsymbol{z}}_j}\Big]\Big\}.\end{aligned} \tag{A.3}$$

Hence $\hat{\boldsymbol{\beta}} = \boldsymbol{A}^{-\frac{1}{2}}\tilde{\boldsymbol{\beta}}$, in which $\tilde{\boldsymbol{\beta}}$ is the regression outcome of the Cox method applied to the *zero-average, uncorrelated and normalized* covariates $\{\tilde{\boldsymbol{z}}_i\}$, i.e.

$$\tilde{\boldsymbol{\beta}} = \mathrm{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^N \Big\{\boldsymbol{\beta}\cdot\tilde{\boldsymbol{z}}_i - \log\Big[\sum_{j=1}^N \theta(t_j-t_i)\mathrm{e}^{\boldsymbol{\beta}\cdot\tilde{\boldsymbol{z}}_j}\Big]\Big\}. \tag{A.4}$$

Similarly, for the base hazard rate we find:

$$\hat{\lambda}(t) = \frac{\sum_{i=1}^N \delta(t-t_i)}{\sum_{i=1}^N \theta(t_i-t)\mathrm{e}^{\hat{\boldsymbol{\beta}}\cdot\bar{\boldsymbol{z}}+\hat{\boldsymbol{\beta}}\cdot\boldsymbol{A}^{\frac{1}{2}}\tilde{\boldsymbol{z}}_i}} = \mathrm{e}^{-\hat{\boldsymbol{\beta}}\cdot\bar{\boldsymbol{z}}}\frac{\sum_{i=1}^N \delta(t-t_i)}{\sum_{i=1}^N \theta(t_i-t)\mathrm{e}^{\tilde{\boldsymbol{\beta}}\cdot\boldsymbol{z}_i}}. \tag{A.5}$$

Hence $\hat{\lambda}(t) = \tilde{\lambda}(t)\exp(-\tilde{\boldsymbol{\beta}}\cdot\boldsymbol{A}^{-\frac{1}{2}}\bar{\boldsymbol{z}})$, in which $\tilde{\lambda}(t)$ is given by Breslow's formula (the regression outcome for the base hazard rate of the Cox method) applied once more to the zero-average uncorrelated and normalised covariates $\{\tilde{\boldsymbol{z}}_i\}$, i.e.

$$\tilde{\lambda}(t) = \frac{\sum_{i=1}^N \delta(t-t_i)}{\sum_{i=1}^N \theta(t_i-t)\mathrm{e}^{\tilde{\boldsymbol{\beta}}\cdot\tilde{\boldsymbol{z}}_i}}. \tag{A.6}$$

We conclude that for the Cox model one can always express the regression outcomes for any choice of covariate vectors in terms of the regression outcomes for zero-average, normalized and uncorrelated covariates, where $\langle z_\mu\rangle = 0$ and $\langle z_\mu z_\nu\rangle = \delta_{\mu\nu}$.

Appendix B. Deriviation of the replica symmetric equations

Assuming replica symmetry to hold converts our problem into calculating

$$E_{\gamma}(S,\lambda_0)=\frac{\partial}{\partial\gamma}\mathrm{extr}_{C,c,c_0;\lambda}\Psi_{\mathrm{RS}}[C,c,c_0;\lambda] \tag{B.1}$$

$$\begin{aligned}\Psi_{\mathrm{RS}}[C,c,c_0;\lambda]=\lim_{n\to0}\frac{1}{n}\Big\{&\frac{1}{2}\log\mathrm{Det}\boldsymbol{C}-\frac{1}{2}\zeta\log\mathrm{Det}\boldsymbol{C}'\\ &-\log\int\frac{\mathrm{d}\mathbf{y}}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}\cdot\boldsymbol{C}^{-1}\boldsymbol{y}}\int\mathrm{d}t\,p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda)}{p(t|y_0,\lambda_0)}\Big]^{\gamma}\Big\}.\end{aligned} \tag{B.2}$$

To proceed we need the determinant and inverse of the $(n+1)\times(n+1)$ covariance matrix $\boldsymbol{C}$, and the determinant of the $n\times n$ matrix $\boldsymbol{C}'$. Both $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ will inherit the assumed replica-symmetric (RS) structure of the saddle-point. Hence they must have the respective forms

$$\boldsymbol{C}=\begin{pmatrix}S^2 & c_0 & \cdots & \cdots & c_0\\ c_0 & C & c & \cdots & c\\ \vdots & c & C & \cdots & c\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ c_0 & c & \cdots & c & C\end{pmatrix}\qquad \boldsymbol{C}^{-1}=\begin{pmatrix}d_{00} & d_0 & \cdots & \cdots & d_0\\ d_0 & D & d & \cdots & d\\ \vdots & d & D & \cdots & d\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ d_0 & d & \cdots & d & D\end{pmatrix}. \tag{B.3}$$

The RS eigenvectors $\boldsymbol{x}$ and eigenvalues μ of $\boldsymbol{C}$ are calculated easily:

$$\boldsymbol{x}=(u,v,\ldots,v):\mu_{\pm}=\frac{1}{2}\Big\{C+(n-1)c+S^2\pm\sqrt{[C+(n-1)c-S^2]^2+4nc_0^2}\Big\} \tag{B.4}$$

$$\boldsymbol{x}=(0,v_1,\ldots,v_n):\quad\sum_{\alpha=1}^{n}v_\alpha=0,\quad\mu=C-c\quad(\text{multiplicity } n-1). \tag{B.5}$$

It follows that

$$\begin{aligned}\log\mathrm{Det}\boldsymbol{C}&=\log[(C-c)^{n-1}\mu_{+}\mu_{-}]\\ &=\log\Big[S^2(C-c)^{n-1}\Big(C-c+n(c-c_0^2/S^2)\Big)\Big]\end{aligned} \tag{B.6}$$

$$=\log S^2+n\log(C-c)+n\frac{c-c_0^2/S^2}{C-c}+\mathcal{O}(n^2). \tag{B.7}$$

We obtain the parameters (D,d,d_{00},d_0) by multiplying the two matrices in (B.3) and demanding that this gives the identity matrix. After some simple algebra this results in:

$$d_{00}=\frac{C+(n-1)c}{S^2(C+(n-1)c)-nc_0^2},\qquad d_0=-\frac{c_0}{S^2(C+(n-1)c)-nc_0^2} \tag{B.8}$$

$$d=\frac{1}{C-c}\,\frac{c_0^2-cS^2}{S^2(C+(n-1)c)-nc_0^2},\quad D=d+\frac{1}{C-c}. \tag{B.9}$$

It is now a trivial matter to calculate also the quantity $\log\mathrm{Det}\boldsymbol{C}'$, since the RS form of $\boldsymbol{C}$ implies that for $\alpha,\rho=1\dots n$ we have $C'_{\alpha\rho}=\delta_{\alpha\rho}(C-c)+c-(c_0/S)^2$. It has one eigenvector $(1,\dots,1)$ with eigenvalue $C-c-nc_0^2/S^2+nc$, and an $(n-1)$-fold degenerate eigenspace with eigenvalue $C-c$. Hence

$$
\begin{aligned}
\log\mathrm{Det}\boldsymbol{C}' &= (n-1)\log(C-c)+\log\Big(C-c+n[c-c_0^2/S^2]\Big)\\
&= n\log(C-c)+\log\Big(1+n\frac{c-c_0^2/S^2}{C-c}\Big)\\
&= n\Big[\log(C-c)+\frac{c-c_0^2/S^2}{C-c}\Big]+\mathcal{O}(n^2).
\end{aligned}
\tag{B.10}
$$

Inserting these results into (B.2) gives, with the short-hand $\mathrm{D}y=(2\pi)^{-1/2}\mathrm{e}^{-\frac{1}{2}y^2}\mathrm{d}y$, and upon carrying out successive Taylor expansions for small n:

$$
\begin{aligned}
\Psi_{\mathrm{RS}}[C,c,c_0;\lambda] &= \lim_{n\to 0}\Big\{\frac{1}{2}(1-\zeta)\Big[\log(C-c)+\frac{c-c_0^2/S^2}{C-c}\Big]+\frac{1}{n}\log S\\
&\quad-\frac{1}{n}\log\int\frac{\mathrm{d}\boldsymbol{y}}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}d_{00}y_0^2-\frac{1}{2}(D-d)\sum_{\alpha=1}^{n}y_\alpha^2-\frac{1}{2}d(\sum_{\alpha=1}^{n}y_\alpha)^2-d_0y_0\sum_{\alpha=1}^{n}y_\alpha}\\
&\quad\times\int\mathrm{d}t\,p(t|y_0,\lambda_0)\prod_{\alpha=1}^{n}\Big[\frac{p(t|y_\alpha,\lambda)}{p(t|y_0,\lambda_0)}\Big]^\gamma\Big\}\\
&= \lim_{n\to 0}\Big\{\frac{1}{2}(1-\zeta)\Big[\log(C-c)+\frac{c-c_0^2/S^2}{C-c}\Big]+\frac{1}{2n}\log(S^2d_{00})\\
&\quad-\frac{1}{n}\log\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\,p(t|\frac{y_0}{\sqrt{d_{00}}},\lambda_0)\\
&\quad\times\Big[\int\mathrm{d}y\,\mathrm{e}^{-\frac{1}{2}(D-d)y^2-y(d_0y_0/\sqrt{d_{00}}+\mathrm{i}z\sqrt{d})}\Big(\frac{p(t|y,\lambda)}{p(t|\frac{y_0}{\sqrt{d_{00}}},\lambda_0)}\Big)^\gamma\Big]^n\Big\}\\
&= \frac{1}{2}(1-\zeta)\Big[\log(C-c)+\frac{c-c_0^2/S^2}{C-c}\Big]\\
&\quad+\lim_{n\to 0}\frac{1}{2n}\log\Big[\frac{1+nc/(C-c)}{1+n[c-c_0^2/S^2]/(C-c)}\Big]\\
&\quad-\lim_{n\to 0}\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\,p(t|\frac{y_0}{\sqrt{d_{00}}},\lambda_0)\\
&\quad\times\log\int\mathrm{d}y\,\mathrm{e}^{-\frac{1}{2}y^2/(C-c)-y(d_0y_0/\sqrt{d_{00}}+\mathrm{i}z\sqrt{d})}\Big(\frac{p(t|y,\lambda)}{p(t|\frac{y_0}{\sqrt{d_{00}}},\lambda_0)}\Big)^\gamma\\
&= \frac{1}{2}(1-\zeta)\Big[\log(C-c)+\frac{c-c_0^2/S^2}{C-c}\Big]+\frac{1}{2}\frac{c_0^2/S^2}{C-c}-\frac{1}{2}\log(C-c)\\
&\quad-\frac{1}{2}\log(2\pi)-\int\mathrm{D}z\mathrm{D}y_0\int\mathrm{d}t\,p(t|Sy_0,\lambda_0)\\
&\quad\times\log\int\mathrm{D}y\,\mathrm{e}^{y[y_0c_0/S\sqrt{C-c}+z\sqrt{(c-c_0^2/S^2)/(C-c)}]}\Big(\frac{p(t|y\sqrt{C-c},\lambda)}{p(t|Sy_0,\lambda_0)}\Big)^\gamma.
\end{aligned}
\tag{B.11}
$$

This expression takes a simpler form if we introduce the following transformation of the trio $\{C,c,c_0\}$ to new non-negative variables $\{u,v,w\}$:

$$u=\sqrt{C-c},\qquad v=\sqrt{c-c_0^2/S^2},\qquad w=c_0/S \tag{B.12}$$

with inverse transformation

$$c_0=Sw,\qquad c=v^2+w^2,\qquad C=u^2+v^2+w^2. \tag{B.13}$$

With these definitions, and upon removing terms that vanish upon differentiation by γ, we can summarise the current state of our RS calculations for the stochastic generalization of the Cox model, in the limit of large data sets, by the following compact expression:

$$\begin{aligned}E_\gamma(S,\lambda_0)=\frac{\partial}{\partial\gamma}\mathrm{extr}_{u,v,w;\lambda}\Big\{\frac{1}{2}(1-\zeta)v^2/u^2+\frac{1}{2}w^2/u^2-\zeta\log u\\ -\int \mathrm{D}z\mathrm{D}y_0\int \mathrm{d}t\,p(t|Sy_0,\lambda_0)\log\int \mathrm{D}y\,\mathrm{e}^{y(wy_0+vz)/u}\Big(\frac{p(t|uy,\lambda)}{p(t|Sy_0,\lambda_0)}\Big)^\gamma\Big\}.\end{aligned} \tag{B.14}$$

If we transform $y\rightarrow y+(wy_0+vz)/u$, we can write this result equivalently as

$$\begin{aligned}E_\gamma(S,\lambda_0)=\int \mathrm{D}y_0\int \mathrm{d}t\,p(t|Sy_0,\lambda_0)\log p(t|Sy_0,\lambda_0)\\ -\frac{\partial}{\partial\gamma}\mathrm{extr}_{u,v,w;\lambda}\Big\{\zeta\Big(\frac{v^2}{2u^2}+\log u\Big)\\ +\int \mathrm{D}z\mathrm{D}y_0\int \mathrm{d}t\,p(t|Sy_0,\lambda_0)\log\int \mathrm{D}y\,p^\gamma(t|uy+wy_0+vz,\lambda)\Big\}.\end{aligned} \tag{B.15}$$

At the relevant saddle point, the order parameter derivative of the function that is being extremized will by definition be zero, so

$$\begin{aligned}E_\gamma(S,\lambda_0)=\int \mathrm{D}y_0\int \mathrm{d}t\,p(t|Sy_0,\lambda_0)\,\{\log p(t|Sy_0,\lambda_0)\\ -\int \mathrm{D}z\left[\frac{\int \mathrm{D}y\,p^\gamma(t|uy+wy_0+vz,\lambda)\log p(t|uy+wy_0+vz,\lambda)}{\int \mathrm{D}y\,p^\gamma(t|uy+wy_0+vz,\lambda)}\right]\Big\}\end{aligned} \tag{B.16}$$

in which the order parameters $\{u,v,w;\lambda\}$ are to be evaluated at the saddle point of

$$\begin{aligned}\Psi_{\mathrm{RS}}(u,v,w;\lambda)=\zeta\Big(\frac{v^2}{2u^2}+\log u\Big)\\ +\int \mathrm{D}z\mathrm{D}y_0\int \mathrm{d}t\,p(t|Sy_0,\lambda_0)\log\int \mathrm{D}y\,p^\gamma(t|uy+wy_0+vz,\lambda).\end{aligned} \tag{B.17}$$

Appendix C. The limits $\zeta\rightarrow 0$ and $\zeta\rightarrow 1$

For $\zeta\rightarrow 0$, the limit of no overfitting, we immediately find from (66) and (70) that $\tilde{u},v\rightarrow 0$. To find also w and $\lambda(t)$ we need to go to the next order in ζ, using $W(z)=z+\mathcal{O}(z^2)$. This results in

$$\frac{\zeta v^2}{\tilde{u}^4}=\int \mathrm{D}z\mathrm{D}y_0\int \mathrm{d}t\,p(t|Sy_0,\lambda_0)\Big[1-\mathrm{e}^{wy_0+vz}\Lambda(t)\Big]^2+\mathcal{O}(\tilde{u}^2) \tag{C.1}$$

$$\frac{\zeta}{\tilde{u}^2} = \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0)\mathrm{e}^{wy_0+vz}\Lambda(t) + \mathcal{O}(\tilde{u}^2) \tag{C.2}$$

$$0 = \int \mathrm{D}z\mathrm{D}y_0\, y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0)\mathrm{e}^{wy_0+vz}\Lambda(t) + \mathcal{O}(\tilde{u}^2) \tag{C.3}$$

$$\frac{p(t)}{\lambda(t)} = \int \mathrm{D}z\mathrm{D}y_0 \int_t^\infty \mathrm{d}t'\, p(t'|Sy_0, \lambda_0)\mathrm{e}^{wy_0+vz} + \mathcal{O}(\tilde{u}^2). \tag{C.4}$$

It follows that $v = \mathcal{O}(\tilde{u})$ and $\tilde{u} = \mathcal{O}(\sqrt{\zeta})$ for $\zeta \to 0$, and that $\lim_{\zeta\to 0} w$ and $\lim_{\zeta\to 0} \lambda(t)$ are to be solved from the following two coupled equations:

$$0 = \int \mathrm{D}y_0\, y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0)\mathrm{e}^{wy_0}\Lambda(t) \tag{C.5}$$

$$\frac{p(t)}{\lambda(t)} = \int \mathrm{D}y_0 \int_t^\infty \mathrm{d}t'\, p(t'|Sy_0, \lambda_0)\mathrm{e}^{wy_0}. \tag{C.6}$$

After some simple rewriting and integration by parts over time, they take the alternative forms

$$0 = \int \mathrm{D}y_0\, y_0\mathrm{e}^{(w-S)y_0} \int \mathrm{d}t\, p(t|Sy_0, \lambda_0)\frac{\lambda(t)}{\lambda_0(t)} \tag{C.7}$$

$$p(t) = \int \mathrm{D}y_0\, \mathrm{e}^{(w-S)y_0}p(t|Sy_0, \lambda_0)\frac{\lambda(t)}{\lambda_0(t)}\mathrm{e}^{wy_0}. \tag{C.8}$$

From this we immediately confirm the correct solution $\lim_{\zeta\to 0} w = S$ and $\lim_{\zeta\to 0} \lambda(t) = \lambda_0(t)$, which describes perfect inference, as expected for $\zeta \to 0$. From the pair (47) and (48) we also find the correct corresponding value for $\lim_{\zeta\to 0}\lim_{\gamma\to\infty} E_\gamma(S, \lambda_0)$:

$$\lim_{\zeta\to 0}\lim_{\gamma\to\infty} \mathcal{P}_\gamma(x, x', t) = \int \mathrm{D}y_0\, p(t|Sy_0, \lambda_0)\delta[x - Sy_0]\delta[x' - Sy_0] \tag{C.9}$$

$$\lim_{\zeta\to 0}\lim_{\gamma\to\infty} E_\gamma(S, \lambda_0) = 0. \tag{C.10}$$

Next we turn to the limit $\zeta \to 1$. Here it follows from (70) that $\tilde{u} \to \infty$, and we need the expansion of $W(z)$ for large arguments, i.e. $W(z) = \log z - \log(\log z) + \ldots$. With a modest amount of foresight we make the ansatz $\tilde{u} = \kappa/\sqrt{1-\zeta} + \mathcal{O}(\log(1/(1-\zeta))$ and $v, w = \mathcal{O}(\log(1/(1-\zeta))$ for $\zeta \to 1$. Using

$$W\big(\tilde{u}^2\mathrm{e}^{\tilde{u}^2+wy_0+vz}\Lambda(t)\big) = \frac{\kappa^2}{1-\zeta} + \mathcal{O}(\log(\frac{1}{1-\zeta})) \tag{C.11}$$

our $\gamma \to \infty$ order parameter equations then give

$$\zeta v^2 = \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0)\Big[\mathcal{O}(\log(\frac{1}{1-\zeta}))\Big]^2 \tag{C.12}$$

$$\zeta = \int \mathrm{D}z\mathrm{D}y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0)[1 - \mathcal{O}(1-\zeta)] \tag{C.13}$$

$$0 = \int \mathrm{D}z\mathrm{D}y_0\, y_0 \int \mathrm{d}t\, p(t|Sy_0, \lambda_0)\; \mathcal{O}\Big((1-\zeta)\log(\frac{1}{1-\zeta})\Big) \tag{C.14}$$

$$\frac{p(t)}{\lambda(t)} = \int \mathrm{D}z\mathrm{D}y_0 \int_t^\infty \mathrm{d}t'\, p(t'|Sy_0,\lambda_0)\frac{1}{\Lambda(t')} \\ \times \Big[1+\mathcal{O}\Big((1-\zeta)\log(\frac{1}{1-\zeta})\Big)\Big]. \tag{C.15}$$

Our scaling ansatz is seen to be consistent with the three scalar order parameter equations. Hence $\tilde{u}$, v and w all diverge at a phase transition point $\zeta = 1$, whereas for the functional order parameter equation we find in the limit $\zeta \to 1$:

$$\frac{p(t)}{\lambda(t)} = \int_t^\infty \mathrm{d}t'\, \frac{p(t')}{\Lambda(t')}. \tag{C.16}$$

From this it follows after differentiation that $\frac{\mathrm{d}}{\mathrm{d}t}[p(t)\Lambda(t)/\lambda(t)] = 0$, and after some further manipulations one arrives at the following degenerate solution for $\Lambda(t)$:

$$\lim_{\zeta\uparrow 1}\lim_{\gamma\to\infty} \Lambda(t) = \begin{cases} 0 & \text{for} \quad t < \tau \\ 1 & \text{for} \quad t = \tau \\ \infty & \text{for} \quad t > \tau. \end{cases} \tag{C.17}$$

Apparently, as one varies the ratio ζ of the number of covariates over the number of samples in the deterministic Cox model, the integrated inferred base hazard rate changes from the correct shape $\Lambda_0(t)$ at $\zeta = 0$ to a step function at the phase transition point $\zeta = 1$, with the discontinuity at some time point τ that should follow from inspecting sub-leading orders in $1-\zeta$. Moreover, at this transition (if not even earlier) one expects to find breaking of the assumed replica symmetry.

Appendix D. Asymptotic form of the event time distribution

Here we calculate the asymptotic form of the function $g(x) = \int \mathrm{D}y\ \mathrm{e}^{Sy - x\exp(Sy)}$ for $x \to \infty$, and derive expression (88). Working out the definition gives

$$\begin{aligned} \log g(x) &= \frac{1}{2}S^2 + \log \int \frac{\mathrm{d}y}{\sqrt{2\pi}}\ \mathrm{e}^{-\frac{1}{2}y^2 - x\exp(S^2+Sy)} \\ &= \frac{1}{2}S^2 + \log \int \frac{\mathrm{d}y}{\sqrt{2\pi}}\ \mathrm{e}^{-\varphi(y,\mathrm{e}^{S^2}x)} \end{aligned} \tag{D.1}$$

with

$$\varphi(y,\eta) = \frac{1}{2}y^2 + \eta\mathrm{e}^{Sy}. \tag{D.2}$$

Differentiation shows that the function $\varphi(y,\eta)$ is mimimal at $y = -W(\eta S^2)$, where $W(z)$ is Lambert's W-function [35]. Expansion of $\varphi(y,\eta)$ around its minimum gives:

$$\begin{aligned} \varphi(y,\eta) = {} & \frac{1}{2S^2}\Big(W(\eta S^2)+1\Big)^2 - \frac{1}{2S^2} + \frac{1}{2}\Big[W(\eta S^2)+1\Big]\Big(y + \frac{1}{S}W(\eta S^2)\Big)^2 \\ & + \mathcal{O}(\big[W(\eta S^2)+1\big]\Big(y+\frac{1}{S}W(\eta S^2)\Big)^3. \end{aligned} \tag{D.3}$$

This leads to the following Gaussian approximation of the integral over y:

$$\log \int \frac{\mathrm{d}y}{\sqrt{2\pi}}\,\mathrm{e}^{-\varphi(y,\eta)} = \frac{1}{2S^2} - \frac{1}{2S^2}\Big(W(\eta S^2)+1\Big)^2 \\ + \mathcal{O}\Big(\log\Big[W(\eta S^2)+1\Big]\Big). \tag{D.4}$$

Application to $\eta = x\mathrm{e}^{S^2}$ then gives:

$$\log g(x) = -\frac{1}{2S^2}\left[W(xS^2\mathrm{e}^{S^2})+1\right]^2 - \frac{1}{2}\log W(xS^2\mathrm{e}^{S^2}) + \mathcal{O}(1). \tag{D.5}$$

Finally, for $x\to\infty$ we can use $W(z) = \log z - \log\log z + \mathcal{O}(\log\log z/\log z)$ to obtain

$$\log g(x) = -\frac{1}{2S^2}(\log x)^2 + \frac{1}{S^2}\log x.\log\log x + \mathcal{O}(\log x). \tag{D.6}$$

References

[1] Hougaard P 2001 *Analysis of Multivariate Survival Data* (New York: Springer)
[2] Klein J P and Moeschberger M L 2003 *Survival Analysis—Techniques for Censored and Truncated Data* (New York: Springer)
[3] Ibrahim J G, Chen M H and Sinha D 2010 *Bayesian Survival Analysis* (New York: Springer)
[4] Crowder M 2012 *Multivariate Survival Analysis and Competing Risks* (London: CRC Press)
[5] Cox D R 1972 *J. R. Stat. Soc.* B **34** 187
[6] Witten D M and Tibshirani R 2009 *J. R. Stat. Soc.* B **71** 615
[7] Witten D M and Tibshirani R 2010 *Stat. Meth. Med. Res.* **19** 29
[8] Keiding N, Andersen P K and Klein J P 1997 *Stat. Med.* **16** 215
[9] Vaida F and Xu R 2000 *Stat. Med.* **19** 3309
[10] Duchateau L and Jansen P 2008 *The Frailty Model* (*Statistics for Biology and Health*) (New York: Springer)
[11] Wienke A 2010 *Frailty Models in Survival Analysis* (*CRC Biostatistics Series*) (Boca Raton: Chapman and Hall)
[12] Rowley M *et al* 2017 *Stat. Med.* **37** 2100
[13] Grigoriadis A *et al* 2017 unpublished
[14] Concato J, Feinstein A R and Holford T R 1993 *Ann. Intern. Med.* **118** 201
[15] Babyak M A 2004 *Psychosomatic Med.* **66** 411
[16] Breslow N E 1972 Discussion section of the paper [5] by DR Cox
[17] MacKay D J C 2003 *Information Theory, Inference and Learning Algorithms* (Cambridge: Cambridge University Press)
[18] Coolen A C C, Kühn R and Sollich P 2005 *Theory of Neural Information Processing Systems* (Oxford: Oxford University Press)
[19] Peduzzi P, Concato J, Feinstein A R and Holford T 1995 *J. Clin. Epidemiol.* **48** 1503
[20] Kawada T 2011 *Int. J. Cardiol.* **153** 110
[21] Dobbin K K and Song X 2013 *Biostatistics* **14** 639
[22] Gardner E 1987 *Europhys. Lett.* **4** 481
[23] Sherrington D and Kirkpatrick S 1975 *Phys. Rev. Lett.* **35** 1792
[24] Parisi G 1979 *Phys. Lett.* A **73** 203
[25] Mézard M, Parisi G and Virasoro M A 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
[26] Monasson R 1998 *J. Phys. A: Math. Gen.* **31** 513
[27] Van Mourik J and Coolen A C C 2001 *J. Phys. A: Math. Gen.* **34** L111
[28] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing* (Oxford: Oxford University Press)
[29] Amit D J, Gutfreund H and Sompolinsky H 1985 *Phys. Rev.* A **32** 1007
[30] Rabello S, Coolen A C C, Pérez-Vicente C J and Fraternali F 2008 *J. Phys. A: Math. Theor.* **41** 285004

[31] Agliari E, Annibale A, Barra A, Coolen A C C and Tantari D 2013 *J. Phys. A: Math. Theor.* **46** 415003
[32] Challet D, Marsili M and Zecchina R 2000 *Phys. Rev. Lett.* **84** 1824
[33] Marsili M and Challet D 2001 *Phys. Rev.* E **64** 056138
[34] Cover T M and Thomas J A 1991 *Elements of Information Theory* (New York: Wiley)
[35] Corless R M, Gonnet G H, Hare D E G, Jeffrey D J and Knuth D E 1996 *Adv. Comput. Math.* **5** 329
[36] Asmussen S, Jensen J L and Rojas-Nandayapa L 2015 *Methodol. Comput. Appl. Prob.* **18** 441
[37] Derrida B 1981 *Phys. Rev.* B **24** 2613
[38] Gradshteyn I S and Rhyzik I M 1979 *Table of Integrals, Series and Products* (London: Academic)