
SQUARE ROOT COX’S SURVIVAL ANALYSIS BY THE FITTEST LINEAR AND NEURAL NETWORKS MODEL

Maxime van Cutsem

Department of Mathematics, University of Geneva

`maxime.vancutsem@unige.ch`

Sylvain Sardy

Department of Mathematics, University of Geneva

`sylvain.sardy@unige.ch`

ABSTRACT

We revisit Cox’s proportional hazard models and LASSO in the aim of improving feature selection in survival analysis. Unlike traditional methods relying on cross-validation or BIC, the penalty parameter λ is directly tuned for feature selection and is asymptotically pivotal thanks to taking the square root of Cox’s partial likelihood. Substantially improving over both cross-validation LASSO and BIC subset selection, our approach has a phase transition on the probability of retrieving all and only the good features, like in compressed sensing. The method can be employed by linear models but also by artificial neural networks.

Keywords Cox model · model selection · pivotal statistic · quantile universal threshold · survival analysis.

1 Introduction

Survival analysis is a branch of Statistics concerned with the study of time-to-event data. Contrarily to a classical binary classification problem, the outcome of interest is not simply whether an event occurs, but when it occurs. If the endpoint is the death of a patient, the resulting data are literally survival times. Data of similar form also arise in other fields, such as engineering, where the survival time may represent the lifespan of a machine component. A distinctive challenge is that, at the time of analysis, the event of interest may not yet have been observed for every individual. This phenomenon, known as censoring, arises when the event has not occurred by the end of the study or observation period.

In addition to time-to-event and censoring information, most survival studies also collect covariates that may influence survival. A typical example is a clinical trial in oncology, where factors such as tumor stage, patient age, smoking status, or gene expression can have a decisive impact on patient outcomes. In

modern high-dimensional datasets such as genomic or imaging data, the number of potential covariates is often large and careful variable selection becomes crucial. Selecting the most relevant variables not only improves model interpretability and reduces overfitting, but also enhances the ability to identify meaningful prognostic and predictive factors that can guide clinical decision-making.

Formally, to account for censoring and the influence of covariates, survival analysis assumes that two random variables underlie the data: the event time T and the censoring time R . The observed data consist of triplets $(y_i, c_i, \mathbf{x}_i)_{i=1, \dots, n}$ that are sampled from a possibly right-censored event time $Y_i = \min(T_i, R_i) \geq 0$ (say, death or machine failure), the censoring indicator $C_i = 1_{\{T_i \leq R_i\}} \in \{0, 1\}$ (with $C_i = 1$ if uncensored), $\mathbf{x}_i \in \mathbb{R}^p$ are covariates possibly influencing the survival time of the i -th individual. The assumption is that T_i and R_i are conditionally independent given $\mathbf{x}_i$. We denote by $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{c} \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ the vectorized form of the data.

The primary objective is to characterize the distribution of the event time T , the survival function $S(t|\mathbf{x}) := \mathbb{P}(T > t|\mathbf{x})$, which represents the probability that an individual with covariates $\mathbf{x}$ survives beyond time t , and the hazard function

$$h(t|\mathbf{x}) := \lim_{\delta t \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + \delta t | T \geq t)}{\delta t}$$

that describes the instantaneous risk of experiencing the event at time t , given survival up to that time. While the survival function captures the long-term probability of survival, the hazard function captures the instantaneous failure rate. The two are tightly connected:

$$S(t|\mathbf{x}) = \exp\left(-\int_0^t h(u|\mathbf{x}) du\right) =: \exp\{-H(t|\mathbf{x})\}. \quad (1)$$

1.1 Cox model

The proportional hazard model for survival data, also known as the Cox model [Cox, 1972], assumes that the hazard at time t depends on the covariates through

$$h(t|\mathbf{x}) = h_0(t)\eta(\mathbf{x}),$$

where h_0 is the unknown baseline hazard function and η equals one when the covariates have no influence on survival. The hazard rate η being positive by definition, the common approach is to write that $\eta(\mathbf{x}) = \exp\{\mu_\theta(\mathbf{x})\}$, where μ_θ is a function of the covariates $\mathbf{x}$ with parameters θ . Most practitioners assume a linear model $\mu_\theta(\mathbf{x}) = \mathbf{x}^\top \theta = \sum_{j=1}^p \theta_j x_j$, with no intercept to avoid a non-identifiability issue with the baseline hazard function h_0 .

Fitting the Cox model entails estimating θ from observed data. Remarkably, Cox showed that estimation of θ can be separated from estimation of $h_0(t)$ by introducing the partial likelihood, which eliminates the baseline hazard. Specifically, assuming censoring is independent of the covariates, the log-partial

likelihood is

$$\ell^{\text{partial}}(\mu_{\boldsymbol{\theta}}(X); \mathbf{y}, \mathbf{c}) = \sum_{i=1}^n c_i [\mu_{\boldsymbol{\theta}}(\mathbf{x}_i) - \log \sum_{j: y_j \geq y_i} \exp\{\mu_{\boldsymbol{\theta}}(\mathbf{x}_j)\}], \quad (2)$$

where the outer sum is taken only over individuals who experienced the event ($c_i = 1$), and the inner sum ranges over the risk set at time y_i , i.e., all subjects still under observation just prior to y_i . Unlike the full likelihood, the partial likelihood does not depend on $h_0(t)$ and can be maximized directly with respect to $\boldsymbol{\theta}$. It is, in fact, a profile likelihood [Murphy and Van der Vaart, 2000]. Letting $\hat{\boldsymbol{\theta}}$ maximize (2), the baseline hazard function can then be estimated through the Breslow estimator of the cumulative baseline hazard:

$$\hat{H}_0(t) = \sum_{i: y_i \leq t} \frac{c_i}{\sum_{j: y_j \geq y_i} \exp\{\mu_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_j)\}}.$$

This estimator increases in steps at each observed event time. Recalling (1), the corresponding baseline survival function is then $\hat{S}_0(t) = \exp\{-\hat{H}_0(t)\}$. Prediction for a subject with covariates $\mathbf{x}$ follows from the proportional hazards assumption:

$$\hat{S}(t | \mathbf{x}) = \hat{S}_0(t)^{\exp\{\mu_{\hat{\boldsymbol{\theta}}}(\mathbf{x})\}}.$$

1.2 Model selection for survival analysis

The linearity assumption for $\mu_{\boldsymbol{\theta}}$ in the Cox model is convenient for at least three reasons. First, the negative log-likelihood function is convex in the parameters $\boldsymbol{\theta} \in \mathbb{R}^p$, and hence easy to optimize. Second, the linear model can be interpreted as it captures the linear main effects in each input. Third, the set of influential inputs

$$\mathcal{S} := \{j \in \{1, \dots, p\} : x_j \text{ has an impact on survival}\} \quad (3)$$

becomes the indices of entries of $\boldsymbol{\theta}$ different from zero: $\mathcal{S} \equiv \{j \in \{1, \dots, p\} : \theta_j \neq 0\}$.

The search of $\mathcal{S}$ is a central objective in survival analysis. From a practical standpoint, determining which covariates significantly influence the risk of death is crucial for scientific and medical applications. Beside univariate tests that neglect dependence in covariates, the classical approach to search which entries of $\boldsymbol{\theta}$ are different from zero is subset selection. An information criterion like AIC [Akaike, 1974, 1998] or BIC [Schwarz, 1978] that penalize Cox's negative log-partial likelihood by an additional term proportional to the number of non-zero entries in $\hat{\boldsymbol{\theta}}$ is used to fit the data. Minimizing these information criteria is an NP-hard discrete optimization problem, however. Instead, Tibshirani [1997] proposed a continuous optimization, moreover convex, by penalizing the likelihood term with the sparsity inducing LASSO penalty [Tibshirani, 1996] and by solving

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} -\ell^{\text{partial}}\{\mu_{\boldsymbol{\theta}}(X); \mathbf{y}, \mathbf{c}\} + \lambda \|\boldsymbol{\theta}\|_1, \quad (4)$$

where λ controls the complexity using cross-validation. Owing to the isotropic nature of the penalty, covariates must be mean-centered and rescaled to have unit variance.

Selection of the model complexity is crucial. Focusing on LASSO, there may exist (if the signal to noise ratio is large enough) an ideal λ for which LASSO discovers all and only the important variables, that is $\hat{\mathcal{S}} = \mathcal{S}$, called exact support recovery. If λ is selected too small, too many false discoveries are present, while, if too large, some important variables are not discovered. Currently the selection of λ is based on cross-validation [Tibshirani, 1997], a method prone to a high false discovery rate. LASSO is also known to have many false discoveries because of a strong shrinkage effect [Su et al., 2015], so that, no ideal λ may exist with LASSO, that is $\hat{\mathcal{S}} \neq \mathcal{S}$ regardless the value of λ .

The goal of this paper is to improve model selection in survival analysis by achieving higher probability of exact support recovery. The method applies to linear models and shallow to deep artificial neural networks.

2 Our Proposal

To enhance estimation of $\mathcal{S}$ and prediction, we change (4) to rather solve

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \sqrt{-\ell^{\text{partial}}\{\mu_{\boldsymbol{\theta}}(X); \mathbf{y}, \mathbf{c}\} + \lambda P(\boldsymbol{\theta})}, \quad (5)$$

where three changes are made. First we use the square root of Cox's negative partial likelihood, for a reason that will become clear. Second we propose a new selection rule for λ to decrease false detections and a new sparsity inducing penalty $P(\boldsymbol{\theta})$ different than LASSO's ℓ_1 penalty. Third we generalize survival analysis to fit artificial neural networks models so as to retrieve $\mathcal{S}$ even when the association $\mu_{\boldsymbol{\theta}}$ is nonlinear.

2.1 Selection of λ

Our method to select λ applies to LASSO's penalty, but also to a larger class of penalties that better retrieve $\mathcal{S}$ thanks to less shrinkage. We consider penalty functions of the form

$$P_{\nu}(\boldsymbol{\theta}) = \sum_{j=1}^p \rho_{\nu}(\theta_j) \quad \text{with} \quad \rho_{\nu}(\theta) = \frac{|\theta|}{1 + |\theta|^{1-\nu}}, \quad (6)$$

where $\nu \in (0, 1)$ [van Cutsem et al., 2025]. When $\nu = 1$, this reduces to the convex ℓ_1 penalty used in LASSO (up to a scaling factor of $1/2$). As $\nu \rightarrow 0$, the penalty approaches the discrete ℓ_0 penalty for large $|\theta|$. Thus, the family of penalties spans a continuum between ℓ_0 and ℓ_1 : moving closer to ℓ_0 reduces shrinkage of relevant coefficients at the cost of non-convexity. In practice we use $\nu = 0.1$ to keep away from the near discrete ℓ_0 penalty. Using this penalty leads to the following proposition:

Proposition 1 (Zero-thresholding function). *Assuming a linear learner $\mu_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\theta}$ and $P(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ or $P(\boldsymbol{\theta}) = P_{\nu}(\boldsymbol{\theta})$ as in (6), there exist finite values of the penalty λ such that a minimum to (5) is created at $\boldsymbol{\theta} = \mathbf{0}$. The smallest such λ is given by the zero-thresholding function:*

$$\lambda_0(\mathbf{y}, \mathbf{c}; X) = \frac{\|\nabla_{\boldsymbol{\theta}} \ell^{\text{partial}}\{\mu_{\boldsymbol{\theta}}(X); \mathbf{y}, \mathbf{c}\}|_{\boldsymbol{\theta}=\mathbf{0}}\|_{\infty}}{2\sqrt{-\ell^{\text{partial}}\{\mu_{\mathbf{0}}(X); \mathbf{y}, \mathbf{c}\}}}. \quad (7)$$

The proof of Proposition 1 stems from Taylor expansion of the unpenalized term $\phi(\boldsymbol{\theta}) = \sqrt{-\ell^{\text{partial}}\{\mu_{\boldsymbol{\theta}}(X); \mathbf{y}, \mathbf{c}\}}$ in (5) at $\boldsymbol{\theta} = \mathbf{0}$. The left and right derivatives of the penalty term at $\boldsymbol{\theta} = \mathbf{0}$ being $\pm\lambda$ coordinatewise, then if λ is larger than the amplitude of the gradient of ϕ at $\boldsymbol{\theta} = \mathbf{0}$, a minimum exists at $\boldsymbol{\theta} = \mathbf{0}$. For a detailed proof, see van Cutsem et al. [2025]. The first order term of Taylor expansion leads to a denominator in λ_0 above thanks to the square-root applied to the partial likelihood, and this denominator will be crucial to balance the variance of the numerator and make a statistic pivotal (see Definition 1 below).

Inspired by the universal threshold [Donoho and Johnstone, 1994] and theoretical results on thresholding estimators [Bühlmann and van de Geer, 2011], the quantile universal threshold [Giacobino et al., 2017] calibrates the selection of λ so as to retrieve zero covariate (that is, $\hat{\mathcal{S}} = \emptyset$) with high probability when there is indeed no association ($\mathcal{S} = \emptyset$) between covariates and survival time. Achieving this desired probabilistic property requires λ to be sufficiently large to set $\hat{\boldsymbol{\theta}}_{\lambda} = \mathbf{0}$. The question of how large is answered by Proposition 1. Indeed, define the random variable $\Lambda = \lambda_0(\mathbf{Y}_0, \mathbf{C}_0; X)$, where $\mathbf{Y}_0$ and $\mathbf{C}_0$ are random outcomes under the pure noise assumption (that is, $\boldsymbol{\theta} = \mathbf{0}$), Proposition 1 tells us that $\mathbb{P}(\hat{\boldsymbol{\theta}}_{\lambda} = \mathbf{0}) = \mathbb{P}(\lambda \geq \lambda_0(\mathbf{Y}_0, \mathbf{C}_0; X))$. So if one sets λ to $\lambda_{\alpha}^{\text{QUT}} = F_{\Lambda}^{-1}(1 - \alpha)$, then $\mathbb{P}(\hat{\boldsymbol{\theta}}_{\lambda} = \mathbf{0}) = 1 - \alpha$. This probabilistic property leads to the definition of the quantile universal threshold.

Definition 1 (QUT). *Let $H_0 : \mathcal{S} = \emptyset$ be the null hypothesis that no input has an impact on survival time and let X be the fixed $n \times p$ matrix of inputs. Define the random variable*

$$\Lambda = \lambda_0(\mathbf{Y}_0, \mathbf{C}_0; X), \quad (8)$$

where $\mathbf{Y}_0$ is the random vector of recorded time events and $\mathbf{C}_0$ is the random vector of censor indicators, both under H_0 for the Cox model. Given a small probability α , the quantile universal threshold (QUT) is defined by $\lambda_{\alpha}^{\text{QUT}} = F_{\Lambda}^{-1}(1 - \alpha)$.

The probability α is the false discovery rate under the null model and should therefore be chosen small, say $\alpha = 0.05$. Choosing the QUT penalty leads to the property that $\mathbb{P}_{H_0}(\hat{\mathcal{S}} = \emptyset) = 1 - \alpha$. The statistic Λ satisfies the following asymptotic result.

Proposition 2. *Assuming the n inputs are i.i.d. with covariance matrix $\Sigma_{\mathbf{X}}$ with p fixed, the law of $\sqrt{\log n} \Lambda$ is asymptotically pivotal.*

Proof. The statistic Λ in (8) can be written as A_n/B_n , where

$$A_n = \left\| \sum_{i=1}^n C_i \left(\mathbf{x}_i - \frac{1}{N_i} \sum_{j: Y_j \geq Y_i} \mathbf{x}_j \right) \right\|_{\infty} / \sqrt{n} \quad \text{and} \quad B_n = 2 \sqrt{\sum_{i=1}^n C_i \log N_i / \sqrt{n}}$$

with $N_i = \sum_{j: Y_j \geq Y_i} 1$. Since C_i and N_i are discrete random variables, $\mathbf{Y}$ only plays a role through the order of its entries and the X matrix is fixed, the random variable Λ takes a finite number of values. Under $H_0 : \mathcal{S} = \emptyset$, letting $p_0 = \mathbb{E}_{H_0}(C_i)$ for $i \in \{1, \dots, n\}$, we have that $A_n \rightarrow_d \|\mathbf{Z}_{p_0}\|_{\infty}$ with

$\mathbf{Z}_{p_0} \sim \mathcal{N}(\mathbf{0}, p_0 \Sigma_{\mathbf{X}})$ and $B_n^2 = 4(p_0 \log n + O_p(1))$ [Li et al., 2018]. So invoking the continuous mapping and Slutsky's theorems, one gets that $2\sqrt{\log n} \Lambda \rightarrow_d \|\mathbf{Z}\|_\infty$, where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{X}})$. $\square$

The role of the square root function applied to the partial likelihood is now clear with Proposition 2: without it, the numerator alone would not be pivotal as it would asymptotically depend on the nuisance parameter $p_0 = \mathbb{E}_{H_0}(C)$, and this censoring probability is unknown in practice; recall that most of the time some inputs have an impact on the survival rate, so that the data set is not under H_0 , which makes p_0 difficult to estimate.

Thanks to the asymptotic pivotal property of Λ , one can estimate the theoretical quantile $\lambda_\alpha^{\text{QUT}}$ of Λ of Definition 1 by using the Gaussianity of $\mathbf{Z}$ in Proposition 2; since the law of Λ is conditional on X , one can simulate m realizations from $\mathcal{N}(\mathbf{0}, X^T X/n)$, take the sup-norms to get m realizations from Λ before taking the upper α -quantiles (the larger m the more precise the desired quantile of Λ). For small sample size, an alternative is by bootstrapping the data. The matrix X of covariates being fixed, one can bootstrap m times the pairs (y_i, c_i) to preserve their dependence to get $(\mathbf{y}, \mathbf{c})_{k=1, \dots, m}^{\text{boot}}$, and then calculate $\lambda^{(k)} = \lambda_0((\mathbf{y}, \mathbf{c})_k^{\text{boot}}; X)$, $k = 1, \dots, m$ to get a bootstrapped sample from Λ . An empirical upper α -quantile of the bootstrapped $\lambda^{(k)}$'s provides an estimate of the quantile universal threshold $\lambda_\alpha^{\text{QUT}}$. Appropriate under $H_0 : \mathcal{S} = \emptyset$, this bootstrap estimate is not affected by the possible shift in the censoring probability under $H_1 : \mathcal{S} \neq \emptyset$ thanks to the pivotal property of the statistic Λ in (8). And under H_1 , the bootstrap brakes the possible link between survival time, censoring and covariates, as under the null model. We observe a close match between the Gaussian and bootstrap approaches.

2.2 Extension to neural networks

We have so far assumed a linear model for $\mu_\theta(\mathbf{x})$ in (2), as in Cox [1972]. In the same spirit as Lemhadri et al. [2021] who extended the linear survival model of Tibshirani [1997] to artificial neural networks (ANNs), we now extend our methodology to ANNs. So we let $\mu_\theta(\mathbf{x})$ be a fully connected ANN. Neural networks are dense in function space when letting its number of neurons and layers grow large (which requires many more parameters). The motivation for allowing nonlinear models remains our original goal: good estimation of the influential inputs on the survival rate, that is, good estimation of $\mathcal{S}$ in (3). Indeed a linear model may miss important inputs that do not have a monotone (e.g., linear) and additive impact on survival. With ANN, relevant features are identified as the columns of the first layer weight matrix W_1 (directly connected to the inputs) that are non-zero.

The adaptation of the methodology developed above from linear models to ANNs is straightforward. The optimization problem 5 becomes

$$\min_{\theta \in \mathbb{R}^\gamma} \sqrt{-\ell^{\text{partial}}\{\mu_\theta(X; \mathbf{y}, \mathbf{c})\}} + \lambda P(W_1), \quad (9)$$

where γ denotes the total number of network parameters and W_1 is the weight matrix of the input layer. The choice of λ follows the same principle as in the linear case, but must be rescaled according to the

network structure:

$$\lambda_{\alpha, \text{ANN}}^{\text{QUT}} = \kappa^{L-1} \pi_L \lambda_{\alpha}^{\text{QUT}}, \quad (10)$$

where L is the total number of layers, $\kappa = \sup_t |\sigma'(t)|$ depends on the activation function σ , and $\pi_L = \sqrt{\prod_{j=3}^L p_j}$, $\pi_1 = \pi_2 = 1$ with p_j denoting the width of layer j . For a linear model in particular, $L = 1$ and we have indeed $\lambda_{\alpha, \text{ANN}}^{\text{QUT}} = \lambda_{\alpha}^{\text{QUT}}$. For details and a proof of (10), see van Cutsem et al. [2025]. What makes the extension straightforward is that the calculation of the gradient of the unpenalized cost is the only ingredient needed for selecting λ and for solving the optimization problem.

2.3 Optimization scheme

The cost function (9) is no longer convex either due to the nonlinear form of the ANN function, due to the non-convex penalty (6) when $\nu = 0.1$, or due to both. To reduce the risk of convergence to poor local minima, we adopt a conservative training strategy inspired by simulated annealing. Specifically, we solve (5) (or (9) for ANNs) over a sequence of (λ, ν) values. The regularization path is defined by $\lambda_i = \frac{\exp(i-1)}{1+\exp(i-1)} \lambda_{\alpha}^{\text{QUT}}$ for $i \in \{0, \dots, 5, +\infty\}$, while ν is varied over $\{0.9, 0.7, 0.4, 0.3, 0.2, 0.1\}$. At each stage, the solution obtained for λ_i is used as initialization for λ_{i+1} , producing a sequence of progressively sparser models until reaching $\lambda_{\alpha}^{\text{QUT}}$ at the final step. Optimization in the intermediate phases is carried out with steepest descent, whereas the final step employs the FISTA algorithm [Beck and Teboulle, 2009]. After completing all training phases, we reduce the penalized parameter by keeping only the nonzero coefficients. A final training phase without regularization is then applied to the reduced model.

Except for the first step of the annealing schedule for $i = 0$, all the other steps (including the last FISTA step) use the warm start, meaning that the initial values of the parameters are not far from the solution since the successive (λ_i, ν_i) are close to one another. So convergence is quickly reached. The final unpenalized refit is likewise quick, since it begins from a well-aligned, sparsified model.

3 Evaluation of the method by phase transition

This section illustrates and quantifies the advantage of using our survival analysis method for practitioners in hospitals and in manufacturing or maintenance industries. In Sections 3.1 (linear models) and 3.2 (nonlinear models), we compare methods on simulated data through the prism of a phase transition in the probability of exact support recovery $\text{PESR} = \mathbb{P}(\hat{\mathcal{S}} = \mathcal{S})$ as a function of the sparsity level $s = |\mathcal{S}|$. This concept, originally studied by Candès and Tao [2005] and Donoho [2006] for compressed sensing, describes a sharp transition: the probability of retrieving $\mathcal{S}$ is high when s is low, and suddenly drops to zero when s gets large. The longer PESR remains near one for larger s , the better the detection method. Looking at the true positive rate $\text{TPR} := \mathbb{E}(|\hat{\mathcal{S}} \cap \mathcal{S}|/|\mathcal{S}|)$ and the false discovery rate $\text{FDR} := \mathbb{E}(|\hat{\mathcal{S}} \cap \bar{\mathcal{S}}|/|\hat{\mathcal{S}}|)$ as a function of s also helps understand the reason why exact support recovery fails. To assess predictive accuracy, we additionally report the concordance index (C-index) [Uno et al., 2011], a widely used metric for censored survival data. Recall that a C-index of 0.5 corresponds to random prediction, whereas 1.0 indicates perfect concordance with observed event times.

To compute our different performance metrics, we generate $m = 200$ independent datasets for each sparsity level s . Each data set is obtained from an $n \times p$ Gaussian input matrix X (i.i.d. standard Gaussian entries) by randomly selecting s columns to form the $n \times s$ submatrix $X_s^{(k)}$ for $k = 1, \dots, m$. Given a true log-risk function $\mu(\cdot) : \mathbb{R}^s \rightarrow \mathbb{R}$, survival times are simulated under the exponential Cox model:

$$T_s^{(k)} \sim \text{Exp}[h_0 \cdot \exp\{\mu(X_s^{(k)})\}], \quad k = 1, \dots, m$$

with $h_0 = 1$. For each sample, we choose a censoring time that yields 50% of censored data. In the following, we consider linear and nonlinear log-risk functions μ .

Since the training sets are generated from a known s -sparse model, both the true support $\mathcal{S}$ and the true association risk function $h(\cdot)$ are known. So all four criteria can be evaluated. The primary focus of our analysis is the phase transition phenomenon: we investigate how PESR behaves as a function of sparsity parameter s . PESR is a stringent criterion; exact recovery is reached only when the estimated support matches the true support exactly, with no missing or extra variables. So PESR measures success as a binary notion of optimality rather than near correctness, making it a highly demanding indicator that nonetheless exposes methods truly capable of achieving optimal feature selection.

We call `HarderLASSO_QUT` our methodology based on the square-root of Cox's partial likelihood, combined with the quantile universal threshold $\lambda_\alpha^{\text{QUT}}$ for $\alpha = 0.05$ derived in Section 2, and the non-convex penalty in (6); its ℓ_1 -penalized analogue is denoted `LASSO_QUT`. Both approaches are evaluated under linear and nonlinear models. We benchmark our approach against well established feature selection methods detailed in the following sections.

3.1 Linear models

We compare our method against three established baselines: AIC, BIC and `LASSO_CV`. The first two apply best subset selection via forward stepwise search, with model choice guided by either AIC or BIC. In contrast, `LASSO_CV` employs ℓ_1 -regularization to induce sparsity [Tibshirani, 1997], with the penalty parameter λ selected by 5-fold cross-validation under default settings. For the BIC penalty, we follow Volinsky and Raftery [2000] and use the number of observed deaths rather than the total sample size, which has been shown to yield more reliable results. The information criterion-based approaches are implemented with the `CoxPHFitter` class from the `lifelines` Python package, while `LASSO_CV` is implemented with `CoxnetSurvivalAnalysis` from `scikit-survival`.

We fix the problem dimensions at $(n, p) = (150, 100)$ and vary the sparsity index $s \in \{0, 1, \dots, 20\}$. This corresponds to a challenging high-dimensional regime: the number of uncensored events is smaller than the number of features p , and the overall sample size n is relatively limited, making model estimation more difficult. The true log-risk function is $\mu(X_s^{(k)}) = X_s^{(k)}\beta$ where $\beta \in \mathbb{R}^s$ is the true coefficient vector, with each entry randomly drawn from $\{-3, -2, -1, 1, 2, 3\}$.

Figure 1 presents the results. The proposed `HarderLASSO_QUT` consistently outperforms the convex `LASSO_QUT`, benefiting from its non-convex penalty, which applies weaker shrinkage. Both approaches are

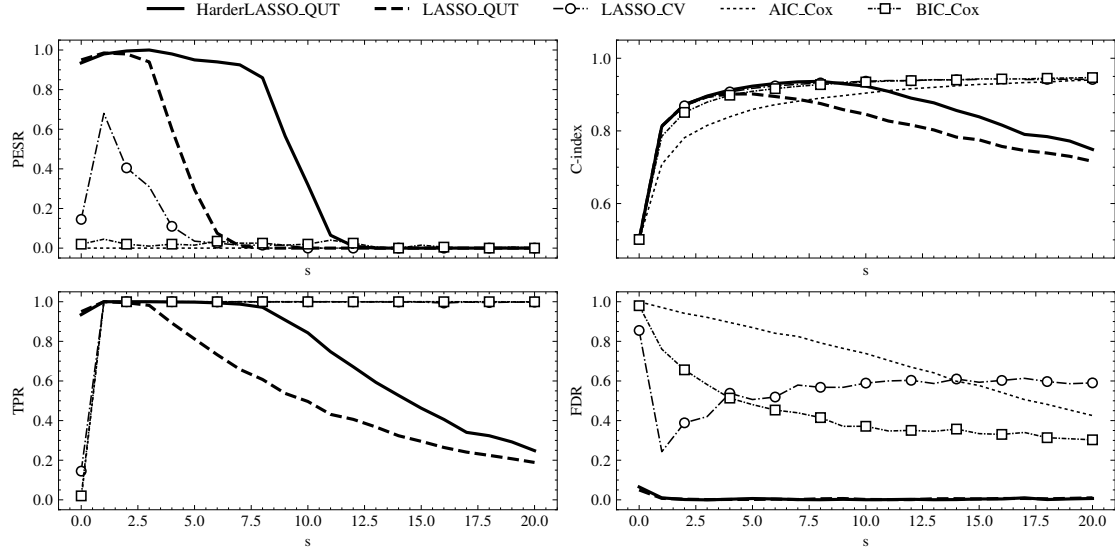


Figure 1: Linear Monte-Carlo simulation results in terms of probability of exact support recovery, C-index, true positive rate and false discovery rate as a function of the sparsity index s .

the only ones to display a clear phase transition in PESR, driven by an impressively low false discovery rate thanks to the quantile universal threshold specifically designed for model selection. Even on the unfavorable side of the transition, the FDR remains low, showing that our method favors selecting no features over selecting incorrect ones. By contrast, the other three models show no phase transition in PESR, but tend to over-select features, which yields high TPR but also inflates FDR. Their performance in terms of PESR and C-index makes it clear that these models prioritize minimizing predictive error rather than identifying the correct features.

3.2 Nonlinear models

We now benchmark our deep learning methodology against two widely used nonlinear methods: GradientBoosting and LassoNet. The former employs gradient-boosted Cox proportional hazard loss with regression trees as base learners, implemented using `scikit-survival` and `GradientBoostingSurvivalAnalysis`. We adopt the default hyperparameters and apply the Boruta algorithm [Kursa et al., 2010] to identify the most relevant features. The model is then retrained on this reduced feature set to ensure optimal fitting. LassoNet combines feature selection with model fitting via a neural network. We use the 5-fold cross-validation procedure implemented in `LassoNetCoxRegressorCV` [Lemhadri et al., 2021]. All neural network models, including ours, are configured with a single hidden layer of 20 neurons and the ReLU activation function.

For a simple nonlinear simulation, we choose

$$\mu(X_s^{(k)}) = \sum_{i=1}^h 10 \cdot |\mathbf{x}_{2i} - \mathbf{x}_{2i-1}|,$$

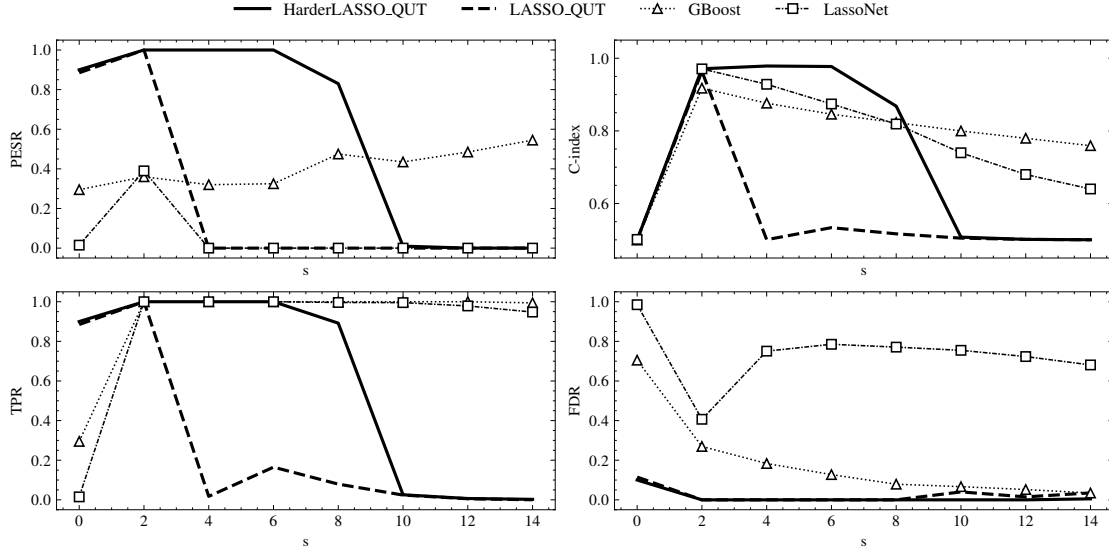


Figure 2: Nonlinear Monte-Carlo simulation results in terms of probability of exact support recovery, C-index, true positive rate and false discovery rate as a function of the sparsity index s .

where $\mathbf{x}_i$ is the i -th column of $X_s^{(k)}$, and $h = s/2$ for a sparsity index s varying over the set $\{0, 2, 4, \dots, 14\}$. This specification introduces a piecewise linear, non-monotone component via the absolute value, making variable selection and estimation more challenging than in standard linear models. To allow detection of the nonlinearity, we choose $p = 50$ and a large $n = 750$ so that dimensionality does not dominate the difficulty of retrieving the non-monotone dependence. Remember that only 50% of those $n = 750$ are uncensored.

Figure 2 reports the results mirroring those of the linear setting. Both QUT-based approaches clearly prioritize correct feature recovery, often favoring sparsity over the inclusion of incorrect variables. The proposed HarderLASSO_QUT again delivers the strongest performance in terms of support recovery, with a phase transition in PESR and a consistently low FDR, while LASSO_QUT exhibits the same phase transition but with slightly lower power, reflecting the more aggressive shrinkage of the convex penalty. By contrast, GradientBoosting and LassoNet demonstrate a markedly different behavior. They achieve high TPR but at the cost of substantially inflated FDR, indicating a tendency to over-select features rather than isolate the correct nonlinear structure. This trade-off results in good predictive accuracy (high C-index), but weak exact support recovery.

4 Real survival data experiments

We analyze the performance of our proposed method across five real-world datasets: the Study to Understand Prognoses Preferences Outcomes and Risks of Treatment (SUPPORT), the German Breast Cancer Study Group 2 (GBSG2), the 70-gene signature dataset from the Netherlands Cancer Institute (NKI70), the Norway/Stanford Breast Cancer Dataset (NSBCD), and the Mantle Cell Lymphoma cleaned dataset (MCLcleaned). These datasets are selected to cover low- to high-dimensional settings. For each dataset,

Dataset	Size n	Covariates p	Proportion of uncensored
SUPPORT	8832	26	0.68
GBSG2	686	9	0.44
NKI70	144	75	0.33
NSBCD	115	549	0.33
MCL	92	574	0.70

Table 1: Key characteristics of datasets used.

categorical variables are encoded using dummy variables when applicable, and missing values are imputed using mean imputation. Table 1 summarizes their key characteristics.

Our objective is again to compare the different methods (the same as described in previous section) in terms of their trade-off between model complexity (measured by the number of selected variables $\hat{s}$) and generalization performance (assessed via the C-index on unseen data). To this end, we conduct 100 simulation runs. In each run, the data are randomly split into a training set (two-thirds of the samples) and a test set (remaining one-third), with stratification on the censoring indicator to preserve a consistent censoring rate across both subsets.

Figure 3 and Table 2 summarize the results. The findings are consistent with those presented in Sections 3.1 and 3.2. The cross-validated methods, LassoNet and LASSO_CV, tend to select an excessive number of features, prioritizing predictive performance over model parsimony. In contrast, the information criterion-based approaches, AIC and BIC, consistently select smaller subsets of features while maintaining competitive predictive accuracy. For low-dimensional datasets, our methods, HarderLASSO_QUT and LASSO_QUT, perform comparably to BIC. However, in high-dimensional settings ($p \gg n$), they clearly outperform the stepwise methods, achieving both smaller subset sizes and superior predictive performance. Finally, our neural network-based variant performs well when the sample size n is large achieving same predictive performances with fewer features, benefiting from the expressive power of deeper architectures, but its performance degrades for smaller n , a behavior consistent with the data-hungry nature of neural networks compared to linear models.

Dataset	LASSO_Q		Hlasso_Q		LASSO_CV		AIC		BIC		Hlasso_Q_net		LassoNet	
	$\hat{s}$	C-idx	$\hat{s}$	C-idx	$\hat{s}$	C-idx	$\hat{s}$	C-idx	$\hat{s}$	C-idx	$\hat{s}$	C-idx	$\hat{s}$	C-idx
SUPPORT	7.02	0.60	7.37	0.60	18.43	0.60	12.17	0.60	7.27	0.60	4.78	0.59	26.00	0.60
GBSG2	2.73	0.67	2.56	0.67	6.73	0.67	4.49	0.67	2.89	0.67	2.34	0.67	7.16	0.67
NKI70	1.61	0.65	1.42	0.64	24.32	0.68	27.03	0.64	14.38	0.65	1.22	0.61	45.74	0.70
NSBCD	2.57	0.68	1.42	0.64	29.55	0.66	17.28	0.59	15.45	0.59	1.37	0.64	165.85	0.67
MCL	1.39	0.64	1.04	0.66	50.20	0.65	21.73	0.61	18.95	0.61	0.96	0.62	371.97	0.68

Table 2: Results on the real-world datasets. The first value is the number $\hat{s}$ of selected features, and the second value is the C-index on the test set, averaged over 100 resamplings. Method abbreviations are as follows: LASSO_Q (our method with ℓ_1 penalization), Hlasso_Q (our method with non convex P_ν penalty), LASSO_CV (LASSO tuned via cross-validation), AIC and BIC (models selected via Akaike and Bayesian information criteria), Hlasso_Q_net (neural network version of Hlasso_Q), and LassoNet (LassoNet neural architecture). Both neural networks use a single hidden layer with 20 neurons.

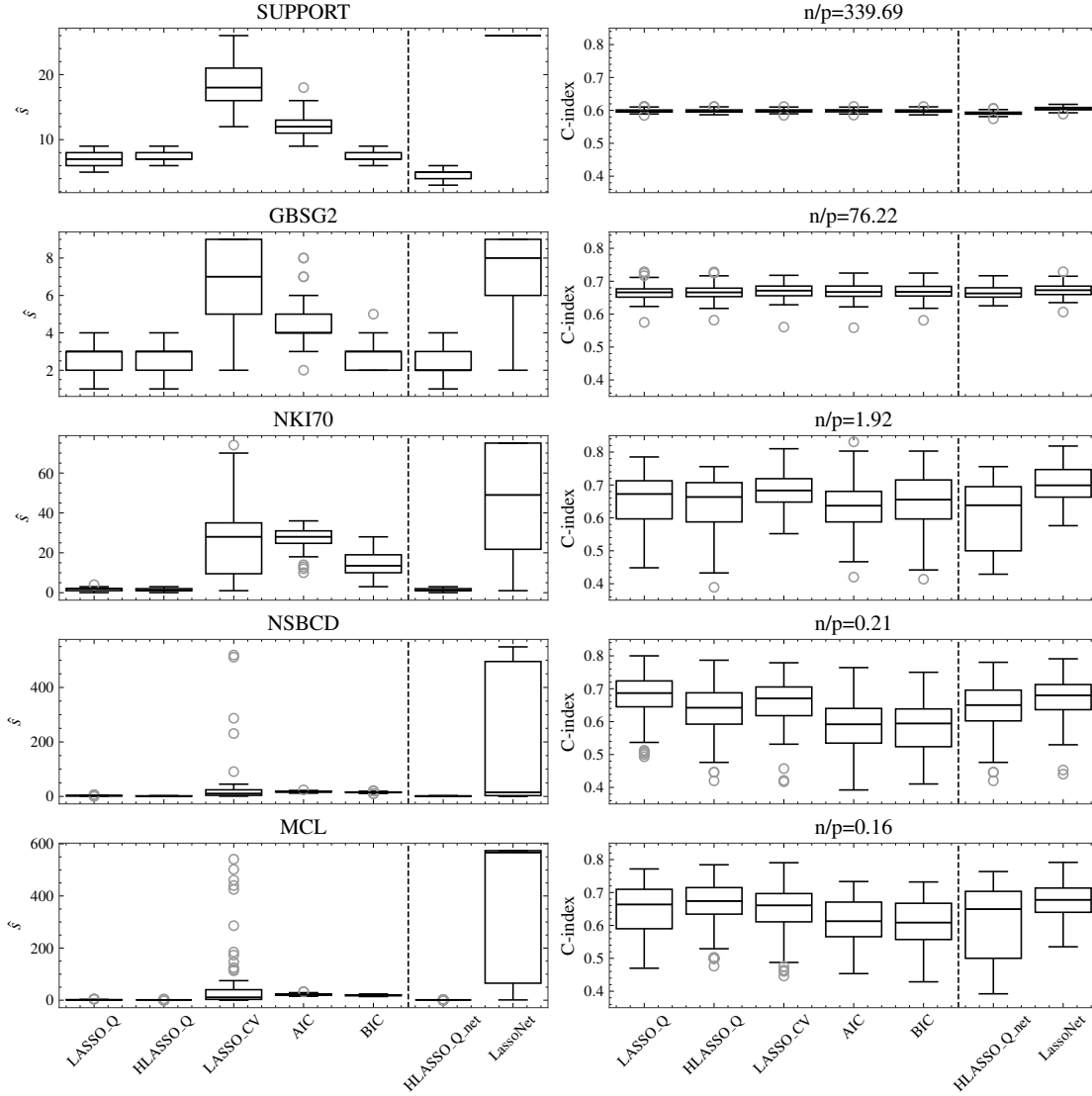


Figure 3: Trade-off between feature selection and predictive performance for different learners. Boxplots display the number of selected variables ($\hat{s}$, left panel) and the concordance index (C-index, right panel) for five datasets (linewise). Method abbreviations are as follows: LASSO_Q (our method with ℓ_1 penalization), HLAGSO_Q (our method with non convex P_ν penalty), LASSO_CV (LASSO tuned via cross-validation), AIC and BIC (models selected via Akaike and Bayesian information criteria), HLAGSO_Q_net (neural network version of HLAGSO_Q), and Lasso_Net (LassoNet neural architecture). Both neural network architectures use a single hidden layer with 20 neurons. The vertical dashed line separates linear models (left) from neural network-based models (right).

5 Primary Biliary Cirrhosis Study

We now perform a comprehensive study of the Mayo Clinic trial on primary biliary cirrhosis (PBC) of the liver. Data is provided by Therneau and Grambsch [2000] and contain $p = 17$ covariates: `age` (in years), `alb` (albumin in g/dl), `alk` (alkaline phosphatase in units/litre), `bil` (serum bilirubin in mg/dl), `cho1` (serum cholesterol in mg/dl), `cop` (urine copper in $\mu\text{g/day}$), `plat` (platelets per cubic ml/1000), `prot` (prothrombin time in seconds), `sgot` (liver enzyme in units/ml), `trig` (triglycerides in mg/dl), `asc` (0 denotes absence of ascites and 1 denotes presence of ascites), `oed` (0 denotes no oedema, 0.5 denotes untreated or successfully treated oedema and 1 denotes unsuccessfully treated oedema), `hep` (0 denotes absence of hepatomegaly and 1 denotes presence of hepatomegaly), `sex` (0 denotes male and 1 denotes female), `spid` (0 denotes absence of spiders and 1 denotes presence of spiders), `stage` (histological stage of disease, graded 1, 2, 3 or 4) and `trt` (1 for control and 2 for treatment). Data consist of 418 individuals but we restrict our study to the $n = 276$ observations without missing values. Out of those 276 individuals, 111 died before the end of the study. The categorical covariates are not one-hot encoded, and each variable is standardized by subtracting the mean and dividing by the standard deviation. Tibshirani [1997] and Zhang and Lu [2007] applied LASSO and adaptive LASSO (using 5-fold cross-validation) to find a set of 9 and 8 variables, respectively, a result much similar to the 8 variables obtained with stepwise selection. From the seventeen variables, `age`, `oed`, `bil`, `alb`, `cop`, `sgot`, `prot` and `stage` are consistently selected. Our method retain 8 variables for LASSO_QUT and 7 for HarderLASSO_QUT: `age`, `oed`, `bil`, `alb`, `cop`, `prot` and `stage`. The estimated coefficients for all methods are presented in Table 3. Figure 4 illustrates our model performance with baseline hazard and baseline survival function. Despite the reduced model complexity, the underlying risk profiles remain similar to the full model. The figure also presents the estimated effects of the four retained key predictors, obtained by fixing all other variables at their mean values and vary one predictor at a time across four representative values given by its empirical quantiles. For each setting, we compute the corresponding survival function from the fitted model.

To investigate the consistency of variable selection and the predictive performance of the sparse models, we conduct a Monte-Carlo simulation with 100 repetitions by randomly splitting the data set into training and testing sets, with the test set comprising 1/3 of the data. The splits are stratified to maintain similar proportions of censored observations in both sets. Table 4 reports the selection frequency of each variable across simulations, along with the average C-index measured on the test sets for different sparse models.

The results demonstrate that the QUT-based approaches exhibit the highest degree of consistency in variable selection among the competing methods. Across 100 Monte-Carlo replications, these two procedures predominantly selected the same variables identified when training on the entire dataset. This remarkable stability under repeated train/test partitions highlights their robustness to sampling variability. Equally noteworthy, several covariates were systematically excluded, never being selected in any of the replications, which indicates that the QUT-based procedures are not only parsimonious but also capable of reliably distinguishing between informative and non-informative predictors.

Variable	MLE	AIC	BIC	LASSO_CV	LASSO_QUT	HarderLASSO_QUT
trt	-0.064	—	—	—	—	—
age	0.303	0.329	0.329	0.165	0.283	0.277
sex	0.120	—	—	—	—	—
asc	0.022	—	—	0.025	0.035	—
hep	0.013	—	—	—	—	—
spid	0.047	—	—	—	—	—
oed	0.273	0.221	0.221	0.176	0.236	0.217
bil	0.367	0.390	0.390	0.385	0.459	0.455
chol	0.115	—	—	—	—	—
alb	-0.297	-0.290	-0.290	-0.221	-0.317	-0.314
cop	0.220	0.251	0.251	0.243	0.311	0.305
alk	0.003	—	—	—	—	—
sgot	0.231	0.247	0.247	0.067	—	—
trig	-0.065	—	—	—	—	—
plat	0.085	—	—	—	—	—
prot	0.234	0.228	0.228	0.128	0.196	0.197
stage	0.388	0.368	0.368	0.227	0.339	0.336

Table 3: Estimated coefficients from different models. MLE, stepwise selection (AIC and BIC), LASSO (cross validation), LASSO_QUT and HarderLASSO_QUT using the quantile universal threshold λ_{α}^{QUT} .

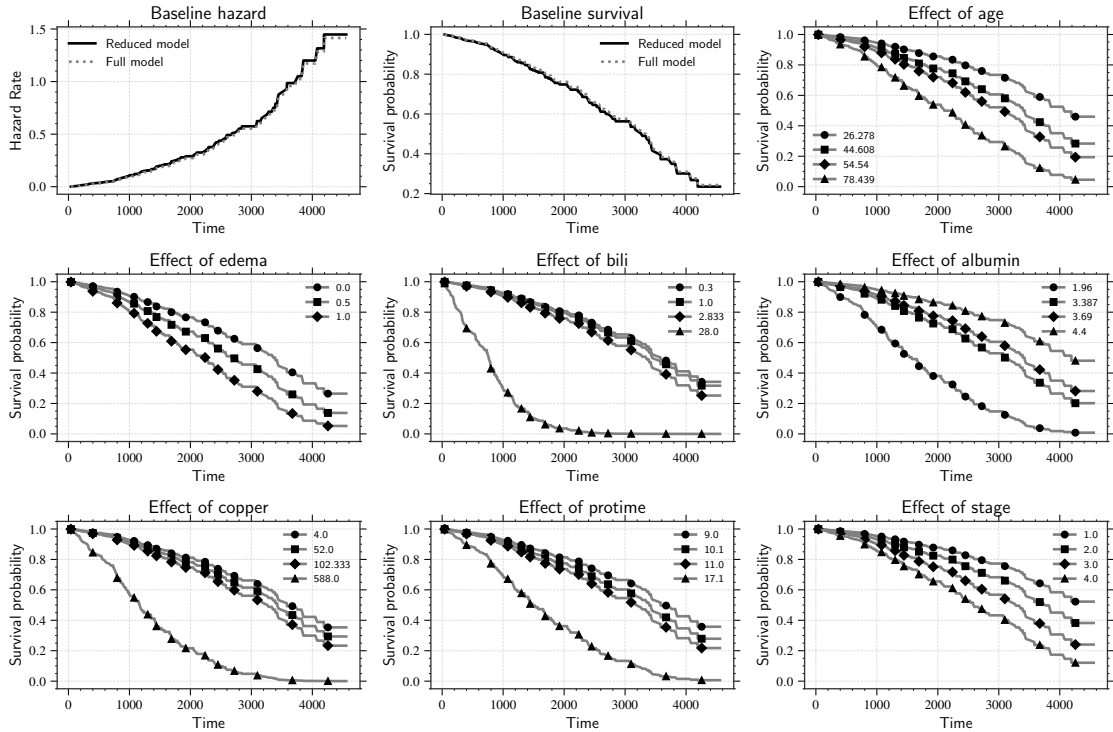


Figure 4: Model comparison and feature effects for the Primary Biliary Cirrhosis dataset. The top row displays baseline cumulative hazard (left) and baseline survival (middle) for both the full model and our HarderLASSO method. The remaining panels plot the estimated effects of the retained predictors on patient survival probability over time.

Model	trt	age	sex	asc	hep	spid	oed	bil	chol	alb	cop	alk	sgot	trig	plat	prot	stage	C-index
HarderLASSO_QUT	0	45	0	19	0	0	65	100	0	77	90	0	0	0	0	35	81	0.816
LASSO_QUT	0	65	0	64	13	5	91	100	0	96	100	0	9	0	0	69	97	0.826
LASSO_CV	21	86	41	62	32	39	90	100	36	97	100	27	69	33	17	89	96	0.829
AIC	9	73	29	19	5	10	68	100	22	85	85	9	66	19	7	66	92	0.821
BIC	0	60	9	14	2	3	44	100	1	70	75	2	24	4	2	31	86	0.814

Table 4: Selection frequency of variables and average C-index across models. The table reports how often each variable is selected across 100 Monte-Carlo simulations for five estimators. The final column shows the mean C-index on the test sets. Bold values indicate variables that are selected when training the model on the full dataset, as shown in Table 3.

By contrast, stepwise selection based on AIC and BIC, as well as the cross-validated LASSO, displayed a markedly less stable behavior. These methods frequently included variables beyond the stable core set, suggesting a tendency toward overfitting and reduced stability to data perturbation. While such procedures may yield acceptable predictive accuracy, their variable selection patterns reveal limited reliability in consistently identifying the truly relevant covariates.

Finally, the QUT-based methods achieve this parsimony with little loss of predictive performance. Their C-indices remain consistently high and competitive with those of the denser alternatives, underscoring their ability to balance sparsity and prediction.

6 Conclusions

We proposed a new method in survival analysis, where a pivotal penalty parameter calibrated at the detection edge and a penalty that shrinks less highly significant coefficients avoid the pitfalls of existing model selection methods that have a high false discovery rate. While old approaches lead to too many false discoveries, practitioners employing our method and Science will now become less attracted by false beliefs that can wrongly trigger new expensive and fruitless studies. The code is fully automatic and can be downloaded at <https://github.com/VcMaxouuu/HarderLASSO> for the linear and neural networks models. Future work will investigate whether the same approach can be applied to extensions of Cox proportional hazard model, where proportionality and some independence assumptions are relaxed.

7 Acknowledgement

We thank Professor Eva Cantoni for useful comments. This work was partially supported by the Swiss National Science Foundation grant 200021E_213166.

References

- H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.
- A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, 2009.

- P. Bühlmann and S. van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, Heidelberg, 2011.
- E. J. Candès and T. Tao. Decoding by linear programming. IEEE Transactions on Information Theory, 51: 4203–4215, 2005.
- D. R. Cox. Regression Models and Life Tables. Journal of the Royal Statistical Society, Series B, B(34): 187–202, 1972.
- D. L. Donoho. Compressed sensing. IEEE Transactions on Information Theory, 52(4):1289–1306, 2006.
- D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81(3): 425–455, 1994.
- C. Giacobino, S. Sardy, J. Diaz Rodriguez, and N. Hengardner. Quantile universal threshold. Electronic Journal of Statistics, 11(2):4701–4722, 2017.
- M. Kurs, A. Jankowski, and W. Rudnicki. Boruta - a system for feature selection. Fundam. Inform., 101: 271–285, 01 2010.
- I. Lemhadri, F. Ruan, L. Abraham, and R. Tibshirani. LassoNet: a neural network with feature sparsity. The Journal of Machine Learning Research, 22(1), 2021.
- R. Li, J. J. Ren, G. Yang, and Y. Yu. Asymptotic behavior of Cox's partial likelihood and its application to variable selection. Statistica Sinica, 28(5):2713–2731, 2018.
- S. A. Murphy and A. W. Van der Vaart. On Profile Likelihood. Journal of the American Statistical Association, 95(450):449–465, 2000.
- G. Schwarz. Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464, 1978.
- W. Su, M. Bogdan, and E. Candès. False discoveries occur early on the lasso path. The Annals of Statistics, 45:2133–2150, 11 2015.
- T. M. Therneau and P. M. Grambsch. Modeling Survival Data: Extending the Cox Model. Springer, New York, 2000. ISBN 0-387-98784-3.
- R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58(1):267–288, 1996.
- R. Tibshirani. The LASSO method for variable selection in the Cox model. Statistics in medicine, 16(4): 385–95, 1997.
- H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine, 30(11): 1105–1117, 2011.
- M. van Cutsem, S. Sardy, and X. Ma. Validation-free sparse learning: A phase transition approach to feature selection. arXiv:2411.17180, 2025.

- C. T. Volinsky and A. Raftery. Bayesian information criterion for censored survival models. Biometrics, 56:256–602, 2000.
- H. H. Zhang and W. Lu. Adaptive Lasso for cox's proportional hazards model. Biometrika, 94(3):691–703, 2007.