Binacox: automatic cut-point detection in high-dimensional Cox model with applications in genetics

Simon Bussy, Mokhtar Z. Alaya, Anne-Sophie Jannot, Agathe Guilloux

Binacox: automatic cut-point detection in high-dimensional Cox model with applications in genetics

Simon Bussy
LPSM, UMR 8001, CNRS, Sorbonne University, Paris, France
email: simon.bussy@gmail.com

Mokhtar Z. Alaya
Modal'X, UPL, Univ Paris Nanterre, F92000 Nanterre, France
email: mokhtarzahdi.alaya@gmail.com

Anne-Sophie Jannot
Biomedical Informatics and Public Health Department, EGPH, APHP
and INSERM, UMRS 1138, Centre de Recherche des Cordeliers, Paris, France
email: annesophie.jannot@aphp.fr

Agathe Guilloux
LaMME, UEVE and UMR 8071, Paris Saclay University, Evry, France
email: agathe.guilloux@math.cnrs.fr

Abstract

We introduce the *binacox*, a prognostic method to deal with the problem of detecting multiple cut-points per features in a multivariate setting where a large number of continuous features are available. The method is based on the Cox model and combines one-hot encoding with the *binarsity* penalty, which uses total-variation regularization together with an extra linear constraint, and enables feature selection. Original nonasymptotic oracle inequalities for prediction (in terms of Kullback-Leibler divergence) and estimation with a fast rate of convergence are established. The statistical performance of the method is examined in an extensive Monte Carlo simulation study, and then illustrated on three publicly available genetic cancer datasets. On these high-dimensional datasets, our proposed method significantly outperforms state-of-the-art survival models regarding risk prediction in terms of the C-index, with a computing time orders of magnitude faster. In addition, it provides powerful interpretability from a clinical perspective by automatically pinpointing significant cut-points in relevant variables.

Keywords. Cox model; Cut-point; Feature binarization; Nonasymptotic oracle inequality; Proximal methods; Survival analysis; Total variation

1 Introduction

Determining significant prognostic biomarkers is of increasing importance in many areas of medicine. Scores used in clinical practice often categorize continuous features into binary ones using expert-driven cut-points. For instance, the Wells score, which categorizes patients into low, moderate and high risk groups for pulmonary embolism [Wells et al., 2000], is one of the most extensively validated predictive scores. One of the categorized feature used in this score is "having a heart rate of over 100 beats per minute, or not".

When used in routine care, this type of threshold makes a score more interpretable from a clinical point of view. In this particular example, it means that experts consider that heart rate has a nonlinear effect: there is reasonable agreement that above this threshold, patients have higher risk of unfavourable outcome. Despite this choice of threshold, there is little agreement on the exact nature of the relationship between heart rate and prognosis.

With the increasing availability of high-dimensional datasets, data-driven predictive scores are becoming increasingly important, e.g., in genetic oncology studies, where similar questions occur because the effect of certain genes' expression on survival times are often non-linear. Therefore, to develop such scores, one has to deal with a two-sided problem: first to select relevant features, and second to find relevant thresholds – also called *cut-off* values or *cut-points* – for these selected continuous features, without prior or expert knowledge.

The cut-point detection problem. Solving this problem means applying non-linearities to feature effects that most models cannot detect. This also offers the ability to classify patients into several groups in terms of their continuous feature values relative to the cut-points. More importantly, this can also lead to a better understanding of the features' effects on the outcome of interest; this strategy might uncover biological thresholds as well as potential criteria for new prospective studies, help diagnose diseases, and make treatment recommendations. A convenient tool for finding optimal cut-points is therefore of high interest.

Indeed, good cut-point detection is a common issue in medical studies, and numerous methods have been proposed for determining a single cut-point for a given feature. This ranges from choosing the mean or median, to methods based on distribution of values, or association with clinical outcomes, e.g., the minimal p-value from multiple log-rank tests, see Camp et al. [2004], Moul et al. [2007], Rota et al. [2015] among many others. However, the choice of the actual cut-points is not a straightforward problem, even for a single cut-point [Lausen and Schumacher, 1992, Klein and Wu, 2003, Contal and O'Quigley, 1999]. Recently, Icuma et al. [2018] proposed a Bayesian approach with accelerated failure time modeling, but still only allowing one cut-point per feature.

Indeed, while many studies have been devoted to find one optimal cut-point, there is often need in medical settings to determine not only one but multiple cut-points. For instance, prognoses are generally worst at both ends of the body mass index, i.e., for obese and underweight individuals [Oreopoulos et al., 2008]. Methods exist to deal with multiple cut-point detection for one-dimensional signals (see for instance Bleakley and Vert [2011] and Harchaoui and Lévy-Leduc [2010] that use a group fused lasso or total-variation penalty, respectively), and for multivariate time series (see Cho and Fryzlewicz [2015]). Though cut-point detection is also a paramount issue in survival analysis [Faraggi and Simon, 1996], methods that have been developed in this setting only look at a single feature at a time (e.g., Motzer et al. [1999] and LeBlanc and Crowley [1993] which use survival trees, or more recently Chang et al. [2019]). To our knowledge, a multivariate survival analysis method well-suited to detect multiple cut-points per feature in a high-dimensional setting has not been previously proposed.

General framework. Let us consider the usual survival analysis framework. Following Andersen et al. [2012], let non-negative random variables T and C stand for the time of the event of interest and censoring time respectively, and X denote the p-dimensional vector of features (e.g., patient characteristics, therapeutic strategy, omics features). The event of interest could be for instance survival time, re-hospitalization, relapse or dis-

ease progression. Conditionally on X, T and C are assumed to be independent, which is classical in survival analysis [Klein and Moeschberger, 2005]. We then denote Z the right-censored time and Δ the censoring indicator, defined as

$$Z = T \wedge C \quad \text{and} \quad \Delta = \mathbb{1}(T \leq C)$$

respectively, where $a \wedge b$ denotes the minimum between two numbers a and b, and $\mathbb{1}(\cdot)$ the indicator function taking the value 1 if the condition in $(\cdot)$ is satisfied and 0 otherwise.

The Cox proportional hazards model [Cox, 1972] is by far the most widely used in survival analysis. It describes the relation between the hazard function and the features by

$$\lambda(t|X = x) = \lambda_0(t)e^{x^\top \beta^{\text{cox}}},$$

where λ_0 is a baseline hazard function describing how the event risk changes over time at baseline levels of features, and $\beta^{\text{cox}} \in \mathbb{R}^p$ a vector quantifying the multiplicative impact on the hazard ratio of each feature.

High-dimensional survival analysis. High-dimensional settings are becoming increasingly frequent, in particular for genetic data applications where cut-point estimation is a common problem (see for instance Harvey et al. [1999], Shirota et al. [2001], Cheang et al. [2009]), but also in other contexts where the number of available features to consider as potential risk factors is tremendous, particularly with the development of electronic health records. A penalized version of the Cox model well-suited for such settings is proposed in Simon et al. [2011], but it cannot model nonlinearity. Theory for using lasso-type methods in the Cox model was developed in Huang et al. [2013]. Other methods have been put forward to deal with this problem in similar settings, like boosting Cox models [Li and Luan, 2005] and random survival forests [Ishwaran et al., 2008]. However, none of these identify cut-point values, which is of major interest for both interpretation and clinical benefit.

Main contribution. In this paper, we propose a method called *binacox* that estimates multiple cut-points in a Cox model with high-dimensional features. First, the binacox one-hot encodes the continuous input features [Wu and Coggeshall, 2012] through a mapping to a new binarized space of much higher dimension, and then trains the Cox model in this space, regularized with the *binarsity* penalty [Alaya et al., 2017] which combines total-variation regularization with an extra sum-to-zero constraint, and enables feature selection. Cut-points of the initial continuous input features are then detected by the jumps in the regression coefficient vectors, which the binarsity penalty forces to be piecewise-constant. The main contribution of this paper is twofold. First we introduce the idea of using a total-variation penalty with an extra linear constraint on the weights of a Cox model trained on a binarization of the raw continuous features. This leads to a procedure that automatically detects relevant features and allows multiple cut-points per feature. Secondly the oracle inequality in prediction of Section 3 (see Theorem 1) is stated in terms of Kullback-Leibler divergence, as opposed to the results in Huang et al. [2013] (for the lasso penalty) expressed in Breiman divergence, the arguments are consequently different.

Organization of the paper. A precise description of the model is given in Section 2. Section 3 highlights the good theoretical properties of the binacox by establishing fast oracle inequalities for prediction and for estimation. Section 4 presents the simulation procedure used to evaluate the performance of our method and compares it with existing

ones. In Section 5, we apply our method to high-dimensional genetic datasets. Finally, we discuss the obtained results in Section 6.

Notation. Throughout the paper, for every $q > 0$, we denote by $\|v\|_q$ the usual ℓ_q-quasi norm of a vector $v \in \mathbb{R}^m$, namely $\|v\|_q = (\sum_{k=1}^m |v_k|^q)^{1/q}$, and $\|v\|_\infty = \max_{1\leq k\leq m} |v_k|$. We write $\mathbf{1}$ (resp. $\mathbf{0}$) the vector having all coordinates equal to one (resp. zero). We also denote $|A|$ the cardinality of a finite set A. If I is an interval, $|I|$ stands for its Lebesgue measure. Then, for any $u \in \mathbb{R}^m$ and any $L \subset \{1, \dots, m\}$, we denote u_L the vector of $\mathbb{R}^m$ satisfying $(u_L)_k = u_k$ for $k \in L$ and $(u_L)_k = 0$ for $k \in L^\complement := \{1, \dots, m\} \backslash L$. Finally, for a matrix M of size $k \times k'$, $M_{j,\bullet}$ denotes its jth row and $M_{\bullet,l}$ its lth column.

2 Model and method

2.1 Cox model with cut-points.

Consider an independent and identically distributed (i.i.d.) sample

$$(X_1, Z_1, \Delta_1), \dots, (X_n, Z_n, \Delta_n) \in [0,1]^p \times \mathbb{R}_+ \times \{0,1\},$$

where the condition $X_i \in [0,1]^p$ for all $i = 1, \dots, n$ is always true after an appropriate rescaling preprocessing step, without loss of generality. Let $\boldsymbol{X} = [X_{i,j}]_{1\leq i\leq n; 1\leq j\leq p}$ be the $n \times p$ design matrix vertically stacking the n samples of p raw features so that $\boldsymbol{X}_{i,\bullet} = X_i$. In order to simplify the presentation of our results, we assume in the paper that the raw features $\boldsymbol{X}_{\bullet,j}$ are continuous for all $j = 1, \dots, p$, but this is not a limitation in practice. Assume that the hazard function for patient i is given by

$$\lambda^\star(t|X_i) = \lambda_0^\star(t) e^{f^\star(X_i)},$$

where $\lambda_0^\star(t)$ is the baseline hazard function, and

$$f^\star(X_i) = \sum_{j=1}^p f_j^\star(X_{i,j}) = \sum_{j=1}^p \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \mathbb{1}(X_{i,j} \in I_{j,k}^\star), \tag{1}$$

with $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ for $k = 1, \dots, K_j^\star + 1$ and where $\beta_{j,k}^\star \neq \beta_{j,k+1}^\star$ for $k = 1, \dots, K_j^\star$. We impose that

$$\sum_{i=1}^n f_j^\star(X_{i,j}) = 0 \quad \text{for all } j = 1, \dots, p$$

to ensure identifiability (see [Meier et al., 2009] for a similar constraint in generalized additive models), which can also be written as a sum-to-zero constraint in each $\beta^\star$'s block, that is:

$$\sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star n_{j,k}^\star = 0 \quad \text{for all } j = 1, \dots, p \tag{2}$$

where $n_{j,k}^\star = |\{i = 1, \dots, n : X_{i,j} \in I_{j,k}^\star\}|$. For each feature $j = 1, \dots, p$, the $\mu_{j,k}^\star$s $(k = 1, \dots, K_j^\star)$ are the so-called cut-points, and are such that

$$\mu_{j,1}^\star < \mu_{j,2}^\star < \cdots < \mu_{j,K_j^\star}^\star,$$

with the conventions $\mu^\star_{j,0} = 0$ and $\mu^\star_{j,K^\star_j+1} = 1$. Denoting $K^\star = \sum_{j=1}^p K^\star_j$, the vector of regression coefficients $\beta^\star \in \mathbb{R}^{K^\star+p}$ is given by

$$\beta^\star = ({\beta^\star_{1,\bullet}}^\top, \ldots, {\beta^\star_{p,\bullet}}^\top)^\top = (\beta^\star_{1,1}, \ldots, \beta^\star_{1,K^\star_1+1}, \ldots, \beta^\star_{p,1}, \ldots, \beta^\star_{p,K^\star_p+1})^\top,$$

and the cut-points vector $\mu^\star \in \mathbb{R}^{K^\star}$ by

$$\mu^\star = ({\mu^\star_{1,\bullet}}^\top, \ldots, {\mu^\star_{p,\bullet}}^\top)^\top = (\mu^\star_{1,1}, \ldots, \mu^\star_{1,K^\star_1}, \ldots, \mu^\star_{p,1}, \ldots, \mu^\star_{p,K^\star_p})^\top.$$

Our goal is to simultaneously estimate $\mu^\star$ and $\beta^\star$, which also requires estimation of the unknown $K^\star_j$ for all $j = 1, \ldots, p$. Towards this end, the first step of our proposed method is to map the feature space to a much higher space of binarized features.

2.2 Binarization.

Let $\boldsymbol{X}^B$ be the sparse binarized matrix with an extended number $p + d$ of columns, typically with $d \gg p$, where continuous input features have been one-hot encoded [Wu and Coggeshall, 2012, Liu et al., 2002]. The jth column $\boldsymbol{X}_{\bullet,j}$ is then replaced by $d_j + 1 \geq 2$ columns $\boldsymbol{X}^B_{\bullet,j,1}, \ldots, \boldsymbol{X}^B_{\bullet,j,d_j+1}$ containing only zeros and ones, where the ith row $X^B_i \in \mathbb{R}^{p+d}$ with $d = \sum_{j=1}^p d_j$ is written

$$X^B_i = (X^B_{i,1,1}, \ldots, X^B_{i,1,d_1+1}, \ldots, X^B_{i,p,1}, \ldots, X^B_{i,p,d_p+1})^\top.$$

We consider a partition of intervals $I_{j,1}, \ldots, I_{j,d_j+1}$ such that

$$\bigcup_{k=1}^{d_j+1} I_{j,k} = [0, 1]$$

and $I_{j,k} \cup I_{j,k'} = \varnothing$ for all $k \neq k'$ with $k, k' = 1, \ldots, d_j + 1$. Now for $i = 1, \ldots, n$ and $l = 1, \ldots, d_j + 1$, we define

$$X^B_{i,j,l} = \begin{cases} 1 & \text{if } X_{i,j} \in I_{j,l}, \\ 0 & \text{otherwise.} \end{cases}$$

We then denote $I_{j,l} = (\mu_{j,l-1}, \mu_{j,l}]$ for $l = 1, \ldots, d_j + 1$, with the convention $\mu_{j,0} = 0$ and $\mu_{j,d_j+1} = 1$. A natural choice for the $\mu_{j,l}$ is given by the quantiles, namely $\mu_{j,l} = q_j\big(l/(d_j+1)\big)$, where $q_j(\alpha)$ denotes a quantile of order $\alpha \in [0, 1]$ for $\boldsymbol{X}_{\bullet,j}$. If training data also contains unordered qualitative features, one-hot encoding with ℓ_1-penalization can be used, for instance.

To each binarized feature $\boldsymbol{X}^B_{\bullet,j,l}$ corresponds a parameter $\beta_{j,l}$, and the vectors associated with the binarization of the jth feature are naturally denoted $\beta_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,d_j+1})^\top$ and $\mu_{j,\bullet} = (\mu_{j,1}, \ldots, \mu_{j,d_j})^\top$. Hence, we define a candidate for the estimation of $f^\star$ defined in (1) as

$$f_\beta(X_i) = \beta^\top X^B_i = \sum_{j=1}^p f_{\beta_{j,\bullet}}(X_{i,j}) = \sum_{j=1}^p \sum_{l=1}^{d_j+1} \beta_{j,l} \mathbb{1}(X_{i,j} \in I_{j,l}). \tag{3}$$

The full parameter vectors of size $p + d$ and d respectively are finally obtained by concatenation of the vectors $\beta_{j,\bullet}$ and $\mu_{j,\bullet}$, i.e.,

$$\beta = (\beta^\top_{1,\bullet}, \ldots, \beta^\top_{p,\bullet})^\top = (\beta_{1,1}, \ldots, \beta_{1,d_1+1}, \ldots, \beta_{p,1}, \ldots, \beta_{p,d_p+1})^\top,$$

and

$$\mu = (\mu^\top_{1,\bullet}, \ldots, \mu^\top_{p,\bullet})^\top = (\mu_{1,1}, \ldots, \mu_{1,d_1}, \ldots, \mu_{p,1}, \ldots, \mu_{p,d_p})^\top.$$

2.3 Estimation procedure.

In the following, for a fixed vector μ of quantization, we define the binarized partial negative log-likelihood (rescaled by $1/n$) as follows:

$$\ell_n(f_\beta) = -\frac{1}{n}\sum_{i=1}^{n}\Delta_i\Big\{f_\beta(X_i) - \log\sum_{i':Z_{i'}\geq Z_i} e^{f_\beta(X_{i'})}\Big\}. \tag{4}$$

Our approach consists in minimizing the function ℓ_n plus the binarsity penalization term introduced in Alaya et al. [2017]. The resulting optimization problem is written

$$\hat{\beta} \in \operatorname{argmin}_{\beta\in\mathscr{B}_{p+d}(R)}\big\{\ell_n(f_\beta) + \mathrm{bina}(\beta)\big\}, \tag{5}$$

where $\mathscr{B}_{p+d}(R) = \{\beta\in\mathbb{R}^{p+d} : \sum_{j=1}^{p}\|\beta_{j,\bullet}\|_\infty \leq R\}$ and

$$\mathrm{bina}(\beta) = \sum_{j=1}^{p}\Big(\sum_{l=2}^{d_j+1}\omega_{j,l}|\beta_{j,l} - \beta_{j,l-1}| + \delta_j(\beta_{j,\bullet})\Big), \tag{6}$$

with

$$\delta_j(u) = \begin{cases} 0 & \text{if} \quad n_{j,\bullet}^\top u = 0, \\ \infty & \text{otherwise,} \end{cases}$$

and where $n_{j,\bullet} = (n_{j,1},\ldots,n_{j,d_j+1})^\top \in \mathbb{N}^{d_j+1}$ with $n_{j,l} = |\{i=1,\ldots,n : X_{i,j}\in I_{j,l}\}|$ for all $j=1,\ldots,p$ and $l=1,\ldots,d_j+1$. The constraint over $\mathscr{B}_{p+d}(R)$ is standard in the literature for obtaining proofs of oracle inequalities for sparse generalized linear models [Van de Geer, 2008], and is discussed in detail below. The weights $\omega_{j,l}$ are of order

$$\omega_{j,l} = \mathcal{O}\Big(\sqrt{\frac{\log(p+d)}{n}}\Big),$$

see Appendix B.1 for their explicit form.

It turns out that the binarsity penalty is well-suited to our problem. First, it tackles the problem that $\boldsymbol{X}^B$ is not full rank by construction, since $\sum_{l=1}^{d_j+1} X^B_{i,j,l} = 1$ for all $j=1,\ldots,p$, which means that the columns in each block sum to $\mathbf{1}$. This problem is solved since the penalty imposes the linear constraint $\sum_{l=1}^{d_j+1} n_{j,l}\beta_{j,l} = 0$ in each block with the $\delta_j(\cdot)$ term. Note that if the $I_{j,l}$ are taken as the interquantiles intervals, we have that $n_{j,l}$ are all equal for $l=1,\ldots,d_j+1$, and we get the standard sum-to-zero constraint $\sum_{l=1}^{d_j+1}\beta_{j,l} = 0$. Then, the other term in the penalty consists of a within-block weighted total variation penalty:

$$\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} = \sum_{l=2}^{d_j+1}\omega_{j,l}|\beta_{j,l} - \beta_{j,l-1}|, \tag{7}$$

that takes advantage of the fact that within each block, binarized features are ordered. The effect is then to keep the number of different values taken by $\beta_{j,\bullet}$ to a minimum, which makes significant cut-points appear, as detailed hereafter.

For all $\beta\in\mathbb{R}^{p+d}$, let $\mathcal{A}(\beta) = \big[\mathcal{A}_1(\beta),\ldots,\mathcal{A}_p(\beta)\big]$ be the concatenation of the support sets relative to the total-variation penalization, namely

$$\mathcal{A}_j(\beta) = \big\{l : \beta_{j,l}\neq\beta_{j,l-1},\ \text{for } l=2,\ldots,d_j+1\big\}$$

for all $j = 1, \dots, p$. Similarly, we denote $\mathcal{A}^{\complement}(\beta) = \left[\mathcal{A}_1^{\complement}(\beta), \dots, \mathcal{A}_p^{\complement}(\beta)\right]$ the complementary set of $\mathcal{A}(\beta)$. We then write

$$\mathcal{A}_j(\hat{\beta}) = \{\hat{l}_{j,1}, \dots, \hat{l}_{j,s_j}\}, \tag{8}$$

where $\hat{l}_{j,1} < \cdots < \hat{l}_{j,s_j}$ and $s_j = |\mathcal{A}_j(\hat{\beta})|$. Finally, we obtain the following $\mu^\star_{j,\bullet}$'s estimator

$$\widehat{\mu}_{j,\bullet} = (\mu_{j,\hat{l}_{j,1}}, \dots, \mu_{j,\hat{l}_{j,s_j}})^\top \tag{9}$$

for all $j = 1, \dots, p$. By construction, $K_j^\star$ is estimated by $\widehat{K}_j = s_j$. Some details on the algorithm used to solve the regularization problem (5) are given in Appendix A.1.

3 Theoretical guarantees

3.1 Oracle inequality for prediction

This section is devoted to a first theoretical result. In order to evaluate the prediction error, we first define the (empirical) Kullback-Leibler divergence [Senoussi, 1990] KL_n between the true function $f^\star$ and any candidate f as

$$KL_n(f^\star, f) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \log\left\{\frac{e^{f^\star(X_i)}\sum_{i=1}^n Y_i(t)e^{f(X_i)}}{e^{f(X_i)}\sum_{i=1}^n Y_i(t)e^{f^\star(X_i)}}\right\} Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}\mathrm{d}t, \tag{10}$$

where we denote $Y_i(t) = \mathbb{1}(Z_i \geq t)$ the at-risk process, and $\tau > 0$ is to be defined later.

We seek to establish an oracle inequality expressed in terms of a compatibility factor [Van de Geer and Bühlmann, 2009] satisfied by the following non-negative symmetric matrix:

$$\Sigma_n(f^\star, \tau) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left(X_i^B - \bar{X}_n(s)\right)\left(X_i^B - \bar{X}_n(s)\right)^\top y_i(s)e^{f^\star(X_i)}\lambda_0^\star(s)\mathrm{d}s, \tag{11}$$

where

$$\bar{X}_n(s) = \frac{\sum_{i=1}^n X_i^B y_i(s)e^{f^\star(X_i)}}{\sum_{i=1}^n y_i(s)e^{f^\star(X_i)}}$$

and

$$y_i(s) = \mathbb{E}[Y_i(s)|X_i] \text{ for all } 0 \leq s \leq t \text{ and all } i = 1, \dots, n.$$

For any concatenation of index subsets $L = [L_1, \dots, L_p]$, we define the compatibility factor

$$\kappa_\tau(L) = \inf_{\beta \in \mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}} \frac{\sqrt{\beta^\top \Sigma_n(f^\star, \tau)\beta}}{\|\beta_L\|_2}, \tag{12}$$

where

$$\mathscr{C}_{\mathrm{TV},\omega}(L) = \left\{\beta \in \mathscr{B}_{p+d}(R) : \sum_{j=1}^p \|(\beta_{j,\bullet})_{L_j^\complement}\|_{\mathrm{TV},\omega_{j,\bullet}} \leq 3\sum_{j=1}^p \|(\beta_{j,\bullet})_{L_j}\|_{\mathrm{TV},\omega_{j,\bullet}}\right\}$$

is a cone composed of all vectors with similar support L.

Assumption 1 *τ is hereafter assumed to satisfy*

$$\max_{1\leq i\leq n}\int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t < \infty \quad \textit{and} \quad \min_{1\leq i\leq n}\mathbb{P}(C_i > \tau|X_i) > 0.$$

Such assumptions on τ are common in survival analysis, see e.g., Andersen et al. [2012] and Lemler [2016]. We refer the reader to Gill [1983] for a discussion on the role of τ. In addition, we define $c_Z := \min_{1\leq i\leq n} y_i(\tau)$ and remark that

$$c_Z \geq \exp\big(-\max_{1\leq i\leq n}\int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t\big)\min_{1\leq i\leq n}\mathbb{P}(C_i>\tau|X_i)>0.$$

For the sake of simplicity, we introduce the additional notation:

$$f^\star_\infty = \max_{1\leq i\leq n}|f^\star(X_i)|, \quad s^{(0)}(\tau)=n^{-1}\sum_{i=1}^n y_i(\tau)e^{f^\star(X_i)}, \quad \text{and} \quad \Lambda^\star_0(\tau)=\int_0^\tau \lambda^\star_0(s)\mathrm{d}s.$$

Assumption 2 *Let $\varepsilon\in(0,1)$ and define $t_{n,p,d,\varepsilon}$ as the solution of*

$$2.221(p+d)^2\exp\{-nt^2_{n,p,d,\varepsilon}/(2+2t_{n,p,d,\varepsilon}/3)\}=\varepsilon.$$

For any concatenation set $L=[L_1,\ldots,L_p]$ such that $\sum_{j=1}^p|L_j|\leq K^\star$, assume that

$$\kappa^2_\tau(L)>\Xi_\tau(L),$$

where

$$\Xi_\tau(L)=4|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\Big\{\big(1+e^{2f^\star_\infty}\Lambda^\star_0(\tau)\big)\sqrt{(2/n)\log\big(2(p+d)^2/\varepsilon\big)} + \big(2e^{2f^\star_\infty}\Lambda^\star_0(\tau)/s^{(0)}(\tau)\big)t^2_{n,p,d,\varepsilon}\Big\}.$$

Note that $\kappa^2_\tau(L)$ is the smallest eigenvalue of a population integrated covariance matrix defined in (11), so it is reasonable to treat it as a constant. Moreover, $t^2_{n,p,d,\varepsilon}$ is of order

$$\frac{1}{n}\log\frac{(p+d)^2}{\varepsilon},$$

so if $|L|\log(p+d)/n$ is sufficiently small, Assumption 2 is verified. With these preparations made, let us now state the oracle inequality for prediction satisfied by our estimator of $f^\star$ which is, by construction, given by $\hat{f}=f_{\hat{\beta}}$ (see (3)).

Theorem 1 *The inequality*

$$KL_n(f^\star,f_{\hat{\beta}})\leq\inf_\beta\left\{3KL_n(f^\star,f_\beta)+\frac{1024(f^\star_\infty+R+2)}{\kappa^2_\tau\big(\mathcal{A}(\beta)\big)-\Xi_\tau\big(\mathcal{A}(\beta)\big)}|\mathcal{A}(\beta)|\max_{1\leq j\leq p}\|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|^2_\infty\right\} \tag{13}$$

holds with a probability greater than $1-28.55e^{-c}-e^{-ns^{(0)}(\tau)^2/8e^{2f^\star_\infty}}-3\varepsilon$ for some $c>0$, where the infimum is over the set of vectors $\beta\in\mathscr{B}_{p+d}(R)$ such that $n_{j,\bullet}^\top\beta_{j,\bullet}=0$ for all $j=1,\ldots,p$, and such that $|\mathcal{A}(\beta)|\leq K^\star$.

The proof of Theorem 1 is postponed to Appendix B.3. The second term in the right-hand side of (13) can be viewed as a "variance" (or "complexity") term, and its dominant term satisfies

$$\frac{|\mathcal{A}(\beta)|\max_j\|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|^2_\infty}{\kappa^2_\tau\big(\mathcal{A}(\beta)\big)-\Xi_\tau\big(\mathcal{A}(\beta)\big)}\lesssim\frac{|\mathcal{A}(\beta)|}{\kappa^2_\tau\big(\mathcal{A}(\beta)\big)-\Xi_\tau\big(\mathcal{A}(\beta)\big)}\frac{\log(p+d)}{n},$$

where the symbol $\lesssim$ means that the inequality holds up to a multiplicative constant. Then, one obtains the expected fast convergence rate $\mathcal{O}\big(\log(p+d)/n\big)$ for the estimator $\hat{f}$. Note that, in the proof of Theorem 1, the fact that the true $f^\star$ lies in the true Cox model with cut-points is not necessary. Hence Theorem 1 can be applied to any $f^\star$.

The value $|\mathcal{A}(\beta)|$ characterizes the sparsity of the vector β, since it counts the number of non-equal consecutive values of β. If β is block-sparse, namely whenever $|\mathscr{A}(\beta)| \ll p$ where $\mathscr{A}(\beta) = \{j = 1, \dots, p : \beta_{j,\bullet} \neq \mathbf{0}\}$ (meaning that few raw features are useful for prediction), then $|\mathcal{A}(\beta)| \leq |\mathscr{A}(\beta)| \max_{j \in \mathscr{A}(\beta)} |\mathcal{A}_j(\beta)|$, which means that $|\mathcal{A}(\beta)|$ is controlled by the block sparsity $|\mathscr{A}(\beta)|$. Also, the oracle inequality still holds for vectors such that $n_{j,\bullet}^\top \beta_{j,\bullet} = 0$, which is natural since the binarsity penalization imposes these extra linear constraints.

The assumption $\beta \in \mathscr{B}_{p+d}(R)$ is a technical one, allowing a connection, via the notion of self-concordance [Bach, 2010], between the empirical squared ℓ_2-norm and the empirical Kullback-Leibler (see Lemma 3). Also, note that

$$\max_{1 \leq i \leq n} |\beta^\top X_i^B| \leq \sum_{j=1}^{p} \|\beta_{j,\bullet}\|_\infty \leq |\mathscr{A}(\beta)| \times \|\beta\|_\infty, \tag{14}$$

where $\|\beta\|_\infty = \max_{1 \leq j \leq p} \|\beta_{j,\bullet}\|_\infty$. The first inequality in (14) comes from the fact that the entries of $\boldsymbol{X}^B$ are in $\{0, 1\}$, and entails that $\max_{1 \leq i \leq n} |\beta^\top X_i^B| \leq R$ whenever $\beta \in \mathscr{B}_{p+d}(R)$.

The second inequality in (14) shows that R can be upper bounded by $|\mathscr{A}(\beta)| \times \|\beta\|_\infty$, and therefore the constraint $\beta \in \mathscr{B}_{p+d}(R)$ becomes merely a box constraint on β, which depends on the dimensionality of the features through $|\mathscr{A}(\beta)|$ only. The fact that the procedure depends on R, and that the oracle inequality stated in Theorem 1 depends linearly on R, is commonly found in the literature on sparse generalized linear models, see Van de Geer [2008], Bach [2010], Ivanoff et al. [2016]. However, the constraint $\mathscr{B}_{p+d}(R)$ is a technicality which is not used in the numerical experiments in Sections 4 and 5.

Notice in addition that our proof is different from that of Huang et al. [2013] and could be applied in their setting (lasso in the Cox model with time-dependent covariates). Alternative oracle inequalities, in terms of the Kullback-Leibler divergence instead of the symmetric Bregman divergence, could hence be proven.

3.2 Oracle inequality for estimation

Approximation of $f^\star$. Since $\beta^\star \in \mathbb{R}^{p+K^\star}$ and $\hat{\beta} \in \mathbb{R}^{p+d}$, we define in this section an approximation of $f^\star$ denoted $f_{b^\star}$ with $b^\star \in \mathbb{R}^{p+d}$. We choose d_j such that

$$\min_{1 \leq k \leq K_j^\star + 1} |I_{j,k}^\star| \geq \max_{1 \leq l \leq d_j + 1} |I_{j,l}| \text{ for all } j = 1, \dots, p.$$

This choice ensures that for all features $j = 1, \dots, p$, there exists a unique interval $I_{j,l}$ containing cut-point $\mu_{j,k}^\star$, which we denote

$$I_{j,l_{j,k}^\star} = (\mu_{j,l_{j,k}^\star - 1}, \mu_{j,l_{j,k}^\star}] \tag{15}$$

for all $k = 1, \dots, K_j^\star$. Note that in practice, this requirement is met by increasing d_j. For each single jth block, let us recall that as defined in (1), we associate with $\beta_{j,\bullet}^\star$ the $\mu_{j,\bullet}^\star$-piecewise constant function

$$f_j^\star : x \mapsto \sum_{k=1}^{K_j^\star + 1} \beta_{j,k}^\star \mathbb{1}(x \in I_{j,k}^\star)$$

defined for all $x \in [0,1]$. Now, let us define the $\mu_{j,\bullet}$-piecewise constant function

$$\tilde{f}_j : x \mapsto \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star} \mathbb{1}(x \in I_{j,l}), \tag{16}$$

for $x \in [0,1]$, where $l_{j,k}^\star$ is defined in (15), and with the conventions $l_{j,0}^\star = 0$ and $l_{j,K_j^\star+1}^\star = d_j + 1$ for all $j = 1, \dots, p$. With this definition, $\tilde{f}_j$ has the same number of jumps and amplitudes thereof as $f_j^\star$. The only difference between these two functions is the location of the jumps: $f_j^\star$ jumps once for each cut-point $\mu_{j,k}^\star$ for all $k = 1, \dots, K_j^\star + 1$, while $\tilde{f}_j$ jumps once for each $\mu_{j,l}$ closest (on the right hand side) to $\mu_{j,k}^\star$ for all $k = 1, \dots, K_j^\star + 1$. This choice of approximation is discussed at the beginning of Appendix C.

In the jth block, the vector associated with $\tilde{f}_j$ now lives in $\mathbb{R}^{d_j+1}$ as expected, but the extra linear constraint required to apply Theorem 1 is not fulfilled. We then define

$$f_{b_{j,\bullet}^\star} : x \mapsto \tilde{f}_j(x) - \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_j(X_{i,j}) \tag{17}$$

for $x \in [0,1]$, which gives rise to $n_{j,\bullet}^\top b_{j,\bullet}^\star = 0$ for all $j = 1, \dots, p$, where $b_{j,\bullet}^\star \in \mathbb{R}^{d_j+1}$ is the vector associated with $f_{b_{j,\bullet}^\star}$.

Denoting $b^\star = \left((b_{1,\bullet}^\star)^\top, \dots, (b_{p,\bullet}^\star)^\top\right)^\top$, our approach to prove the oracle inequality for estimation relies on the application of Theorem 1 to the approximate candidate $b^\star \in \mathbb{R}^{p+d}$ of $\beta^\star$. Figure 1 gives a clearer view of the different quantities involved so far in the estimation procedure on a toy example. See also the upper part of Figure 4 in Section 4.4. Note that, in addition, if $\beta^\star$ is block-sparse, then it is also the case for $b^\star$, and the following holds:

$$|\mathscr{A}(b^\star)| \leq |\mathscr{A}(\beta^\star)|.$$

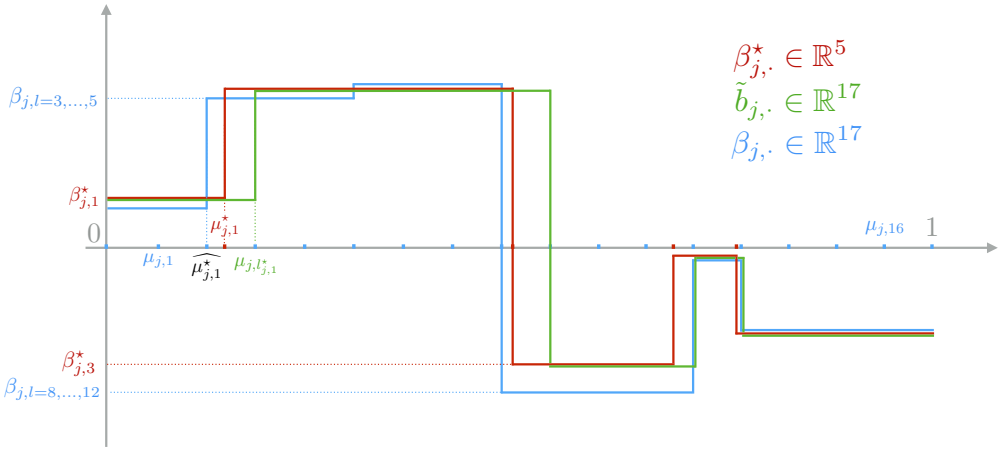


Fig. 1: Illustration of the different vectors for the jth block, with $d_j = 17$. In this scenario, the algorithm detects an extra cut-point and $\widehat{K}_j = 5 = s_j$, while $K_j^\star = 4$.

Let us introduce some further notation. We define

$$\pi_n = \frac{|\{i = 1, \dots, n : N_i(\tau) = 1\}|}{n}, \tag{18}$$

and let in addition

$$R^\star = \sum_{j \in \mathscr{A}(\beta^\star)} \|b^\star_{j,\bullet}\|_\infty,$$

$$\mathbf{I} = 2\big(|\mathscr{A}(\beta^\star)| + K^\star\big)\Big(1 + 3\frac{\psi(f^\star_\infty + R^\star + 2)}{f^\star_\infty + R^\star + 2}\Big)\pi_n \max_{j \in \mathscr{A}(\beta^\star)} \|\beta_{j,\bullet}\|^2_\infty \max_{j \in \mathscr{A}(\beta^\star)} \|n_{j,\bullet}/n\|^2_\infty \Big(1 + \frac{4e^{2f^\star_\infty}}{c_Z}\Big),$$

where $\psi(x) = e^x - x - 1$, and

$$\mathbf{II} = \frac{2048(f^\star_\infty + R^\star + 2)^2 K^\star \max_{1 \le j \le p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(b^\star)}\|^2_\infty}{\kappa^2_\tau\big(\mathcal{A}(b^\star)\big) - \Xi_\tau\big(\mathcal{A}(b^\star)\big)}.$$

Theorem 2 *The inequality*

$$\|(\hat{\beta} - b^\star)_{\mathcal{A}(b^\star)}\|_1 \le \frac{\sqrt{K^\star(\mathbf{I} + \mathbf{II})}}{\kappa_\tau\big(\mathcal{A}(b^\star)\big)} \tag{19}$$

holds with probability greater than $1 - 28.55e^{-c} - e^{-ns^{(0)}(\tau)^2/8e^{2f^\star_\infty}} - 3\varepsilon - 2e^{-nc_Z^2/2}$ *for some* $c > 0$.

A proof of Theorem 2 is presented in Appendix C. The term $\mathbf{I}$ is a bias term and, if all $d_j \to \infty$ as $n \to \infty$ and under mild conditions on the distributions of the $X_{i,j}$, it goes to 0 as $n \to \infty$. The order of magnitude in the inequality of Theorem 2 is then given, for n and d_j large enough, by

$$\frac{\sqrt{K^\star(\mathbf{I} + \mathbf{II})}}{\kappa_\tau\big(\mathcal{A}(b^\star)\big)} \lesssim \frac{K^\star\sqrt{\log(p+d)/n}}{\kappa_\tau\big(\mathcal{A}(b^\star)\big)\sqrt{\kappa^2_\tau\big(\mathcal{A}(b^\star)\big) - \Xi_\tau\big(\mathcal{A}(b^\star)\big)}},$$

which is the expected fast rate in oracle inequalities for estimation, see for instance Bickel et al. [2009].

4 Performance evaluation

4.1 Practical details

Let us now give some details about the binacox's use in practice. First, as already mentioned, we naturally choose the estimated quantiles for the $\mu_{j,l}$. This choice provides two major practical advantages: *i*) the resulting grid is data-driven and follows the distribution of $\boldsymbol{X}_{\bullet,j}$, and *ii*) there is no need to tune hyper-parameters d_j (number of bins for the one-hot encoding of raw feature j). Indeed, if d_j is "large enough" (we take $d_j = 50$ for all $j = 1, \dots, p$ in practice), increasing d_j barely changes the results since the cut-points selected by the penalization no longer change, and the size of each block automatically adapts itself to the data; depending on the distribution of $\boldsymbol{X}_{\bullet,j}$, ties may appear in the corresponding empirical quantiles (for more details on this last point, see Alaya et al. [2017]).

Note also that the binacox is proposed in the `tick` library [Bacry et al., 2017], and that all the code used in this paper is open-sourced at `https://github.com/SimonBussy/binacox` ; we provide sample code for its use in Figure 2. For practical convenience, we take all weights $\omega_{j,l} = \gamma$ and select the hyper-parameter γ using a V-fold cross-validation procedure with $V = 10$, taking the negative partial log-likelihood defined in (4) as a score computed after a refit of the model on the binary space obtained by the estimated

cut-points, and with the sum-to-zero constraint only (without the TV penalty, which actually gives a fair estimate of $\beta^\star$ in practice), which intuitively makes sense. Figure 10 in Appendix A.2 gives the learning curves obtained with this cross-validation procedure on an example.

We also add a simple de-noising step in the cut-point detection phase, which is useful in practice. Indeed, it is usual to observe two consecutive $\hat{\beta}$'s jumps in the neighbourhood of a true cut-point, leading to an over-estimation of $K^\star$. This can be viewed as a clustering problem. We tried different clustering methods but in practice, nothing works better than this simple routine: if $\hat{\beta}$ has three consecutive different coefficients within a block, then only the largest jump is considered as a "true" jump. Figure 11 in Appendix A.2 illustrates this routine.

```python
from tick.simulation import SimuCoxRegWithCutPoints
from tick.preprocessing.features_binarizer import FeaturesBinarizer
from tick.inference import CoxRegression

# Generate data
simu = SimuCoxRegWithCutPoints(n_samples=1000, n_features=20)
X, Y, delta = simu.simulate()

# Binarize features
binarizer = FeaturesBinarizer(n_cuts=50)
X_bin = binarizer.fit_transform(X)

# Fit the model with a penalty strength equal to `C`
learner = CoxRegression(penalty='binarsity',
                        blocks_start=binarizer.blocks_start,
                        blocks_length=binarizer.blocks_length,
                        C=10)
learner.fit(X_bin, Y, delta)

# Obtain the estimated vector
beta = learner.coeffs
```

Fig. 2: Sample python code for the use of the binacox in the `tick` library, using the `FeaturesBinarizer` transformer for feature binarization.

4.2 Simulation

In order to assess the methods, we run an extensive Monte Carlo simulation study. Let us first present the design used in the following.

4.2.1 Design.

We first take $[X_{i,j}] \in \mathbb{R}^{n\times p} \sim \mathcal{N}\big(0, \Sigma(\rho)\big)$, with $\Sigma(\rho)$ a $(p \times p)$ Toeplitz covariance matrix [Mukherjee and Maiti, 1988] with correlation $\rho \in (0, 1)$. For each feature $j = 1, \ldots, p$, we sample the cut-points $\mu^\star_{jk}$ uniformly without replacement from the estimated quantiles $q_j(u/10)$ for $u = 1, \ldots, 9$ and $k = 1, \ldots, K^\star_j$. In this way, we avoid having undetectable cut-points (with very few examples above the cut-point value) or pairs of overly close together indissociable cut-points. We choose the same $K^\star_j$ values for all $j = 1, \ldots, p$. Now that the true cut-points vector $\mu^\star$ has been generated, one can compute the corresponding binarized version of the features, which we denote $x_i^{B^\star}$ for the ith example. Then, we generate

$$c_{jk} \sim (-1)^k |\mathcal{N}(1, 0.5)|$$

Table 1: Hyper-parameter choices for simulation.

n	p	ρ	$K_j^\star$	ν	ς	r_c	r_s
(200, 4000)	50	0.5	$\{1, 2, 3\}$	2	0.1	0.3	0.2

for all $k = 1, \ldots, K_j^\star + 1$ and $j = 1, \ldots, p$ to make sure we create "real" cut-points, and take

$$\beta_{jk}^\star = c_{jk} - (K_j^\star + 1)^{-1} \sum_{k=1}^{K_j^\star+1} c_{jk}$$

in order to impose the sum-to-zero constraint of the true coefficients in each block. We also induce a sparsity aspect by uniformly selecting a proportion r_s of features $j \in \mathcal{S}$ with no cut-point effect, i.e., features for which we enforce $\beta_{jk}^\star = 0$ for all $k = 1, \ldots, K_j^\star + 1$. Lastly, we generate survival times using Weibull distributions, which is a common choice in survival analysis [Klein and Moeschberger, 2005]:

$$T_i \sim \nu^{-1} \big[-\log(U_i) \exp\big(-(x_i^{B^\star})^\top \beta_i^\star \big) \big]^{1/\varsigma}$$

with $\nu > 0$ and $\varsigma > 0$ the scale and shape parameters respectively, and $U_i \sim \mathcal{U}([0, 1])$, where $\mathcal{U}([a, b])$ stands for the uniform distribution on a segment $[a, b]$. The distribution of the censoring variable C_i is the geometric distribution $\mathcal{G}(\alpha_c)$, where $\alpha_c \in (0, 1)$ is empirically tuned to maintain a desired censoring rate $r_c \in [0, 1]$. The choice of all hyper-parameters is driven by the applications on real data presented in Section 5, and summarized in Table 1. Figure 3 gives an example of data generated according to the design we have just described.

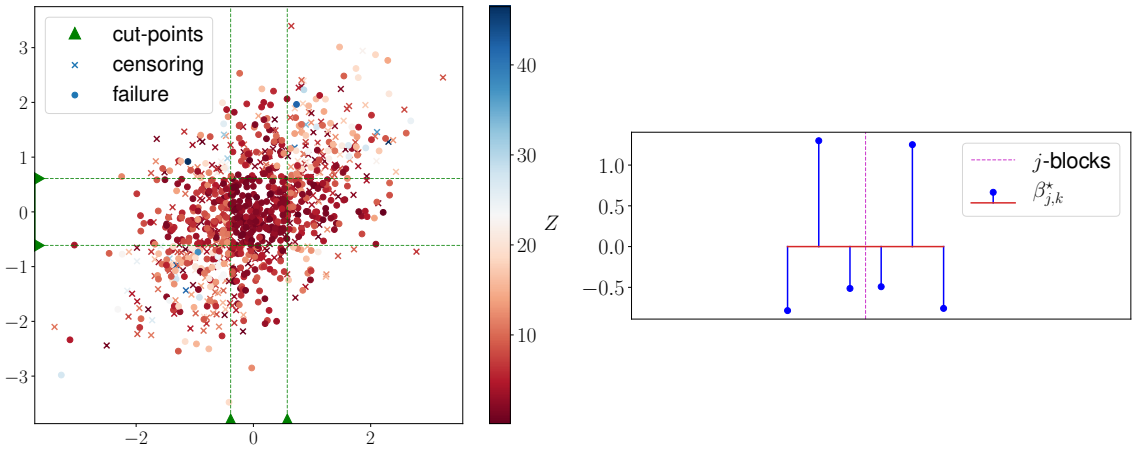


Fig. 3: Left: illustration of data simulated with $p = 2$, $K_1^\star = K_2^\star = 2$, and $n = 1000$. Dots represent failure times ($z_i = t_i$) while crosses represent censoring times ($z_i = c_i$), and the colour gradient represents the z_i values (red for low and blue for high). Right: $\beta^\star$ is plotted, with a dotted line to demarcate the two blocks (since $p = 2$).

4.2.2 Metrics.

We evaluate the methods being analysed using two metrics. The first assesses the estimation of the cut-points values by

$$m_1 = |\mathcal{S}'|^{-1} \sum_{j \in \mathcal{S}'} \mathcal{H}(\mathcal{M}_j^\star, \widehat{\mathcal{M}}_j),$$

where $\mathcal{M}_j^\star = \{\mu_{j,1}^\star, \ldots, \mu_{j,K_j^\star}^\star\}$ (resp. $\widehat{\mathcal{M}}_j = \{\hat{\mu}_{j,1}, \ldots, \hat{\mu}_{j,\widehat{K}_j}\}$) is the set of true (resp. estimated) cut-points for feature j, $\mathcal{S}' = \left\{j,\ j \notin \mathcal{S} \cap \{l,\ \widehat{\mathcal{M}}_l = \emptyset\}\right\}$ the indexes corresponding to features with at least one true cut-point and one detected cut-point, and $\mathcal{H}(A, B)$ the Hausdorff distance between the sets A and B, defined as

$$\mathcal{H}(A, B) = \max\left(\mathcal{E}(A||B), \mathcal{E}(B||A)\right),$$

where $\mathcal{E}(A||B) = \sup_{b \in B} \inf_{a \in A} |a - b|$. This is inspired by Harchaoui and Lévy-Leduc [2010], except that in our case, both $\mathcal{M}_j^\star$ and $\widehat{\mathcal{M}}_j$ can be empty, which explains the use of $\mathcal{S}'$. The second metric we use is precisely focused on the sparsity aspect; it assesses the ability for each method to detect features with no cut-points, and is defined by

$$m_2 = |\mathcal{S}|^{-1} \sum_{j \in \mathcal{S}} \widehat{K}_j.$$

4.3 Competing methods

To the best of our knowledge, all existing algorithms and methods are based on multiple log-rank tests in univariate models. These methods are widely used, and recent implementations include the web applications `Cutoff Finder` and `Findcutoffs` described in Budczies et al. [2012] and Chang et al. [2017] respectively.

We describe in what follows the principle of these univariate log-rank tests. Consider one of the initial variables $\boldsymbol{X}_{\bullet,j} = (x_{1,j}, \ldots, x_{n,j})^\top$, and denote its 10th and 90th quantiles as $x_{10th,j}$ and $x_{90th,j}$. Then, define a grid $\{g_{j,1}, \ldots, g_{j,\kappa_j}\}$. In most implementations, the $g_{j,k}$'s are chosen at the original observation points and are such that $x_{10th,j} \leq g_{j,k} \leq x_{90th,j}$. For each $g_{j,k}$, the p-value $\mathrm{pv}_{j,k}$ of the log-rank test associated with the univariate Cox model defined by

$$\lambda_0(t) \exp\left(\beta^j \mathbb{1}(x \leq g_{j,k})\right)$$

is computed (via the `python` package `lifelines` in our implementation). For each initial variable $\boldsymbol{X}_{\bullet,j}$, κ_j p-values are available at this stage. The choice of the size κ_j of the grid depends on the implementation, and ranges for several dozen to all observed values between $x_{10th,j}$ and $x_{90th,j}$.

In Figure 4, the values $-\log(\mathrm{pv}_{j,k})$ for $k = 1, \ldots, \kappa_j$ (denoted by "MT" for "Multiple Testing") are represented, for the simulated example illustrated in Figure 3. Notice that the level $-\log(\alpha) = -\log(0.05)$ is exceeded for numerous $g_{j,k}$'s values, and of course this procedure allows us to detect only a single cut-point per feature. A common approach is to consider the maximal value $-\log(\mathrm{pv}_{j,\hat{k}})$ and then define the cut-point for variable j as $g_{j,\hat{k}}$. As argued in Altman et al. [1994], this is obviously "associated with an inflation of type I error", and for this reason we do not consider this approach.

To cope with the multiple testing (MT) problem at hand, multiple testing corrections have to be applied, of which we consider two. The first is the well-known Bonferroni p-value correction, referred to as MT-B in the following. We insist on the fact that although commonly used, this method is not correct in this situation since the p-values

are correlated. Note also that in this context, the Benjamini–Hochberg (BH) procedure would result in the same cut-points being detected as MT-B (with FDR=α), since we only consider as a cut-point candidate the points with minimal p-value. Indeed, applying the classical BH procedure would select far too many cut-points. The second correction, denoted MT-LS, is the correction proposed in Lausen and Schumacher [1992], based on asymptotic theoretical considerations. Figure 4 also illustrates how these corrections behave on the simulated example illustrated in Figure 3. A third correction we could imagine would be a bootstrap-based MaxT procedure (or MinP) as proposed in Dudoit and Van Der Laan [2007] or Westfall et al. [1993], but this would be intractable in our high-dimensional setting (see Figure 5(a) that compares the computing times for a single feature only; a bootstrap procedure based on MT would dramatically increase the required computing time).

4.4 Simulation results

4.4.1 Example.

Figure 4 illustrates how the methods considered behave on the data shown in Figure 3. With the help of this example, we can clearly see the good performance of the binacox method: the position, strength and number of cut-points are well estimated. The MT-B and MT-LS methods can only detect one cut-point by construction. Both methods detect "the most significant" cut-point for each of the 2 features, namely those corresponding to the highest jumps in $\beta^\star_{j,\bullet}$ (see Figure 3): $\mu^\star_{1,1}$ and $\mu^\star_{2,2}$.

With regards to the shape of the "p-value curves", one can see that for each of the two features, the two "main" local maxima correspond to the true cut-points. One could then imagine creating a method for detecting such maxima, but this is beyond the scope of this paper (plus it would still be based on MT methods, which have high computational costs, as detailed hereafter).

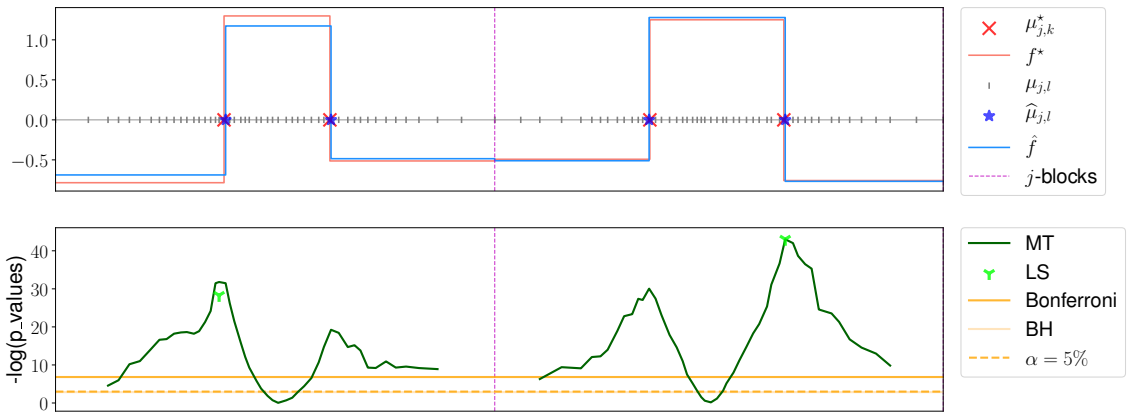


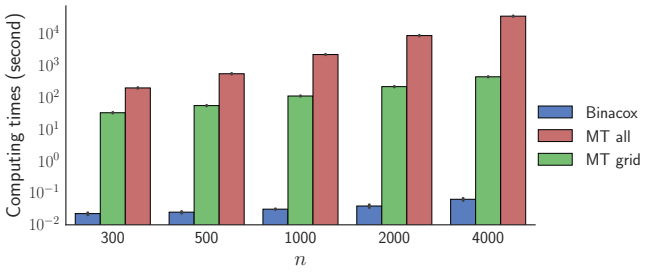
Fig. 4: Top: Illustration of the main quantities involved in the binacox, with estimations obtained for the data represented in Figure 3. Our algorithm detects the correct number of cut-points $\widehat{K}_j = 2$, and estimates their positions accurately, as well as their amplitudes. Bottom: results obtained using the multiple testing-related methods introduced in Section 4.3. Here the BH threshold lines overlap that corresponding to $\alpha = 5\%$. The BH procedure would consider as cut-points all $\mu_{j,l}$ values for which the corresponding dark green (MT) line's values are above this, thus detecting far too many cut-points.

4.4.2 Computing times.

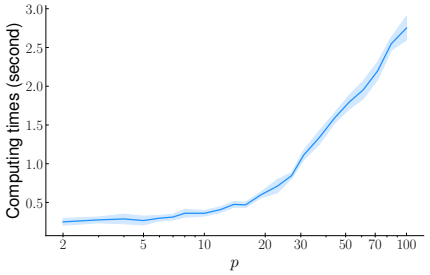
Now let us look at the computing time required for the methods considered. As the multiple testing-related methods are univariate, we can directly parallelize their computations across dimensions (which is what we did in the applications), so let us consider here a single feature X ($p = 1$). Following the method explained in Section 4.3, we have to compute all log-rank test p-values computed on the populations $\{y_i : x_i > \mu\}$ and $\{y_i : x_i \leq \mu\}$ for $i = 1, \ldots, n$, for μ taking all x_i values between the 10th and 90th empirical quantiles of X. We denote "MT all" this method in Figure 5(a), and compare its computing times with the binacox for various values of n. We also show the "MT grid" method that only computes the p-values for candidates $\mu_{j,l}$ used in the binacox method.

Since the number of candidates does not change with n for the "MT grid" method, the computing time ratio between "MT all" and "MT grid" naturally increases, going roughly from one to two orders of magnitude higher when n goes from 300 to 4000. Hence to make computations much faster, we will use the "MT grid" for all multiple testing-related methods in the following. The resulting loss of precision in the MT-related methods is negligible for a high enough d_j ($= 50$ in practice).

Next, we emphasize the fact that the binacox is still roughly 5 times faster than the "MT grid" method, and it remains very fast when we increase the dimension, as shown in Figure 5(b). It turns out that the computational time grows roughly logarithmically with p.



(a) Average computing times in seconds (with the black lines representing ± the standard deviation) obtained on 100 simulated datasets (according to Section 4.2 with $p = 1$ and $K^\star = 2$) for training the binacox versus the multiple testing methods, where cut-point candidates are either all x_i values between the 10th and 90th empirical quantiles of X ("MT all"), or the same candidates as the grid considered by the binacox ("MT grid").



(b) Average (bold) computing times in seconds and standard deviation (bands) obtained on 100 simulated datasets (according to Section 4.2 with $K_j^\star = 2$) for training the binacox when increasing the dimension p up to 100. The method remains very fast in high-dimensional settings.

Fig. 5: Computing time for the methods considered.

4.4.3 Performance comparison.

Let us compare now the results of simulations in terms of the m_1 and m_2 metrics introduced in Section 4.2. Figure 6 gives a comparison of the methods considered for the cut-point estimation aspect, i.e., in terms of the m_1 score. It appears that the binacox outperforms the MT-related methods when $K_j^\star > 1$, and is competitive when $K_j^\star = 1$ except for small values of n. This is due to an overestimation in the number of cut-points by the binacox (see Figure 7), especially when p is high and n is small, which gives higher m_1 values, even if the "true" cut-point is actually well-estimated. Note that for such values of p, the binacox runs much faster than the MT-related methods.

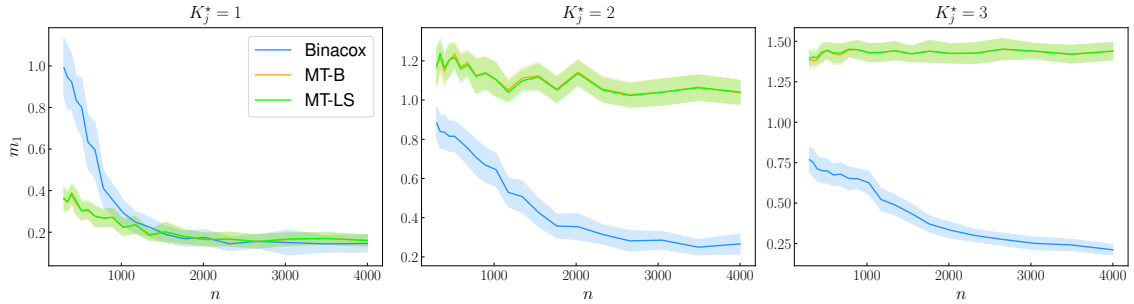


Fig. 6: Average (bold) m_1 scores and standard deviation (bands) obtained on 100 datasets simulated according to Section 4.2 with $p = 50$ and $K_j^\star$ equal to 1, 2 and 3 (for all $j = 1, \ldots, p$) for the left, center and right sub-figures respectively) for varying n. The lower the value of m_1, the better the result; the binacox clearly outperforms the other methods when there is more than one cut-point, and is competitive with other methods when there is only one cut-point, but performs worse when n is small because it overestimates $K_j^\star$.

Figure 7, on the other hand, assesses the ability of each method to detect features with no cut-points using the m_2 metric, i.e., the ability to estimate $\hat{K}_j^\star = 0$ for $j \in \mathcal{S}$. The binacox appears to be quite effective at detecting features with no cut-point when n takes a high enough value compared to p, which is not the case for the MT-related methods.

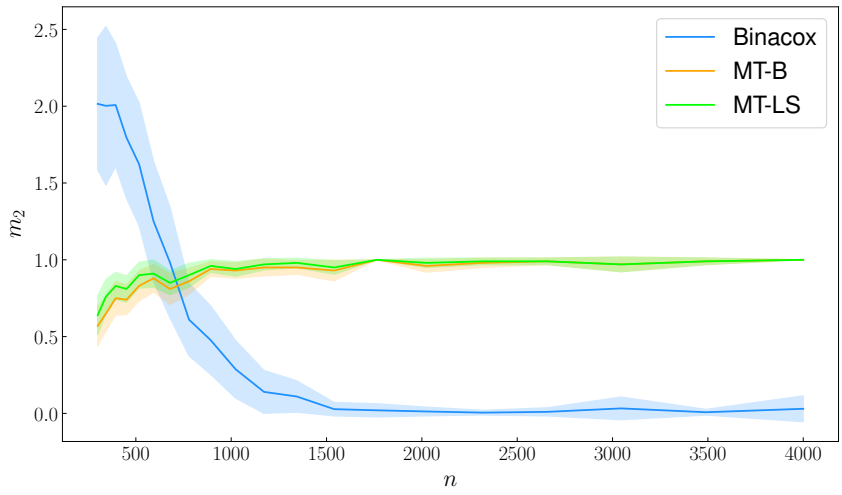


Fig. 7: Average (bold) m_2 scores and standard deviation (bands) obtained on 100 datasets simulated according to Section 4.2 with $p = 50$ for varying n. MT-B and MT-LS tend to detect a cut-point when there is none (no matter the value of n), while binacox overestimates the number of cut-points for small values of n but detects $\mathcal{S}$ well for $p = 50$ on the simulated data when $n > 1000$.

5 Application on genetic data

In this section, we apply our method to three biomedical datasets. We extracted normalized expression data and survival times Z in days from breast invasive carcinoma (BRCA, $n = 1211$), glioblastoma multiforme (GBM, $n = 168$) and kidney renal clear cell carcinoma (KIRC, $n = 605$). These datasets are available on *The Cancer Genome Atlas* (TCGA)

Table 2: Estimated cut-point values for each method on the top 10 genes presented in Figure 8 for GBM. Dots (·) mean "no cut-point detected".

Genes	Binacox	MT-B	MT-LS
SOD3 6649	200.87, 326.40, 606.48	·	·
LOC 400752	31.46, 62.50	·	34.04
C11orf63 79864	40.30, 109.67	19.65	19.65
KTI12 112970	219.60, 305.70	219.60	219.60
HOXC8 3224	3.30, 15.75	3.30	3.30
DDX5 1655	10630.11, 13094.89	·	·
FKBP9L 360132	111.72	·	·
HOXA1 3198	67.28	·	·
MOSC2 54996	107.53	107.53	107.53
ZNF680 340252	385.85, 638.06	385.85	385.85

platform, which aims to accelerate the understanding of the molecular basis of cancer with the help of genomic technology, including large-scale genome sequencing. For each patient, 20,531 features corresponding to normalized gene expression values are available.

As we saw in Section 4.4, the MT-related methods are intractable in such high-dimensional cases. We therefore include a screening step to select the portion of features most relevant to our problem from the 20,531 available. To do so, we fit the binacox on each jth block separately and take the resulting $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ as a score that roughly assess the propensity for feature j to have one (or more) relevant cut-point(s). We then select the features corresponding to the top P values with $P = 50$, this choice being suggested by the distribution of the obtained scores given in Figure 12 of Appendix A.3.

5.1 Estimation results.

In Figure 8 we present the results obtained by the methods considered on the GBM cancer dataset for the top 10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values. We observe that all cut-points detected by the univariate multiple testing methods with Bonferroni (MT-B) or Lausen and Schumacher (MT-LS) corrections are also detected by the multivariate binacox (which detects more cut-points); see Table 2. The binacox identifies many more cut-points than the univariate MT-B and MT-LS methods. Further, all cut-points detected by these two methods are also detected by the binacox. Furthermore, it turns out that these top 10 genes (from the original 20,531) are quite relevant to GBM, the most aggressive cancer that begins in the brain.

For instance, the first gene, SOD3, is relevant from a physiopathological point of view since its polymorphisms are already known as GBM risk factors [Rajaraman et al., 2008]. Other genes in the top 10 (C11orf63 or the HOX genes) are also known to be directly related to brain development [Canu et al., 2009], and are already known as potential GBM prognosis marker [Duan et al., 2015, Guan et al., 2019].

Relevant results were also obtained on the KIRC and BRCA datasets; these are postponed to Appendix A.4.

5.2 Risk prediction.

Let us now investigate how performances are impacted in terms of risk prediction when detected cut-points are taken into account; namely, comparing predictions when training

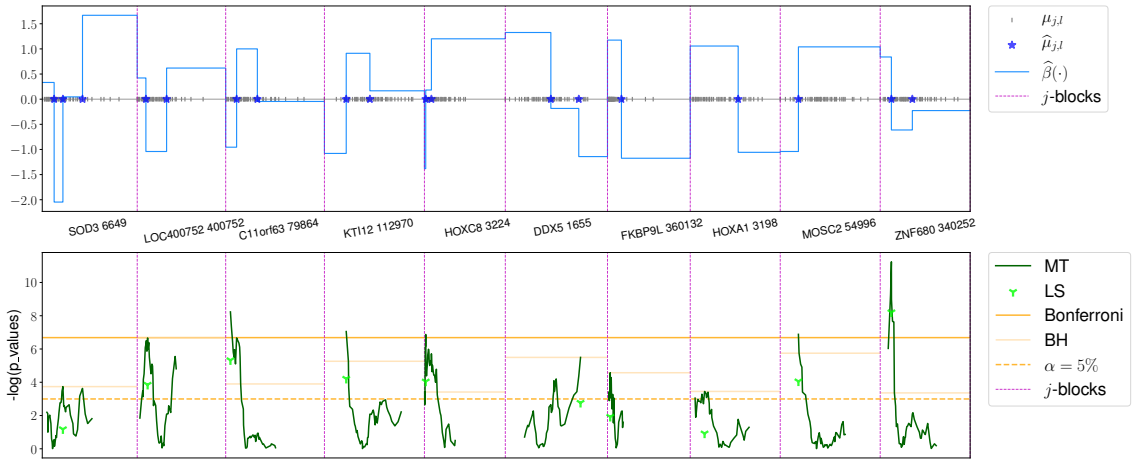


Fig. 8: Illustration of the results obtained on the top 10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the GBM dataset. The binacox detects multiple cut-points and sheds light on non-linear effects for various genes. The BH thresholds are shown, but are unusable in practice.

a Cox model on the original continuous feature space versus on the $\widehat{\mu}$-binarized space constructed with the cut-point estimates.

In a classical Cox model, $R_i = \exp(X_i^\top \hat{\beta})$ is known as the predicted risk for patient i measured at $t = 0$. A common metric to evaluate risk prediction performances in this type of survival setting is the C-index [Heagerty and Zheng, 2005], which is defined by

$$\mathcal{C}_\tau = \mathbb{P}[R_i > R_j | Z_i < Z_j, Z_i < \tau],$$

with $i \neq j$ two independent patients and τ the follow-up period. A Kaplan-Meier estimator for the censoring distribution leads to a nonparametric and consistent estimator of $\mathcal{C}_\tau$ [Uno et al., 2011], which is already implemented in the `python` package `lifelines`.

We randomly split the three datasets 100 times into training and validation sets (30% for testing) and compare the average C-index on the validation sets in Table 5.2 when the $\widehat{\mu}$-binarized space is constructed based on the $\widehat{\mu}$'s obtained either from the binacox, MT-B, or MT-LS. We also compare performances obtained by two nonlinear multivariate methods known to perform well in high-dimensional settings: boosted Cox (CoxBoost) [Li and Luan, 2005] used with 300 boosting steps (this number being fine-tuned by cross-validation), and random survival forests (RSF) [Ishwaran et al., 2008] used with 200 trees (also cross-validated), respectively implemented in the `R` packages `CoxBoost` and `randomForestSRC`. Note that for a fair comparison, and to avoid selection bias [Ambroise and McLachlan, 2002], the screening step is re-run on each training set, using the C-index obtained by univariate Cox models (not to confer advantage to our method), namely Cox PH models fitted on each covariate separately.

The binacox method clearly improves risk prediction compare to classical Cox, as well as with respect to the MT-B and MT-LS methods. Moreover, it also significantly outperforms both CoxBoost and RSF. To the best of our knowledge, no better performances have been achieved on this data in the literature [Yousefi et al., 2017]. Figure 9 compares the computing times of the methods. Clearly the binacox is by far the most computationally efficient.

Table 3: Comparison of average C-indexes (and standard deviation in parentheses) on 100 random train/test splits for the Cox model trained on continuous features versus on its binarized version constructed using the considered methods' cut-point estimates, and the CoxBoost and RSF methods. On the three datasets, the binacox method gives by far the best results (in bold).

Cancer	Continuous	Binacox	MT-B	MT-LS	CoxBoost	RSF
GBM	0.563 (0.037)	**0.603 (0.048)**	0.579 (0.049)	0.577 (0.043)	0.569 (0.037)	0.564 (0.036)
KIRC	0.675 (0.028)	**0.709 (0.022)**	0.682 (0.022)	0.682 (0.022)	0.683 (0.029)	0.695 (0.026)
BRCA	0.592 (0.050)	**0.669 (0.047)**	0.626 (0.055)	0.621 (0.061)	0.598 (0.053)	0.659 (0.037)

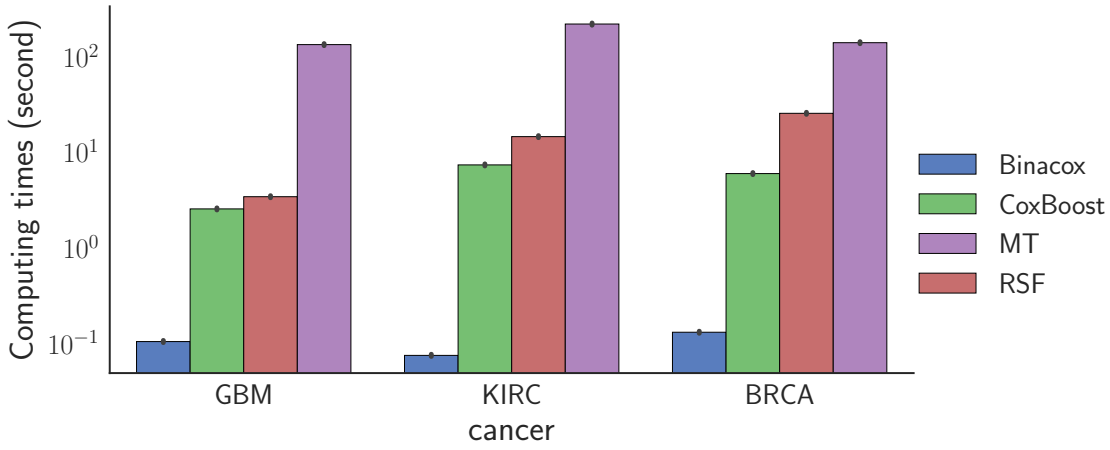


Fig. 9: Average computing times (in seconds) required by each method on the three datasets (with the black lines representing ± the standard deviation) obtained on 100 random train/test split. The binacox method is at least one and up to several orders of magnitude faster.

6 Conclusion

In this paper, we introduced the binacox method, designed for estimating multiple cut-points in a Cox model with high-dimensional features. We illustrated the good theoretical properties of the model by establishing nonasymptotic oracle inequalities for prediction and estimation. An extensive Monte Carlo simulation study was then carried out to evaluate the method's performance. It showed that our approach outperforms existing methods, with computing times orders of magnitude faster. Moreover, in addition to the raw feature selection ability of the binacox, it succeeds in detecting multiple cut-points per feature. We also applied the binacox to three publicly available high-dimensional genetics datasets. Furthermore, several genes pinpointed by the model turn out to be biologically relevant (e.g., the gene SOD3 for GBM), whilst others require further investigation in the genetics research community. More importantly, our method provides powerful interpretation aspects that could be useful in both clinical research and daily practice. Indeed, the estimated cut-points could be directly considered in clinical practice. Thus, the method could be an interesting alternative to more classical methods found in the medical literature to deal with prognosis studies in high-dimensional frameworks, providing a new way to model nonlinear feature associations, and giving rise to new data-driven risk scores. Our study lays the groundwork for the development of powerful methods which could one day help provide improved personalized care.

Acknowledgments

Mokhtar Z. Alaya is grateful for a grant from DIM Math Innov Région Ile-de-France http://www.dim-mathinnov.fr. Agathe Guilloux's work has been supported by the INCA-DGOS grant PTR-K 2014. The results shown in this paper are based upon data generated by the TCGA Research Network and freely available from http://cancergenome.nih.gov. *Conflict of Interest*: None declared.

Software

All methodology discussed in the paper is implemented in `Python/C++` and `R`. The code that generates all figures is available from https://github.com/SimonBussy/binacox in the form of annotated programs, together with notebook tutorials.

Appendix A Additional details

A.1 Algorithm.

To solve regularization problem (5), we first look at the proximal operator of the binarsity penalty [Alaya et al., 2017]. It turns out that it can be computed very efficiently, using an algorithm introduced in Condat [2013] that we modify in order to include the weights $\omega_{j,k}$. It basically applies – in each block – the proximal operator of the total variation (since the binarsity penalty is block separable), followed by a centering within each block to satisfy the constraint, see Algorithm 1 below. We refer to Alaya et al. [2015] for the weighted total variation proximal operator.

Algorithm 1 Proximal operator of bina(β), see [Alaya et al., 2017]

Input: vector $\beta \in \mathscr{B}_{p+d}(R)$ and weights $\omega_{j,l}$ for $j = 1, \ldots, p$ and $l = 1, \ldots, d_j + 1$
Output: vector $\eta = \mathrm{prox}_{\mathrm{bina}}(\beta)$
for $j = 1$ **to** p **do**
  $\theta_{j,\bullet} \leftarrow \mathrm{prox}_{\|\cdot\|_{\mathrm{TV}},\omega_{j,\bullet}}(\beta_{j,\bullet})$ (TV-weighted in block j, see (7))
  $\eta_{j,\bullet} \leftarrow \theta_{j,\bullet} - \frac{n_{j,\bullet}^\top \theta_{j,\bullet}}{\|n_{j,\bullet}\|_2^2} n_{j,\bullet}$ (projection onto $\mathrm{span}(n_{j,\bullet})^\perp$)
end for
Return: η

A.2 Implementation

Figure 10 gives the learning curves obtained during the V-fold cross-validation procedure presented in Section 4.3 with $V = 10$ for the fine-tuning of parameter γ, which is the strength of the binarsity penalty. We randomly split the data into training and validation sets (30% for validation, cross-validation being done on the training). Recall that the score we use is the negative partial log-likelihood defined in (4) computed after a refit of the model on the binary space obtained by the estimated cut-points, with the sum-to-zero constraint in each block but without the TV penalty.

Figure 11 illustrates the de-noising step for the cut-point detection when looking at the $\hat{\beta}$ support relative to the TV norm. The $\hat{\beta}$ vector plotted here corresponds to the data generated in Figure 3 of Section 4.2, where the final estimation results were presented in Figure 4 of Section 4.4. Since it is usual to observe three consecutive $\hat{\beta}$'s jumps in the neighbourhood of a true cut-point, which is the case in Figure 11 for the first and the last jumps, this could lead to an over-estimation of $K^\star$. To bypass this problem, we then use the following rule: if $\hat{\beta}$ has three consecutive different coefficients within a block, then only the largest jump is considered as a "true" one.

A.3 TCGA gene screening

Figure 12 illustrates the screening procedure followed to reduce the high-dimensionality of the TCGA datasets to make the multiple testing related methods tractable. We then fit a univariate binacox on each block j separately and compute the resulting $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ to assess the propensity for feature j to obtain one (or more) relevant cut-point(s). It appears that taking the top P features with $P = 50$ is a reasonable choice for each dataset considered.

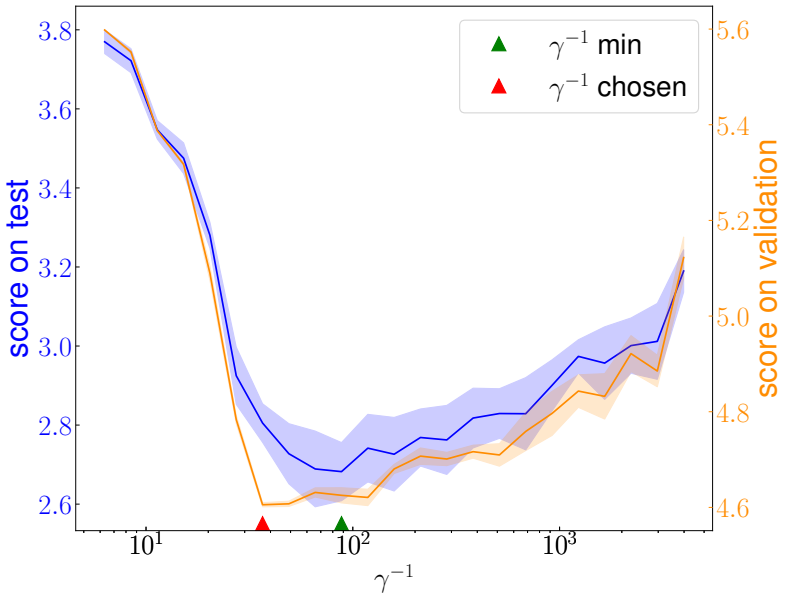


Fig. 10: Learning curves obtained for various γ, in blue on the different test sets during cross-validation, and in orange on the validation set. Bold lines represent average scores on the folds, and bands represent 95% Gaussian confidence intervals. The green triangle points out the value of γ^{-1} that gives the minimum score (best training score), while the γ^{-1} value we automatically select (the red triangle) is the smallest value such that the score is within one standard error of the minimum, which is a classical trick [Simon et al., 2011] that favors a slightly higher penalty strength (smaller γ^{-1}) to avoid over-estimation of $K^\star$ in our case.

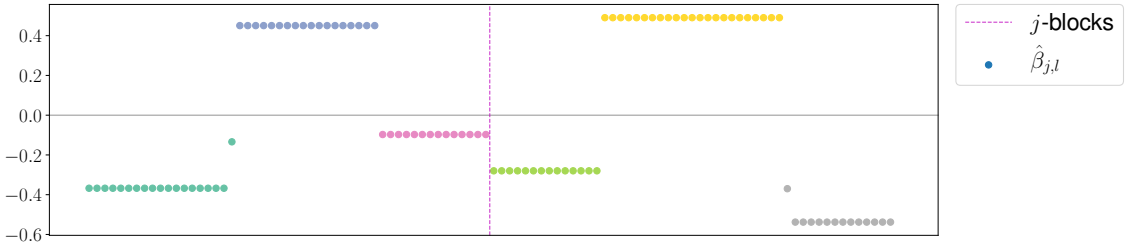


Fig. 11: Illustration of the de-noising step in the cut-point detection phase on the simulated date of Figure 3. Within each block (separated with the dotted pink line), the different colors represent $\hat{\beta}_{j,l}$ with corresponding $\mu_{j,l}$ in distinct estimated $I^\star_{j,k}$. The following rule is applied: when a $\hat{\beta}_{j,l}$ is "isolated", it is assigned to its "closest" group.

A.4 Results on BRCA and KIRC data

Figure 13 illustrates the results obtained by all methods we consider on the BRCA cancer dataset for the top 10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values. Table 4 summarizes the detected cut-point values for each method. It turns out that the selected genes are quite relevant from a clinical point of view (for instance, NPRL2 is a tumor suppressor gene [Huang et al., 2016]), and in particular for BRCA (breast) cancer. For instance, HBS1L expression is known for being predictive of breast cancer survival [Antonov et al., 2014, Antonov, 2011, BioProfiling, 2009], while FOXA1 and PPFIA1 are highly related to breast cancer, see Badve et al. [2007] and Dancau et al. [2010] respectively.

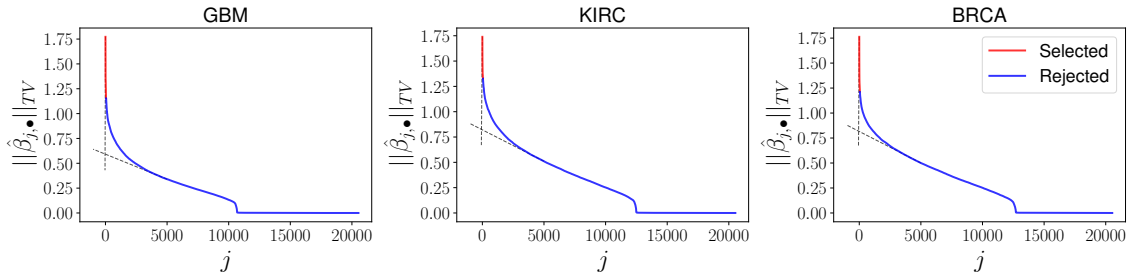


Fig. 12: $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ obtained for univariate binacox fits for the three datasets considered. The top P selected features appear in red, and it turns out that taking $P = 50$ coincides with the elbow (represented with the dotted grey lines) in each of the three curves.

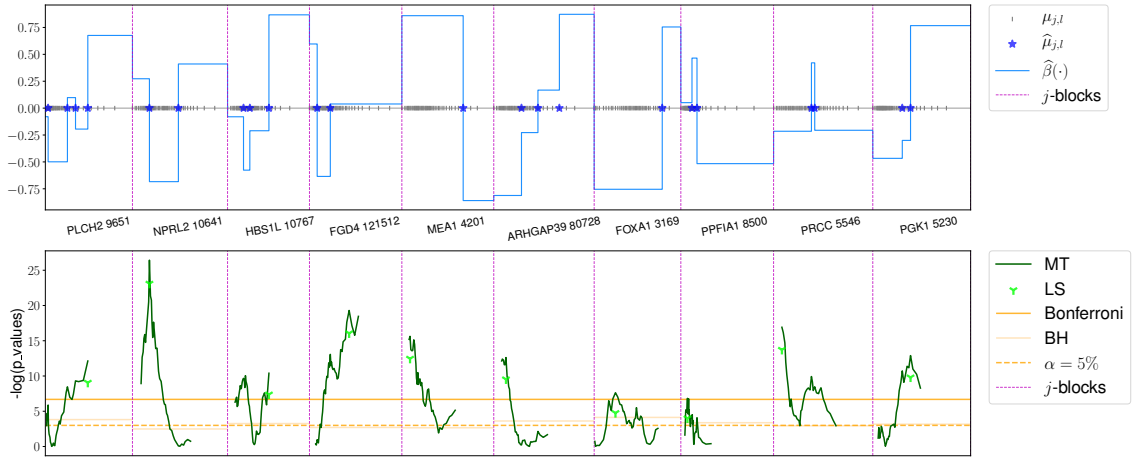


Fig. 13: Illustration of the results obtained on the top 10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the BRCA dataset.

Lastly, Figure 14 gives the results obtained by the various methods on the KIRC cancer dataset for the top 10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values, and Table 5 summarizes the detected cut-point values for each method. Once again, the selected genes are relevant for cancer studies including KIRC. For instance, EIF4EBP2 is related to cancer proliferation [Mizutani et al., 2016]), RGS17 is known to be overexpressed in various cancers [James et al., 2009], and both COL7A1 and NUF2 are known to be related to renal cell carcinoma (see [Csikos et al., 2003] and [Kulkarni et al., 2012] respectively). Moreover, the first two genes MARS 4141 and STRADA 92335 already appear as relevant KIRC prognosis markers in Bussy et al. [2019] .

Appendix B Proof of Theorem 1

In this section, we provide the proof of Theorem 1. First, we derive some preliminary results which will be required in the following.

B.1 Preliminary results

Additional notation. For $u, v \in \mathbb{R}^m$, we denote by $u \odot v$ the Hadamard product defined by $u \odot v = (u_1 v_1, \ldots, u_m v_m)^\top$. We denote by $\mathrm{sign}(u)$ the subdifferential of the function

Table 4: Estimated cut-point values for each method on the top 10 genes presented in Figure 13 for BRCA.

Genes	Binacox	MT-B	MT-LS
PLCH2 9651	28.43, 200.74, 273.04, 382.87	382.87	382.87
NPRL2 10641	330.64, 568.06	330.64	330.64
HBS1L 10767	1023.91, 1212.54, 1782.77	1782.77	1782.77
FGD4 121512	163.59, 309.24	517.90	517.90
MEA1 4201	2199.21	786.29	786.29
ARHGAP39 80728	493.01, 734.37, 1049.04	265.26	265.26
FOXA1 3169	11442.32	3586.03	3586.03
PPFIA1 8500	1500.02, 1885.27	1152.98	1152.98
PRCC 5546	2091.16, 2194.08	1165.49	1165.49
PGK1 5230	10205.72, 12036.29	12036.29	12036.29

Table 5: Estimated cut-point values for each method on the top 10 genes illustrated in Figure 14 for KIRC.

Genes	Binacox	MT-B	MT-LS
MARS 4141	1196.21, 1350.00	1350.00	1350.00
STRADA 92335	495.24, 553.73	586.88	586.88
PTPRH 5794	3.32	3.32	3.32
EIF4EBP2 1979	6504.80	5455.59	5455.59
RGS17 26575	4.30	4.30	4.30
COL7A1 1294	44.19	113.08	113.08
HJURP 55355	99.83	134.31	134.31
NUF2 83540	42.18	63.09	63.09
NDC80 10403	91.39	107.53	107.53
CDCA3 83461	52.03	110.18	110.18

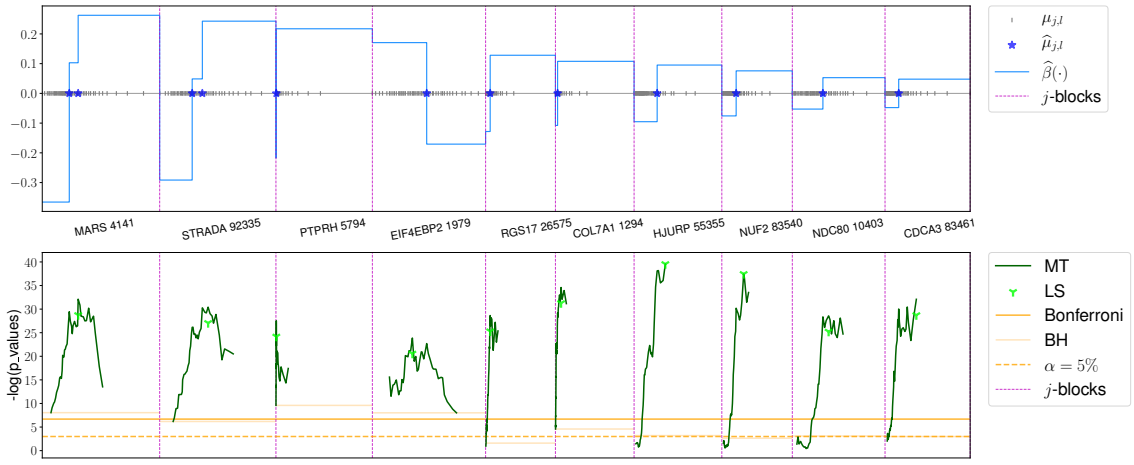


Fig. 14: Illustration of the results obtained on the top 10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the KIRC dataset.

$u \mapsto |u|$, i.e.,

$$\mathrm{sign}(u) = \begin{cases} \{1\} & \text{if } u > 0, \\ [-1,1] & \text{if } u = 0, \\ \{-1\} & \text{if } u < 0. \end{cases}$$

We write $\partial(\phi)$ for the subdifferential mapping of a convex functional ϕ. We adopt in the proofs counting process notation. We then define the observed-failure counting process $N_i(t) = \mathbb{1}(Z_i \leq t, \Delta_i = 1)$, the at-risk process $Y_i(t) = \mathbb{1}(Z_i \geq t)$, and $\bar{N}(t) = \frac{1}{n}\sum_{i=1}^n N_i(t)$. For every vector v, let us denote $v^{\otimes 0} = 1$, $v^{\otimes 1} = v$, and $v^{\otimes 2} = vv^\top$ (outer product). Recall finally that $\tau > 0$ denotes the finite study duration.

Weights. For a given numerical constant $c > 0$, the weights $\omega_{j,l}$ have an explicit form given by

$$\omega_{j,l} = 5.64\sqrt{\frac{c + \log(p+d) + \mathscr{L}_{n,c}}{n}} + 18.62\frac{(c + \log(p+d) + 1 + \mathscr{L}_{n,c})}{n}, \tag{20}$$

where $\mathscr{L}_{n,c} = 2\log\log\big((2en + 24ec) \vee e\big)$.

Properties of the binarsity penalty. We define $\omega = (\omega_{1,\bullet}, \ldots, \omega_{p,\bullet})$ the weights vector, with $\omega_{j,1} = 0$ for all $j = 1, \ldots, p$. Then, we rewrite the total variation part in the binarsity penalty as follows. Let us define the $(d_j + 1) \times (d_j + 1)$ matrix D_j by

$$D_j = \begin{bmatrix} 1 & 0 & & 0 \\ -1 & 1 & & \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{d_j+1} \times \mathbb{R}^{d_j+1}.$$

We then remark that for all $\beta_{j,\bullet} \in \mathbb{R}^{d_j+1}$, one has $\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} = \|\omega_{j,\bullet} \odot D_j\beta_{j,\bullet}\|_1$. Moreover, note that the matrix D_j is invertible. We denote its inverse T_j, which is defined

by the $(d_j+1)\times(d_j+1)$ lower triangular matrix with entries $(T_j)_{r,s}=0$ if $r<s$ and $(T_j)_{r,s}=1$ otherwise. We set

$$\mathbf{D}=\mathrm{diag}(D_1,\dots,D_p) \quad \text{and} \quad \mathbf{T}=\mathrm{diag}(T_1,\dots,T_p). \tag{21}$$

Lemma 1 then states that binarsity is a sub-additive penalty [Kutateladze, 2013].

Lemma 1 *For all $\beta,\beta'\in\mathbb{R}^{p+d}$, we have that*

$$\mathrm{bina}(\beta+\beta')\le\mathrm{bina}(\beta)+\mathrm{bina}(\beta') \quad \textit{and} \quad \mathrm{bina}(-\beta)\le\mathrm{bina}(\beta).$$

Proof of Lemma 1. The hyperplane $\mathrm{span}\{u\in\mathbb{R}^{d_j+1}: n_{j,\bullet}^\top u=0\}$ is a convex cone, then the indicator function δ_j is sublinear (i.e., positively homogeneous and sub-additive [Kutateladze, 2013]). Furthermore, the total variation penalization satisfies the triangle inequality, which gives the first statement of Lemma 1. To prove the second, we use the fact that $\delta_j(\beta_{j,\bullet})+\delta_j(-\beta_{j,\bullet})\ge 0$ to obtain:

$$\mathrm{bina}(-\beta)=\sum_{j=1}^p\Big(\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}+\delta_j(-\beta_{j,\bullet})\Big)\le\sum_{j=1}^p\Big(\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}+\delta_j(\beta_{j,\bullet}))\Big)=\mathrm{bina}(\beta),$$

which concludes the proof of Lemma 1. □

Additional useful quantities. The Doob-Meyer decomposition [Aalen, 1978] implies that, for all $i=1,\dots,n$ and all $t\ge 0$,

$$\mathrm{d}N_i(t)=Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}\mathrm{d}t+\mathrm{d}M_i(t),$$

where the martingales M_i are square integrable and orthogonal. With this notation, we define, for all $t\ge 0$ and any f, the process

$$S_n^{(r)}(f,t)=\sum_{i=1}^n Y_i(t)e^{f(X_i)}(X_i^B)^{\otimes r}$$

for $r\in\{0,1,2\}$, where X_i^B is the ith row of the binarized matrix $\boldsymbol{X}^B$. The empirical loss ℓ_n can then be rewritten as

$$\ell_n(f)=-\frac{1}{n}\sum_{i=1}^n\int_0^\tau\big\{f(X_i)-\log\big(S_n^{(0)}(f,t)\big)\big\}\mathrm{d}N_i(t).$$

Together with this loss, we introduce the loss

$$\begin{aligned}\ell(f)&=-\frac{1}{n}\sum_{i=1}^n\int_0^\tau\big\{f(X_i)-\log\big(S_n^{(0)}(f,t)\big)\big\}Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}\mathrm{d}t\\&=-\frac{1}{n}\sum_{i=1}^n\int_0^\tau\log\Big(\frac{e^{f(X_i)}}{S_n^{(0)}(f,t)}\Big)Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}\mathrm{d}t.\end{aligned}$$

We will use the fact that for a function f_β of the form $f_\beta(X_i)=\beta^\top X_i^B=\sum_{j=1}^p f_{\beta_{j,\bullet}}(X_i)$, the Doob-Meyer decomposition implies that

$$\begin{aligned}\nabla\ell_n(f_\beta)&=-\frac{1}{n}\sum_{i=1}^n\int_0^\tau\Big\{X_i^B-\frac{S_n^{(1)}(f_\beta,t)}{S_n^{(0)}(f_\beta,t)}\Big\}\mathrm{d}N_i(t)\\&=\nabla\ell(f_\beta)+H_n(f_\beta),\end{aligned}\tag{22}$$

where $H_n(f_\beta)$ is an error term defined by

$$H_n(f_\beta) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{X_i^B - \frac{S_n^{(1)}(f_\beta,t)}{S_n^{(0)}(f_\beta,t)}\Big\}\mathrm{d}M_i(t). \tag{23}$$

We also introduce the empirical ℓ_2-norm defined for any function f as

$$\|f\|_n^2 = \int_0^\tau \sum_{i=1}^{n}\big(f(X_i) - \bar{f}(t)\big)^2 \frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)}\mathrm{d}\bar{N}(t), \tag{24}$$

with

$$\bar{f}(t) = \sum_{i=1}^{n}\frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)}f(X_i).$$

In the following section, we state some lemmas required for proving our theorems. Their proofs are postponed to Section B.4.

B.2 Lemmas

First, Lemma 2 is a consequence of the Karush-Kuhn-Tucker (KKT) optimality conditions [Boyd and Vandenberghe, 2004] for a convex optimization and the monotony of subdifferential mappings.

Lemma 2 *Let $\beta \in \mathscr{B}_{p+d}(R)$ such that $n_{j,\bullet}^\top \beta_{j,\bullet} = 0$, and $h = (h_{1,\bullet}^\top, \dots, h_{p,\bullet}^\top)^\top$ with $h_{j,\bullet} \in \partial\big(\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}\big)$ for all $j = 1,\dots,p$. Then the following holds:*

$$(\hat{\beta}-\beta)^\top \nabla\ell(f_{\hat{\beta}}) \le -(\hat{\beta}-\beta)^\top H_n(f_{\hat{\beta}}) - (\hat{\beta}-\beta)^\top h.$$

Next, Lemma 3 is derived from the self-concordance definition and Lemma 1 in Bach [2010]. It connects the empirical ℓ_2-norm defined in (24) to our empirical divergence defined in (10).

Lemma 3 *Let $\hat{\beta}$ be defined by Equation (5) and $\beta \in \mathscr{B}_{p+d}(R)$. Then the following inequalities hold almost surely:*

$$KL_n(f^\star, f_\beta) - KL_n(f^\star, f_{\hat{\beta}}) + (\hat{\beta}-\beta)^\top \nabla\ell(f_{\hat{\beta}}) \ge 0, \tag{25}$$

and

$$\|f^\star - f_\beta\|_n^2 \frac{\psi(-\|f^\star - f_\beta\|_\infty)}{\|f^\star - f_\beta\|_\infty^2} \le KL_n(f^\star, f_\beta) \le \|f^\star - f_\beta\|_n^2 \frac{\psi(\|f^\star - f_\beta\|_\infty)}{\|f^\star - f_\beta\|_\infty^2}, \tag{26}$$

where we recall that $\psi(x) = e^x - x - 1$.

Let us now define the non-negative definite matrix

$$\widehat{\Sigma}_n(f^\star,\tau) = \sum_{i=1}^{n}\int_0^\tau \big(X_i^B - \breve{X}_n(t)\big)^{\otimes 2}\frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)}\mathrm{d}\bar{N}(t),$$

where

$$\breve{X}_n(t) = \frac{S_n^{(1)}(f^\star,t)}{S_n^{(0)}(f^\star,t)}.$$

This matrix is linked to our empirical norm via the relation $\|f_\beta\|_n^2 = \beta^\top \widehat{\Sigma}_n(f^\star,\tau)\beta$. The proof of Theorem 1 requires the matrix $\widehat{\Sigma}_n(f^\star,\tau)$ to fulfill a compatibility condition. The following lemma shows that such a condition is true with large probability as long as Assumption 2 holds.

Lemma 4 *Let $\zeta \in \mathbb{R}_+^{p+d}$ be a given vector of non-negative weights and $L = [L_1, \dots, L_p]$ a concatenation of index subsets. Set for all $j = 1, \dots, p$,*

$$L_j = \{a_j^1, \dots, a_j^{b_j}\} \subset \{1, \dots, d_j + 1\}, \tag{27}$$

with the convention that $a_j^0 = 0$ and $a_j^{b_j+1} = d_j + 2$. Then, with a probability greater than $1 - e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}} - 3\varepsilon$, one has

$$\inf_{u \in \mathscr{C}_{1,\omega}(L)\setminus\{\mathbf{0}\}} \frac{(\mathbf{T}u)^\top \widehat{\Sigma}_n(f^\star, \tau)\mathbf{T}u}{|\|u_L \odot \zeta_L\|_1 - \|u_{L^\complement} \odot \zeta_{L^\complement}\|_1|^2} \geq \big(\kappa_\tau^2(L) - \Xi_\tau(L)\big)\kappa_{\mathbf{T},\zeta}^2(L),$$

where

$$\Xi_\tau(L) = 4|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\Big\{\big(1 + e^{2f_\infty^\star}\Lambda_0^\star(\tau)\big)\sqrt{2/n\log(2(p+d)^2/\varepsilon)} + (2e^{2f_\infty^\star}\Lambda_0^\star(\tau)/s^{(0)}(\tau))t_{n,p,d,\varepsilon}^2\Big\},$$

$$\kappa_{\mathbf{T},\zeta}(L) = \Big(32\sum_{j=1}^p\sum_{l=1}^{d_j+1}|\zeta_{j,l+1} - \zeta_{j,l}|^2 + (b_j+1)\|\zeta_{j,\bullet}\|_\infty^2\big\{\min_{1\leq b\leq b^j}|a_j^b - a_j^{b-1}|\big\}^{-1}\Big)^{-\frac{1}{2}},$$

and

$$\mathscr{C}_{1,\omega}(L) = \Big\{u \in \mathscr{B}_{p+d}(R) : \sum_{j=1}^p \|(u_{j,\bullet})_{L_j^\complement}\|_{1,\omega_{j,\bullet}} \leq 3\sum_{j=1}^p \|(u_{j,\bullet})_{L_j}\|_{1,\omega_{j,\bullet}}\Big\}.$$

We now state a technical result connecting the norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on $\mathscr{C}_{\mathrm{TV},\omega}(L)$.

Lemma 5 *Let Σ and $\tilde{\Sigma}$ be two non-negative matrices of the same size. For any concatenation $L = [L_1, \dots, L_p]$ of index subsets, one has*

$$\inf_{\beta\in\mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}}\frac{\beta^\top\tilde{\Sigma}\beta}{\|\beta_L\|_2^2} \geq \inf_{\beta\in\mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}}\frac{\beta^\top\Sigma\beta}{\|\beta_L\|_2^2} - |L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\max_{j,l}|\Sigma_{j,l} - \tilde{\Sigma}_{j,l}|.$$

B.3 Proof of Theorem 1

Combining Lemmas 2 and 3, we get

$$\begin{aligned} KL_n(f^\star, f_{\hat{\beta}}) &\leq KL_n(f^\star, f_\beta) + (\hat{\beta} - \beta)^\top\nabla\ell(f_{\hat{\beta}}) \\ &\leq KL_n(f^\star, f_\beta) - (\hat{\beta} - \beta)^\top H_n(f_{\hat{\beta}}) - (\hat{\beta} - \beta)^\top h. \end{aligned}$$

Then, if $-(\hat{\beta} - \beta)^\top H_n(f_{\hat{\beta}}) - (\hat{\beta} - \beta)^\top h < 0$, the theorem holds. Let us assume for now that $-(\hat{\beta} - \beta)^\top H_n(f_{\hat{\beta}}) - (\hat{\beta} - \beta)^\top h \geq 0$.

Bound for $-(\hat{\beta} - \beta)^\top H_n(f_{\hat{\beta}}) - (\hat{\beta} - \beta)^\top h$. From the definition of the sub-gradient $\hat{h} = (\hat{h}_{1,\bullet}^\top, \dots, \hat{h}_{p,\bullet}^\top)^\top \in \partial\big(\|\hat{\beta}\|_{\mathrm{TV},\omega}\big)$, one can choose h such that

$$h_{j,l} = \begin{cases} 2D_j^\top\big(\omega_{j,\bullet} \odot \mathrm{sign}(D_j\beta_{j,\bullet})\big) & \text{if } l \in \mathcal{A}_j(\beta), \\ 2D_j^\top\big(\omega_{j,\bullet} \odot \mathrm{sign}\big(D_j(\hat{\beta}_{j,\bullet} - \beta_{j,\bullet})\big)\big) & \text{if } l \in \mathcal{A}_j^\complement(\beta). \end{cases}$$

This gives

$$
\begin{aligned}
-(\hat{\beta}-\beta)^\top h &= -\sum_{j=1}^{p}(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})^\top h_{j,\bullet} \\
&= \sum_{j=1}^{p}\big((-h_{j,\bullet})_{\mathcal{A}_j(\beta)}\big)^\top(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)} - \sum_{j=1}^{p}\big((h_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\big)^\top(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)} \\
&= 2\sum_{j=1}^{p}\big((-\omega_{j,\bullet}\odot \operatorname{sign}(D_j\beta_{j,\bullet}))_{\mathcal{A}_j(\beta)}\big)^\top D_j(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)} \\
&\quad - 2\sum_{j=1}^{p}\big((\omega_{j,\bullet}\odot \operatorname{sign}\big(D_j(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})\big)_{\mathcal{A}_j^{\complement}(\beta)}\big)^\top D_j(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}.
\end{aligned}
$$

Using the fact that $u^\top \operatorname{sign}(u) = \|u\|_1$, we have that

$$
\begin{aligned}
-(\hat{\beta}-\beta)^\top h &\leq 2\sum_{j=1}^{p}\|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\odot D_j(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_1 \\
&\quad - 2\sum_{j=1}^{p}\|(\omega_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\odot D_j(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_1 \\
&= 2\sum_{j=1}^{p}\|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} - 2\sum_{j=1}^{p}\|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}. \qquad (28)
\end{aligned}
$$

Inequality (28) therefore gives

$$
\begin{aligned}
KL_n(f^\star, f_{\hat{\beta}}) &\leq KL_n(f^\star, f_\beta) - (\hat{\beta}-\beta)^\top H_n(f_{\hat{\beta}}) + 2\sum_{j=1}^{p}\|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} \\
&\quad - 2\sum_{j=1}^{p}\|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}.
\end{aligned}
$$

Using the fact that $\mathbf{TD} = \mathbf{I}$ (see their definitions in Equation (21)), we get

$$
\begin{aligned}
KL_n(f^\star, f_{\hat{\beta}}) &\leq KL_n(f^\star, f_\beta) - \big(\mathbf{D}(\hat{\beta}-\beta)\big)^\top \mathbf{T}^\top H_n(f_{\hat{\beta}}) \\
&\quad + 2\sum_{j=1}^{p}\|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} - 2\sum_{j=1}^{p}\|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}.
\end{aligned}
$$

On the event

$$
\mathscr{E}_n := \Big\{|\mathbf{T}^\top H_n(f_{\hat{\beta}})| \leq (\omega_{1,1},\ldots,\omega_{p,d_p+1})\Big\} \qquad (29)
$$

(the vector comparison has to be understood elementwise), we have

$$
\begin{aligned}
KL_n(f^\star, f_{\hat{\beta}}) &\leq KL_n(f^\star, f_\beta) + \sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\omega_{j,l}\big|\big(\mathbf{D}(\hat{\beta}-\beta)\big)_{j,l}\big| \\
&\quad + 2\sum_{j=1}^{p}\|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} - 2\sum_{j=1}^{p}\|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}.
\end{aligned}
$$

Hence,

$$KL_n(f^\star, f_{\hat\beta}) \le KL_n(f^\star, f_\beta) + \sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} + \sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}$$
$$+ 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} - 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}$$
$$\le KL_n(f^\star, f_\beta) + 3\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} - \sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}.$$

One therefore has

$$KL_n(f^\star, f_{\hat\beta}) \le KL_n(f^\star, f_\beta) + 3\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}. \tag{30}$$

On the event $\mathscr{E}_n$, the following also holds

$$\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} \le 3\sum_{j=1}^p \|(\hat\beta_{j,\bullet} - \beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}},$$

which means that $\hat\beta - \beta \in \mathscr{C}_{\mathrm{TV},\omega}\big(\mathcal{A}(\beta)\big)$ and $\mathbf{D}(\hat\beta - \beta) \in \mathscr{C}_{1,\omega}\big(\mathcal{A}(\beta)\big)$. Now returning to (30), by Lemma 4 and under Assumption 2, we get

$$KL_n(f^\star, f_{\hat\beta}) \le KL_n(f^\star, f_\beta) + \frac{\|f_{\hat\beta} - f_\beta\|_n}{\sqrt{\kappa_\tau^2\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)}\,\kappa_{\mathbf{T},\hat\zeta}\big(\mathcal{A}(\beta)\big)}, \tag{31}$$

where

$$\hat\zeta_{j,l} = \begin{cases} 3\omega_{j,l} & \text{if } l \in \mathcal{A}(\beta), \\ 0 & \text{if } l \in \mathcal{A}^{\complement}(\beta). \end{cases}$$

The second term in the right-hand side of (31) fulfills

$$\frac{\|f_{\hat\beta} - f_\beta\|_n}{\sqrt{\kappa_\tau^2\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)}\,\kappa_{\mathbf{T},\hat\zeta}\big(\mathcal{A}(\beta)\big)} \le \frac{\|f^\star - f_{\hat\beta}\|_n + \|f^\star - f_\beta\|_n}{\sqrt{\kappa_\tau^2\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)}\,\kappa_{\mathbf{T},\hat\zeta}\big(\mathcal{A}(\beta)\big)}.$$

By (26) in Lemma 3, we get that

$$\|f^\star - f_\beta\|_n \le \sqrt{\frac{\|f^\star - f_\beta\|_\infty^2}{\psi(-\|f^\star - f_\beta\|_\infty)} KL_n(f^\star, f_\beta)}.$$

Introducing $g(x) = x^2/\psi(-x) = x^2/(e^{-x} + x + 1)$, we note that

$$g(x) \le x + 2 \text{ for any } x > 0. \tag{32}$$

Then

$$\|f^\star - f_\beta\|_n \le \sqrt{(\|f^\star - f_\beta\|_\infty + 2) KL_n(f^\star, f_\beta)}.$$

In addition, one can easily check that $\max_{1\le i\le n} \sup_{\beta\in\mathscr{B}_{p+d}(R)} |f_\beta(X_i)| \le R$. Hence,

$$\|f^\star - f_\beta\|_\infty \le \max_{1\le i\le n} \big\{|f^\star(X_i)| + |f_\beta(X_i)|\big\} \le f_\infty^\star + R.$$

This implies that

$$\|f^\star - f_\beta\|_n \leq \sqrt{(f^\star_\infty + R + 2) KL_n(f^\star, f_\beta)}.$$

With these bounds, inequality (31) yields

$$KL_n(f^\star, f_{\hat\beta}) \leq KL_n(f^\star, f_\beta) + \sqrt{(f^\star_\infty + R + 2)} \frac{\sqrt{KL_n(f^\star, f_\beta)} + \sqrt{KL_n(f^\star, f_{\hat\beta})}}{\sqrt{\kappa^2_\tau\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)}\kappa_{\mathbf{T},\hat\zeta}\big(\mathcal{A}(\beta)\big)}.$$

We now use the elementary inequality $2uv \leq \varrho u^2 + v^2/\varrho$ with $\varrho > 0$. We get

$$KL_n(f^\star, f_{\hat\beta}) \leq KL_n(f^\star, f_\beta) + \frac{\varrho(f^\star_\infty + R + 2)}{2\Big(\kappa^2_\tau\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)\Big)\kappa^2_{\mathbf{T},\hat\zeta}\big(\mathcal{A}(\beta)\big)} + \frac{1}{2\varrho}\Big(\sqrt{KL_n(f^\star, f_\beta)} + \sqrt{KL_n(f^\star, f_{\hat\beta})}\Big)^2.$$

Hence

$$\big(1 - \frac{1}{\varrho}\big) KL_n(f^\star, f_{\hat\beta}) \leq \big(1 + \frac{1}{\varrho}\big) KL_n(f^\star, f_\beta) + \frac{\varrho(f^\star_\infty + R + 2)}{2\Big(\kappa^2_\tau\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)\Big)\kappa^2_{\mathbf{T},\hat\zeta}\big(\mathcal{A}(\beta)\big)}.$$

By choosing $\varrho = 2$, we obtain

$$KL_n(f^\star, f_{\hat\beta}) \leq 3KL_n(f^\star, f_\beta) + \frac{2(f^\star_\infty + R + 2)}{\Big(\kappa^2_\tau\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)\Big)\kappa^2_{\mathbf{T},\hat\zeta}\big(\mathcal{A}(\beta)\big)}.$$

On the other hand, by definition of $\kappa^2_{\mathbf{T},\zeta}$ (see Lemma 4), we know that

$$\frac{1}{\kappa^2_{\mathbf{T},\hat\zeta}\big(\mathcal{A}(\beta)\big)} \leq 512|\mathcal{A}(\beta)| \max_{1\leq j\leq p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|^2_\infty.$$

Finally,

$$KL_n(f^\star, f_{\hat\beta}) \leq 3KL_n(f^\star, f_\beta) + \frac{1024(f^\star_\infty + R + 2)|\mathcal{A}(\beta)| \max_{1\leq j\leq p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|^2_\infty}{\kappa^2_\tau\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)}.$$

Therefore, on the event $\mathscr{E}_n$, we obtain the desired result.

Computation of $\mathbb{P}[\mathscr{E}_n^\complement]$. From the definition of H_n in Equation (23), $\mathbf{T}^\top H_n(f_{\hat\beta})$ is written:

$$\mathbf{T}^\top H_n(f_{\hat\beta}) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big\{\mathbf{T}^\top X_i^B - \mathbf{T}^\top \frac{S_n^{(1)}(f_{\hat\beta}, t)}{S_n^{(0)}(f_{\hat\beta}, t)}\Big\} \mathrm{d}M_i(t).$$

Hence, each component of this vector has the form required to apply Theorem 3 from Gaïffas and Guilloux [2012]. We recall that H_n and $\mathbf{T}^\top H_n$ have a block structure: they are vectors of p blocks of length $d_j + 1$ for all $j = 1, \ldots, p$. We then denote by $\big(\mathbf{T}^\top H_n\big)_{j,l}$ the lth component of the jth block.

In addition, due to the definition of X_i^B, we know that each coefficient of $\mathbf{T}^\top X_i^B$ takes a value lower than 1. As a consequence, for all $t \leq \tau$, one has

$$\left|(\mathbf{T}^\top X_i^B - \mathbf{T}^\top \frac{S_n^{(1)}(f_{\hat\beta},t)}{S_n^{(0)}(f_{\hat\beta},t)})_{j,k}\right| \leq \left|(\mathbf{T}^\top X_i^B)_{j,k}\right| + \left|(\mathbf{T}^\top \frac{S_n^{(1)}(f_{\hat\beta},t)}{S_n^{(0)}(f_{\hat\beta},t)})_{j,k}\right| \leq 2.$$

We now use Theorem 3 from Gaïffas and Guilloux [2012] to obtain

$$\mathbb{P}\Big[\big|\big(\mathbf{T}^\top H_n(f_{\hat\beta},t)\big)_{j,l}\big| \geq 5.64\sqrt{\frac{c+\mathscr{L}_{n,c}}{n}} + 18.62\frac{(c+1+\mathscr{L}_{n,c})}{n}\Big] \leq 28.55e^{-c},$$

and by choosing the weights $\omega_{j,l}$ as defined in (20), we conclude that $\mathbb{P}[\mathscr{E}_n^\complement] \leq 28.55e^{-c}$ for some $c > 0$.

□

B.4 Proofs of the lemmas

B.4.1 Proof of Lemma 2

To characterize the solution of Problem (5), the following result can be sraightforwardly obtained using the Karush-Kuhn-Tucker (KKT) optimality conditions [Boyd and Vandenberghe, 2004] for a convex optimization problem. A vector $\hat\beta \in \mathbb{R}^{p+d}$ is an optimum of the objective function in (5) if and only if there exists the following three sequences of subgradient:

$$\begin{cases} \hat h = (\hat h_{j,\bullet})_{j=1,\dots,p} \text{ with } \hat h_{j,\bullet} \in \partial\big(\|\hat\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}\big), \\ \hat g = (\hat g_{j,\bullet})_{j=1,\dots,p} \text{ with } \hat g_{j,\bullet} \in \partial\big(\delta_j(\hat\beta_{j,\bullet})\big), \\ \hat k \in \partial\big(\delta_{\mathscr{B}_{p+d}(R)}(\hat\beta)\big) \end{cases}$$

such that

$$(\nabla \ell_n(f_{\hat\beta}))_{j,\bullet} + \hat h_{j,\bullet} + \hat g_{j,\bullet} + \hat k_{j,\bullet} = \mathbf{0}, \tag{33}$$

for all $j = 1,\dots,p$, and where

$$\hat h_{j,l} \begin{cases} = \Big(D_j^\top\big(\omega_{j,\bullet} \odot \operatorname{sign}(D_j\hat\beta_{j,\bullet})\big)\Big)_l & \text{if } l \in \mathcal{A}_j(\hat\beta), \\ \in \Big(D_j^\top\big(\omega_{j,\bullet} \odot [-1,+1]^{d_j+1}\big)\Big)_l & \text{if } l \in \mathcal{A}_j^\complement(\hat\beta), \end{cases}$$

where $\mathcal{A}(\hat\beta)$ is the active set of $\hat\beta$, see (8). The subgradient $\hat g_{j,\bullet}$ belongs to

$$\partial\big(\delta_j(\hat\beta_{j,\bullet})\big) = \big\{v \in \mathbb{R}^{d_j+1} : (\hat\beta_{j,\bullet} - \beta_{j,\bullet})^\top v \geq 0 \text{ for all } \beta_{j,\bullet} \text{ such that } n_{j,\bullet}^\top \beta_{j,\bullet} = 0\big\},$$

and $\hat k$ to

$$\partial\big(\delta_{\mathscr{B}_{p+d}(R)}(\hat\beta)\big) = \Big\{v \in \mathbb{R}^{p+d} : (\hat\beta - \beta)^\top v \geq 0 \text{ for all } \beta \text{ such that } \sum_{j=1}^p \|\beta_{j,\bullet}\|_\infty \leq R\Big\}.$$

From Equation (33), and considering any vector $\beta \in \mathbb{R}^{p+d}$, we obtain

$$(\hat\beta - \beta)^\top \nabla \ell_n(f_{\hat\beta}) + (\hat\beta - \beta)^\top(\hat h + \hat g + \hat k) = 0, \tag{34}$$

and Equation (22) gives

$$(\hat\beta - \beta)^\top \nabla \ell(f_{\hat\beta}) + (\hat\beta - \beta)^\top H_n(f_{\hat\beta}) + (\hat\beta - \beta)^\top(\hat h + \hat g + \hat k) = 0.$$

Consider now a vector $\beta \in \mathscr{B}_{p+d}(R)$ such that $n_{j,\bullet}^\top \beta_{j,\bullet} = 0$ for all $j = 1,\dots,p$, and $h \in \partial\big(\|\beta\|_{\mathrm{TV},\omega}\big)$. Then, the monotony of sub-differential mappings (which is an immediate consequence of their definition, see Rockafellar [1970]) gives the result.

□

B.4.2 Proof of Lemma 3

Let us consider the function $G : \mathbb{R} \to \mathbb{R}$ defined by $G(\eta) = \ell(f_1 + \eta f_2)$, i.e.,

$$G(\eta) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (f_1+\eta f_2)(X_i)Y_i(t)e^{f^\star(X_i)}\lambda_0^\star(t)\mathrm{d}t + \frac{1}{n}\int_0^\tau \log\left\{S_n^{(0)}(f_1+\eta f_2,t)\right\}S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t.$$

By differentiating G with respect to the variable η, we get

$$G'(\eta) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau f_2(X_i)Y_i(t)e^{f^\star(X_i)}\lambda_0^\star(t)\mathrm{d}t + \frac{1}{n}\int_0^\tau \frac{\sum_{i=1}^n f_2(X_i)Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)}{\sum_{i=1}^n Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)}S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t,$$

and

$$G''(\eta) = \frac{1}{n}\int_0^\tau \frac{\sum_{i=1}^n f_2^2(X_i)Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)}{\sum_{i=1}^n Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)}S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t - \int_0^\tau \left(\frac{\sum_{i=1}^n f_2(X_i)Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)}{\sum_{i=1}^n Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)}\right)^2 S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t.$$

For a given $t \geq 0$, we now consider the discrete random variable U_t that takes the value $f_2(X_i)$ with probability

$$\mathbb{P}[U_t = f_2(X_i)] = \pi_{t,f_1,f_2,\eta}(i) = \frac{Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)}{\sum_{i=1}^n Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)}.$$

We observe that for all $k \in \mathbb{N}$, one has

$$\frac{\sum_{i=1}^n f_2^k(X_i)Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)}{\sum_{i=1}^n Y_i(t)\exp\left(f_1(X_i)+\eta f_2(X_i)\right)} = \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t^k].$$

Then

$$G'(\eta) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau f_2(X_i)Y_i(t)e^{f^\star(X_i)}\lambda_0^\star(t)\mathrm{d}t + \frac{1}{n}\int_0^\tau \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t]S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t,$$

and

$$G''(\eta) = \frac{1}{n}\int_0^\tau \left(\mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t^2] - \left(\mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t]\right)^2\right)S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t = \frac{1}{n}\int_0^\tau \mathbb{V}_{\pi_{t,f_1,f_2,\eta}}[U_t]S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t.$$

Differentiating again, we obtain

$$G'''(\eta) = \frac{1}{n}\int_0^\tau \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}\left[\left(U_t - \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t]\right)^3\right]S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t.$$

Therefore, we have

$$\begin{aligned} G^{'''}(\eta) &\leq \frac{1}{n}\int_0^\tau \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}\Big[\big|U_t - \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t]\big|^3\Big] S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t \\ &\leq \frac{1}{n}2\|f_2\|_\infty \int_0^\tau \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}\Big[\big(U_t - \mathbb{E}_{\pi_{t,f_1,f_2,\eta}}[U_t]\big)^2\Big] S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t \\ &\leq 2\|f_2\|_\infty G^{''}(\eta), \end{aligned}$$

where $\|f_2\|_\infty := \max_{1\leq i\leq n}|f_2(X_i)|$. Applying now Lemma 1 in Bach [2010] to G, we obtain for all $\eta \geq 0$,

$$G^{''}(0)\frac{\psi(-\|f_2\|_\infty)}{\|f_2\|_\infty^2} \leq G(\eta) - G(0) - \eta G^{'}(0) \leq G^{''}(0)\frac{\psi(\|f_2\|_\infty)}{\|f_2\|_\infty^2}. \tag{35}$$

We will apply inequalities in (35) in the following two situations:

- Case #1: $\eta = 1$, $f_1 = f_{\hat\beta}$ and $f_2 = f_\beta - f_{\hat\beta}$.
- Case #2: $\eta = 1$, $f_1 = f^\star$ and $f_2 = f_\beta - f^\star$.

In case #1,

$$\begin{aligned} G'(0) &= -(\beta - \hat\beta)^\top \frac{1}{n}\sum_{i=1}^n \Big\{ \int_0^\tau X_i^B Y_i(t) e^{f^\star(X_i)}\lambda_0^\star(t)\mathrm{d}t \\ &\qquad\qquad - \int_0^\tau X_i^B Y_i(t) e^{f_{\hat\beta}(X_i)} \frac{S_n^{(0)}(f^\star,t)}{S_n^{(0)}(f_{\hat\beta},t)}\lambda_0^\star(t)\mathrm{d}t \Big\} \\ &= (\beta - \hat\beta)^\top \nabla \ell(f_{\hat\beta}), \end{aligned}$$

and then

$$G(1) - G(0) - G^{'}(0) = \ell(f_\beta) - \ell(f_{\hat\beta}) + (\hat\beta - \beta)^\top \nabla\ell(f_{\hat\beta}).$$

With the left bound of the self-concordance inequality (35), we obtain (25) in Lemma 3.

In case # 2, one gets

$$\begin{aligned} &G^{'}(0) = 0, \\ \text{and}\quad &G^{''}(0) = \frac{1}{n}\int_0^\tau \frac{\sum_{i=1}^n \big(f_\beta(X_i) - f^\star(X_i)\big)^2 Y_i(t) e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i(t) e^{f^\star(X_i)}} S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t \\ &\qquad - \frac{1}{n}\int_0^\tau \left(\frac{\sum_{i=1}^n (f_\beta(X_i) - f^\star(X_i)) Y_i(t) e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i(t) e^{f^\star(X_i)}}\right)^2 S_n^{(0)}(f^\star,t)\lambda_0^\star(t)\mathrm{d}t \\ &\qquad = \|f^\star - f_\beta\|_n^2, \end{aligned}$$

which gives (26) in Lemma 3.

□

B.4.3 Proof of Lemma 4

For any concatenation of index sets $L = [L_1, \ldots, L_p]$, we define

$$\hat\kappa_\tau(L) = \inf_{\beta \in \mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}} \frac{\sqrt{\beta^\top \hat\Sigma_n(f^\star,\tau)\beta}}{\|\beta_L\|_2}.$$

To prove Lemma 4, we will first establish the following lemma, which asssures us that if Assumption 2 is fulfilled, our random bound $\hat\kappa_\tau(L)$ is bounded away from 0 with large probability.

Lemma 6 *Let $L = [L_1, \ldots, L_p]$ be a concatenation of index sets. Then,*

$$\hat{\kappa}_\tau^2(L) \geq \kappa_\tau^2(L) - 4|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2 \times \Big\{\big(1+e^{2f_\infty^\star}\Lambda_0^\star(\tau)\big)\sqrt{2/n\log(2(p+d)^2/\varepsilon)} + (2e^{2f_\infty^\star}\Lambda_0^\star(\tau)/s^{(0)}(\tau))t_{n,p,d,\varepsilon}^2\Big\}$$

holds with probability at least $1 - e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}} - 3\varepsilon$.

Proof of Lemma 6. The proof is adapted from Theorem 4.1 in Huang et al. [2013], with the difference that we work here in a fixed design setting. We break down the proof into three steps.

Step 1. By replacing $\mathrm{d}\bar{N}(t)$ by its compensator $n^{-1}S_n^0(f^\star,t)\lambda_0^\star(t)\mathrm{d}t$, an approximation of $\widehat{\Sigma}_n(f^\star,\tau)$ can be defined by

$$\bar{\Sigma}_n(f^\star,\tau) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \big(X_i^B - \breve{X}_n(s)\big)^{\otimes 2} Y_i(s)e^{f^\star(X_i)}\lambda_0^\star(s)\mathrm{d}s.$$

The (m,m')th component of

$$\sum_{i=1}^n \big(X_i^B - \breve{X}_n(s)\big)^{\otimes 2}\frac{Y_i(s)e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i(s)e^{f^\star(X_i)}}$$

is given by

$$\sum_{i=1}^n \big[(X_i^B)_m - \big(\breve{X}_n(s)\big)_m\big]\big[(X_i^B)_{m'} - \big(\breve{X}_n(s)\big)_{m'}\big]\frac{Y_i(s)e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i(s)e^{f^\star(X_i)}},$$

which is bounded by 4 in our case. Moreover, we know that

$$\int_0^\tau Y_i(t)\mathrm{d}N_i(t) \leq 1 \quad \text{for all} \quad i = 1,\ldots,n.$$

Thus, Lemma 3.3 in Huang et al. [2013] applies and

$$\mathbb{P}\big[\big(\widehat{\Sigma}_n(f^\star,\tau) - \bar{\Sigma}_n(f^\star,\tau)\big)_{m,m'} > 4x\big] \leq 2e^{-nx^2/2}.$$

Next, using an union bound, we get

$$\mathbb{P}\big[\max_{m,m'}\big(\widehat{\Sigma}_n(f^\star,\tau) - \bar{\Sigma}_n(f^\star,\tau)\big)_{m,m'} > 4\sqrt{2/n\log\big(2(p+d)^2/\varepsilon\big)}\big] \leq \varepsilon.$$

Let

$$\bar{\kappa}_\tau^2(L) = \inf_{\beta\in\mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}}\frac{\sqrt{\beta^\top\bar{\Sigma}_n(f^\star,\tau)\beta}}{\|\beta_L\|_2}.$$

Lemma 5 implies that

$$\mathbb{P}\Big[\hat{\kappa}_\tau^2(L) \geq \bar{\kappa}_\tau^2(L) - 4|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\sqrt{2/n\log\big(2(p+d)^2/\varepsilon\big)}\Big] \geq 1-\varepsilon. \tag{36}$$

Step 2. Let

$$\widetilde{\Sigma}_n(f^\star,\tau) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \big(X_i^B - \bar{X}_n(s)\big)^{\otimes 2} Y_i(s)\; e^{f^\star(X_i)}\lambda_0^\star(s)\mathrm{d}s$$

and

$$\tilde{\kappa}_\tau(L) = \inf_{\beta \in \mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}} \frac{\sqrt{\beta^\top \widetilde{\Sigma}_n(f^\star,\tau)\beta}}{\|\beta_L\|_2}.$$

We will now compare $\bar{\kappa}_\tau^2(L)$ and $\tilde{\kappa}_\tau^2(L)$. Straightforward computations lead to the following equality:

$$\begin{aligned}&\sum_{i=1}^n \big(X_i^B - \bar{X}_n(s)\big)^{\otimes 2} Y_i(s) e^{f^\star(X_i)} - \sum_{i=1}^n \big(X_i^B - \breve{X}_n(s)\big)^{\otimes 2} Y_i(s) e^{f^\star(X_i)} \\ &\quad = S_n^{(0)}(f^\star, s)\big(\breve{X}_n(s) - \bar{X}_n(s)\big)^{\otimes 2}.\end{aligned}$$

Hence,

$$\bar{\Sigma}_n(f^\star,\tau) = \widetilde{\Sigma}_n(f^\star,\tau) - \frac{1}{n}\int_0^\tau S_n^{(0)}(f^\star,s)\big(\breve{X}_n(s) - \bar{X}_n(s)\big)^{\otimes 2}\lambda_0^\star(s)\mathrm{d}s. \tag{37}$$

We first bound the second term on the right-hand side of (37). Let

$$\Delta_n(s) = \frac{1}{n} S_n^{(0)}(f^\star,s)\big(\breve{X}_n(s) - \bar{X}_n(s)\big),$$

so that for each (m,m'), we get

$$\Big(\frac{1}{n}\int_0^\tau S_n^{(0)}(f^\star,s)\big(\breve{X}_n(s) - \bar{X}_n(s)\big)^{\otimes 2}\lambda_0^\star(s)\mathrm{d}s\Big)_{m,m'} \le \Big(\frac{\int_0^\tau \Delta_n(s)^{\otimes 2}\lambda_0^\star(s)\mathrm{d}s}{n^{-1}S_n^{(0)}(f^\star,\tau)}\Big)_{m,m'}.$$

In our setting, for each i and all $t \le \tau$, $Y_i(t)e^{f^\star(X_i)} \le e^{f_\infty^\star}$. By Hoeffding's inequality, we then obtain

$$\mathbb{P}[\frac{1}{n}S_n^{(0)}(f^\star,\tau) < s^{(0)}(\tau)/2] \le e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}}.$$

Furthermore, we have

$$\mathbb{E}[\Delta_n(s)|X] = \frac{1}{n}\sum_{i=1}^n y_i(s)e^{f^\star(X_i)}\Big(X_i^B - \frac{\sum_{i=1}^n X_i^B y_i(s)e^{f^\star(X_i)}}{\sum_{i=1}^n y_i(s)e^{f^\star(X_i)}}\Big) = \mathbf{0},$$

and the (m,m')th component of $\Delta_n(s)^{\otimes 2}$ is given by

$$\begin{aligned}\big(\Delta_n(s)^{\otimes 2}\big)_{m,m'} = \frac{1}{n^2}\sum_{i=1}^n\sum_{i'=1}^n &Y_i(s)Y_{i'}(s)e^{f^\star(X_i)}e^{f^\star(X_{i'})} \\ &\times \big[(X_i^B)_m - \big(\bar{X}_n(s)\big)_m\big]\big[(X_{i'}^B)_{m'} - \big(\bar{X}_n(s)\big)_{m'}\big].\end{aligned}$$

Therefore, $\int_0^\tau \big(\Delta_n(s)^{\otimes 2}\big)_{m,m'}\lambda_0^\star(s)\mathrm{d}s$ is a V-statistic for all (m,m'). Moreover,

$$\int_0^\tau \big|\big(\Delta_n(s)^{\otimes 2}\big)_{m,m'}\big|\lambda_0^\star(s)\mathrm{d}s \le 4e^{2f_\infty^\star}\Lambda_0^\star(\tau),$$

where $\Lambda_0^\star(\tau) = \int_0^\tau \lambda_0^\star(s)\mathrm{d}s$. By Lemma 4.2 in Huang et al. [2013], we obtain that

$$\mathbb{P}\Big[\max_{1\le m,m'\le p+d} \pm\int_0^\tau \big|\big(\Delta_n(s)^{\otimes 2}\big)_{m,m'}\big|\lambda_0^\star(s)\mathrm{d}s > 4e^{2f_\infty^\star}\Lambda_0^\star(\tau)x^2\Big] \le 2.221(p+d)^2\exp\Big(\frac{-nx^2/2}{1+x/3}\Big).$$

Thanks to (37), Lemma 5, and the above two probability bounds, we obtain

$$\bar{\kappa}_\tau^2(L) \geq \tilde{\kappa}_\tau^2(L) - 8e^{2f_\infty^\star}\Lambda_0^\star(\tau)|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\frac{t_{n,p,d,\varepsilon}^2}{s^{(0)}(\tau)} \tag{38}$$

holds with probability $1 - e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}} - \varepsilon$.

Step 3. Next, $\widetilde{\Sigma}_n(f^\star,\tau)$ is an average of independent matrices with mean $\Sigma_n(f^\star,\tau)$ and $\big(\widetilde{\Sigma}_n(f^\star,\tau)\big)_{m,m'}$ which are uniformly bounded by $4e^{2f_\infty^\star}\Lambda_0^\star(\tau)$, so Hoeffding's inequality ensures that

$$\mathbb{P}\big[\max_{m,m'}\big|\big(\widetilde{\Sigma}_n(f^\star,\tau)\big)_{m,m'} - \big(\Sigma_n(f^\star,\tau)\big)_{m,m'}\big| > 4e^{2f_\infty^\star}\Lambda_0^\star(\tau)x\big] \leq (p+d)^2e^{-nx^2/2}.$$

Again, Lemma 5 implies that with probability larger than $1-\varepsilon$, one has

$$\tilde{\kappa}_\tau^2(L) \geq \kappa_\tau^2(L) - 4e^{2f_\infty^\star}\Lambda_0^\star(\tau)|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\sqrt{2/n\log\big(2(p+d)^2/\varepsilon\big)}. \tag{39}$$

Finally, the result follows from (36), (38) and (39).

□

Going back to the proof of Lemma 4, following Lemma 5 in Alaya et al. [2017], for any u in

$$\mathscr{C}_{1,\omega}(K) = \Big\{u\in\mathbb{R}^d : \sum_{j=1}^p \|(u_{j,\bullet})_{K_j^\complement}\|_{1,\omega_{j,\bullet}} \leq 3\sum_{j=1}^p\|(u_{j,\bullet})_{K_j}\|_{1,\omega_{j,\bullet}}\Big\}, \tag{40}$$

the following holds:

$$\frac{(\mathbf{T}u)^\top\widehat{\Sigma}_n(f^\star,\tau)\mathbf{T}u}{\|\|u_L\odot\zeta_L\|_1 - \|u_{L^\complement}\odot\zeta_{L^\complement}\|_1|^2} \geq \kappa_{\mathbf{T},\zeta}^2(L)\frac{(\mathbf{T}u)^\top\widehat{\Sigma}_n(f^\star,\tau)\mathbf{T}u}{(\mathbf{T}u)^\top\mathbf{T}u}.$$

Then, note that if $u\in\mathscr{C}_{1,\omega}(K)$, $\mathbf{T}u\in\mathscr{C}_{\mathrm{TV},\omega}(K)$. Hence, by the definition of $\hat{\kappa}_\tau(L)$ and Lemma 6, we obtain the desired result.

□

B.4.4 Proof of Lemma 5

First, we have that

$$|\beta^\top\tilde{\Sigma}\beta - \beta^\top\Sigma\beta| \leq \|\beta\|_1^2\max_{j,l}|\tilde{\Sigma}_{j,l} - \Sigma_{j,l}|.$$

Hence, we get

$$\beta^\top\tilde{\Sigma}\beta \geq \beta^\top\Sigma\beta - \|\beta\|_1^2\max_{j,l}|\tilde{\Sigma}_{j,l} - \Sigma_{j,l}|.$$

Thus, to obtain the desired result, it is sufficient to control $\|\beta\|_1$ using the cone $\mathscr{C}_{\mathrm{TV},\omega}$. Recall that for all $j = 1, \dots, p$, we have $T_j D_j = I$. Then, for any β we have that

$$\begin{aligned}
\|\beta\|_1 &= \sum_{j=1}^{p} \|T_j D_j \beta_{j,\bullet}\| \\
&= \sum_{j=1}^{p} \sum_{l=1}^{d_j+1} \Big| \sum_{r=1}^{l} (D_j \beta_{j,\bullet})_r \Big| \\
&\leq \sum_{j=1}^{p} (d_j + 1) \sum_{l=1}^{d_j+1} \big| (D_j \beta_{j,\bullet})_l \big| \\
&\leq \frac{\max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \sum_{j=1}^{p} \sum_{l=1}^{d_j+1} \omega_{j,l} \big| (D_j \beta_{j,\bullet})_l \big| \\
&\leq \frac{\max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \sum_{j=1}^{p} \|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}.
\end{aligned}$$

For any concatenation of index subsets $L = [L_1, \dots, L_p] \subset \{1, \dots, p+d\}$, we then get

$$\|\beta\|_1 \leq \frac{\max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \Big(\sum_{j=1}^{p} \|(\beta_{j,\bullet})_{L_j}\|_{\mathrm{TV},\omega_{j,\bullet}} + \sum_{j=1}^{p} \|(\beta_{j,\bullet})_{L_j^{\complement}}\|_{\mathrm{TV},\omega_{j,\bullet}} \Big).$$

Now, if $\beta \in \mathscr{C}_{\mathrm{TV},\omega}(L)$, we obtain

$$\|\beta\|_1 \leq \frac{4 \max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \sum_{j=1}^{p} \|(\beta_{j,\bullet})_{L_j}\|_{\mathrm{TV},\omega_{j,\bullet}}.$$

Further, we have that $\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} \leq 2 \max_{j,l} \omega_{j,l} \|\beta_{j,\bullet}\|_1$. Hence, we obtain

$$\begin{aligned}
\|\beta\|_1 &\leq \frac{8 \max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \max_{j,l} \omega_{j,l} \sum_{j=1}^{p} \|(\beta_{j,\bullet})_{L_j}\|_1 \\
&= \frac{8 \max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \max_{j,l} \omega_{j,l} \|\beta_L\|_1 \qquad (41) \\
&\leq \sqrt{|L|} \frac{8 \max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \max_{j,l} \omega_{j,l} \|\beta_L\|_2.
\end{aligned}$$

□

Appendix C Proof of Theorem 2

On the definition of $b^\star$. Let us first make a remark concerning the choice we made to approximate $f^\star$ using $b^\star$. Instead of what we did in (16) and (17), it may be tempting to define $b^\star$ such that

$$\tilde{f}_{j,\bullet} \in \mathrm{argmin}_{f_{\beta_{j,\bullet}} \in \mathcal{P}^{\mu_{j,\bullet}}} \|f^\star_{j,\bullet} - f_{\beta_{j,\bullet}}\|_{\mathcal{Q}}$$

for all $j = 1, \dots, p$, with $\mathcal{P}^{\mu_{j,\bullet}}$ the set of $\mu_{j,\bullet}$-piecewise-constant functions defined on $[0,1]$, and $\mathcal{Q}$ denoting either the Hilbert space over $[0,1]$ endowed by the norm $\|f\|^2 = \int_0^1 f^2(x)\mathrm{d}x$, or the complete normed vector space of real integrable functions in the

Lebesgue sense. In the first case $\big(\mathcal{Q} = L^2([0,1])\big)$, $\tilde{f}_{j,\bullet}$ could be viewed as an orthogonal projection. However, the resulting approximated vector $b^\star$ would almost surely have a support set relative to the total variation penalty double the size of $\beta^\star$'s one, which is not intuitive. In the second case $\big(\mathcal{Q} = L^1([0,1])\big)$, both $\beta^\star$ and $b^\star$ would have the same cardinality of their respective support sets relative to the total variation penalty. But for a given cut-point $\mu^\star_{j,k}$, the corresponding $b^\star$ cut-point would be $\mu_{j,l^\star_{j,k}-1}$ if $\mu^\star_{j,k}$ was closer to $\mu_{j,l^\star_{j,k}-1}$ than to $\mu_{j,l^\star_{j,k}}$ and vice versa, which would make the writing more cumbersome. To get around this difficulty, we defined $\tilde{f}_{j,\bullet}$ in (16) such that the corresponding cut-point is always the right bound of $I_{j,l^\star_{j,k}}$, i.e., $\mu_{j,l^\star_{j,k}}$.

On the approximation bias. Let us now state an initial lemma concerning the "bias" existing between the true function $f^\star$ and its approximation $f_{b^\star}$ defined in (17). We state the following result bounding $\|f^\star - f_{b^\star}\|_n^2$ with large probability. Towards this end, we define

$$\hat{\pi}_{j,k} = \frac{|\{i = 1, \ldots, n : X_{i,j} \in \mathcal{I}^\star_{j,k}\}|}{n},$$

where we denote

$$\mathcal{I}^\star_{j,k} = \big(I^\star_{j,k} \cap I_{j,l^\star_{j,k-1}}\big) \bigcup \big((I^\star_{j,k})^c \cap I_{j,l^\star_{j,k}}\big)$$

for all $j = 1, \ldots, n$ and $k = 1, \ldots, K^\star_j + 1$.

Lemma 7 *The inequality*

$$\|f^\star - f_{b^\star}\|_n^2 \leq \Big\{ \sum_{j \in \mathscr{A}(\beta^\star)} \sum_{k=1}^{K^\star_j+1} |\beta^\star_{j,k}| \frac{n_{j,l^\star_{j,k}}}{n} \Big\}^2 \pi_n + \frac{2\pi_n e^{2f^\star_\infty}}{c_Z} \sum_{j \in \mathscr{A}(\beta^\star)} \sum_{k=1}^{K^\star_j+1} \hat{\pi}_{j,k} |\beta^\star_{j,k}|^2$$

holds with probability at least $1 - 2e^{-nc_Z^2/2}$.

Proof of Lemma 7. We have

$$\|f^\star - f_{b^\star}\|_n^2 = \int_0^\tau \sum_{i=1}^n \big[(f^\star - f_{b^\star})(X_i) - \big(\bar{f}^\star(t) - \bar{f}_{b^\star}(t)\big)\big]^2 \frac{Y_i(t) e^{f^\star(X_i)}}{S^{(0)}_n(f^\star, t)} \mathrm{d}\bar{N}(t)$$

and

$$\bar{f}^\star(t) - \bar{f}_{b^\star}(t) = \sum_{i=1}^n (f^\star - f_{b^\star})(X_i) \frac{Y_i(t) e^{f^\star(X_i)}}{S^{(0)}_n(f^\star, t)}.$$

It is obvious that

$$\|f^\star - f_{b^\star}\|_n^2 = \int_0^\tau \sum_{i=1}^n \big((f^\star - f_{b^\star})(X_i)\big)^2 \frac{Y_i(t) e^{f^\star(X_i)}}{S^{(0)}_n(f^\star, t)} \mathrm{d}\bar{N}(t) - \int_0^\tau \big(\bar{f}^\star(t) - \bar{f}_{b^\star}(t)\big)^2 \mathrm{d}\bar{N}(t),$$

which means that

$$\|f^\star - f_{b^\star}\|_n^2 \leq \int_0^\tau \sum_{i=1}^n \big((f^\star - f_{b^\star})(X_i)\big)^2 \frac{Y_i(t) e^{f^\star(X_i)}}{S^{(0)}_n(f^\star, t)} \mathrm{d}\bar{N}(t). \tag{42}$$

Next, we control the right-hand-side of (42). For all $i = 1, \dots, n$, we have that

$$
\begin{aligned}
&(f_j^\star - f_{b^\star_{j,\bullet}})(X_i) \\
&\quad = \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \big(\mathbb{1}(X_{i,j} \in I_{j,k}^\star) - \sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star} \mathbb{1}(X_{i,j} \in I_{j,l}) \big) + \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star} \frac{n_{j,l}}{n} \\
&\quad = \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \big\{ \mathbb{1}(X_{i,j} \in I_{j,k}^\star \cap I_{j,l_{j,k-1}^\star} - \mathbb{1}(X_{i,j} \in (I_{j,k}^\star)^c \cap I_{j,l_{j,k}^\star}) \big\} + \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star} \frac{n_{j,l}}{n}.
\end{aligned}
$$

Then, we obtain

$$
|f^\star(X_i) - f_{b^\star}(X_i)| \le \sum_{j=1}^{p} \sum_{k=1}^{K_j^\star+1} \big| \beta_{j,k}^\star | \mathbb{1}(X_{i,j} \in \mathcal{I}_{j,k}^\star) + \Big| \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star} \frac{n_{j,l}}{n} \Big|.
$$

Let us rewrite constraint (2) such that

$$
0 = \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star n_{j,k}^\star = \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \Big(\sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star-1} n_{j,l} + |\{ i : X_{i,j} \in \big(I_{j,k}^\star \cap I_{j,l_{j,k}^\star} \big) \cup \big(I_{j,k}^\star \cap I_{j,l_{j,k-1}^\star} \big) \}| \Big)
$$

(see Figure 1) to obtain

$$
\begin{aligned}
\sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star} n_{j,l} &= \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \Big(\sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star-1} n_{j,l} + n_{j,l_{j,k}^\star} \Big) \\
&= \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \big(n_{j,l_{j,k}^\star} - |\{ i : X_{i,j} \in (I_{j,k}^\star \cap I_{j,l_{j,k}^\star}) \cup (I_{j,k}^\star \cap I_{j,l_{j,k-1}^\star}) \}| \big).
\end{aligned}
$$

Hence,

$$
\Big| \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star} n_{j,l} \Big| \le \sum_{k=1}^{K_j^\star+1} |\beta_{j,k}^\star| n_{j,l_{j,k}^\star}
$$

and

$$
|f^\star(X_i) - f_{b^\star}(X_i)| \le \sum_{j=1}^{p} \sum_{k=1}^{K_j^\star+1} |\beta_{j,k}^\star| \Big(\mathbb{1}(X_{i,j} \in \mathcal{I}_{j,k}^\star) + \frac{n_{j,l_{j,k}^\star}}{n} \Big).
$$

Bringing this all together, we have that

$$\begin{aligned}
&\int_0^\tau \sum_{i=1}^n \big((f^\star - f_{b^\star})(X_i)\big)^2 \frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)}\mathrm{d}\bar{N}(t) \\
&\leq \int_0^\tau \sum_{i=1}^n \Big\{ \sum_{j=1}^p \sum_{k=1}^{K_j^\star+1} |\beta_{j,k}^\star| \big(\mathbb{1}(X_{i,j} \in \mathcal{I}_{j,k}^\star) + \frac{n_{j,l_{j,k}^\star}}{n}\big) \Big\}^2 \frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)}\mathrm{d}\bar{N}(t) \\
&\leq \underbrace{2\int_0^\tau \sum_{i=1}^n \sum_{j=1}^p \sum_{k=1}^{K_j^\star+1} |\beta_{j,k}^\star|^2 \mathbb{1}(X_{i,j} \in \mathcal{I}_{j,k}^\star) \frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)}\mathrm{d}\bar{N}(t)}_{(i)} \\
&\quad + \underbrace{2\int_0^\tau \sum_{i=1}^n \Big\{ \sum_{j=1}^p \sum_{k=1}^{K_j^\star+1} |\beta_{j,k}^\star| \frac{n_{j,l_{j,k}^\star}}{n} \Big\}^2 \frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)}\mathrm{d}\bar{N}(t)}_{(ii)},
\end{aligned}$$

where we used the fact that the indicator functions are orthogonal. On the one hand, we have

$$\begin{aligned}
(ii) &= \Big\{ \sum_{j \in \mathscr{A}(\beta^\star)} \sum_{k=1}^{K_j^\star+1} |\beta_{j,k}^\star| \frac{n_{j,l_{j,k}^\star}}{n} \Big\}^2 \pi_n \\
&\leq \frac{\max_{j\in\mathscr{A}(\beta^\star)} \|\beta_{j,\bullet}\|_\infty^2 \max_{j\in\mathscr{A}(\beta^\star)} \|n_{j,\bullet}\|_\infty^2}{n} \big(|\mathscr{A}(\beta^\star)| + K^\star\big)\pi_n. \qquad (43)
\end{aligned}$$

On the other, using the fact that $e^{f^\star(X_i)} \leq e^{f_\infty}$ and $Y_i(t) \leq 1$ for all $t \in [0,\tau]$, we get

$$\begin{aligned}
(i) &\leq e^{f_\infty^\star} \sum_{j=1}^p \sum_{k=1}^{K_j^\star+1} \frac{1}{n} \sum_{i=1}^n \mathbb{1}(X_{i,j} \in \mathcal{I}_{j,k}^\star) |\beta_{j,k}^\star|^2 \int_0^\tau \frac{1}{n^{-1}S_n^{(0)}(f^\star,t)}\mathrm{d}\bar{N}(t) \\
&\leq \frac{\pi_n e^{f_\infty^\star}}{\inf_{t\in[0,\tau]} n^{-1}S_n^{(0)}(f^\star,t)} \sum_{j=1}^p \sum_{k=1}^{K_j^\star+1} \hat{\pi}_{j,k} |\beta_{j,k}^\star|^2 \\
&\leq \frac{\pi_n e^{f_\infty^\star}}{\inf_{t\in[0,\tau]} n^{-1}S_n^{(0)}(f^\star,t)} \max_{j\in\mathscr{A}(\beta^\star)} \|\beta_{j,\bullet}\|_\infty^2 \max_{j\in\mathscr{A}(\beta^\star)} \|\hat{\pi}_{j,\bullet}\|_\infty \big(|\mathscr{A}(\beta^\star)| + K^\star\big).
\end{aligned}$$

Moreover, remember that $n^{-1}S_n^{(0)}(f^\star,t) = n^{-1}\sum_{i=1}^n \mathbb{1}(Z_i \geq t)e^{f^\star(X_i)}$, and observe that for all $t \leq \tau$, we have $\{Z_i \geq \tau\} \subset \{Z_i \geq t\}$. Hence,

$$\frac{1}{n}S_n^{(0)}(f^\star,t) \geq e^{-f_\infty^\star}\frac{1}{n}\sum_{i=1}^n \mathbb{1}(Z_i \geq \tau) \text{ for all } t \leq \tau.$$

Using the Dvoretzky-Kiefer-Wolfowitz inequality [Massart, 1990], we get that:

$$\begin{aligned}
&\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n \mathbb{1}(Z_i \geq \tau) \geq \frac{1}{2}\mathbb{P}[Z_1 \geq \tau]\Big] \\
&\qquad \geq \mathbb{P}\Big[\sqrt{n}\sup_{t\in[0,\tau]} \Big|\frac{1}{n}\sum_{i=1}^n \mathbb{1}(Z_i \geq t) - \mathbb{P}[Z_1 \geq t]\Big| \geq \frac{\sqrt{n}}{2}\mathbb{P}[Z_1 \geq \tau]\Big] \\
&\qquad \geq 1 - 2e^{-nc_Z^2/2}.
\end{aligned}$$

Then, we have

$$\mathbb{P}\Big[\inf_{t\in[0,\tau]}\frac{1}{n}S_n^{(0)}(f^\star,t)\geq e^{-f_\infty^\star}\frac{c_Z}{2}\Big]\geq\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n\mathbb{1}(Z_i\geq\tau)\geq\frac{c_Z}{2}\Big]\geq 1-2e^{-nc_Z^2/2}. \tag{44}$$

Combining (43) and (44), we obtain the desired result.

□

Proof of Theorem 2. Using the triangle inequality, we have that

$$\|f_{b^\star}-f_{\hat\beta}\|_n^2\leq(\|f_{b^\star}-f^\star\|_n+\|f^\star-f_{\hat\beta}\|_n)^2\leq 2(\|f_{b^\star}-f^\star\|_n^2+\|f^\star-f_{\hat\beta}\|_n^2).$$

Inequality (26) in Lemma 3 yields

$$\|f^\star-f_{\hat\beta}\|_n^2\leq\frac{\|f^\star-f_{\hat\beta}\|_\infty^2}{\psi(-\|f^\star-f_{\hat\beta}\|_\infty)}KL_n(f^\star,f_{\hat\beta})\leq(f_\infty^\star+R+2)KL_n(f^\star,f_{\hat\beta}),$$

where we use inequality (32). The construction of the approximation $f_{b^\star}$ of $f^\star$ gives $|\mathcal{A}(b^\star)|=K^\star$, so an application of Theorem 1 to $b^\star$ combined with inequality (26) in Lemma 3 ensures that with a probability greater than $1-28.55e^{-c}-e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}}-3\varepsilon$,

$$\begin{aligned}KL_n(f^\star,f_{\hat\beta})&\leq 3KL_n(f^\star,f_{b^\star})+\frac{1024(f_\infty^\star+R^\star+2)K^\star\max_{1\leq j\leq p}\|(\omega_{j,\bullet})_{\mathcal{A}_j(b^\star)}\|_\infty^2}{\kappa_\tau^2\big(\mathcal{A}(b^\star)\big)-\Xi_\tau\big(\mathcal{A}(b^\star)\big)}\\&\leq 3\|f^\star-f_{b^\star}\|_n^2\frac{\psi(f_\infty^\star+R^\star+2)}{(f_\infty^\star+R^\star+2)^2}+\frac{1024(f_\infty^\star+R^\star+2)K^\star\max_{1\leq j\leq p}\|(\omega_{j,\bullet})_{\mathcal{A}_j(b^\star)}\|_\infty^2}{\kappa_\tau^2\big(\mathcal{A}(b^\star)\big)-\Xi_\tau\big(\mathcal{A}(b^\star)\big)},\end{aligned}$$

where we used the fact that $u\mapsto\psi(u)/u^2$ is increasing. Therefore, with a probability greater than $1-28.55e^{-c}-e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}}-3\varepsilon$, the following holds:

$$\begin{aligned}\|f_{b^\star}-f_{\hat\beta}\|_n^2\leq 2\|f_{b^\star}-f^\star\|_n^2\Big(1+3\frac{\psi(f_\infty^\star+R+2)}{f_\infty^\star+R+2}\Big)\\+\frac{2048(f_\infty^\star+R+2)^2K^\star\max_{1\leq j\leq p}\|(\omega_{j,\bullet})_{\mathcal{A}_j(b^\star)}\|_\infty^2}{\kappa_\tau^2\big(\mathcal{A}(b^\star)\big)-\Xi_\tau\big(\mathcal{A}(b^\star)\big)}.\end{aligned}$$

By Lemma 7, we obtain

$$\|f_{b^\star}-f_{\hat\beta}\|_n^2\leq\mathbf{I}+\mathbf{II}$$

with a probability larger than $1-28.55e^{-c}-e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}}-3\varepsilon-2e^{-nc_Z^2/2}$. Now using the definitions of $\|\cdot\|_n$ and κ_τ in (12), we have

$$\|f_{b^\star}-f_{\hat\beta}\|_n^2=(b^\star-\hat\beta)^\top\widehat{\Sigma}_n(f^\star,\tau)(b^\star-\hat\beta)\geq\kappa_\tau^2\big(\mathcal{A}(b^\star)\big)\|(b^\star-\hat\beta)_{\mathcal{A}(b^\star)}\|_2^2.$$

We therefore have that

$$\|(\hat\beta-b^\star)_{\mathcal{A}(b^\star)}\|_1\leq\frac{\sqrt{K^\star(\mathbf{I}+\mathbf{II})}}{\kappa_\tau\big(\mathcal{A}(b^\star)\big)},$$

with a probability larger than $1-28.55e^{-c}-e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}}-3\varepsilon-2e^{-nc_Z^2/2}$.

□

References

O. Aalen. Nonparametric inference for a family of counting processes. *Ann. Statist.*, 6(4): 701–726, 1978.

M. Z. Alaya, S. Gaïffas, and A. Guilloux. Learning the intensity of time events with change-points. *Information Theory, IEEE Transactions on*, 61(9):5148–5171, 2015.

M. Z. Alaya, S. Bussy, S. Gaïffas, and A. Guilloux. Binarsity: a penalization for one-hot encoded features. *preprint*, 2017.

D.G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *JNCI: Journal of the National Cancer Institute*, 86(11):829–835, 1994.

Christophe Ambroise and Geoffrey J McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the national academy of sciences*, 99(10):6562–6566, 2002.

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

A.V. Antonov. Bioprofiling. de: analytical web portal for high-throughput cell biology. *Nucleic acids research*, 39(suppl_2):W323–W327, 2011.

A.V. Antonov, M. Krestyaninova, R.A. Knight, I. Rodchenkov, G. Melino, and N.A. Barlev. Ppisurv: a novel bioinformatics tool for uncovering the hidden role of specific genes in cancer survival outcome. *Oncogene*, 33(13):1621, 2014.

F. Bach. Self-concordant analysis for logistic regression. *Electron. J. Statist.*, 4:384–414, 2010.

E. Bacry, M. Bompaire, S. Gaïffas, and S. Poulsen. tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling. *ArXiv e-prints*, July 2017.

S. Badve, D. Turbin, M.A. Thorat, A. Morimiya, T.O. Nielsen, C.M. Perou, S. Dunn, D.G. Huntsman, and H. Nakshatri. Foxa1 expression in breast cancer—correlation with luminal subtype a and survival. *Clinical cancer research*, 13(15):4415–4421, 2007.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

BioProfiling. Hbs1l ppisurv, 2009. URL http://www.bioprofiling.de/cgi-bin/GEO/DRUGSURV/display_GENE_GEO.pl?ID=GSE2034&affy=209314_S_AT&ncbi=10767&geneA=HBS1L.

K. Bleakley and J. P. Vert. The group fused lasso for multiple change-point detection. 2011.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

J. Budczies, F. Klauschen, B. V. Sinn, B. Győrffy, W. D. Schmitt, S. Darb-Esfahani, and C. Denkert. Cutoff finder: a comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PloS one*, 7(12):e51862, 2012.

Simon Bussy, Agathe Guilloux, Stéphane Gaïffas, and Anne-Sophie Jannot. C-mix: A high-dimensional mixture model for censored durations, with applications to genetic data. *Statistical methods in medical research*, 28(5):1523–1539, 2019.

R. L. Camp, M. Dolled-Filhart, and D. L. Rimm. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clinical cancer research*, 10(21):7252–7259, 2004.

E. Canu, M. Boccardi, R. Ghidoni, L. Benussi, S. Duchesne, C. Testa, G. Binetti, and G. B. Frisoni. Hoxa1 a218g polymorphism is associated with smaller cerebellar volume in healthy humans. *Journal of Neuroimaging*, 19(4):353–358, 2009.

C. Chang, M. Hsieh, W. Chang, A. Chiang, and J. Chen. Determining the optimal number and location of cutoff points with application to data of cervical cancer. *PloS one*, 12 (4):e0176231, 2017.

C. Chang, M. Hsieh, A. J. Chiang, Y. H. Tsai, C. Liu, and J. Chen. Methods for estimating the optimal number and location of cut points in multivariate survival analysis: a statistical solution to the controversial effect of bmi. *Computational Statistics*, pages 1–26, 2019.

M. C. U. Cheang, S. K. Chia, D. Voduc, D. Gao, S. Leung, J. Snider, M. Watson, S. Davies, P. S. Bernard, J. S. Parker, et al. Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. *JNCI: Journal of the National Cancer Institute*, 101(10): 736–750, 2009.

H. Cho and P. Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015.

L. Condat. A Direct Algorithm for 1D Total Variation Denoising. *IEEE Signal Processing Letters*, 20(11):1054–1057, 2013.

C. Contal and J. O'Quigley. An application of changepoint methods in studying the effect of age on survival in breast cancer. *Computational statistics & data analysis*, 30(3): 253–270, 1999.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

M. Csikos, Z. Orosz, G. Bottlik, H. Szöcs, Z. Szalai, Z. Rozgonyi, J. Hársing, E. Török, L. Bruckner-Tuderman, A. Horváth, et al. Dystrophic epidermolysis bullosa complicated by cutaneous squamous cell carcinoma and pulmonary and renal amyloidosis. *Clinical and experimental dermatology*, 28(2):163–166, 2003.

A. Dancau, L. Wuth, M. Waschow, F. Holst, A. Krohn, M. Choschzick, L. Terracciano, S. Politis, S. Kurtz, A. Lebeau, et al. Ppfia1 and ccnd1 are frequently coamplified in breast cancer. *Genes, Chromosomes and Cancer*, 49(1):1–8, 2010.

Ran Duan, Lei Han, Qixue Wang, Jianwei Wei, Luyue Chen, Jianning Zhang, Chunsheng Kang, and Lei Wang. Hoxa13 is a potential gbm diagnostic marker and promotes glioma invasion by activating the wnt and tgf-β pathways. *Oncotarget*, 6(29):27778, 2015.

S. Dudoit and M. J. Van Der Laan. *Multiple testing procedures with applications to genomics*. Springer Science & Business Media, 2007.

D. Faraggi and R. Simon. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Statistics in medicine*, 15(20):2203–2213, 1996.

S. Gaïffas and A. Guilloux. High-dimensional additive hazards models and the Lasso. *Electron. J. Stat.*, 6:522–546, 2012.

R. Gill. Large sample behaviour of the product-limit estimator on the whole line. *The annals of statistics*, 11(1):49–58, 1983.

Yong Guan, Yajie He, Shaoping Lv, Xiaoqun Hou, Luo Li, and Jianjun Song. Overexpression of hoxc10 promotes glioblastoma cell progression to a poor prognosis via the pi3k/akt signalling pathway. *Journal of drug targeting*, 27(1):60–66, 2019.

Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.*, 105(492):1480–1493, 2010.

J. M. Harvey, G. M. Clark, C. K. Osborne, D. C. Allred, et al. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. *Journal of clinical oncology*, 17(5): 1474–1481, 1999.

P. J. Heagerty and Y. Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

J. Huang, T. Sun, Z. Ying, Y. Yu, and C. H. Zhang. Oracle inequalities for the lasso in the cox model. *Ann. Statist.*, 41(3):1142–1165, 06 2013.

N. Huang, S. Cheng, X. Mi, Q. Tian, Q. Huang, F. Wang, Z. Xu, Z. Xie, J. Chen, and Y. Cheng. Downregulation of nitrogen permease regulator like-2 activates pdk1-akt1 and contributes to the malignant growth of glioma cells. *Molecular carcinogenesis*, 55 (11):1613–1626, 2016.

T. R. Icuma, J. A. Achcar, E. Z. Martinez, and N. Davarzani. Determination of optimum medical cut points for continuous covariates in lifetime regression models. *Model Assisted Statistics and Applications*, 13(2):141–159, 2018.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

S. Ivanoff, F. Picard, and V. Rivoirard. Adaptive lasso and group-lasso for functional poisson regression. *The Journal of Machine Learning Research*, 17(1):1903–1948, 2016.

M. A. James, Y. Lu, Y. Liu, H. G. Vikis, and M. You. Rgs17, an overexpressed gene in human lung and prostate cancer, induces tumor cell proliferation through the cyclic amp-pka-creb pathway. *Cancer research*, 69(5):2108–2116, 2009.

J. P. Klein and M. L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

J. P. Klein and J. Wu. Discretizing a continuous covariate in survival studies. *Handbook of Statistics*, 23:27–42, 2003.

P. Kulkarni, T. Shiraishi, K. Rajagopalan, R. Kim, S. M. Mooney, and R. H. Getzenberg. Cancer/testis antigens and urological malignancies. *Nature Reviews Urology*, 9(7):386, 2012.

S. S. Kutateladze. *Fundamentals of functional analysis*, volume 12. Springer Science & Business Media, 2013.

B. Lausen and M. Schumacher. Maximally selected rank statistics. *Biometrics*, pages 73–85, 1992.

M. LeBlanc and J. Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467, 1993.

S. Lemler. Oracle inequalities for the lasso in the high-dimensional aalen multiplicative intensity model. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 981–1008. Institut Henri Poincaré, 2016.

H. Li and Y. Luan. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21(10):2403–2409, 2005.

H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: an enabling technique. *Data Min. Knowl. Discov.*, 6(4):393–423, 2002.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 07 1990.

L. Meier, S. Van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B):3779–3821, 2009.

R. Mizutani, N. Imamachi, Y. Suzuki, H. Yoshida, N. Tochigi, T. Oonishi, and N. Akimitsu. Oncofetal protein igf2bp3 facilitates the activity of proto-oncogene protein eif4e through the destabilization of eif4e-bp2 mrna. *Oncogene*, 35(27):3495, 2016.

R. J. Motzer, M. Mazumdar, J. Bacik, W. Berg, A. Amsterdam, and J. Ferrara. Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma. *Journal of clinical oncology*, 17(8):2530–2530, 1999.

J. W. Moul, L. Sun, J. M. Hotaling, N. J. Fitzsimons, T. J. Polascik, C. N. Robertson, P. Dahm, M. S. Anscher, V. Mouraviev, P. A. Pappas, et al. Age adjusted prostate specific antigen and prostate specific antigen velocity cut points in prostate cancer screening. *The Journal of urology*, 177(2):499–504, 2007.

B. N. Mukherjee and S. S. Maiti. On some properties of positive definite toeplitz matrices and their possible applications. *Linear algebra and its applications*, 102:211–240, 1988.

A. Oreopoulos, R. Padwal, K. Kalantar-Zadeh, G. C. Fonarow, C. M. Norris, and F. A. McAlister. Body mass index and mortality in heart failure: a meta-analysis. *American heart journal*, 156(1):13–22, 2008.

P. Rajaraman, A. Hutchinson, N. Rothman, P. M. Black, H. A. Fine, J. S. Loeffler, R. G. Selker, W. R. Shapiro, M. S. Linet, and P. D. Inskip. Oxidative response gene polymorphisms and risk of adult brain tumors. *Neuro-oncology*, 10(5):709–715, 2008.

R. T. Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

M. Rota, L. Antolini, and M. G. Valsecchi. Optimal cut-point definition in biomarkers: the case of censored failure time outcome. *BMC medical research methodology*, 15(1): 24, 2015.

R. Senoussi. Problème d'identification dans le modèle de cox. *Ann. Inst. Henri Poincaré*, 26:45–64, 1990.

Y. Shirota, J. Stoehlmacher, J. Brabender, Y. Xiong, H. Uetake, K. D. Danenberg, S. Groshen, D. D. Tsao-Wei, P. V. Danenberg, and H. J. Lenz. Ercc1 and thymidylate synthase mrna levels predict survival for colorectal cancer patients receiving combination oxaliplatin and fluorouracil chemotherapy. *Journal of Clinical Oncology*, 19(23): 4298–4304, 2001.

N. Simon, J. Friedman, T. Hastie, R. Tibshirani, et al. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39 (5):1–13, 2011.

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(10):1105–1117, 2011.

S. Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

S. Van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, 3:1360–1392, 2009.

P. S. Wells, D. R. Anderson, M. Rodger, J. S. Ginsberg, C. Kearon, M. Gent, A. G. G. Turpie, J. Bormanis, J. Weitz, and M. Chamberlain. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the simplired d-dimer. *Thrombosis and haemostasis*, 83(03):416–420, 2000.

P. H. Westfall, S. S. Young, and S. P. Wright. On adjusting p-values for multiplicity. *Biometrics*, 49(3):941–945, 1993.

J. Wu and S. Coggeshall. *Foundations of Predictive Analytics (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 1st edition, 2012.

S. Yousefi, F. Amrollahi, M. Amgad, C. Dong, J. E. Lewis, C. Song, D. A. Gutman, S. H. Halani, J. E. V. Vega, and D. J. Brat. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):11707, 2017.