Received: 11 June 2020 | Revised: 8 July 2021 | Accepted: 15 July 2021

DOI: 10.1111/biom.13547

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

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

Simon Bussy[1,2] | **Mokhtar Z. Alaya**[3] | **Anne-Sophie Jannot**[4] | **Agathe Guilloux**[5]

[1] LPSM, UMR 8001, CNRS, Sorbonne University, Paris, France

[2] LOPF, Califrais' Machine Learning Lab, Paris, France

[3] LMAC EA 2222, Université de Technologie de Compiègne, Compiègne, France

[4] Biomedical Informatics and Public Health Department, APHP and Université de Paris, INSERM, Centre de Recherche des Cordeliers, Paris, France

[5] LaMME, UEVE and UMR 8071, Paris Saclay University, Evry, France

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

Funding information
DIM Math Innov Région Ile-de-France; INCA-DGOS, Grant/Award Number: PTR-K 2014

Abstract

We introduce *binacox*, a prognostic method to deal with the problem of detecting multiple cut-points per feature 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 data sets. On these high-dimensional data sets, our proposed method 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

feature binarization, genetic cancer data, 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 features 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: above this threshold, patients have higher risk of unfavorable outcome. Another well-known clinical example where finding optimal thresholds is paramount is the platelet transfusion decision, based on a platelet-count threshold (Curley *et al.*, 2019) ; or the piece-wise constant effect assumption made on the body mass index where prognoses are generally worst for obese and underweight individuals (Bédat *et al.*, 2015) (hence with two cut-points here).

With the increasing availability of high-dimensional data sets, data-driven predictive scores are becoming increasingly important. A convenient tool for finding multiple cut-points in a multivariate and high-dimensional setting—and then automatically building interpretable predictive scores—is therefore of high interest. For instance, in genetic oncology studies, similar questions occur because the effect of certain genes' expression on survival times is often nonlinear. 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.

Solving this problem means applying nonlinearities 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.

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, for example, the minimal p-value from multiple log-rank tests, see, for instance, Rota *et al.* (2015). However, the choice of the actual cut-points is not a straightforward problem, even for a single cut-point (Lausen and Schumacher, 1992; Contal and O'Quigley, 1999; Klein and Wu, 2003). Recently, Icuma *et al.* (2018) proposed a Bayesian approach with accelerated failure time modeling, but still only allowing one cut-point per feature.

Indeed, although many studies have been devoted to finding one optimal cut-point, there is often need in medical settings to determine not only one but multiple cut-points (e.g., the body mass index example discussed in the introduction). 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 (Cho and Fryzlewicz, 2015). Although 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) that 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.

Let us consider the usual survival analysis framework. Following Andersen *et al.* (2012), let nonnegative 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, and atomics features). The event of interest could be, for instance, death, re-hospitalization, relapse, or disease 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 settings are becoming increasingly frequent, in particular for genetic data applications where cut-point estimation is a common problem (see, for instance, 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.

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 method 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.*, 2019) that 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. Second 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 Bregman divergence. The arguments used to obtain our results are then different, and also differ from the ones used in Kong and Nan (2014). A precise description of the model is given in Section 2. Section 3 highlights the good theoretical properties of the binacox method 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 data sets. Finally, we discuss the obtained results in Section 6.

2 | MODEL AND METHOD

Consider an independent and identically distributed 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 $\boldsymbol{X} = [X_{i,j}]_{1 \le i \le n; 1 \le j \le p}$ be the $n \times p$ fixed 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

$$\lambda^\star(t|X_i) = \lambda_0^\star(t) \mathrm{e}^{f^\star(X_i)}, \tag{1}$$

where $\lambda_0^\star(t)$ is the baseline hazard function. Our goal is to estimate $f^\star$.

2.1 | 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). The jth column $\boldsymbol{X}_{\bullet,j}$ is then replaced by $d_j + 1 \ge 2$ columns $\boldsymbol{X}^B_{\bullet,j,1}, \dots, \boldsymbol{X}^B_{\bullet,j,d_j+1}$ containing only zeros and ones, where the ith row $X_i^B \in \mathbb{R}^{p+d}$ with $d = \sum_{j=1}^p d_j$ is written

$$X_i^B = (X^B_{i,1,1}, \dots, X^B_{i,1,d_1+1}, \dots, X^B_{i,p,1}, \dots, X^B_{i,p,d_p+1})^\top.$$

We consider the intervals $I_{j,1}, \dots, I_{j,d_j+1}$ defining a partition of $[0,1]$, that is

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

and $I_{j,k} \cap I_{j,k'} = \varnothing$ for all $k \neq k'$ with $k, k' = 1, \dots, d_j + 1$. Now for $i = 1, \dots, n$ and $l = 1, \dots, d_j + 1$, 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 $l = 1, \dots, d_j + 1$, 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 $\mu_{j,l} = q_j(l/(d_j+1))$, where $q_j(\alpha)$ denotes a quantile of order $\alpha \in [0,1]$ for $\boldsymbol{X}_{\bullet,j}$. If training data also contain 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 $\beta_{j,\bullet} = (\beta_{j,1}, \dots, \beta_{j,d_j+1})^\top$ and $\mu_{j,\bullet} = (\mu_{j,1}, \dots, \mu_{j,d_j})^\top$. Hence, we define a candidate for the estimation of $f^\star$ defined in (1) as

$$f_\beta(X_i) = \beta^\top X_i^B = \sum_{j=1}^p f_{\beta_{j,\bullet}}(X_{i,j}) = \sum_{j=1}^p \sum_{l=1}^{d_j+1} \beta_{j,l} \mathbb{1}(X_{i,j} \in I_{j,l}). \tag{2}$$

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

$$\begin{aligned} \beta &= (\beta_{1,\bullet}^\top, \dots, \beta_{p,\bullet}^\top)^\top \\ &= (\beta_{1,1}, \dots, \beta_{1,d_1+1}, \dots, \beta_{p,1}, \dots, \beta_{p,d_p+1})^\top, \end{aligned}$$

and

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

2.2 | 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:

$$\ell_n(f_\beta) = -\frac{1}{n}\sum_{i=1}^{n}\Delta_i\left\{f_\beta(X_i) - \log\sum_{i':Z_{i'}\geq Z_i} e^{f_\beta(X_{i'})}\right\}. \quad (3)$$

Our approach consists in minimizing the function ℓ_n plus the binarsity penalization term introduced in Alaya *et al.* (2019). The resulting optimization problem is written

$$\hat{\beta} \in \operatorname{argmin}_{\beta\in\mathcal{B}_{p+d}(R)}\left\{\ell_n(f_\beta) + \mathrm{bina}(\beta)\right\}, \quad (4)$$

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

$$\mathrm{bina}(\beta) = \sum_{j=1}^{p}\left(\sum_{l=2}^{d_j+1}\omega_{j,l}|\beta_{j,l} - \beta_{j,l-1}| + \delta_j(\beta_{j,\bullet})\right), \quad (5)$$

with

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

and where $n_{j,\bullet} = (n_{j,1},\dots,n_{j,d_j+1})^\top \in \mathbb{N}^{d_j+1}$ with $n_{j,l} = |\{i = 1,\dots,n : X_{i,j} \in I_{j,l}\}|$ for all $j = 1,\dots,p$ and $l = 1,\dots,d_j+1$. The constraint over $\mathcal{B}_{p+d}(R)$ is standard in the literature for obtaining proofs of oracle inequalities for sparse generalized linear models (Van de Geer *et al.*, 2008), and is discussed in detail in Section 3 right after Theorem 1. For a given numerical constant $c > 0$, the weights $\omega_{j,l}$ have an explicit form given by

$$\begin{aligned}\omega_{j,l} &= 11.32\sqrt{\frac{c + \log(p+d) + \mathscr{L}_{n,c}}{n}\hat{V}_{j,l}} \\ &\quad + 18.62\frac{c + 1 + \log(p+d) + \mathscr{L}_{n,c}}{n} \\ &= \mathcal{O}\left(\sqrt{\frac{\log(p+d)}{n}\hat{V}_{j,l}}\right),\end{aligned}$$

where

$$\mathscr{L}_{n,c} = 2\log\log\frac{2n\hat{V}_{j,l} + 18,66e(c + \log(p+d))}{8}$$

and with

$$\hat{V}_{j,l} = \frac{\left|\left\{i = 1,\dots,n : X_{i,j} \in \bigcup_{u=l}^{d_j+1} I_{j,u}\right\}\right|}{n}.$$

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 $j = 1,\dots,p$, 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 interquantile intervals, we have that $n_{j,l}$ are all equal for $l = 1,\dots,d_j+1$, 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:

$$\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} = \sum_{l=2}^{d_j+1}\omega_{j,l}|\beta_{j,l} - \beta_{j,l-1}|, \quad (6)$$

which 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 $\mathcal{A}(\beta) = [\mathcal{A}_1(\beta),\dots,\mathcal{A}_p(\beta)]$ be the concatenation of the support sets relative to the total-variation penalization, namely

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

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

$$\mathcal{A}_j(\hat{\beta}) = \{\hat{l}_{j,1},\dots,\hat{l}_{j,s_j}\}, \quad (7)$$

where $\hat{l}_{j,1} < \cdots < \hat{l}_{j,s_j}$ and $s_j = |\mathcal{A}_j(\hat{\beta})|$. Some details on the algorithm used to solve the regularization problem (4) are given in Appendices A.5 (with, among others, some explanations about an ad hoc de-noising step) and A.6 in the Supplementary Material.

3 | THEORETICAL GUARANTEES

3.1 | A fast 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

$$\begin{aligned}KL_n(f^\star, f) = &\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\left\{\frac{e^{f^\star(X_i)}\sum_{i=1}^{n}Y_i(t)e^{f(X_i)}}{e^{f(X_i)}\sum_{i=1}^{n}Y_i(t)e^{f^\star(X_i)}}\right\} \\ &\times Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}\mathrm{d}t, \quad (8)\end{aligned}$$

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 nonnegative symmetric matrix:

$$\Sigma_n(f^\star, \tau) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(X_i^B - \bar{X}_n(s)\right)\left(X_i^B - \bar{X}_n(s)\right)^\top \times y_i(s)e^{f^\star(X_i)}\lambda_0^\star(s)\mathrm{d}s, \tag{9}$$

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 $y_i(s) = \mathbb{E}[Y_i(s)|X_i]$ for all $0 \leq s \leq t$ and all $i = 1, \dots, n$. For any concatenation of index subsets $L = [L_1, \dots, L_p]$, we define the compatibility factor

$$\kappa_\tau(L) = \inf_{\beta \in \mathscr{C}_{\mathrm{TV},\omega}(L)\setminus\{\mathbf{0}\}} \frac{\sqrt{\beta^\top \Sigma_n(f^\star,\tau)\beta}}{\|\beta_L\|_2}, \tag{10}$$

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*

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

Such assumptions on τ are common in survival analysis, see, for example, Andersen *et al.* (2012). In addition, we define $c_Z := \min_{1\leq i\leq n} y_i(\tau)$ and remark that

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

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

$$f_\infty^\star = \max_{1\leq i\leq n}|f^\star(X_i)|, \quad s^{(0)}(\tau) = n^{-1}\sum_{i=1}^n y_i(\tau)e^{f^\star(X_i)},$$

$$\text{and} \quad \Lambda_0^\star(\tau) = \int_0^\tau \lambda_0^\star(s)\mathrm{d}s.$$

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

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

For any concatenation set $L = [L_1, \dots, L_p]$ such that $\sum_{j=1}^p |L_j| \leq K^\star$, assume that $\kappa_\tau^2(L) > \Xi_\tau(L)$, where

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

Note that $\kappa_\tau^2(L)$ is the smallest eigenvalue of a population integrated covariance matrix defined in (9), so it is reasonable to treat it as a constant. Moreover, $t_{n,p,d,\varepsilon}^2$ is of order $n^{-1}\log((p+d)^2/\varepsilon)$, 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 (2)).

Theorem 1. *The inequality*

$$KL_n(f^\star, f_{\hat{\beta}}) \leq \inf_\beta\left\{3KL_n(f^\star, f_\beta) + \frac{1024(f_\infty^\star + R + 2)}{\kappa_\tau^2(\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta))}|\mathcal{A}(\beta)|\max_{1\leq j\leq p}\|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_\infty^2\right\} \tag{11}$$

holds with a probability greater than $1 - 57.1e^{-c} - e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}} - 3\varepsilon$ 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 $j = 1, \dots, p$, and such that $|\mathcal{A}(\beta)| \leq K^\star$.

The proof of Theorem 1 is postponed to Appendix B in the Supplementary Material. The second term in the right-hand side of (11) can be viewed as a "variance" (or "complexity") term, and its dominant term satisfies

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

where the symbol $\lesssim$ means that the inequality holds up to a multiplicative constant. Then, one obtains the expected fast convergence rate for the variance $\mathcal{O}(\log(p+d)/n)$ for the estimator $\hat{f}$. In Section 3.2, we adapt Theorem 1 to the case

where the true $f^\star$ lies in a Cox model with cut-points, while the case of a Cox generalized additive model with Lipschitz components is considered in Appendix A.4, with a rate of convergence of order $\mathcal{O}(|\mathscr{A}^\star|n^{-1/4})$.

The value $|\mathcal{A}(\beta)|$ characterizes the sparsity of the vector β, since it counts the number of nonequal 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

$$\max_{1\leq i\leq n} |\beta^\top X_i^B| \leq \sum_{j=1}^{p} \|\beta_{j,\bullet}\|_\infty \leq |\mathscr{A}(\beta)| \times \|\beta\|_\infty, \tag{12}$$

where $\|\beta\|_\infty = \max_{1\leq j\leq p} \|\beta_{j,\bullet}\|_\infty$. The first inequality in (12) 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 (12) 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 *et al.* (2008), Bach (2010), and Ivanoff *et al.* (2016). However, the constraint $\mathscr{B}_{p+d}(R)$ is a technicality that is not used in the numerical experiments in Sections 4 and 5.

Note 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.

Our proof and result also differ from the ones of Kong and Nan (2014), which follows the lines of Van de Geer *et al.* (2008) adapting it to the Cox model. In particular, their bound is given for the excess risk from an expected partial likelihood (integrated also for the covariates distribution), whereas in our paper we bound an empirical Kullback divergence introduced in Senoussi (1990) and which has the properties of a divergence.

3.2 | Piecewise constant case

We now assume that the true $f^\star$ lies in the cut-points model.

Assumption 3. *Assume that $f^\star$ has the following form*

$$f^\star(X_i) = \sum_{j=1}^{p} f_j^\star(X_{i,j}) = \sum_{j=1}^{p} \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \mathbb{1}(X_{i,j} \in I_{j,k}^\star), \tag{13}$$

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

$$\sum_{i=1}^{n} f_j^\star(X_{i,j}) = 0 \quad \textit{for all } j = 1, \dots, p \tag{14}$$

to ensure identifiability.

The identifiability condition is common, see, for example, 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

$$\sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star n_{j,k}^\star = 0 \quad \text{for all } j = 1, \dots, p, \tag{15}$$

where $n_{j,k}^\star = |\{i = 1, \dots, n : X_{i,j} \in I_{j,k}^\star\}|$. This constraint could be replaced by $\mathbb{E}[f_j^\star(X_{i,j})] = 0$ for all $j = 1, \dots, p$. In Section 4, we use (14) to generate the data for the simulation study. 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 $\mu_{j,0}^\star = 0$ and $\mu_{j,K_j^\star+1}^\star = 1$. Denoting $K^\star = \sum_{j=1}^p K_j^\star$, the vector of regression coefficients $\beta^\star \in \mathbb{R}^{K^\star+p}$ is given by

$$\begin{aligned}\beta^\star &= ({\beta_{1,\bullet}^\star}^\top, \dots, {\beta_{p,\bullet}^\star}^\top)^\top \\ &= (\beta_{1,1}^\star, \dots, \beta_{1,K_1^\star+1}^\star, \dots, \beta_{p,1}^\star, \dots, \beta_{p,K_p^\star+1}^\star)^\top,\end{aligned}$$

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

$$\begin{aligned}\mu^\star &= ({\mu_{1,\bullet}^\star}^\top, \dots, {\mu_{p,\bullet}^\star}^\top)^\top \\ &= (\mu_{1,1}^\star, \dots, \mu_{1,K_1^\star}^\star, \dots, \mu_{p,1}^\star, \dots, \mu_{p,K_p^\star}^\star)^\top.\end{aligned}$$

Under this assumption, our goal is now to simultaneously estimate $\mu^\star$ and $\beta^\star$, which also requires estimation of the unknown $K_j^\star$ for all $j = 1, \dots, p$. We obtain the following

$\mu^\star_{j,\bullet}$'s estimator

$$\widehat{\mu}_{j,\bullet} = \left(\mu_{j,\hat{l}_{j,1}}, \dots, \mu_{j,\hat{l}_{j,s_j}}\right)^\top \tag{16}$$

for all $j = 1, \dots, p$. By construction, $K^\star_j$ is estimated by $\widehat{K}_j = s_j$, see the lines following Equation (7) for a definition of the $\hat{l}_{j,k}$ and s_j. Finally, an estimator of $\beta^\star$ is obtained from (4).

Theorem 2. *Fix $\epsilon > 0$ and consider that intervals $I_{j,l}$ ($j = 1, \dots, p$, $l = 1, \dots, d_j + 1$) are now chosen as interquantiles (as in Subsection 2.1), that is $I_{j,l} = (\mu_{j,l-1}, \mu_{j,l}]$ with $\mu_{j,l} = q_j(l/(D+1))$ where D is the integer part of*

$$\frac{6\Delta^2_{\beta,\max} K^\star C_{K^\star, f^\star_\infty} (\max_{1\le i\le n} \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t + \epsilon)}{K^\star |\mathscr{A}^\star| C_{cut}} \cdot n^2,$$

for some positive constant C_{cut} and with $\Delta_{\beta^\star,\max} = \max_{1\le j\le p} \max_{1\le k,k'\le K^\star+1} |\beta^\star_{j,k} - \beta^\star_{j,k'}|$, $C_{K^\star, f^\star_\infty} = (1 + K^\star) f^\star_\infty$. Under Assumption 3, the following holds with high probability

$$\begin{aligned} KL_n(f^\star, f_{\hat{\beta}}) &\le \left(6\Delta^2_{\beta,\max} K^\star C_{K^\star, f^\star_\infty} \left(\max_{1\le i\le n} \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t + \epsilon\right)\right. \\ &\quad \left. + C_{cut} K^\star |\mathscr{A}^\star|\right) n^{-1}. \end{aligned}$$

The proof of Theorem 2 is postponed to Appendix C in the Supplementary Material. A fast rate oracle inequality for estimation in this context is established in Appendix A.3 in the Supplementary Material, and the special case where we consider a Cox generalized additive model with Lipschitz components for $f^\star$ is considered in Appendix A.4.

4 | PERFORMANCE EVALUATION

4.1 | 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. We first take $[X_{i,j}] \in \mathbb{R}^{n\times p} \sim \mathcal{N}(0, \Sigma(\rho))$, with $\Sigma(\rho)$ a $(p \times p)$ Toeplitz covariance matrix (Mukherjee and Maiti, 1988) with correlation $\rho \in (0, 1)$, such that $\Sigma(\rho)_{i,j} = \rho^{|i-j|}$. For each feature $j = 1, \dots, p$, we sample the cut-points $\mu^\star_{jk}$ uniformly without replacement from the estimated quantiles $q_j(u/10)$ for $u = 1, \dots, 9$ and $k = 1, \dots, K^\star_j$. 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 $j = 1, \dots, p$. 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 $c_{jk} \sim (-1)^k |\mathcal{N}(1, 0.5)|$ for all $k = 1, \dots, K^\star_j + 1$ and $j = 1, \dots, p$ to make sure we create "real" cut-points, and take

$$\beta^\star_{jk} = c_{jk} - (K^\star_j + 1)^{-1} \sum_{k=1}^{K^\star_j+1} c_{jk}$$

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, that is, features for which we enforce $\beta^\star_{jk} = 0$ for all $k = 1, \dots, K^\star_j + 1$. Finally, we generate survival times using Weibull distributions, which is a common choice in survival analysis (Klein and Moeschberger, 2005): $T_i \sim \nu^{-1}[-\log(U_i)\exp(-(x_i^{B^\star})^\top \beta^\star_i)]^{1/\varsigma}$ 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 1 gives an example of data generated according to the design we have just described.

TABLE 1 Hyper-parameter choices for simulation

n	p	ρ	$K^\star_j$	ν	ς	r_c	r_s
[200, 4000]	[2, 100]	0.5	{1, 2, 3}	2	0.1	0.3	0.2

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

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

where $\mathcal{M}^\star_j = \{\mu^\star_{j,1}, \dots, \mu^\star_{j,K^\star_j}\}$ (respectively, $\widehat{\mathcal{M}}_j = \{\hat{\mu}_{j,1}, \dots, \hat{\mu}_{j,\widehat{K}_j}\}$) is the set of true (respectively, estimated) cut-points for feature j, $\mathcal{S}' = \{j,\ j \notin \mathcal{S} \cap \{l,\ \widehat{\mathcal{M}}_l = \emptyset\}\}$ 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 $\mathcal{H}(A, B) = \max(\mathcal{E}(A||B), \mathcal{E}(B||A))$, 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,

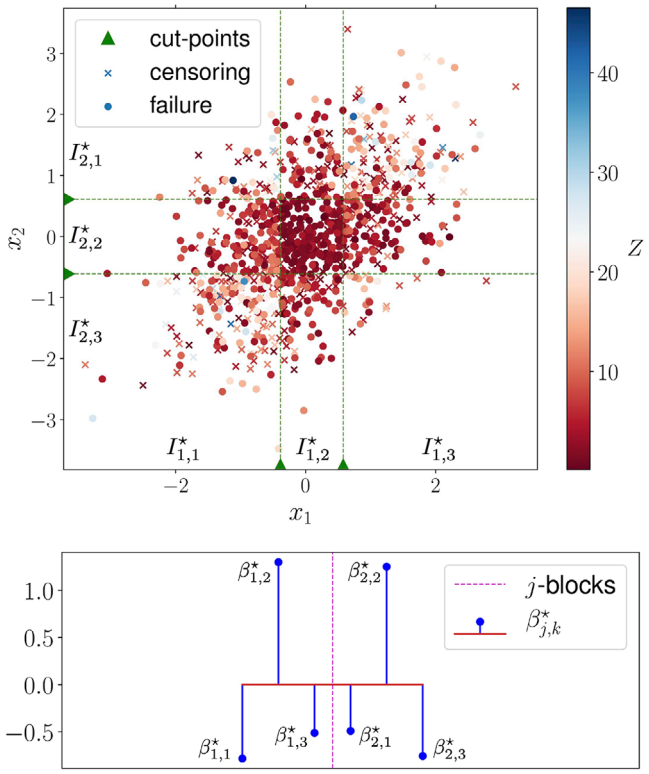

FIGURE 1 Top: illustration of data simulated with $p = 2$, $K^\star_1 = K^\star_2 = 2$, and $n = 1000$. Dots represent failure times ($z_i = t_i$) while crosses represent censoring times ($z_i = c_i$), and the color gradient represents the z_i values (red for low and blue for high). Bottom: $\beta^\star$ is plotted, with a dotted line to demarcate the two blocks (since $p = 2$). To each interval, $I^\star_{j,k}$ corresponds an effect $\beta^\star_{j,k}$. Note that both $\beta^\star_{1,2}$ and $\beta^\star_{2,2}$ are high, so all subjects having $x_1 \in I^\star_{1,2}$ and $x_2 \in I^\star_{2,2}$ at the same time have a high risk for the event to occur quickly. This induces the dark red dots in the small square in the center of the figure. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

both $\mathcal{M}^\star_j$ 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

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

The state-of-the-art competing methods for automatic cut-points detection in a survival analysis context are based on multiple log-rank tests ("MT" for "Multiple testing"), and we consider the Bonferroni correction (see Bland and Altman (1995)), denoted "MT-B," as well as a correction proposed in Lausen and Schumacher (1992), denoted "MT-LS." A precise description of these methods is given in Appendix A.2 of the Supplementary material.

4.2 | Simulation results

Figure 2 illustrates how the methods considered behave on the data shown in Figure 1. 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 two features, namely those corresponding to the highest jumps in $\beta^\star_{j,\bullet}$ (see Figure 1): $\mu^\star_{1,1}$ 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).

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 competing methods, 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 3(a), and compare its computing times with the binacox method 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 method is still roughly five times faster than the "MT grid" method, and it remains very fast when we increase the dimension, as shown in Figure 3(b). It turns out that the computational time grows roughly logarithmically with p.

Let us compare now the results of simulations in terms of the m_1 and m_2 metrics introduced in Section 4.1. Figure 4 gives a comparison of the methods considered for the cut-point estimation aspect, that is, in terms of the m_1 score. It appears that the binacox method outperforms the MT-related methods when $K^\star_j > 1$, and is competitive

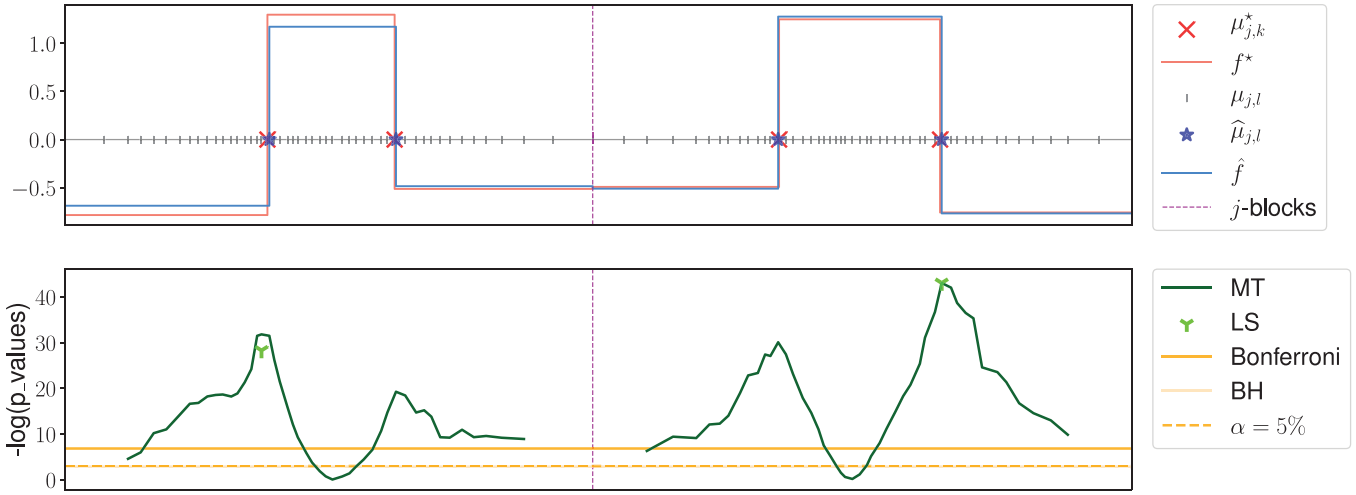

FIGURE 2 Top: Illustration of the main quantities involved in the binacox method, with estimations obtained for the data represented in Figure 1. Our algorithm detects the correct number of cut-points $\widehat{K}_j = 2$, and estimates their positions accurately, as well as their amplitudes. Bottom: results obtained using the multiple testing-related methods introduced in Appendix A.2 of the Supplementary material. Here the Benjamini–Hochberg (BH) (Benjamini and Hochberg, 1995) 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. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

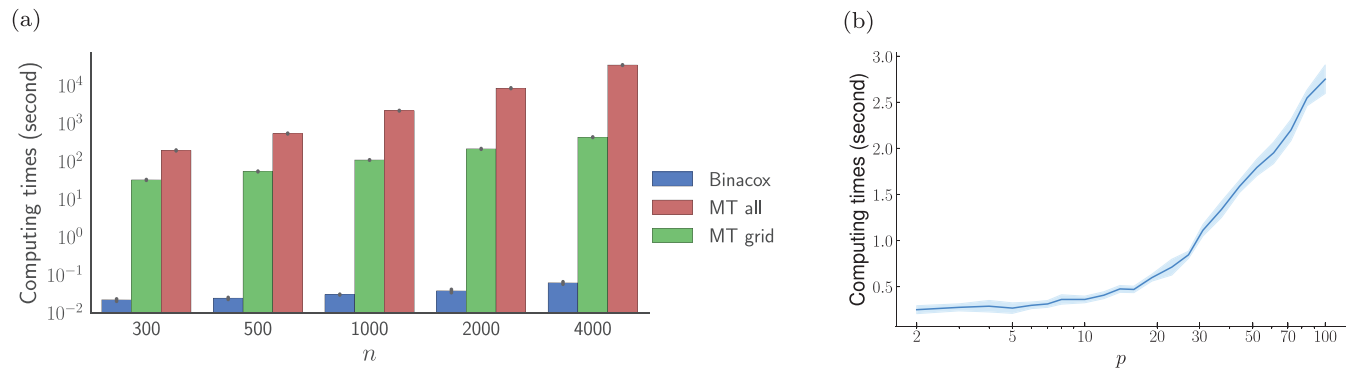

FIGURE 3 Computing time for the methods considered. This figure appears in color in the electronic version of this article, and any mention of color refers to that version (A) Average computing times in seconds (with the black lines representing ± the standard deviation) obtained on 100 simulated datasets (according to Section 4.1 with $p = 1$ and $K^\star = 2$) for training the binacox method 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 method ("MT grid"). (B) Average (bold) computing times in seconds and standard deviation (bands) obtained on 100 simulated datasets (according to Section 4.1 with $K^\star_j = 2$) for training the binacox method when increasing the dimension p up to 100. The method remains very fast in high-dimensional settings

when $K^\star_j = 1$ except for small values of n. This is due to an overestimation in the number of cut-points by the binacox method (see Figure 5), 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 method runs much faster than the MT-related methods.

Figure 5, on the other hand, assesses the ability of each method to detect features with no cut-points using the m_2 metric, that is, the ability to estimate $\hat{K}^\star_j = 0$ for $j \in S$. The binacox method 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.

5 | APPLICATION ON GENETIC DATA

In this section, we apply our method to three biomedical data sets. We extracted normalized expression data and survival times Z in days from breast invasive carcinoma (BRCA, $n = 1211$), glioblastoma multiforme (GBM,

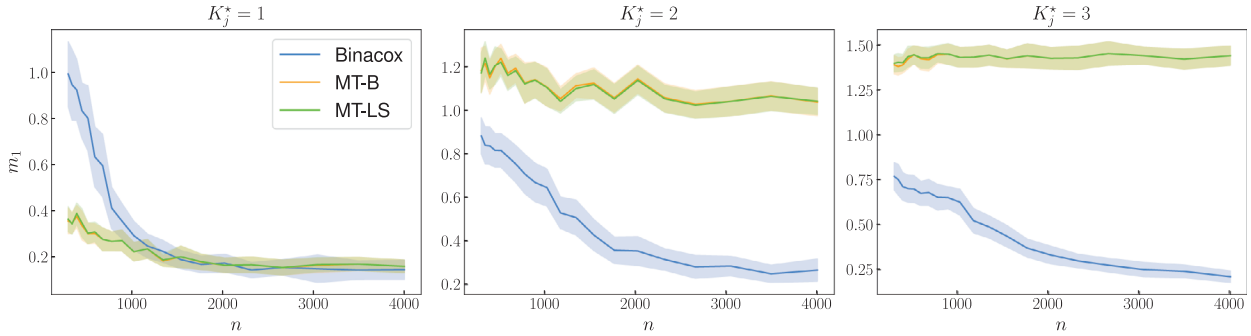

FIGURE 4 Average (bold) m_1 scores and standard deviation (bands) obtained on 100 data sets simulated according to Section 4.1 with $p = 50$ and $K_j^\star$ equal to 1, 2, and 3 (for all $j = 1, \dots, p$) for the left, center, and right sub-figures, respectively) for varying n. The lower the value of m_1, the better the result; the binacox method 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 $K_j^\star$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

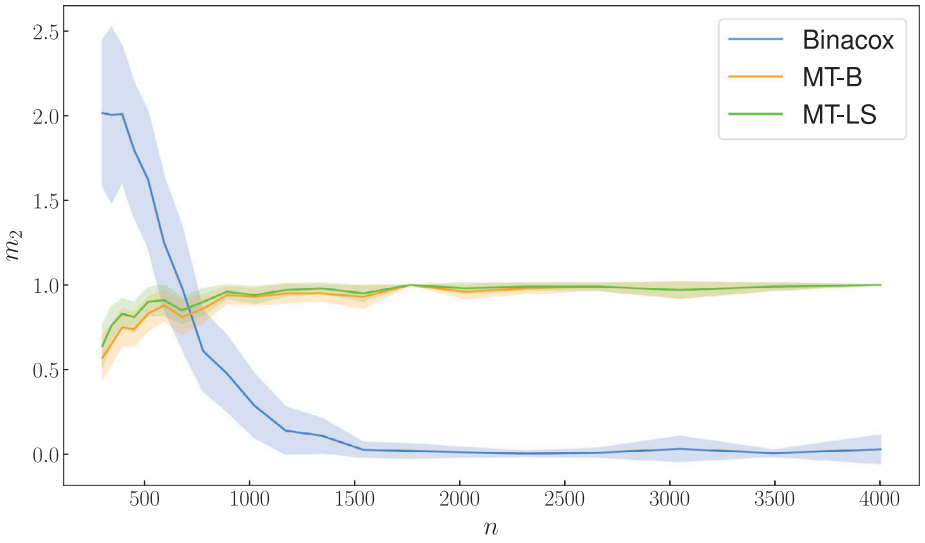

FIGURE 5 Average (bold) m_2 scores and standard deviation (bands) obtained on 100 data sets simulated according to Section 4.1 with $p = 50$ for varying n. MT-B and MT-LS tend to detect a cut-point when there is none (no matter the value of n), while binacox overestimates the number of cut-points for small values of n but detects S well for $p = 50$ on the simulated data when $n > 1000$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

$n = 168$), and kidney renal clear cell carcinoma (KIRC, $n = 605$). These data sets are available on *The Cancer Genome Atlas* (TCGA) 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 4.2, 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. The screening step is used for all competing models. To do so, we fit the binacox method on each jth block separately and take the resulting $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ as a score that roughly assesses 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 6 of Appendix A.8.1 in the Supplementary Material.

5.1 | Estimation results

In Figure 6, we present the results obtained by the methods considered on the GBM cancer data set 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 MT-B or MT-LS are also detected by the multivariate binacox (which detects more cut-points); see Table 2. Furthermore, it turns out that these top 10

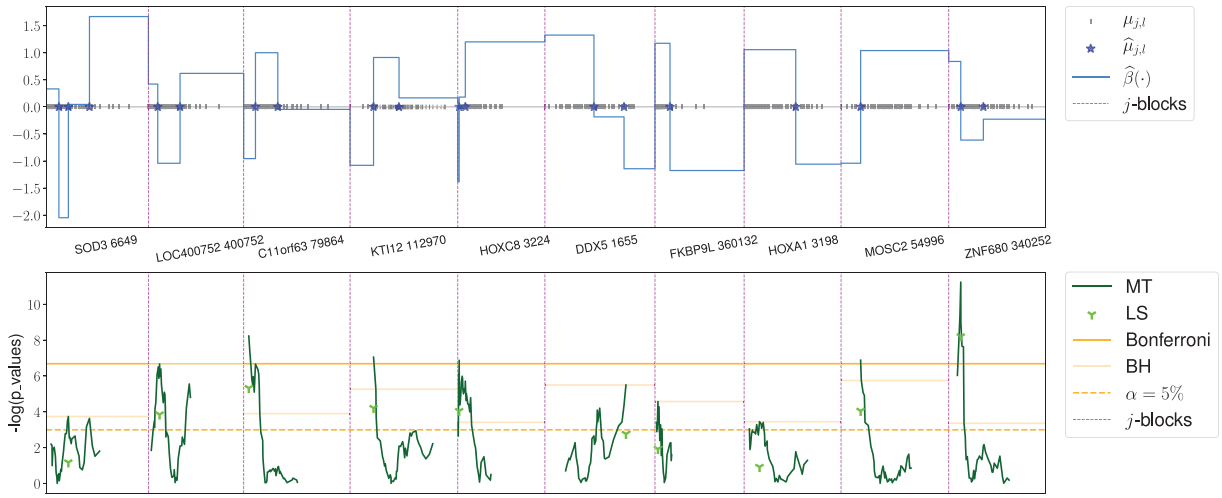

FIGURE 6 Illustration of the results obtained on the top 10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the GBM dataset. The binacox method detects multiple cut-points and sheds light on nonlinear effects for various genes. The BH thresholds are shown, but are unusable in practice. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

TABLE 2 Estimated cut-point values for each method on the top 10 genes presented in Figure 6 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	10,630.11, 13,094.89	·	·
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

genes are 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).

Relevant results were also obtained on the KIRC and BRCA data sets; these are postponed to Appendix A.8.2 in the Supplementary Material.

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 a Cox model on the original continuous feature space versus on the $\widehat{\mu}$-binarized space constructed with the cut-point estimates. We randomly split the three data sets 100 times into training and validation sets (30% for testing) and compare the average C-index on the validation sets in Table 3 when the $\widehat{\mu}$-binarized space is constructed based on the $\widehat{\mu}$'s obtained either from the binacox method, 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) and random survival forests (RSF) (Ishwaran *et al.*, 2008).

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 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). See also Appendix A.8 of the Supplementary Material for additional details and results on computing times.

6 | DISCUSSION

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

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 data sets, the binacox method gives the best results (in bold)

Cancer	Continuous	Binacox	MT-B	MT-LS	CoxBoost	RSF
GBM	0.567 (0.042)	**0.602 (0.050)**	0.574 (0.052)	0.573 (0.048)	0.574 (0.043)	0.568 (0.045)
KIRC	0.669 (0.032)	**0.702 (0.031)**	0.672 (0.034)	0.672 (0.034)	0.679 (0.030)	0.699 (0.034)
BRCA	0.586 (0.052)	**0.671 (0.044)**	0.630 (0.058)	0.626 (0.053)	0.592 (0.054)	0.654 (0.045)

method, it succeeds in detecting multiple cut-points per feature. We also applied binacox to three publicly available high-dimensional genetics data sets. Furthermore, several genes pinpointed by the model turn out to be biologically relevant, while others require further investigation in the genetics research community. More importantly, our method provides powerful and innovative 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 that 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 authors thank professor J.P. Vert for his support and the interesting discussions about selection bias.

CONFLICT OF INTEREST

The authors have no relevant conflicts of interest to disclose.

OPEN RESEARCH BADGES

This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at https://github.com/jjfeng/aACP.

DATA AVAILABILITY STATEMENT

The results shown in this paper are based on data generated by the Cancer Genome Atlas (TCGA) Research Network and freely available at http://cancergenome.nih.gov.

ORCID

Simon Bussy https://orcid.org/0000-0002-6059-0685

REFERENCES

Alaya, M.Z., Bussy, S., Gaıffas, S.& Guilloux, A. (2019) Binarsity: a penalization for one-hot encoded features in linear supervised learning. *Journal of Machine Learning Research*, 20, 1–34.

Andersen, P.K., Borgan, Ø., Gill, R.D. & Keiding, N. (2012) *Statistical Models Based on Counting Processes*. Amsterdam: Springer Science & Business Media.

Bach, F. (2010) Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4, 384–414.

Bacry, E., Bompaire, M., Deegan, P., Gaïffas, S.& Poulsen, S.V. (2017) Tick: a python library for statistical learning, with an emphasis on Hawkes processes and time-dependent models. *Journal of Machine Learning Research*, 18, 7937–7941.

Bédat, B., Niclauss, N., Jannot, A.-S., Andres, A., Toso, C., Morel, P. et al. (2015) Impact of recipient body mass index on short-term and long-term survival of pancreatic grafts. *Transplantation*, 99, 94–99.

Benjamini, Y.& Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57, 289–300.

Bland, J.M.& Altman, D.G. (1995) Multiple significance tests: the Bonferroni method. *BMJ*, 310, 170.

Bleakley, K. & Vert, J.P. (2011) The group fused Lasso for multiple change-point detection. Technical Report HAL-00602121.

Chang, C., Hsieh, M.-K., Chiang, A.J., Tsai, Y.-H., Liu, C.-C.& Chen, J. (2019) 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*, 34, 1649–1674.

Cheang, M.C.U., Chia, S.K., Voduc, D., Gao, D., Leung, S., Snider, J. et al. (2009) Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *JNCI: Journal of the National Cancer Institute*, 101, 736–750.

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

Contal, C.& O'Quigley, J. (1999) An application of changepoint methods in studying the effect of age on survival in breast cancer. *Computational Statistics & Data Analysis*, 30, 253–270.

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

Curley, A., Stanworth, S.J., Willoughby, K., Fustolo-Gunnink, S.F., Venkatesh, V., Hudson, C. et al. (2019) Randomized trial of

platelet-transfusion thresholds in neonates. *New England Journal of Medicine*, 380, 242–251.

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

Harchaoui, Z.& Lévy-Leduc, C. (2010) Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105, 1480–1493.

Huang, J., Sun, T., Ying, Z., Yu, Y.& Zhang, C.H. (2013) Oracle inequalities for the lasso in the Cox model. *The Annals of Statistics*, 41, 1142–1165.

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

Ishwaran, H., Kogalur, U.B., Blackstone, E.H.& Lauer, M.S. (2008) Random survival forests. *The Annals of Applied Statistics*, 2, 841–860.

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

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

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

Kong, S.& Nan, B. (2014) Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. *Statistica Sinica*, 24, 25.

Lausen, B.& Schumacher, M. (1992) Maximally selected rank statistics. *Biometrics*, 48, 73–85.

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

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

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

Motzer, R.J., Mazumdar, M., Bacik, J., Berg, W., Amsterdam, A.& Ferrara, J. (1999) Survival and prognostic stratification of 670 patients with advanced renal cell carcinoma. *Journal of Clinical Oncology*, 17, 2530–2540.

Mukherjee, B.N.& Maiti, S.S. (1988) On some properties of positive definite Toeplitz matrices and their possible applications. *Linear Algebra and its Applications*, 102, 211–240.

Rajaraman, P., Hutchinson, A., Rothman, N., Black, P.M., Fine, H.A., Loeffler, J.S. et al. (2008) Oxidative response gene polymorphisms and risk of adult brain tumors. *Neuro-Oncology*, 10, 709–715.

Rota, M., Antolini, L.& Valsecchi, M.G. (2015) Optimal cut-point definition in biomarkers: the case of censored failure time outcome. *BMC Medical Research Methodology*, 15, 24.

Senoussi, R. (1990) Problème d'identification dans le modèle de Cox. *Annales de l'Institut Henri Poincaré Probabilities et Statistiques*, 26, 45–64.

Simon, N., Friedman, J., Hastie, T.& Tibshirani, R. (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39, 1–13.

Van de Geer, S.& Bühlmann, P. (2009) On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3, 1360–1392.

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

Wells, P.S., Anderson, D.R., Rodger, M., Ginsberg, J.S., Kearon, C., Gent, M. et al. (2000) 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, 416–420.

Wu, J. & Coggeshall, S. (2012) *Foundations of predictive analytics*, 1st edition. Boca Raton, FL: CRC Press.

Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J.E., Song, C. et al. (2017) Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, 7, 11707.

SUPPORTING INFORMATION

Web Appendices referenced in Section 2– 5 are available with this paper at the Biometrics website on Wiley Online Library. The includes details on the algorithm implementation, additional results on genetic data, and the proofs of the theoretical results. Moreover, all methodology discussed in the paper is implemented in `Python/C++` and `R` and also available online at the Biometrics website, with the code that generates all figures (also open-sourced at https://github.com/SimonBussy/binacox) in the form of annotated programs, together with notebook tutorials. The binacox method is implemented in the `tick` library (Bacry *et al.*, 2017).

How to cite this article: Bussy, S., Alaya, M.Z., Jannot, A.-S., Guilloux, A. Binacox: automatic cut-point detection in high-dimensional Cox model with applications in genetics. *Biometrics*. 2021;1–13. https://doi.org/10.1111/biom.13547

Supporting Information for Binacox: automatic cut-point detection in high-dimensional Cox model with applications in genetics by Simon Bussy, Mokhtar Z. Alaya, Anne-Sophie Jannot and Agathe Guilloux

Appendix A Additional details

A.1 Notations

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.

A.2 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 $\boldsymbol{X}_{\bullet,j} = (x_{1,j}, \dots, x_{n,j})^\top$, and denote its 10th and 90th quantiles as $x_{10th,j}$ and $x_{90th,j}$. Then, define a grid $\{g_{j,1}, \dots, g_{j,\kappa_j}\}$. 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 2, the values $-\log(\mathrm{pv}_{j,k})$ for $k = 1, \dots, \kappa_j$ (denoted by "MT" for "Multiple Testing") are represented, for the simulated example illustrated in Figure 1. 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 (Bland and Altman, 1995), 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) (Benjamini and Hochberg, 1995) 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 2 also illustrates how these corrections behave on the simulated example illustrated in Figure 1. 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 3(a) that compares the computing times for a single feature only; a bootstrap procedure based on MT would dramatically increase the required computing time).

A.3 Oracle inequality for estimation

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

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

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 (13), 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

$$\tilde{f}_j : x \mapsto \sum_{k=1}^{K_j^\star+1} \beta_{j,k}^\star \sum_{l=l_{j,k-1}^\star+1}^{l_{j,k}^\star} \mathbb{1}(x \in I_{j,l}), \tag{2}$$

for $x \in [0,1]$, where $l_{j,k}^\star$ is defined in (1), 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 D.

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

$$f_{b^\star_{j,\bullet}} : x \mapsto \tilde{f}_j(x) - \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_j(X_{i,j}) \tag{3}$$

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

Denoting $b^\star = \big((b^\star_{1,\bullet})^\top, \dots, (b^\star_{p,\bullet})^\top\big)^\top$, 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 2 in Section 4.2. Note that, in addition, if $\beta^\star$ is block-sparse, then it is also the case for $b^\star$, and the following

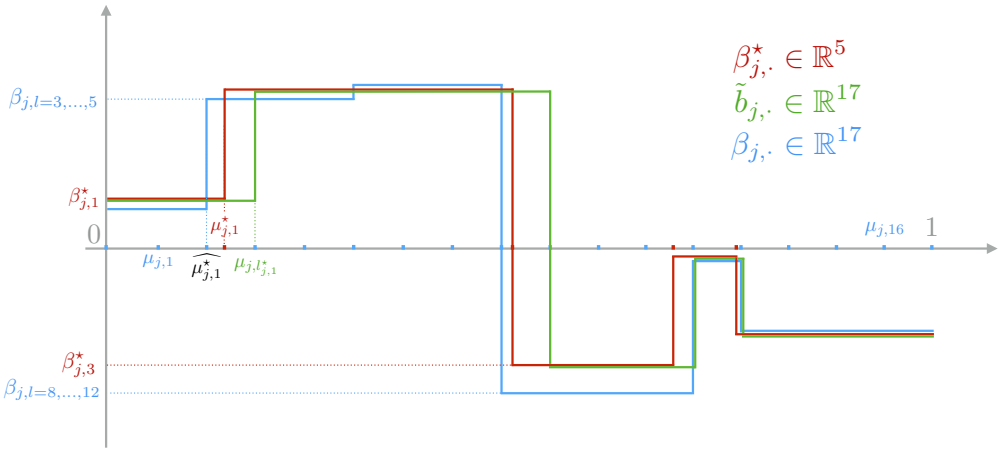

Figure 1: Illustration of the different vectors for the jth block, with $d_j = 17$. In this scenario, the algorithm detects an extra cut-point and $\widehat{K}_j = 5 = s_j$, while $K^\star_j = 4$.

holds:

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

Let us introduce some further notation. We define

$$\pi_n = \frac{|\{i = 1, \dots, n : N_i(\tau) = 1\}|}{n}, \tag{4}$$

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

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

Theorem 3 *The inequality*

$$\|(\hat{\beta} - b^\star)_{\mathcal{A}(b^\star)}\|_1 \leq \frac{\sqrt{K^\star(\mathbf{I}+\mathbf{II})}}{\kappa_\tau\big(\mathcal{A}(b^\star)\big)} \tag{5}$$

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 3 is presented in Appendix D. 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 3 is then given, for n and d_j large enough, by

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

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

A.4 Generalized additive model with Lipschitz components

We study here the case where the true $f^\star$ has a generalized additive model structure and where its components fulfill a Lipschitz assumption.

Assumption 1 *Assume that* $f^\star$ *has the following sparse additive structure* $f^\star(X_i) = \sum_{j\in\mathscr{A}^\star} f_j^\star(X_{i,j})$ *where* $f_j^\star : [0,1] \to \mathbb{R}$ *are* $L-$*Lipschitz functions, namely satisfying* $|f_j^\star(x) - f_j^\star(x')| \leq L|x - x'|$, *for any* $x, x' \in [0,1]$, *and where* $\mathscr{A}^\star \subset \{1,\dots,p\}$ *is a set of active features such that* $|\mathscr{A}^\star| \ll p$.

In that case, the bias term can be upper bounded and we can derive the following rate of convergence.

Theorem 4 *Let Assumption 1 holds and* $d_j = D$ *where* D *is the integer part of*

$$\sqrt{\frac{6L\max_{1\leq i\leq n}\int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t}{2C_{lip}}} \cdot n^{3/4}$$

for some positive constant C_{lip} *and* $I_{j,1} = [0, \frac{1}{D+1}], I_{j,l} = (\frac{l-1}{D+1}, \frac{l}{D+1}]$. *Then the following holds*

$$KL_n(f^\star, f_{\hat{\beta}}) \leq \Big(6L\max_{1\leq i\leq n}\int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t + 2C_{lip}\Big)\frac{|\mathscr{A}^\star|}{n^{1/4}}.$$

with the same probability as in Theorem 1.

The proof of Theorem 4 is postponed to Section E.

A.5 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. (2019), with an illustration of the previous phenomenon given in Figure 5).

Note also that the binacox method 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

```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
```

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

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 (3) 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 4 gives the learning curves obtained with this cross-validation procedure on an example. Note also that one could consider data-driven weights based on Equation (6), but this would require non-trivial technicalities in the coding of the proximal operator, which is beyond the scope of this paper.

Refitting strategies are classical and well studied for the lasso penalty (Chzhen et al., 2019; Lederer, 2013), addressing problems regarding bias of the lasso selection in high-dimension (Zhang et al., 2008). The `linearmodel.OLS.fit_regularized` function of the `python` package `statsmodels` (Seabold and Perktold, 2010) allows for instance to refit the model using only the variables that have non-zero coefficients in the regularized fit. The refitted model is not regularized.

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 5 in Appendix A.7 illustrates

this routine.

Then, in Banerjee et al. (2007), the authors consider the univariate regression problem and propose to approximate the regression function by a piecewise constant function with a single cut-point, using a decision tree estimator with a single terminal node. They prove that the cut-point estimate converges to the best possible cut-point with a $n^{1/3}$ rate. We then added an empirical investigation of the standard error of the estimated cut-points in our case. Hence, we plot in Figure 3 the standard error of the m_1 score (being adapted to our problem) based on experiments from Figure 4.

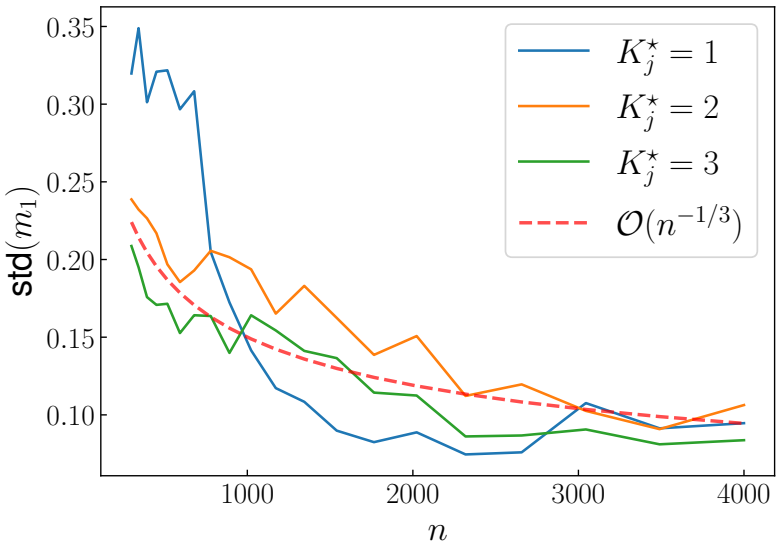

Figure 3: Standard error of the m_1 score obtained on 100 datasets simulated according to Section 5.1 with $p = 50$ and $K_j^\star$ equal to 1, 2 and 3 (for all $j = 1, \dots, p$) for varying n. The red dashed line represents the $n^{-1/3}$ speed.

A.6 Algorithm.

To solve regularization problem (4), we first look at the proximal operator of the binarsity penalty (Alaya et al., 2019). 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.

A.7 Implementation

Figure 4 gives the learning curves obtained during the V-fold cross-validation procedure presented in Section A.2 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 (3) 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.

Algorithm 1 Proximal operator of bina(β), see (Alaya et al., 2019)

Input: vector $\beta \in \mathscr{B}_{p+d}(R)$ and weights $\omega_{j,l}$ for $j = 1, \dots, p$ and $l = 1, \dots, d_j + 1$
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 (6))
  $\eta_{j,\bullet} \leftarrow \theta_{j,\bullet} - \frac{n_{j,\bullet}^\top \theta_{j,\bullet}}{\|n_{j,\bullet}\|_2^2} n_{j,\bullet}$ (projection onto $\mathrm{span}(n_{j,\bullet})^\perp$)
end for
Return: η

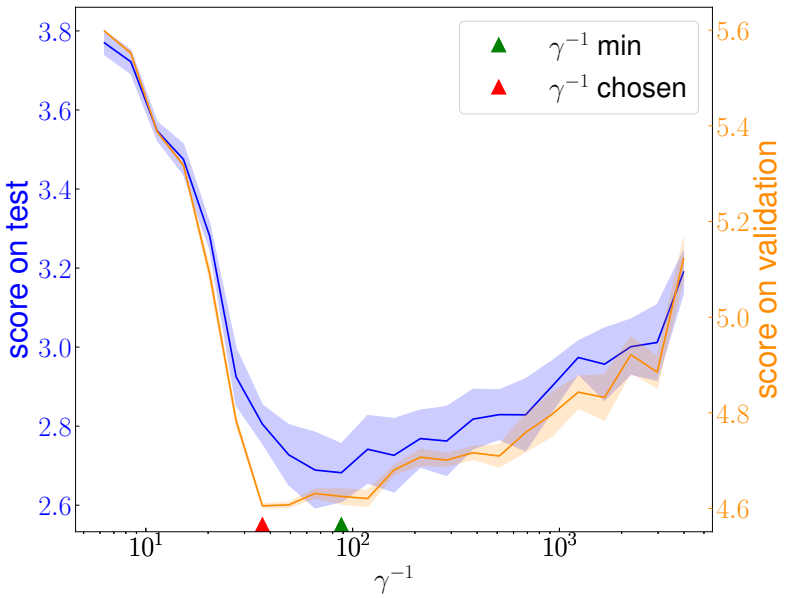

Figure 4: Learning curves obtained for various γ, in blue on the different test sets during cross-validation, and in orange on the validation set. Bold lines represent average scores on the folds, and bands represent 95% Gaussian confidence intervals. The green triangle points out the value of γ^{-1} 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.

Figure 5 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 1 of Section 4.1, where the final estimation results were presented in Figure 2 of Section 4.2. 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 5 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.

Figure 5: Illustration of the de-noising step in the cut-point detection phase on the simulated date of Figure 1. 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.8 Other results on TCGA data

This section gives additional details concerning Section 5 of the paper. Let us first recall that in a classical Cox model, $R_i = \exp(X_i^\top \hat{\beta})$ is known as the risk score for patient i (Therneau and Grambsch, 2000). 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`.

Then, in Section 5.2, CoxBoost is used with 300 boosting steps (this number being fine-tuned by cross-validation), and the RSF are used with 200 trees and the number of random splits to consider for each candidate splitting variable set to 10 (also cross-validated, while the number of variables randomly selected as candidates for splitting a node is set to $\sqrt{p}$), 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.

A.8.1 Gene screening

Figure 6 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.

A.8.2 Results on BRCA and KIRC data

Before giving the results obtained on the BRCA and KIRC data, let us precise that some genes in the top 10 selection for GBM (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 (Guan et al., 2019).

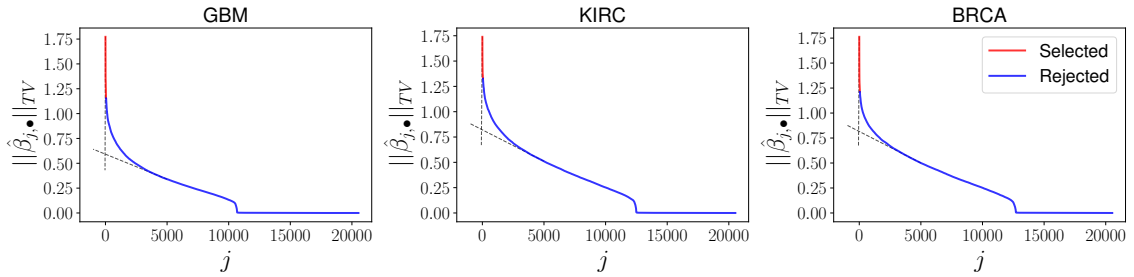

Figure 6: $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ obtained for univariate binacox fits for the three datasets considered. The top P selected features appear in red, and it turns out that taking $P = 50$ coincides with the elbow (represented with the dotted grey lines) in each of the three curves.

Figure 7 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 1 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.

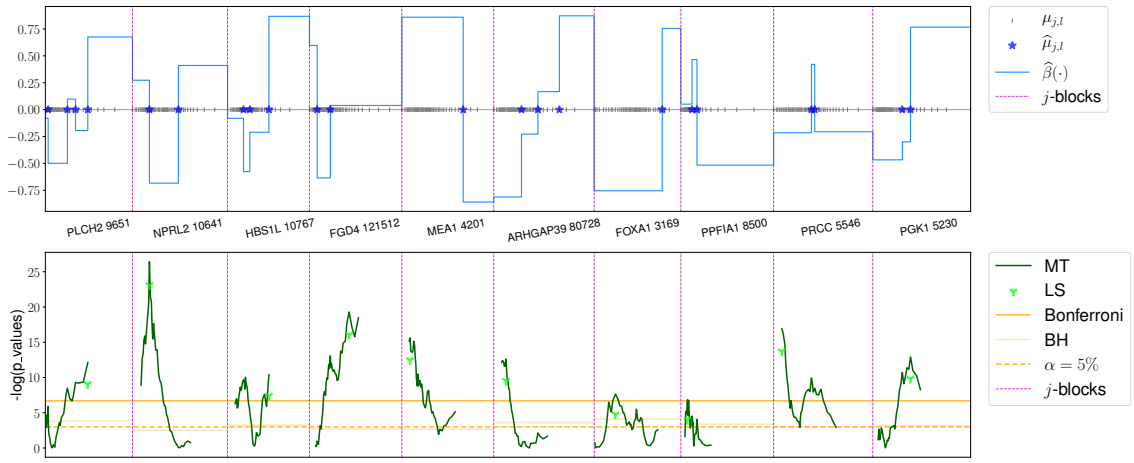

Figure 7: Illustration of the results obtained on the top 10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the BRCA dataset.

Lastly, Figure 8 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 2 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) .

Table 1: Estimated cut-point values for each method on the top 10 genes presented in Figure 7 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 2: Estimated cut-point values for each method on the top 10 genes illustrated in Figure 8 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

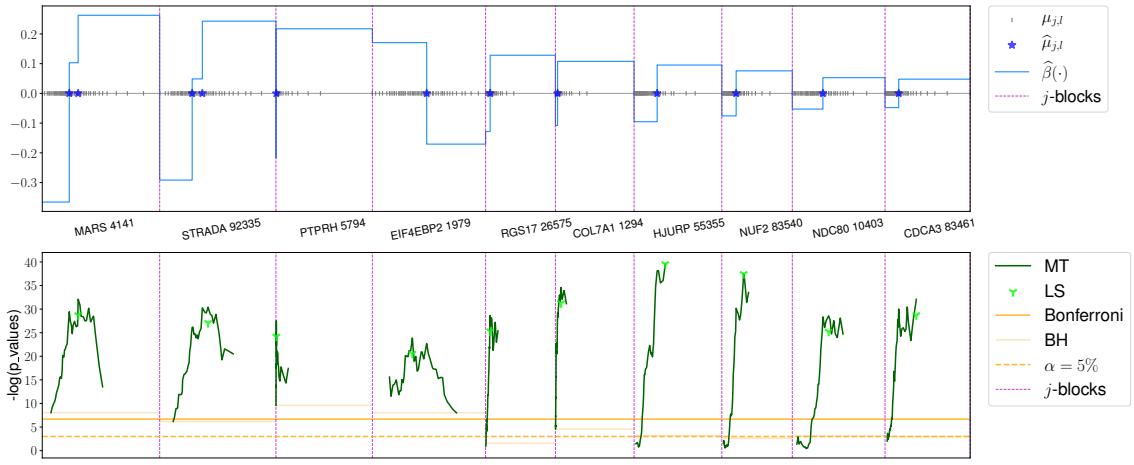

Figure 8: Illustration of the results obtained on the top 10 features ordered according to the binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the KIRC dataset.

A.8.3 Computing times

Figure 9 below compares the computing times of the considered methods on the TCGA data. Clearly the binacox method is by far the most computationally efficient.

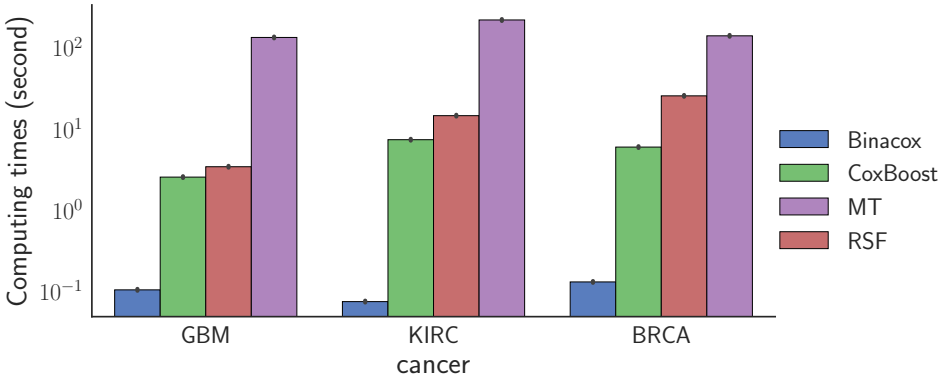

Figure 9: Average computing times (in seconds) required by each method on the three datasets (with the black lines representing $\pm$ the standard deviation) obtained on 100 random train/test split. The binacox method is at least one and up to several orders of magnitude faster.

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

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 $u \mapsto |u|$, 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.

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

$$\omega_{j,l} = 11.32\sqrt{\frac{c + \log(p+d) + \mathscr{L}_{n,c}}{n}\hat{V}_{j,l}} + 18.62\frac{c + 1 + \log(p+d) + \mathscr{L}_{n,c}}{n}, \tag{6}$$

where

$$\mathscr{L}_{n,c} = 2\log\log\frac{2n\hat{V}_{j,l} + 18,66e(c + \log(p+d))}{8}.$$

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

$$\mathbf{D} = \mathrm{diag}(D_1, \ldots, D_p) \quad \text{and} \quad \mathbf{T} = \mathrm{diag}(T_1, \ldots, T_p). \tag{7}$$

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*

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

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 $\delta_j(\beta_{j,\bullet}) + \delta_j(-\beta_{j,\bullet}) \geq 0$ to obtain:

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

which concludes the proof of Lemma 1. □

The Doob-Meyer decomposition (Aalen, 1978) implies that, for all $i = 1, \dots, n$ and all $t \geq 0$,

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

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

$$\begin{aligned}\nabla\ell_n(f_\beta) &= -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big\{X_i^B - \frac{S_n^{(1)}(f_\beta,t)}{S_n^{(0)}(f_\beta,t)}\Big\}\mathrm{d}N_i(t) \\ &= \nabla\ell(f_\beta) + H_n(f_\beta), \end{aligned} \tag{8}$$

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

$$H_n(f_\beta) = -\frac{1}{n}\sum_{i=1}^n \int_0^\tau \Big\{X_i^B - \frac{S_n^{(1)}(f_\beta,t)}{S_n^{(0)}(f_\beta,t)}\Big\}\mathrm{d}M_i(t). \tag{9}$$

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

$$\|f\|_n^2 = \int_0^\tau \sum_{i=1}^n \big(f(X_i) - \bar{f}(t)\big)^2 \frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)}\mathrm{d}\bar{N}(t), \tag{10}$$

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 $h = (h_{1,\bullet}^\top, \ldots, h_{p,\bullet}^\top)^\top$ with $h_{j,\bullet} \in \partial\big(\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}\big)$ for all $j = 1, \ldots, p$. Then the following holds:*

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

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 (10) to our empirical divergence defined in (8).

Lemma 3 *Let $\hat{\beta}$ be defined by Equation (4) and $\beta \in \mathscr{B}_{p+d}(R)$. Then the following inequalities hold almost surely:*

$$KL_n(f^\star, f_\beta) - KL_n(f^\star, f_{\hat{\beta}}) + (\hat{\beta} - \beta)^\top \nabla \ell(f_{\hat{\beta}}) \geq 0, \tag{11}$$

and

$$\|f^\star - f_\beta\|_n^2 \frac{\psi(-\|f^\star - f_\beta\|_\infty)}{\|f^\star - f_\beta\|_\infty^2} \leq KL_n(f^\star, f_\beta) \leq \|f^\star - f_\beta\|_n^2 \frac{\psi(\|f^\star - f_\beta\|_\infty)}{\|f^\star - f_\beta\|_\infty^2}, \tag{12}$$

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 $L = [L_1, \ldots, L_p]$ a concatenation of index subsets. Set for all $j = 1, \ldots, p$,*

$$L_j = \{a_j^1, \ldots, a_j^{b_j}\} \subset \{1, \ldots, d_j + 1\}, \tag{13}$$

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

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

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

and

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

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*

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

B.3 Proof of Theorem 1

Combining Lemmas 2 and 3, we get

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

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 $-(\hat\beta-\beta)^\top H_n(f_{\hat\beta}) - (\hat\beta-\beta)^\top h \ge 0$.

Let us derive now a 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

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

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

$$\begin{aligned}
-(\hat\beta-\beta)^\top h &\le 2\sum_{j=1}^p \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)} \odot D_j(\hat\beta_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_1 \\
&\quad - 2\sum_{j=1}^p \|(\omega_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)} \odot D_j(\hat\beta_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_1 \\
&= 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}} - 2\sum_{j=1}^p \|(\hat\beta_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}. \qquad (14)
\end{aligned}$$

Inequality (14) therefore gives

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

Using the fact that $\mathbf{TD} = \mathbf{I}$ (see their definitions in Equation (7)), we get

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

On the event

$$\mathscr{E}_n := \Big\{ \bigcap_{j=1}^p \bigcap_{l=1}^{d_j+1} |(\mathbf{T}^\top H_n(f_{\hat\beta}))_{j,l}| \le \omega_{j,l} \Big\} \qquad (15)$$

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

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

Hence,

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

One therefore has

$$KL_n(f^\star, f_{\hat\beta}) \le KL_n(f^\star, f_\beta) + 3\sum_{j=1}^p \|(\hat\beta_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\mathrm{TV},\omega_{j,\bullet}}. \qquad (16)$$

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

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

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 (16), by Lemma 4 and under Assumption 2, we get

$$KL_n(f^\star, f_{\hat{\beta}}) \leq KL_n(f^\star, f_\beta) + \frac{\|f_{\hat{\beta}} - f_\beta\|_n}{\sqrt{\kappa_\tau^2\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)}\kappa_{\mathbf{T},\hat{\zeta}}\big(\mathcal{A}(\beta)\big)}, \tag{17}$$

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 (17) fulfills

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

By (12) in Lemma 3, we get that

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

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

$$g(x) \leq x + 2 \text{ for any } x > 0. \tag{18}$$

Then

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

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

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

This implies that

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

With these bounds, inequality (17) yields

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

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

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

Hence

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

By choosing $\varrho=2$, we obtain

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

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

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

Finally,

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

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

Concerning the computation of $\mathbb{P}[\mathscr{E}_n^\complement]$ now. From the definition of H_n in Equation (9), $\mathbf{T}^\top H_n(f_{\hat\beta})$ 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 $j=1,\dots,p$. 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 $t\le\tau$, one has

$$\begin{aligned}
&\Big|\Big(\mathbf{T}^\top X_i^B - \mathbf{T}^\top\frac{S_n^{(1)}(f_{\hat\beta},t)}{S_n^{(0)}(f_{\hat\beta},t)}\Big)_{j,l}\Big| \\
&\quad= \Big|\Big(\mathbf{T}^\top X_i^B - \mathbf{T}^\top\frac{Y_i(t)e^{f_{\hat\beta}(X_i)}X_i^B}{\sum_{i'=1}^n Y_{i'}(t)e^{f_{\hat\beta}(X_{i'})}} - \mathbf{T}^\top\frac{\sum_{i'\neq i}^n Y_{i'}(t)e^{f_{\hat\beta}(X_{i'})}X_{i'}^B}{\sum_{i'=1}^n Y_{i'}(t)e^{f_{\hat\beta}(X_{i'})}}\Big)_{j,l}\Big| \\
&\quad\le \Big|\Big(1-\frac{Y_i(t)e^{f_{\hat\beta}(X_i)}X_i^B}{\sum_{i'=1}^n Y_{i'}(t)e^{f_{\hat\beta}(X_{i'})}}\Big)(\mathbf{T}^\top X_i^B)_{j,l}\Big| + \Big|\frac{\sum_{i'\neq i}^n Y_{i'}(t)e^{f_{\hat\beta}(X_{i'})}}{\sum_{i'=1}^n Y_{i'}(t)e^{f_{\hat\beta}(X_{i'})}}(\mathbf{T}^\top X_{i'}^B)_{j,l}\Big| \\
&\quad\le (\mathbf{T}^\top X_i^B)_{j,l}+1.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{P}[\mathscr{E}_n^c] &\leq \sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{(\mathbf{T}^\top X_i^B)_{j,l}+1\right\}\mathrm{d}M_i(t) > \omega_{j,l}\right] \\
&\leq \sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (\mathbf{T}^\top X_i^B)_{j,l}\mathrm{d}M_i(t) > \frac{\omega_{j,l}}{2}\right] \\
&\quad + \sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \mathrm{d}M_i(t) > \frac{\omega_{j,l}}{2}\right] \\
&\leq \sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \sum_{u=l}^{d_j+1}\mathbb{1}(X_{i,j}\in I_{j,l})\mathrm{d}M_i(t) > \frac{\omega_{j,l}}{2}\right] \\
&\quad + \sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \mathrm{d}M_i(t) > \frac{\omega_{j,l}}{2}\right].
\end{aligned}$$

Using Theorem 3 from Gaïffas and Guilloux (2012) and choosing the weights $\omega_{j,l}$ as defined in (6), we conclude that $\mathbb{P}[\mathscr{E}_n^{\complement}] \leq 57.1e^{-c}$ for some $c > 0$.

□

B.4 Proofs of the lemmas

Let us start with the proof of Lemma 2. To characterize the solution of Problem (4), 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 (4) 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

$$(\nabla \ell_n(f_{\hat{\beta}}))_{j,\bullet} + \hat{h}_{j,\bullet} + \hat{g}_{j,\bullet} + \hat{k}_{j,\bullet} = \mathbf{0}, \tag{19}$$

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

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

where $\mathcal{A}(\hat{\beta})$ is the active set of $\hat{\beta}$, see (7). 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 (19), and considering any vector $\beta \in \mathbb{R}^{p+d}$, we obtain

$$(\hat{\beta} - \beta)^\top \nabla \ell_n(f_{\hat{\beta}}) + (\hat{\beta} - \beta)^\top (\hat{h} + \hat{g} + \hat{k}) = 0, \tag{20}$$

and Equation (8) 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 $j = 1, \ldots, p$, 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.

□

Les us now derive the 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.,

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

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

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

and

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

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

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

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

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

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

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

Differentiating again, we obtain

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

Therefore, we have

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

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

$$G^{''}(0)\frac{\psi(-\|f_2\|_\infty)}{\|f_2\|_\infty^2} \le G(\eta) - G(0) - \eta G^{'}(0) \le G^{''}(0)\frac{\psi(\|f_2\|_\infty)}{\|f_2\|_\infty^2}. \tag{21}$$

We will apply inequalities in (21) 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,

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

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 (21), we obtain (11) in Lemma 3.

In case # 2, one gets

$$G^{'}(0) = 0,$$

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

which gives (12) in Lemma 3.

□

Then, we give a proof for 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,*

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

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

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

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

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

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

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

$$\mathbb{P}\Big[\hat{\kappa}_\tau^2(L) \geq \bar{\kappa}_\tau^2(L) - 4|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\sqrt{2/n\log\big(2(p+d)^2/\varepsilon\big)}\Big] \geq 1-\varepsilon. \tag{22}$$

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,

$$\bar{\Sigma}_n(f^\star,\tau) = \widetilde{\Sigma}_n(f^\star,\tau) - \frac{1}{n}\int_0^\tau S_n^{(0)}(f^\star,s)\big(\breve{X}_n(s) - \bar{X}_n(s)\big)^{\otimes 2}\lambda_0^\star(s)\mathrm{d}s. \tag{23}$$

We first bound the second term on the right-hand side of (23). 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

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

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

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

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 (23), Lemma 5, and the above two probability bounds, we obtain

$$\bar\kappa_\tau^2(L) \ge \tilde\kappa_\tau^2(L) - 8e^{2f_\infty^\star}\Lambda_0^\star(\tau)|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\frac{t_{n,p,d,\varepsilon}^2}{s^{(0)}(\tau)} \tag{24}$$

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

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

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

$$\tilde\kappa_\tau^2(L) \ge \kappa_\tau^2(L) - 4e^{2f_\infty^\star}\Lambda_0^\star(\tau)|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\sqrt{2/n\log\big(2(p+d)^2/\varepsilon\big)}. \tag{25}$$

Finally, the result follows from (22), (24) and (25).

□

Going back to the proof of Lemma 4, following Lemma 5 in Alaya et al. (2019), for any u in

$$\mathscr{C}_{1,\omega}(K) = \Big\{u\in\mathbb{R}^d : \sum_{j=1}^p \|(u_{j,\bullet})_{K_j^\complement}\|_{1,\omega_{j,\bullet}} \le 3\sum_{j=1}^p\|(u_{j,\bullet})_{K_j}\|_{1,\omega_{j,\bullet}}\Big\}, \tag{26}$$

the following holds:

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

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.

□

Finally, we give a proof for Lemma 5. First, we have that

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

Hence, we get

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

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 $j = 1, \ldots, p$, 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 $L = [L_1, \ldots, L_p] \subset \{1, \ldots, p + d\}$, 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

$$\begin{aligned}
\|\beta\|_1 &\leq \frac{8 \max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \max_{j,l} \omega_{j,l} \sum_{j=1}^{p} \|(\beta_{j,\bullet})_{L_j}\|_1 \\
&= \frac{8 \max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \max_{j,l} \omega_{j,l} \|\beta_L\|_1 \qquad (27) \\
&\leq \sqrt{|L|} \frac{8 \max_j (d_j+1)}{\min_{j,l} \omega_{j,l}} \max_{j,l} \omega_{j,l} \|\beta_L\|_2.
\end{aligned}$$

□

Appendix C Proof of Theorem 2

For all $j = 1, \ldots, p$, let $(\bar{k}_{j,l})_{l=0,\ldots,d_j+1} \subset \{1, \ldots, K^\star + 1\}$ be the sequence defined by

$$\bar{k}_{j,0} = 1, \text{ and } \bar{k}_{j,l} = \max\{k = 1, \ldots, K^\star + 1 : I^\star_{j,k} \cap I_{j,l} \neq \varnothing\}, \text{ for } l = 1, \ldots, d_j + 1.$$

Using the sequence $(\bar{k}_{j,l})_{l=0,\ldots,d_j+1}$, one has the expression of the $f^\star$ as follows:

$$f^\star(X_i) = \sum_{j=1}^{p} \sum_{l=1}^{d_j+1} \sum_{k=\bar{k}_{j,l-1}}^{\bar{k}_{j,l}} \beta^\star_{j,k} \mathbb{1}(X_{i,j} \in I^\star_{j,k} \cap I_{j,l}).$$

We have that $\sum_{i=1}^{n} f_j^\star(X_i) = 0$, this implies

$$\sum_{i=1}^{n} \sum_{l=1}^{d_j+1} \sum_{k=\bar{k}_{j,l-1}}^{\bar{k}_{j,l}} \beta_{j,k}^\star \mathbb{1}(X_{i,j} \in I_{j,k}^\star \cap I_{j,l}) = 0. \tag{28}$$

For all $j = 1, \ldots, p$ and $l = 1, \ldots, d_j + 1$, we set

$$|I_{j,l}|_n := \frac{n_{j,l}}{n} = \frac{|\{i = 1, \ldots, n : X_{i,j} \in I_{j,l}\}|}{n}.$$

$|I_{j,l}|_n$ stands for the proportion of the data features in the the discretization interval $I_{j,l}$. Next, we define $f_{\bar{b}}$ as follows:

$$\begin{aligned}
f_{\bar{b}}(X_i) &:= \sum_{j=1}^{p} (f_{\bar{b}})_j(X_i) \\
&:= \sum_{j=1}^{p} \sum_{l=1}^{d_j+1} \bar{b}_{j,l} \mathbb{1}(X_{i,j} \in I_{j,m}) \\
&:= \sum_{j=1}^{p} \sum_{l=1}^{d_j+1} \sum_{k=\bar{k}_{j,l-1}}^{\bar{k}_{j,l}} \beta_{j,k}^\star \frac{|I_{j,k}^\star \cap I_{j,l}|_n}{|I_{j,l}|_n} \mathbb{1}(X_{i,j} \in I_{j,l}).
\end{aligned}$$

The vector parameter $\bar{b}$ verifies the sum-zero constraint, i.e., $\sum_{l=1}^{d_j+1} n_{j,l} \bar{b}_{j,l} = 0$, since

$$\begin{aligned}
\sum_{l=1}^{d_j+1} n_{j,l} \bar{b}_{j,l} &= \sum_{l=1}^{d_j+1} n_{j,l} \sum_{k=\bar{k}_{j,l-1}}^{\bar{k}_{j,l}} \beta_{j,k}^\star \frac{|I_{j,k}^\star \cap I_{j,l}|_n}{|I_{j,l}|_n} \mathbb{1}(X_{i,j} \in I_{j,l}) \\
&= \sum_{l=1}^{d_j+1} n_{j,l} \sum_{k=\bar{k}_{j,l-1}}^{\bar{k}_{j,l}} \beta_{j,k}^\star \frac{|\{i = 1, \ldots, n : X_{i,j} \in I_{j,k}^\star \cap I_{j,l}\}|}{n_{j,l}} \mathbb{1}(X_{i,j} \in I_{j,l}) \\
&= \sum_{l=1}^{d_j+1} \sum_{k=\bar{k}_{j,l-1}}^{\bar{k}_{j,l}} \beta_{j,k}^\star |\{i = 1, \ldots, n : X_{i,j} \in I_{j,k}^\star \cap I_{j,l}\}| \mathbb{1}(X_{i,j} \in I_{j,l}) \\
&= \sum_{l=1}^{d_j+1} \sum_{k=\bar{k}_{j,l-1}}^{\bar{k}_{j,l}} \beta_{j,k}^\star \sum_{i=1}^{n} \mathbb{1}(X_{i,j} \in I_{j,k}^\star \cap I_{j,l}) \mathbb{1}(X_{i,j} \in I_{j,l}) \\
&= \sum_{i=1}^{n} \sum_{l=1}^{d_j+1} \sum_{k=\bar{k}_{j,l-1}}^{\bar{k}_{j,l}} \beta_{j,k}^\star \mathbb{1}(X_{i,j} \in I_{j,k}^\star \cap I_{j,l}) \\
&= 0,
\end{aligned}$$

where the last equality comes from Equation (28).

Thereofore,

$$
\begin{aligned}
&\sum_{i=1}^{n}(f^\star(X_i)-f_{\bar b}(X_i))^2\\
&=\sum_{i=1}^{n}\Big(\sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\sum_{k=\bar k_{j,l-1}}^{\bar k_{j,l}}\Big(\beta^\star_{j,k}-\sum_{k'=\bar k_{j,l-1}}^{\bar k_{j,l}}\beta^\star_{j,k'}\frac{|I^\star_{j,k'}\cap I_{j,l}|_n}{|I_{j,l}|_n}\Big)\mathbb{1}(X_{i,j}\in I^\star_{j,k}\cap I_{j,l})\Big)^2\\
&=\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\sum_{k=\bar k_{j,l-1}}^{\bar k_{j,l}}\Big(\beta^\star_{j,k}-\sum_{k'=\bar k_{j,l-1}}^{\bar k_{j,l}}\beta^\star_{j,k'}\frac{|I^\star_{j,k'}\cap I_{j,l}|_n}{|I_{j,l}|_n}\Big)^2\mathbb{1}(X_{i,j}\in I^\star_{j,k}\cap I_{j,l})\\
&=\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\mathbb{1}(\bar k_{j,l}-\bar k_{j,l-1}>0)\sum_{k=\bar k_{j,l-1}}^{\bar k_{j,l}}\Big(\beta^\star_{j,k}-\sum_{k'=\bar k_{j,l-1}}^{\bar k_{j,l}}\beta^\star_{j,k'}\frac{|I^\star_{j,k'}\cap I_{j,l}|_n}{|I_{j,l}|_n}\Big)^2\mathbb{1}(X_{i,j}\in I^\star_{j,k}\cap I_{j,l})\\
&\le\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\mathbb{1}(\bar k_{j,l}-\bar k_{j,l-1}>0)\sum_{k=\bar k_{j,l-1}}^{\bar k_{j,l}}(\bar k_{j,l}-\bar k_{j,l-1}+1)\\
&\qquad\qquad\times\max_{k,k'=\bar k_{j,l-1},\dots,\bar k_{j,l}}(\beta^\star_{j,k}-\beta^\star_{j,k'})^2\mathbb{1}(X_{i,j}\in I^\star_{j,k}\cap I_{j,l})\\
&\le\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\mathbb{1}(\bar k_{j,l}-\bar k_{j,l-1}>0)(\bar k_{j,l}-\bar k_{j,l-1}+1)\max_{k,k'=\bar k_{j,l-1},\dots,\bar k_{j,l}}(\beta^\star_{j,k}-\beta^\star_{j,k'})^2\mathbb{1}(X_{i,j}\in I_{j,l})\\
&\le\sum_{j=1}^{p}\sum_{l=1}^{d_j+1}\mathbb{1}(\bar k_{j,l}-\bar k_{j,l-1}>0)(\bar k_{j,l}-\bar k_{j,l-1}+1)\max_{k,k'=\bar k_{j,l-1},\dots,\bar k_{j,l}}(\beta^\star_{j,k}-\beta^\star_{j,k'})^2\cdot n_{j,l}.
\end{aligned}
$$

As the intervals $I_{j,l}$ are been chosen as quantile intervals, we have for all $j=1,\dots,p$ and $l=1,\dots,d_j+1$

$$n_{j,l}=\frac{n}{D+1}.$$

We obtain

$$\sum_{i=1}^{n}(f^\star(X_i)-f_{\bar b}(X_i))^2\le 2K^\star\Delta^2_{\beta,\max}\frac{n}{D+1}$$

where $\Delta_{\beta^\star,\max}=\max_{1\le j\le p}\max_{1\le k,k'\le K^\star+1}|\beta^\star_{j,k}-\beta^\star_{j,k'}|$. Applying Theorem 1 for $f_\beta=f_{\bar b}$, we get

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

By Application Lemma (12) (see Equation (12)), we have that

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

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

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

Remark that $\max_{1\le i\le n}|f_{\bar b}(X_i)|\le K^\star f^\star_\infty$ (by construction). Then

$$\|f^\star-f_{\bar b}\|_\infty\le(1+K^\star)f^\star_\infty.$$

Hence, by the fact $\psi(x)/x^2$ is a non-deacrsing function

$$KL_n(f^\star, f_{\bar{b}}) \leq C_{K^\star, f^\star_\infty} \|f^\star - f_{\bar{b}}\|_n^2,$$

where $C_{K^\star, f^\star_\infty} = \frac{\psi((1+K^\star)f^\star_\infty)}{((1+K^\star)f^\star_\infty)^2}$. Recall that

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

where the last inequality is due to the $\frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star,t)} \leq 1, \forall t$. Therefore,

$$KL_n(f^\star, f_{\bar{b}}) \leq 2\Delta^2_{\beta,\max} K^\star C_{K^\star, f^\star_\infty} \bar{N}(\tau) \frac{n}{D+1}$$

and

$$\begin{aligned}
KL_n(f^\star, f_{\hat{\beta}}) \leq\ & 6\Delta^2_{\beta,\max} K^\star C_{K^\star, f^\star_\infty} \bar{N}(\tau) \frac{n}{D+1} \\
&+ \frac{1024(f^\star_\infty + R + 2)}{\kappa_\tau^2\big(\mathcal{A}(\bar{b})\big) - \Xi_\tau\big(\mathcal{A}(\bar{b})\big)} |\mathcal{A}(\bar{b})| \max_{1\leq j\leq p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\bar{b})}\|_\infty^2.
\end{aligned}$$

By construction the active set of $\bar{b}$ verifies

$$|\mathcal{A}(\bar{b})| = \sum_{j=1}^p |\mathcal{A}_j(\bar{b})| \leq 2K^\star$$

Hence,

$$\begin{aligned}
KL_n(f^\star, f_{\hat{\beta}}) &\leq 6\Delta^2_{\beta,\max} K^\star C_{K^\star, f^\star_\infty} \bar{N}(\tau) \frac{n}{D+1} \\
&\quad + \frac{1024(f^\star_\infty + R + 2)}{\kappa_\tau^2\big(\mathcal{A}(\bar{b})\big) - \Xi_\tau\big(\mathcal{A}(\bar{b})\big)} |\mathscr{A}^\star|(D+1) \max_{1\leq j\leq p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\bar{b})}\|_\infty^2 \\
&\leq 6\Delta^2_{\beta,\max} K^\star C_{K^\star, f^\star_\infty} \bar{N}(\tau) \frac{n}{D+1} + C_{\mathrm{cut}} |\mathscr{A}^\star| \frac{K^\star}{n},
\end{aligned}$$

where $C_{\mathrm{cut}} = \mathcal{O}(\frac{1024(f^\star_\infty + R + 2)}{\kappa_\tau^2\big(\mathcal{A}(\bar{b})\big) - \Xi_\tau\big(\mathcal{A}(\bar{b})\big)})$.

Moreover, we have that the cumulative couting process $\bar{N}(\tau) = \frac{1}{n}\sum_{i=1}^n N_i(\tau)$ where $N_i(\tau) \in [0,1]$ for all i. Then, by application of Hoeffding concentration inequality, we have that for all $\epsilon > 0$

$$\mathbb{P}[|\bar{N}(\tau) - \mathbb{E}[\bar{N}(\tau)] > \epsilon] \leq e^{-2n\epsilon^2},$$

with $\mathbb{E}[\bar{N}(\tau)] \leq \max_{i=1}^n \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t$. This implies that $\bar{N}(\tau) \leq \max_{i=1}^n \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t + \epsilon$, with probability larger than $1 - e^{-2n\epsilon^2}$. Choosing D as the integer part of

$$\frac{6\Delta^2_{\beta,\max} K^\star C_{K^\star, f^\star_\infty} (\max_{1\leq i\leq n} \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t + \epsilon)}{K^\star |\mathscr{A}^\star| C_{\mathrm{cut}}} \cdot n^2.$$

leads to the rate of convergence exhibited in Theorem 2.

□

Appendix D Proof of Theorem 3

Let us first make a remark concerning the choice we made to approximate $f^\star$ using $b^\star$. Instead of what we did in (2) and (3), it may be tempting to define $b^\star$ such that

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

for all $j = 1, \ldots, p$, 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 (2) such that the corresponding cut-point is always the right bound of $I_{j,l^\star_{j,k}}$, i.e., $\mu_{j,l^\star_{j,k}}$.

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 (3). 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

$$\|f^\star - f_{b^\star}\|_n^2 \leq \int_0^\tau \sum_{i=1}^n \big((f^\star - f_{b^\star})(X_i)\big)^2 \frac{Y_i(t)e^{f^\star(X_i)}}{S_n^{(0)}(f^\star, t)} \mathrm{d}\bar{N}(t). \tag{29}$$

Next, we control the right-hand-side of (29). For all $i = 1, \ldots, n$, 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

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

Let us rewrite constraint (15) 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,

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

and

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

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

$$\begin{aligned}
(ii) &= \Big\{ \sum_{j \in \mathscr{A}(\beta^\star)} \sum_{k=1}^{K_j^\star+1} |\beta_{j,k}^\star| \frac{n_{j,l_{j,k}^\star}}{n} \Big\}^2 \pi_n \\
&\leq \frac{\max_{j\in\mathscr{A}(\beta^\star)} \|\beta_{j,\bullet}\|_\infty^2 \max_{j\in\mathscr{A}(\beta^\star)} \|n_{j,\bullet}\|_\infty^2}{n} \big(|\mathscr{A}(\beta^\star)| + K^\star\big)\pi_n. \qquad (30)
\end{aligned}$$

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

$$\mathbb{P}\Big[\inf_{t\in[0,\tau]}\frac{1}{n}S_n^{(0)}(f^\star,t)\geq e^{-f_\infty^\star}\frac{c_Z}{2}\Big]\geq\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n\mathbb{1}(Z_i\geq\tau)\geq\frac{c_Z}{2}\Big]\geq 1-2e^{-nc_Z^2/2}. \tag{31}$$

Combining (30) and (31), we obtain the desired result.

□

Let us derive the proof of Theorem 3. 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 (12) 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 (18). 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 (12) 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 (10), 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 $1-28.55e^{-c}-e^{-ns^{(0)}(\tau)^2/8e^{2f_\infty^\star}}-3\varepsilon-2e^{-nc_Z^2/2}$.

□

Appendix E Proof of Theorem 4

Let us introduce

$$\breve{f}(X_i) = \sum_{j=1}^{p} \breve{f}_j(X_{i,j}) = \sum_{j \in \mathscr{A}^\star} \sum_{l=1}^{D+1} \breve{\beta}_{j,l} \mathbb{1}(X_{i,j} \in I_{j,l}),$$

where

$$\breve{\beta}_{j,l} = \frac{\sum_{i=1}^n f_j^\star(X_{i,j}) \mathbb{1}(X_{i,j} \in I_{j,l})}{\sum_{i=1}^n \mathbb{1}(X_{i,j} \in I_{j,l})}, \text{ for all } j \in \mathcal{A}^\star \text{ and } \breve{\beta}_{j,\bullet} = \mathbf{0}_D, \text{ for all } j \notin \mathcal{A}^\star.$$

The vector parameter $\breve{\beta}$ verifies the sum-zero constraint, i.e., $\sum_{l=1}^{d+1} n_{j,l} \breve{\beta}_{j,l} = 0$ since

$$\begin{aligned}
\sum_{l=1}^{D+1} n_{j,l} \breve{\beta}_{j,l} &= \sum_{l=1}^{D+1} n_{j,l} \frac{\sum_{i=1}^n f_j^\star(X_{i,j}) \mathbb{1}(X_{i,j} \in I_{j,l})}{\sum_{i=1}^n \mathbb{1}(X_{i,j} \in I_{j,l})} \\
&= \sum_{l=1}^{D+1} |\{i = 1, \ldots, n : X_{i,j} \in I_{j,l}\}| \frac{\sum_{i=1}^n f_j^\star(X_{i,j}) \mathbb{1}(X_{i,j} \in I_{j,l})}{\sum_{i=1}^n \mathbb{1}(X_{i,j} \in I_{j,l})} \\
&= \sum_{l=1}^{D+1} \sum_{i=1}^{n} \mathbb{1}(X_{i,j} \in I_{j,l}) \frac{\sum_{i=1}^n f_j^\star(X_{i,j}) \mathbb{1}(X_{i,j} \in I_{j,l})}{\sum_{i=1}^n \mathbb{1}(X_{i,j} \in I_{j,l})} \\
&= \sum_{l=1}^{D+1} \sum_{i=1}^{n} f_j^\star(X_{i,j}) \mathbb{1}(X_{i,j} \in I_{j,l}) \\
&= \sum_{i=1}^{n} f_j^\star(X_{i,j}) \underbrace{\sum_{l=1}^{D+1} \mathbb{1}(X_{i,j} \in I_{j,l})}_{=1} \\
&= \sum_{i=1}^{n} f_j^\star(X_{i,j}) \\
&= 0,
\end{aligned}$$

where the last equality comes from the identifiability condition on $f^\star$. On the other hand, we have

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} (f^\star(X_i) - \breve{f}(X_i))^2 &= \frac{1}{n} \sum_{i=1}^{n} \Big(\sum_{j \in \mathscr{A}^\star} \big(f_j^\star(X_{i,j}) - \breve{f}_j(X_{i,j})\big) \Big)^2 \\
&\leq \frac{|\mathscr{A}^\star|}{n} \sum_{i=1}^{n} \sum_{j \in \mathscr{A}^\star} \big(f_j^\star(X_{i,j}) - \breve{f}_j(X_{i,j})\big)^2 \\
&= \frac{|\mathscr{A}^\star|}{n} \sum_{i=1}^{n} \sum_{j \in \mathscr{A}^\star} \Big(\sum_{l=1}^{d+1} (f_j^\star(X_{i,j}) - \breve{\beta}_{j,l}) \mathbb{1}(X_{i,j} \in I_{j,l}) \Big)^2 \\
&= \frac{|\mathscr{A}^\star|}{n} \sum_{i=1}^{n} \sum_{j \in \mathscr{A}^\star} \sum_{l=1}^{d+1} (f_j^\star(X_{i,j}) - \breve{\beta}_{j,l})^2 \mathbb{1}(X_{i,j} \in I_{j,l}).
\end{aligned}$$

It easy to show that

$$\begin{aligned}|f_j^\star(X_{i,j}) - \breve{\beta}_{j,k}| &= \Big|f_j^\star(X_{i,j}) - \frac{\sum_{i'=1}^n f_j^\star(X_{i',j})\mathbb{1}(X_{i',j} \in I_{j,k})}{\sum_{i'=1}^n \mathbb{1}(X_{i',j} \in I_{j,k})}\Big| \\ &\leq \Big|\frac{\sum_{i'=1}^n |f_j^\star(X_{i,j}) - f_j^\star(X_{i',j})|\mathbb{1}(X_{i',j} \in I_{j,k})}{\sum_{i'=1}^n \mathbb{1}(X_{i',j} \in I_{j,k})}\Big| \\ &\leq L|I_{j,k}| \\ &= \frac{L}{D+1}.\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (f^\star(X_i) - \breve{f}(X_i))^2 &\leq \frac{|\mathscr{A}^\star|}{n}\sum_{i=1}^n \sum_{j\in\mathscr{A}^\star}\sum_{l=1}^{d+1} L^2|I_{j,k}|^2\mathbb{1}(X_{i,j} \in I_{j,k}) \\ &\leq \frac{L^2|\mathscr{A}^\star|^2}{D^2}.\end{aligned}$$

Define the empirical maximum norm as $\|f\|_{n,\infty} = \max_{1\leq i\leq n}|f(X_i)|$. The following lemma holds (see Lemma 1.5.4 in Letué (2000)).

Lemma 8 *One has*

$$KL_n(f^\star, f) \leq 2\|f^\star - f\|_{n,\infty}\max_{1\leq i\leq n}\int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t.$$

Proof of Lemma 8. By definition, we have

$$\begin{aligned}KL_n(f^\star, f_\beta) &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \log\left[\frac{e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i(t)e^{f^\star(X_i)}}\frac{\sum_{i=1}^n Y_i e^{f(X_i)}}{e^{f(X_i)}}\right]Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}\mathrm{d}t \\ &= \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left(f^\star(X_i) - f(X_i) + \log\left[\frac{\sum_{i=1}^n Y_i e^{f(X_i)}}{\sum_{i=1}^n Y_i e^{f^\star(X_i)}}\right]\right)Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}\mathrm{d}t.\end{aligned}$$

Moreover,

$$\begin{aligned}0 \leq \frac{\sum_{i=1}^n Y_i e^{f(X_i)}}{\sum_{i=1}^n Y_i e^{f^\star(X_i)}} &= \frac{\sum_{i=1}^n Y_i e^{f(X_i)-f^\star(X_i)}e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i e^{f^\star(X_i)}} \\ &\leq \frac{\sum_{i=1}^n Y_i e^{\|f-f^\star\|_{n,\infty}}e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i e^{f^\star(X_i)}} \leq e^{\|f-f^\star\|_{n,\infty}}.\end{aligned}$$

Hence

$$f^\star(X_i) - f(X_i) + \log\left[\frac{\sum_{i=1}^n Y_i e^{f(X_i)}}{\sum_{i=1}^n Y_i e^{f^\star(X_i)}}\right] \leq f^\star(X_i) - f(X_i) + \|f-f^\star\|_{n,\infty} \leq 2\|f-f^\star\|_{n,\infty}.$$

In another hand, we have

$$\|f-f^\star\|_{n,\infty} \leq \sqrt{\sum_{i=1}^n (f^\star(X_i) - f(X_i))^2} = \sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^n (f^\star(X_i) - f(X_i))^2}.$$

Applying Theorem 1 for $f_\beta = \breve{f}$ leads to

$$\begin{aligned} KL_n(f^\star, f_{\hat{\beta}}) &\leq 3KL_n(f^\star, \breve{f}) + \frac{1024(f^\star_\infty + R + 2)}{\kappa_\tau^2\big(\mathcal{A}(\breve{\beta})\big) - \Xi_\tau\big(\mathcal{A}(\breve{\beta})\big)} |\mathcal{A}(\breve{\beta})| \max_{1\leq j\leq p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\breve{\beta})}\|_\infty^2 \\ &\leq 3KL_n(f^\star, \breve{f}) + C|\mathcal{A}(\breve{\beta})|\frac{\log(p+d)}{n} \\ &\leq 6\|f^\star - \breve{f}\|_{n,\infty} \max_{1\leq i\leq n} \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t + C|\mathcal{A}(\breve{\beta})|\frac{\log(p+d)}{n} \\ &\leq \frac{6L|\mathscr{A}^\star|\sqrt{n}}{D} \max_{1\leq i\leq n} \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t + C|\mathcal{A}(\breve{\beta})|\frac{\log(p+d)}{n}. \end{aligned}$$

In this case, we straightforwardly have

$$|\mathcal{A}(\breve{\beta})| = \sum_{j=1}^{p} |\mathcal{A}_j(\breve{\beta})| \leq |\mathcal{A}(\breve{\beta})|(D+1) \leq |\mathscr{A}^\star|(D+1).$$

Taking into account that $\log(p+d) = \mathcal{O}(1)$, we arrive at

$$KL_n(f^\star, f_{\hat{\beta}}) \leq \frac{6L|\mathscr{A}^\star|\sqrt{n}}{D} \max_{1\leq i\leq n} \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t + 2C_{\mathrm{lip}}|\mathscr{A}^\star|\frac{D}{n},$$

with $C_{\mathrm{lip}} = \mathcal{O}\Big(\frac{1024(f^\star_\infty + R + 2)}{\kappa_\tau^2\big(\mathcal{A}(\breve{\beta})\big) - \Xi_\tau\big(\mathcal{A}(\breve{\beta})\big)}\Big)$

$$KL_n(f^\star, f_{\hat{\beta}}) \leq \frac{6L|\mathscr{A}^\star|\sqrt{n}}{D} \max_{1\leq i\leq n} \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t + 2C_{\mathrm{lip}}|\mathscr{A}^\star|\frac{D}{n}.$$

Choosing D as the integer part of

$$\sqrt{\frac{6L \max_{1\leq i\leq n} \int_0^\tau \lambda^\star(t|X_i)\mathrm{d}t}{2C_{\mathrm{lip}}}} \cdot n^{3/4}$$

leads to the rate of convergence exhibited in Theorem 4.

□

References

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

Alaya, M. Z., Bussy, S., Gaıffas, S., and Guilloux, A. (2019). Binarsity: a penalization for one-hot encoded features in linear supervised learning. *Journal of Machine Learning Research* **20,** 1–34.

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

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

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

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

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

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

Bacry, E., Bompaire, M., Deegan, P., Gaïffas, S., and Poulsen, S. V. (2017). Tick: a python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *The Journal of Machine Learning Research* **18,** 7937–7941.

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

Banerjee, M., McKeague, I. W., et al. (2007). Confidence sets for split points in decision trees. *The Annals of Statistics* **35,** 543–574.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57,** 289–300.

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

BioProfiling (2009). Hbs1l ppisurv.

Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: the bonferroni method. *Bmj* **310,** 170.

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

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

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

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

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

Chzhen, E., Hebiri, M., Salmon, J., et al. (2019). On lasso refitting strategies. *Bernoulli* **25,** 3175–3200.

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

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

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

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

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

Guan, Y., He, Y., Lv, S., Hou, X., Li, L., and Song, J. (2019). Overexpression of hoxc10 promotes glioblastoma cell progression to a poor prognosis via the pi3k/akt signalling pathway. *Journal of drug targeting* **27,** 60–66.

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

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

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

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

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

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

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

Lederer, J. (2013). Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions. *arXiv preprint arXiv:1306.0113* .

Letué, F. (2000). Modele de cox : estimation par selection de modele et modele de chocs bivarie.

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

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

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

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Austin, TX.

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

Therneau, T. M. and Grambsch, P. M. (2000). The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine* **30,** 1105–1117.

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

Zhang, C.-H., Huang, J., et al. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* **36,** 1567–1594.