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

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

Binacox: automatic cut-points detection in high-dimensional Cox model, with applications to genetic data

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

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

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

Anne-Sophie Jannot
Biomedical Informatics and Public Health Department
European Georges Pompidou Hospital, Assistance Publique-Hôpitaux de Paris
and INSERM UMRS 1138, Centre de Recherche des Cordeliers, Paris, France
email: annesophie.jannot@aphp.fr

Abstract

Determining significant prognostic biomarkers is of increasing importance in many areas of medicine. In order to translate a continuous biomarker into a clinical decision, it is often necessary to determine cut-points. There is so far no standard method to help evaluate how many cut-points points are optimal for a given feature in a survival analysis setting. Moreover, most existing methods are univariate, hence not well suited for high-dimensional frameworks. This paper introduces a prognostic method called *Binacox* to deal with the problem of detecting multiple cut-points per features in a multivariate setting where a large number of continuous features are available. It is based on the Cox model and combines one-hot encodings with the binarsity penalty. This penalty uses total-variation regularization together with an extra linear constraint to avoid collinearity between the one-hot encodings and enable feature selection. A non-asymptotic oracle inequality is established. The statistical performance of the method is then examined on an extensive Monte Carlo simulation study, and finally illustrated on three publicly available genetic cancer datasets with high-dimensional features. On this datasets, our proposed methodology significantly outperforms the state-of-the-art survival models regarding risk prediction in terms of C-index, with a computing time orders of magnitude faster. In addition, it provides powerful interpretability by automatically pinpointing significant cut-points on relevant features from a clinical point of view.

Keywords. Cox proportional hazards model; Cut-points; Features Binarization; Non-asymptotic oracle inequalities; Proximal methods; Survival analysis; Total-Variation

1 Introduction

In any medical applications, the effects of certain clinical variables on the prognostic are sometimes known, but their precise roles remain to be clearly established. For instance, in a breast cancer study, there is reasonable agreement that younger patients have higher risk of an unfavourable outcome, but there is little agreement on the exact nature of the relationship between age and prognosis. Similar issues occur in genetic oncology studies where gene expressions effects on survival times are often non-linear.

The cut-points detection problem. A simple and popular way to treat this problem consists in determining cut-off values, or cut-points, of the continuous features (e.g. the age or the gene expressions in the previous examples). This technique brings to light potential non-linearities on feature effects that most models cannot detect. It also offers the ability to classify patients into several groups regarding its features values relatively to the cut-points. More importantly, it can lead to a better understanding of the features effects on the outcome under study. A convenient tool to find optimal cut-points is, therefore, of high interest.

Hence, cut-points detection is a widespread issue in many medical studies and multiple methods have been proposed to determine a single cut-point for a given feature . They range from choosing the mean or median value to methods based on distribution of values or association with clinical outcomes, such as the minimal p-value of multiple log-rank tests, see for instance Camp et al. [2004], Moul et al. [2007], Rota et al. [2015] among many others. However, the choice of the actual cut-points is not a straightforward problem, even for one single cut-point [Lausen and Schumacher, 1992, Klein and Wu, 2003, Contal and O'Quigley, 1999].

While many studies have been devoted to find one optimal cut-point, there is often need in practical medicine to determine not only one, but multiple cut-points. Some method deal with multiple cut-points detection for one-dimensional signals (see for instance Bleakley and Vert [2011] and Harchaoui and Lévy-Leduc [2010] that use a group fused Lasso and total variation penalties respectively) or for multivariate time series (see Cho and Fryzlewicz [2015]). Whereas cut-points detection is known to be a paramount issue in survival analysis also [Faraggi and Simon, 1996], the corresponding developed methods are looking only at a single feature at a time (e.g. Motzer et al. [1999] or LeBlanc and Crowley [1993] with the survival trees). To our knowledge,no multivariate survival analysis method well suited to detect multiple cut-points per feature in a high-dimensional setting has yet been proposed.

General framework. Let us consider the usual survival analysis framework. Following Andersen et al. [1993], let non-negative random variables T and C stand for the times of the event of interest and censoring times respectively, and X denotes the p-dimensional vector of features (e.g. patients characteristics, therapeutic strategy, omics features). The event of interest could be for instance survival time, re-hospitalization, relapse or disease progression. Conditional 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

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

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 (PH) model [Cox, 1972] is by far the most widely used in survival analysis. It is a regression model that describes the relation between intensity of events and features, given by

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

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

High-dimensional survival analysis. High-dimension settings are becoming increasingly frequent, in particular for genetic data applications where the cut-points estimation is a common problem (see for instance Harvey et al. [1999], Shirota et al. [2001], Cheang et al. [2009]), but also in other contexts where the number of available features to consider as potential risk factors is tremendous, especially with the development of electronic health records. A penalized version of the Cox PH model well suited for such settings is proposed in Simon et al. [2011], but it cannot model nonlinearities. Other methods have being developed to deal with this problem in such settings, like boosting Cox PH models [Li and Luan, 2005] or random survival forests [Ishwaran et al., 2008]. But none of them identify cut-points values, which is of major interest for both interpretations and clinical benefits.

The proposed method. In this paper, we propose a method called *Binacox* for estimating multiple cut-points in a Cox PH model with high-dimensional features. First, the Binacox one-hot encodes the continuous input features [Wu and Coggeshall, 2012] through a mapping to a new binarized space of much higher dimension, and then trains the Cox PH model in this space, regularized with the binarsity penalty [Alaya et al., 2017] that combines total-variation regularization with an extra sum-to-zero constraint to avoid collinearity between the one-hot encodings and enable feature selection. Cut-points of the initial continuous input features are then detected by the jumps in the regression coefficient vector, that the binarsity penalty enforces to be piecewise constant.

Organization of the paper. The main contribution of this paper is then the idea of using a total-variation penalization, with an extra linear constraint, on the weights of a Cox PH model trained on a binarization of the raw continuous features, leading to a procedure that selects multiple cut-points per feature, looking at all features simultaneously and that also selects relevant features. A precise description of the model is given in Section 2. Section 3 highlights the good theoretical properties of the Binacox by establishing fast oracle inequalities in prediction. Section 4 presents the simulation procedure used to evaluate the performances and compares it with existing methods. In Section 5, we apply our method to high-dimensional genetic datasets. Finally, we discuss the obtained results in Section 6.

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 $\|v\|_\infty = \max_{k=1,\dots,m} |v_k|$. 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. Finally, for any $u \in \mathbb{R}^m$ and any $L \subset \{1, \dots, m\}$, we denote u_L as the vector in $\mathbb{R}^m$ satisfying $(u_L)_k = u_k$ for $k \in L$ and $(u_L)_k = 0$ for $k \in L^\complement = \{1, \dots, m\} \backslash L$. Let M be a matrix of size $k \times k'$, $M_{j,\bullet}$ denotes its j-th row and $M_{\bullet,l}$ its l-th column.

2 Method

Cox PH model with cut-points. Consider a training dataset of n independent and identically distributed (i.i.d.) examples $(x_1, z_1, \delta_1), \ldots, (x_n, z_n, \delta_n) \in [0,1]^p \times \mathbb{R}_+ \times \{0,1\}$, where the condition $x_i \in [0,1]^p$ for all $i \in \{1, \ldots, n\}$ is always true after an appropriate rescaling preprocessing step, with no loss of generality. Let us denote $\boldsymbol{X} = [x_{i,j}]_{1 \leq i \leq n; 1 \leq j \leq p}$ the $n \times p$ design matrix vertically stacking the n samples of p raw features. Let $\boldsymbol{X}_{\bullet,j}$ be the j-th feature column of $\boldsymbol{X}$ and $\boldsymbol{X}_{i,\bullet}$ the i-th row example. In order to simplify presentation of our results, we assume in the paper that all raw features $\boldsymbol{X}_{\bullet,j}$ are continuous. Assume that intensity of events for patient i is given by

$$\lambda^\star(t | \boldsymbol{X}_{i,\bullet} = x_i) = \lambda_0^\star(t) e^{f^\star(x_i)}, \tag{1}$$

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

$$f^\star(x_i) = \sum_{j=1}^{p} \sum_{k=1}^{K_j^\star + 1} \beta_{j,k}^\star \mathbb{1}(x_{i,j} \in I_{j,k}^\star), \tag{2}$$

with $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ for $k \in \{1, \ldots, K_j^\star + 1\}$. Since our model defined in (1) is not identifiable, we choose to impose a sum-to-zero constraint in each $\beta^\star$'s block, that is

$$\sum_{k=1}^{K_j^\star + 1} \beta_{j,k}^\star = 0 \text{ for all } j \in \{1, \ldots, p\}, \tag{3}$$

then re-defining the baseline in (1) as $\lambda_0^\star(t) := \lambda_0^\star(t) \exp(\sum_{j=1}^{p} \sum_{k=1}^{K_j^\star + 1} \beta_{j,k}^\star)$.

Here, $\mu_{j,k}^\star$ for $k \in \{1, \cdots, K_j^\star\}$ denote the so-called cut-points of feature $j \in \{1, \ldots, p\}$ that 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

$$\beta^\star = ({\beta_{1,\bullet}^\star}^\top, \ldots, {\beta_{p,\bullet}^\star}^\top)^\top = (\beta_{1,1}^\star, \ldots, \beta_{1,K_1^\star+1}^\star, \ldots, \beta_{p,1}^\star, \ldots, \beta_{p,K_p^\star+1}^\star)^\top,$$

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

$$\mu^\star = ({\mu_{1,\bullet}^\star}^\top, \ldots, {\mu_{p,\bullet}^\star}^\top)^\top = (\mu_{1,1}^\star, \ldots, \mu_{1,K_1^\star}^\star, \ldots, \mu_{p,1}^\star, \ldots, \mu_{p,K_p^\star}^\star)^\top.$$

Our goal is to estimate simultaneously $\mu^\star$ and $\beta^\star$, which also requires an estimation of unknown $K_j^\star$ for $j \in \{1, \ldots, p\}$. To this end, the first step of our proposed methodology is to map the features space to a much higher space of binarized features.

Binarization. The binarized matrix $\boldsymbol{X}^B$ is a sparse matrix with an extended number $p + d$ of columns, typically with $d \gg p$, where features are one-hot encoded [Wu and Coggeshall, 2012, Liu et al., 2002]. The j-th column $\boldsymbol{X}_{\bullet,j}$ is then replaced by $d_j + 1 \geq 2$ columns $\boldsymbol{X}_{\bullet,j,1}^B, \ldots, \boldsymbol{X}_{\bullet,j,d_j+1}^B$ containing only zeros and ones and the i-th row $x_i^B \in \mathbb{R}^{p+d}$ is written

$$x_i^B = (x_{i,1,1}^B, \ldots, x_{i,1,d_1+1}^B, \ldots, x_{i,p,1}^B, \ldots, x_{i,p,d_p+1}^B)^\top.$$

To be more precise, we consider a partition of intervals $I_{j,1}, \dots, I_{j,d_j+1}$ where $I_{j,l} = (\mu_{j,l-1}, \mu_{j,l}]$ for $l \in \{1, \dots, d_j + 1\}$, with $\mu_{j,0} = 0$ and $\mu_{j,d_j+1} = 1$ by convention. Then for $i \in \{1, \dots, n\}$ and $l \in \{1, \dots, d_j + 1\}$, we define

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

A natural choice of intervals $I_{j,l}$ is given by a uniform grid $\mu_{j,l} = l/(d_j + 1)$.

To each binarized feature $\boldsymbol{X}^B_{\bullet,j,l}$ corresponds a parameter $\beta_{j,l}$ and the vectors associated to the binarization of the j-th 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

$$f_\beta(x_i) = \beta^\top x_i^B = \sum_{j=1}^{p} f_{\beta_{j,\bullet}}(x_i) \tag{4}$$

where for all $j \in \{1, \dots, p\}$, $f_{\beta_{j,\bullet}}(x_i) = \sum_{l=1}^{d_j+1} \beta_{j,l} \mathbb{1}(x_{i,j} \in I_{j,l})$. Thus, f_β is a candidate for the estimation of $f^\star = f_{\beta^\star}$ defined in (2).

The full parameters vectors of size $p + d$ and d respectively, where $d = \sum_{j=1}^p d_j$, are simply obtained by concatenation of the vectors $\beta_{j,\bullet}$ and $\mu_{j,\bullet}$, that is

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

and

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

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

$$\ell_n(f_\beta) = -\frac{1}{n} \sum_{i=1}^{n} \delta_i \Big\{ f_\beta(x_i) - \log \sum_{i' : z_{i'} \geq z_i} e^{f_\beta(x_{i'})} \Big\}. \tag{5}$$

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

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathscr{B}_{p+d}(R)} \left\{ \ell_n(f_\beta) + \operatorname{bina}(\beta) \right\} \tag{6}$$

where $\mathscr{B}_{p+d}(R) = \{\beta \in \mathbb{R}^{p+d} : \|\beta\|_2 \leq R\}$ is the ℓ_2-ball of radius $R > 0$ in $\mathbb{R}^{p+d}$ and

$$\operatorname{bina}(\beta) = \sum_{j=1}^{p} \Big(\sum_{l=2}^{d_j+1} \omega_{j,l} |\beta_{j,l} - \beta_{j,l-1}| + \delta_1(\beta_{j,\bullet}) \Big), \tag{7}$$

where

$$\delta_1(u) = \begin{cases} 0 & \text{if} \quad \mathbf{1}^\top u = 0, \\ \infty & \text{otherwise} \end{cases}$$

and the weights $\omega_{j,l}$ are of order

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

see Appendix B.1 for their explicit form.

It turns out that the binarsity penalty is well suited for 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 $j \in \{1, \dots, p\}$, which means that the columns of each block sum to $\mathbf{1}$. This problem is solved since the penalty imposes the linear constraint $\sum_{l=1}^{d_j+1} \beta_{j,l} = 0$ in each block with the $\delta_1(\cdot)$ term. Then, the other term in the penalty consists in a within block weighted total-variation penalization

$$\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} = \sum_{l=2}^{d_j+1} \omega_{j,l} |\beta_{j,l} - \beta_{j,l-1}|, \tag{8}$$

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

Let us make a first assumption required for being sure to detect all cut-points.

Assumption 1. *We choose d_j such that $\min_{1 \le k \le K_j^\star + 1} |I^\star_{j,k}| \ge \max_{1 \le l \le d_j+1} |I_{j,l}|$ for all $j \in \{1, \dots, p\}$.*

This assumption ensures that for all features $j \in \{1, \dots, p\}$, there exists a unique interval $I_{j,l}$ containing cut-point $\mu^\star_{j,k}$, which we denote $I_{j,l^\star_{j,k}} = (\mu_{j,l^\star_{j,k}-1}, \mu_{j,l^\star_{j,k}}]$ for all $k \in \{1, \dots, K_j^\star\}$. Note that in practice, one can always work under Assumption 1 by increasing d_j.

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

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

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

$$\mathcal{A}_j(\hat{\beta}) = \{\hat{l}_{j,1}, \dots, \hat{l}_{j,s_j}\}, \tag{9}$$

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

$$\widehat{\mu}_{j,\bullet} = (\mu_{j,\hat{l}_{j,1}}, \dots, \mu_{j,\hat{l}_{j,s_j}})^\top \tag{10}$$

for $\mu^\star_{j,\bullet}$ and $j = 1, \dots, d$. By construction, $K_j^\star$ is estimated by $\widehat{K}_j = s_j$, see Appendix B.1 for its explicit form. Details on the algorithm used to solve the regularization problem (6) are given in Appendix A.1.

3 Theoretical guarantees

This paragraph is devoted to our theoretical result. In order to evaluate the prediction error, we construct an (empirical) Kullback-Leibler divergence KL_n between the true function $f^\star$ and any other candidate f as

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

This divergence has been introduced in Senoussi [1990]. The oracle inequality in Theorem 3 is expressed in terms of compatibility factor [van de Geer and Bühlmann, 2009] satisfied by the following nonnegative symmetric matrix

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

where

$$\bar{X}_n(s)=\frac{\sum_{i=1}^n X_i^B y_i(s)e^{f^\star(X_i)}}{\sum_{i=1}^n y_i(s)e^{f^\star(X_i)}} \text{ and } y_i(s)=\mathbb{E}[Y_i(s)|X_i].$$

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

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

where

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

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

Assumption 2. *Let $\varepsilon\in(0,1)$, and define*

- $f_\infty^\star=\max_{i=1,\ldots,n}|f^\star(X_i)|\le\sum_{j=1}^p\|\beta_{j,\bullet}^\star\|_\infty$,
- $r_\tau=(1/n)\mathbb{E}[\sum_{i=1}^n Y_i(\tau)e^{f^\star(X_i)}]$,
- $\Lambda_0^\star(\tau)=\int_0^\tau\lambda_0^\star(s)ds$,
- $t_{n,p,d,\varepsilon}$ *as the solution of* $(p+d)^2\exp\{-nt^2_{n,p,d,\varepsilon}/(2+2t_{n,p,d,\varepsilon}/3)\}=\varepsilon/2.221$.

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

$$\kappa_\tau^2(L)>\Xi_\tau(L)$$

where

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

Note that $\kappa_\tau^2(L)$ is the smallest eigenvalue of a population integrated covariance matrix defined in (11), so it is reasonable to treat it as a constant. Moreover, $t^2_{n,p,d,\varepsilon}$ is of order $(1/n)\log((p+d)^2/\varepsilon)$. If $|L|\log(p+d)/n$ is sufficiently small, then Assumption 2 is verified. With these preparations, let us now state the oracle inequality satisfied by our estimator of $f^\star$ which is by construction given by $\hat{f}=f_{\hat{\beta}}$ (see (4)).

Theorem. *Let $c_{p,R,K^\star}=\sqrt{p}R+\sum_{j=1}^p\|\beta_{j,\bullet}^\star\|_\infty$, $\psi(u)=e^u-u-1$, $\varrho>2c^2_{p,R,K^\star}/\psi(-c_{p,R,K^\star})$ and $\xi=2/\big(\varrho\psi(-c_{p,R,K^\star})/2c^2_{p,R,K^\star}-1\big)$. The following inequality*

$$KL_n(f^\star,f_{\hat{\beta}})\le(1+\xi)\inf_{\substack{\beta\in\mathscr{B}_{p+d}\\|\mathcal{A}(\beta)|\le K^\star\\\forall j,\mathbf{1}^\top\beta_{j,\bullet}=0}}\Big\{KL_n(f^\star,f_\beta)+\frac{512\varrho}{1-2c^2_{p,R,K^\star}/\varrho\psi(-c_{p,R,K^\star})}\frac{|\mathcal{A}(\beta)|\max_{j=1,\ldots,p}\|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_\infty^2}{\kappa_\tau^2\big(\mathcal{A}(\beta)\big)-\Xi_\tau\big(\mathcal{A}(\beta)\big)}\Big\} \tag{12}$$

holds with probability greater than $1-28.55e^{-c}-e^{-nr_\tau^2/(8e^{2f_\infty^\star})}-3\varepsilon$, for some $c>0$.

The proof of the theorem is postponed to Appendix B.3. The second term in the right-hand side of (12) can be viewed as a variance term, and its dominant term satisfies

$$\frac{|\mathcal{A}(\beta)| \max_{j=1,\dots,p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_\infty^2}{\kappa_\tau^2\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)} \lesssim \frac{|\mathcal{A}(\beta)|}{\kappa_\tau^2\big(\mathcal{A}(\beta)\big) - \Xi_\tau\big(\mathcal{A}(\beta)\big)} \frac{\log(p+d)}{n} \tag{13}$$

where the symbol $\lesssim$ means that the inequality holds up to multiplicative constant. The complexity term in (13) depends on both the sparsity of the vector β relatively to the total-variation penalization (through $|\mathcal{A}(\beta)|$) and the compatibility factor. Finally, the rate of convergence of the estimator $\hat{f} = f_{\hat{\beta}}$ has the expected shape $\log(p+d)/n$.

4 Performance evaluation

4.1 Practical details

Let us give some details about Binacox's use in practice. First, instead of taking the uniform grid for the intervals $I_{j,l}$ that makes theoretical results easier to state, we choose the estimated quantiles $\mu_{j,l} = q_j\big(l/(d_j+1)\big)$ where $q_j(u)$ denotes an empirical quantile of order u for $\boldsymbol{X}_{\bullet,j}$. This choice provides two major practical advantages: 1) the resulting grid is data-driven and follows the distribution of $\boldsymbol{X}_{\bullet,j}$ and 2) 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 $j \in \{1,\dots,p\}$ in practice), increasing d_j barely changes the results since the cut-points selected by the penalization do not change any more, and the size of each block automatically adapts itself to the data: depending on the distribution of $\boldsymbol{X}_{\bullet,j}$, ties may appear in the corresponding empirical quantiles (for more details on this last point, see Alaya et al. [2017]).

Then, let us precise that the Binacox is proposed in the `tick` library [Bacry et al., 2017], we provide sample code for its use in Figure 1. For practical convenience, we take all weights $\omega_{j,l} = \gamma$ and select the hyper-parameter γ using a V-fold cross-validation procedure with $V = 10$, taking the negative partial log-likelihood defined in (5) 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 $\beta^\star$ estimate in practice), which intuitively makes sense. Figure 9 in Appendix A.2 gives the learning curves obtained with this cross-validation procedure on an example.

We also add a simple de-noising step in the cut-point detection phase which is useful in practice. Indeed, it is usual to observe two consecutive $\hat{\beta}$'s jumps in the neighbourhood of a true cut-point, leading to an over-estimation of $K^\star$. This can be viewed as a clustering problem. We tried different clustering methods but in practice, nothing works better than this simple routine: if $\hat{\beta}$ has three consecutive different coefficients within a group, then only the largest jump is considered as a "true" jump. Figure 10 in Appendix A.2 illustrates this last point.

4.2 Simulation

Design. In order to assess the methods, we perform an extensive Monte Carlo simulation study. We first take $[x_{ij}] \in \mathbb{R}^{n\times p} \sim \mathcal{N}\big(0, \Sigma(\rho)\big)$, with $\Sigma(\rho)$ a $(p \times p)$ Toeplitz covariance matrix [Mukherjee and Maiti, 1988] with correlation $\rho \in (0,1)$. For each feature $j \in \{1,\dots,p\}$, we sample the cut-points $\mu_{jk}^\star$ uniformly without replacement among the estimated quantiles $q_j(u/10)$ for $u \in \{1,\dots,9\}$ for $k \in \{1,\dots,K_j^\star\}$. This way, we

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

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

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

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

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

Fig. 1: Sample python code for the use of Binacox in the `tick` library, with the use of the `FeaturesBinarizer` transformer for features binarization.

avoid having undetectable cut-points (with very few examples above the cut-point value) as well as two cut-points indissociable because too close. We choose the same $K_j^\star$ values for all $j \in \{1, \dots, p\}$. Now that the true cut-points $\mu^\star$ are generated, one can compute the corresponding binarized version of the features that we denote $x_i^{B^\star}$ for example i. Then, we generate $c_{jk} \sim (-1)^k |\mathcal{N}(1, 0.5)|$ for $k \in \{1, \dots, K_j^\star + 1\}$ and $j \in \{1, \dots, p\}$ to make sure we create "real" cut-points, and take $\beta_{jk}^\star = c_{jk} - (K_j^\star + 1)^{-1} \sum_{k=1}^{K_j^\star + 1} c_{jk}$ 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 $\beta_{jk}^\star = 0$ for all $k \in \{1, \dots, K_j^\star + 1\}$. Finally, we generate survival times using Weibull distributions, which is a common choice in survival analysis [Klein and Moeschberger, 2005], that is $T_i \sim \nu^{-1} \big[-\log(U_i) \exp\big(-(x_i^{B^\star})^\top \beta_i^\star \big) \big]^{1/\varsigma}$ with $\nu > 0$ and $\varsigma > 0$ the scale and shape parameters respectively, $U_i \sim \mathcal{U}([0, 1])$ and where $\mathcal{U}([a, b])$ stands for the uniform distribution on a segment $[a, b]$. The distribution of the censoring variable C_i is geometric $\mathcal{G}(\alpha_c)$, where $\alpha_c \in (0, 1)$ is empirically tuned to maintain a desired censoring rate $r_c \in [0, 1]$. The choices of the hyper-parameters is driven by the applications on real data presented in Section 5 and are summarized in Table 1. Figure 2 gives an example of data generated according to the design we just described.

Table 1: Hyper-parameters choice for simulation.

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

Metrics. We evaluate the considered methods using two metrics. The first one assesses the estimation of the cut-points values by $m_1 = |\mathcal{S}'|^{-1} \sum_{j \in \mathcal{S}'} \mathcal{H}(\mathcal{M}_j^\star, \widehat{\mathcal{M}}_j)$ where $\mathcal{M}_j^\star = \{\mu_{j,1}^\star, \dots, \mu_{j,K_j^\star}^\star\}$ (resp. $\widehat{\mathcal{M}}_j = \{\hat{\mu}_{j,1}, \dots, \hat{\mu}_{j,\widehat{K}_j}\}$) is the set of true (resp. estimated) cut-points for feature j, $\mathcal{S}' = \big\{j,\ j \notin \mathcal{S} \cap \{l,\ \widehat{\mathcal{M}}_l = \varnothing\}\big\}$ is the indexes corresponding to features with at least one true cut-point and one detected cut-point, and $\mathcal{H}(A, B)$ is the Hausdorff

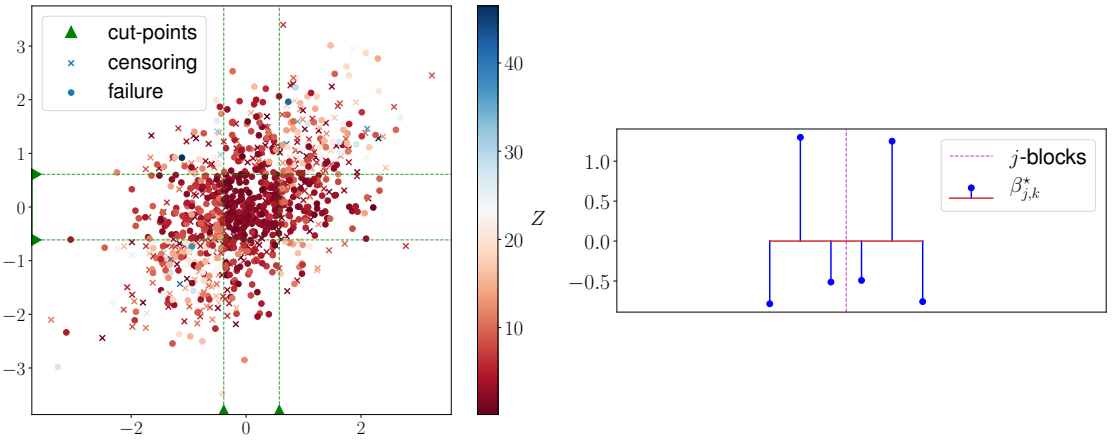


Fig. 2: On the left, illustration of data simulated with $p = 2$, $K_1^\star = K_2^\star = 2$ and $n = 1000$. Dots represent failure times ($z_i = t_i$) while crosses represent censoring times ($z_i = c_i$), and the colour gradient represents the z_i values (red for low and blue for high values). On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

distance between the two sets A and B, that is $\mathcal{H}(A, B) = \max\big(\mathcal{E}(A||B), \mathcal{E}(B||A)\big)$ with $\mathcal{E}(A||B) = \sup_{b\in B} \inf_{a\in A} |a - b|$ for two sets A and B. This is inspired by Harchaoui and Lévy-Leduc [2010], except that in our case, both $\mathcal{M}_j^\star$ and $\widehat{\mathcal{M}}_j$ can be empty, which explain the use of $\mathcal{S}'$. The second metric we use is precisely focused on the sparsity aspect: it assesses the ability for each method to detect features with no cut-points and is defined by $m_2 = |\mathcal{S}|^{-1} \sum_{j\in\mathcal{S}} \widehat{K}_j$.

4.3 Competing methods

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

We describe in what follows the principle of the univariate log-rank tests. Consider one of the initial variable $\boldsymbol{X}_{\bullet,j} = (x_{1,j}, \ldots, x_{n,j}\}$, and define its 10th and 90th quantiles as $x_{10th,j}$ and $x_{90th,j}$. Define then a grid $\{g_{j,1}, \ldots, g_{j,\kappa_j}\}$. In most implementations, the $g_{j,k}$'s are chosen at the original observation points and 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 to the univariate Cox model

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

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 dozens to all observed values between $x_{10th,j}$ and $x_{90th,j}$.

In Figure 3, the values $-\log(\mathrm{pv}_{j,k})$ for $k = 1, \ldots, \kappa_j$ (denoted by "MT" for multiple testing) are represented, for the simulated example described in Figure 2. Notice that the level $-\log(\alpha) = -\log(0.055)$ is exceed at numerous $g_{j,k}$'s. 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", for this reason we do not consider this approach.

To cope with the multiple testing (MT) problem at hand, a multiple test correction has to be applied. We consider two corrections. This first is the well-known Bonferroni p-values correction, referred to as MT-B. We insist on the fact that, although commonly used, this method is not correct in this situation as the p-values are correlated. Note also that in this context, the Benjamini–Hochberg (BH) procedure would result in the same detection 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, denoted MT-LS, is the correction proposed in Lausen and Schumacher [1992], based on asymptotic theoretical considerations. Figure 3 illustrates how these methods behave on a simulated example. A third correction we could think of would be a bootstrap based MaxT procedure (or MinP) developed in Dudoit and Van Der Laan [2007] or Westfall et al. [1993], but this would be intractable in our high-dimensional setting (see Figure 4a that only considers a single feature, and a bootstrap procedure based on MT would dramatically increase the required computing time).

4.4 Results of simulation

Example. Figure 3 illustrates how the considered methods behave on the data illustrated in Figure 2. Through this example, one can visualise the good performances 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 the 2 features, namely the one corresponding to the higher jump in $\beta^\star_{j,\bullet}$ (see Figure 2): $\mu^\star_{1,1}$ and $\mu^\star_{2,2}$.

With regard 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 consider a method for detecting those preponderant maxima, but it is beyond the scope of the article (plus it would still be based on the MT methods, which has high computational cost, as detailed hereafter).

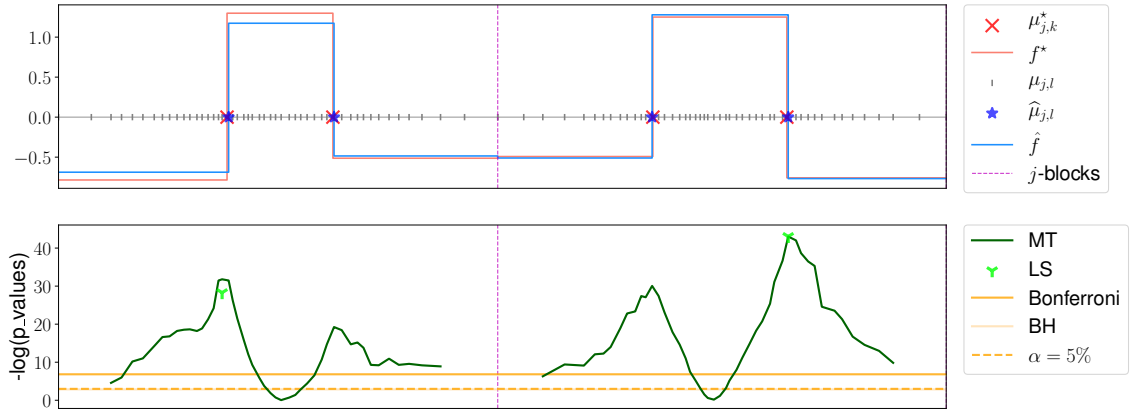
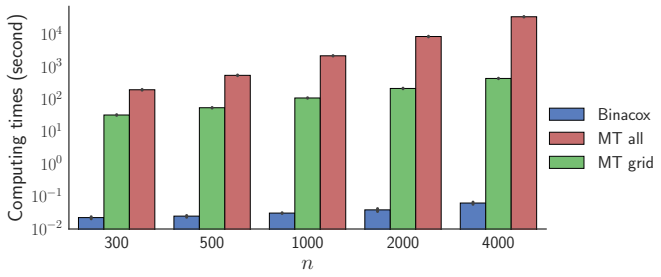


Fig. 3: Illustration of the main quantities involved in the Binacox on top, with estimation obtained on the data presented in Figure 2. Our algorithm detects the correct number of cut-points $\widehat{K}_j = 2$, and estimates their position very accurately, as well as their strength. At the bottom, one observe the results on the same data using the multiple testing related methods presented in Section 4.3. Here the BH threshold lines overlap the one corresponding to $\alpha = 5\%$. The BH procedure would consider as cut-point all $\mu_{j,l}$ value for which the corresponding darkgreen (MT) line value is above, then detecting far too many cut-points.

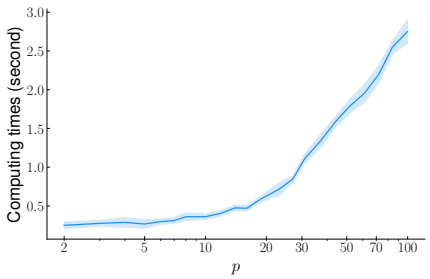
Computing times. Let us focus on the computing times required for the considered methods. The multiple testing related methods being univariate, one can directly parallelize their computation on the dimension p (which is what we did), and we consider here a single feature X ($p = 1$). Following the method explained in Section 4.3, we have to compute all log-rank test p-values computed on the two populations $\{y_i : x_i > \mu\}$ and $\{y_i : x_i \leq \mu\}$ for $i \in \{1, \ldots, n\}$, for μ taking all x_i values between the 10-th and 90-th empirical quantile of X. We denote "MT all" this method in Figure 4a that compares its computing times with the Binacox one for varying n, and where we add 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, and goes roughly from one to two orders of magnitude higher when n goes from 300 to 4000. Hence to make computations much faster, we consider the MT grid for all multiple testing related method in the sequel of the paper without mentioning it. The resulting loss of precision in the MT related methods is negligible for a high enough d_j value (we take 50 in practice).

Then, let us stress the fact that the Binacox is still roughly 5 times faster than the MT grid method, and it remains very fast when we increase the dimension, as shown in Figure 4b. It turns out that the computational time grows roughly logarithmically with p.



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



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

Fig. 4: Illustration of the computing times for the considered methods.

Performances comparison. Let us compare now the results of simulations in terms of m_1 and m_2 metrics introduced in Section 4.2. Figure 5 gives a comparison of the considered methods on the cut-points estimation aspect, hence in terms of m_1 score. It appears that the Binacox outperforms the multiple testing related methods when $K_j^\star > 1$, and is competitive when $K_j^\star = 1$ except for small values of n. This is due to an overestimation by the Binacox in the number of cut-points (see see Figure 6) when p is high for small n, which gives higher m_1 values, even if the "true" cut-point is actually well estimated. Note that for such p value, the Binacox is way faster than the multiple testing related methods.

Figure 6, on the other hand, assesses the ability for each method to detect features with no cut-points using the m_2 metric, that is to estimate $\hat{K}_j^\star = 0$ for $j \in \mathcal{S}$. The Binacox

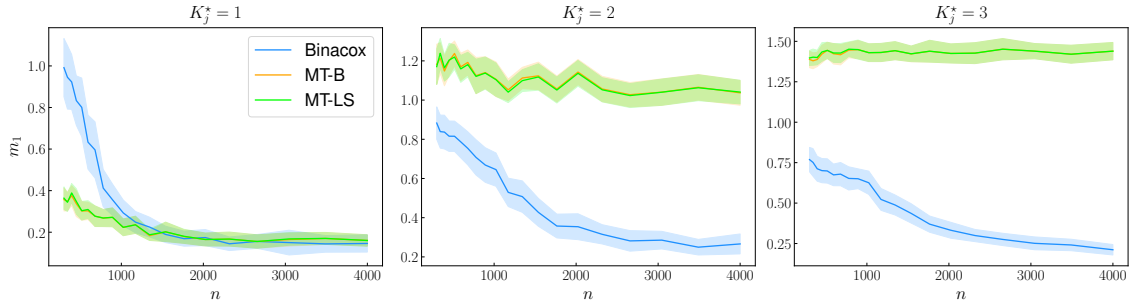


Fig. 5: Average (bold) m_1 scores and standard deviation (bands) obtained on 100 simulated datasets according to Section 4.2 with $p = 50$ and $K_j^\star$ equals to 1, 2 and 3 (for all $j \in \{1, \ldots, p\}$) for the left, center and right sub-figures respectively) for varying n. The lower m_1 the best result: the Binacox outperforms clearly other methods when there are more than one cut-point, and is competitive with other methods when there is only one cut-points with poorer performances when n is small because of an overestimation of $K_j^\star$ in this case.

appears to have a strong ability to detect features with no cut-point when n takes a high enough value compared to p, which is not the case for the multiple testing related methods.

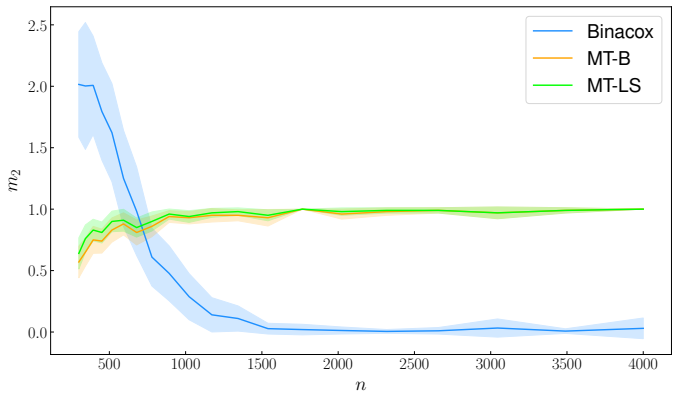


Fig. 6: Average (bold) m_2 scores and standard deviation (bands) obtained on 100 simulated datasets according to Section 4.2 with $p = 50$ for varying n. It turns out that MT-B and MT-LS tend to detect a cut-point while there is not (no matter the value of n), and that the Binacox overestimates the number of cut-points for small n values but detects well $\mathcal{S}$ for $p = 50$ on the simulated data when $n > 1000$.

5 Application to genetic data

In this section, we apply our method on three biomedical datasets. We extracted normalized expression data and survival times Z in days from breast invasive carcinoma (BRCA, $n = 1211$), glioblastoma multiforme (GBM, $n = 168$) and kidney renal clear cell carcinoma (KIRC, $n = 605$). These datasets are available on The Cancer Genome Atlas (TCGA) platform, which aims at accelerating the understanding of the molecular basis of cancer through the application of genomic technologies, including large-scale genome sequencing. For each patients, 20531 features corresponding to the normalized gene expressions are

available.

As we saw in Section 4.4, the multiple testing related methods are intractable in such high dimension. We therefore make a screening step to select the portion of features the most relevant for our problem among the 20531 ones. To do so, we fit our method on each block j separately and we compute the resulting $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ as a score that roughly assess the propensity for feature j to get one (or more) relevant cut-point. 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 11 of Appendix A.3.

Estimation results. Let us present in Figure 7 the results obtained by the considered methods on the GBM cancer dataset for the top-10 features ordered according to the Binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values. One can observe that all cut-points detected by the univariate multiple testing methods with Bonferroni (MT-B) or Lausen and Schumacher (MT-LS) correction are also detected by the multivariate Binacox that detects more cut-points, which is summarized in Table 2. It turns out that among the 20531 initial genes, the resulting top-10 are very relevant for a study on GBM cancer (being the most aggressive cancer that begins within the brain). For instance, the first gene SOD3 is related to an antioxidant enzyme that may protect in particular the brain from oxidative stress, which is believed to play a key role in tumour formation [Rajaraman et al., 2008]. Some other genes (like C11orf63 or HOXA1) are known to be directly related to the brain development [Canu et al., 2009].

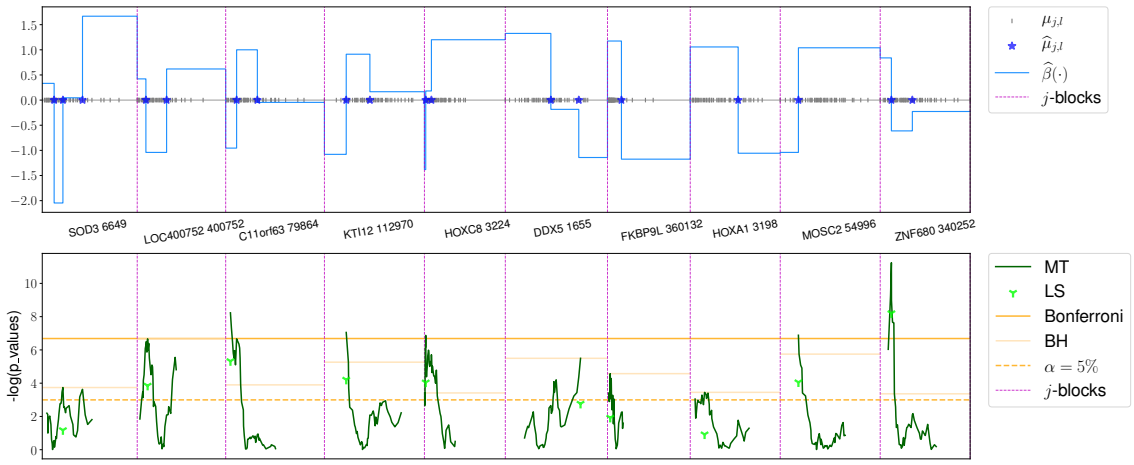


Fig. 7: Illustration of the results obtained on the top-10 features ordered according to the Binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the GBM dataset. The Binacox detects multiple cut-points and sheds light on non-linear effects for various genes. The BH thresholds are plotted for informational purposes, but are unusable in practice.

Similar results are obtained on the KIRC and BRCA cancers and are postponed in Appendix A.4.

Risk prediction. Let us now investigate how performances in terms of risk prediction are impacted when account is taken of the detected cut-points, namely let us compare predictions when training a Cox PH model either on the original continuous feature space versus on the $\widehat{\mu}$-binarized space constructed with the cut-points estimates.
In a classical Cox PH model, $R_i = \exp(X_i^\top \hat{\beta})$ is known as the predicted risk for patient i measured at $t = 0$. A common metric to evaluate risk prediction performances in a

Table 2: Estimated cut-points values for each method on the top−10 genes presented in Figure 7 for the GBM cancer. Dots (·) mean "no cut-point detected". The Binacox identifies much more cut-points than the univariate MT-B and MT-LS methods. But all cut-points detected by those two methods are also detected by the Binacox.

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

survival setting is the C-index [Heagerty and Zheng, 2005]. It 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 non-parametric and consistent estimator of $\mathcal{C}_\tau$ [Uno et al., 2011], already implemented in the `python` package `lifelines`. We randomly split the three datasets into a training and a testing sets (30% for testing) and compare the C-index on the test sets in Table 3 when the $\widehat{\mu}$-binarized space is constructed based on $\widehat{\mu}$ obtained either from the Binacox, MT-B or MT-LS. We also compare performances obtained by two nonlinear multivariate methods known to perform well in high-dimension: the boosting Cox PH (CoxBoost) [Li and Luan, 2005] used with 500 boosting steps, and the random survival forests (RSF) [Ishwaran et al., 2008] used with 500 trees, respectively implemented in the `R` packages `CoxBoost` and `randomForestSRC`. The Binacox method clearly improves risk prediction compare to classical Cox PH, as well as MT-B and MT-LS methods. Moreover, it also significantly outperforms both CoxBoost and RSF methods. Figure 8 compares the computing times of the considered methods. It appears that the Binacox is by far the most computationally efficient.

Table 3: C-index comparison for Cox PH model trained on continuous features vs. on its binarized version constructed using the considered methods cut-points estimates, and the CoxBoost and RSF methods. On the three datasets, the Binacox method gives by far the best results (in bold).

Cancer	Continuous	Binacox	MT-B	MT-LS	CoxBoost	RSF
GBM	0.660	**0.806**	0.753	0.768	0.684	0.691
KIRC	0.682	**0.727**	0.663	0.663	0.679	0.686
BRCA	0.713	**0.849**	0.741	0.738	0.723	0.746

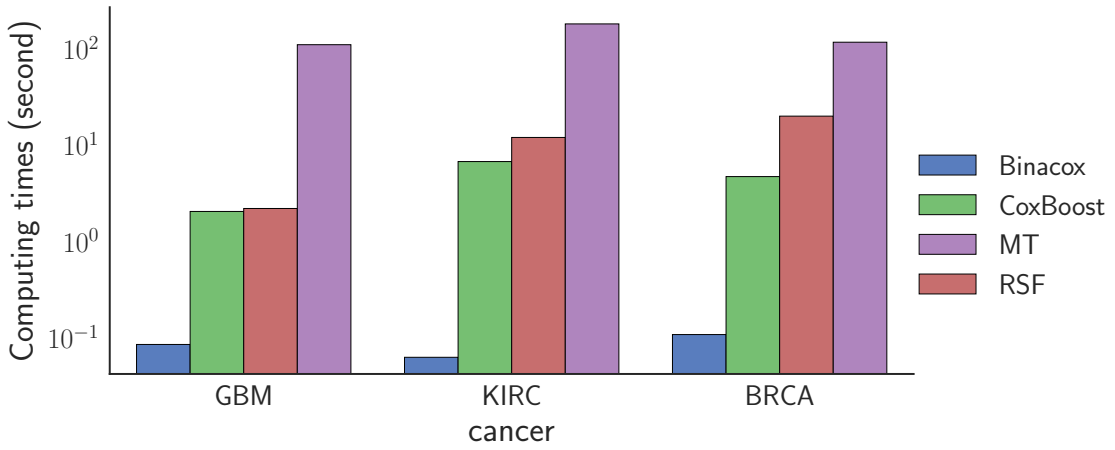


Fig. 8: Comparison of the computing times required by the considered method on the three datasets. The Binacox method is orders of magnitude faster.

6 Conclusion

In this paper, we introduced the Binacox designed for estimating multiple cut-points in a Cox PH model with high-dimensional features. We illustrated the good theoretical properties of the model by establishing non-asymptotic oracle inequality. An extensive Monte Carlo simulation study has been carried out to evaluate the performance of the developed estimation procedure. It showed that our approach outperforms existing methods with a computing time orders of magnitude faster. Moreover, it succeeds in detecting automatically multiple cut-points per feature. The proposed methodology has then been applied on two high-dimensional genetic public datasets. Many gene expressions pinpointed by the model are relevant for medical interpretations (e.g. the gene SOD3 for the GBM cancer), whilst others must involve further investigations in the genetic research community. Furthermore, the Binacox outperformed the classical Cox PH model in terms of risk prediction performances evaluated through the C-index metric. It can then be an interesting alternative to more classical methods found in the medical literature to deal with prognosis studies in a high dimensional framework, providing a new way to model non-linear features associations. More importantly, our method provides powerful interpretation aspects that could be useful in both clinical research and daily practice. Indeed, in addition to its raw feature selection ability, the estimated cut-points could directly be used in clinical routine. For instance, the Binacox directly estimates the impact on the survival risk for a feature (gene expression in our application) to be in a relevant interval through the estimated coefficient corresponding to this interval. Our study lays the groundwork for the development of powerful methods which could help provide 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' work has been supported by the INCA-DGOS grant PTR-K 2014. The results shown in this paper are based upon data generated by the TCGA Research Network and freely available from http://cancergenome.nih.gov. *Conflict of Interest*: None declared.

Software

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

Appendices

A Additional details

A.1 Algorithm.

To solve the regularization problem (6), we are first interested in the proximal operator of binarsity [Alaya et al., 2017]. It turns out that it can be computed very efficiently, using an algorithm [Condat, 2013] that we modify in order to include weights $\omega_{j,k}$. It applies in each group the proximal operator of the total-variation since binarsity penalty is block separable, followed by a centering within each block to satisfy the sum-to-zero constraint, see Algorithm 1 below. We refer to Alaya et al. [2015] for the weighted total-variation proximal operator.

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

Input: vector $\beta \in \mathscr{B}_{p+d}(R)$ and weights $\omega_{j,l}$ for $j = 1, \dots, p$ and $l = 1, \dots, d_j + 1$
Output: vector $\theta = \text{prox}_{\text{bina}}(\beta)$
for $j = 1$ **to** p **do**
  $\theta_{j,\cdot} \leftarrow \text{prox}_{\|\cdot\|_{\text{TV},\omega_{j,\cdot}}}(\beta_{j,\cdot})$ (TV-weighted prox in block j, see (8))
  $\eta_{j,\cdot} \leftarrow \theta_{j,\cdot} - \frac{1}{d_j+1}\sum_{l=1}^{d_j+1}\theta_{j,l}$ (within-block centering)
end for
Return: η

A.2 Implementation

Figure 9 gives the learning curves obtained during the V-fold cross-validation procedure detailed in Section 4.3 with $V = 10$ for fine-tuning parameter γ, being the strength of the binarsity penalty. We randomly split the data into a training and a validation set (30% for validation, cross-validation being done on the training). Recall that the score we use is the negative partial log-likelihood defined in (5) computed after a refit of the model on the binary space obtained by the estimated cut-points, with the sum-to-zero constraint in each block but without the TV penalty.

Figure 10 illustrates the denoising step when detecting the cut-points when looking at the $\hat{\beta}$ support relatively to the TV norm. The $\hat{\beta}$ vector plotted here corresponds to the data generated in Figure 2 of Section 4.2 and where final estimation results are presented in Figure 3 of Section 4.4. Since it is usual to observe three consecutive $\hat{\beta}$'s jumps in the neighbourhood of a true cut-point, which is the case in Figure 10 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 group, then only the largest jump is considered as a "true" jump.

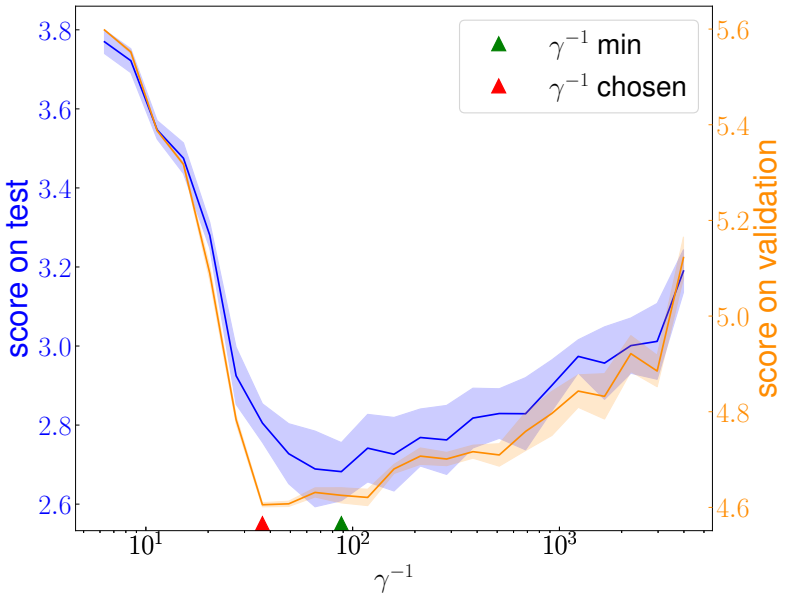


Fig. 9: Learning curves obtained for various γ, in blue on the changing test sets of the cross-validation, and in orange on the validation set. Bold lines represent average scores on the folds and bands represent gaussian 95% confidence intervals. The green triangle points out the value of γ^{-1} that 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, wich is a classical trick [Simon et al., 2011] that favors a slightly higher penalty strength (smaller γ^{-1}), to avoid an over-estimation of $K^\star$ in our case.

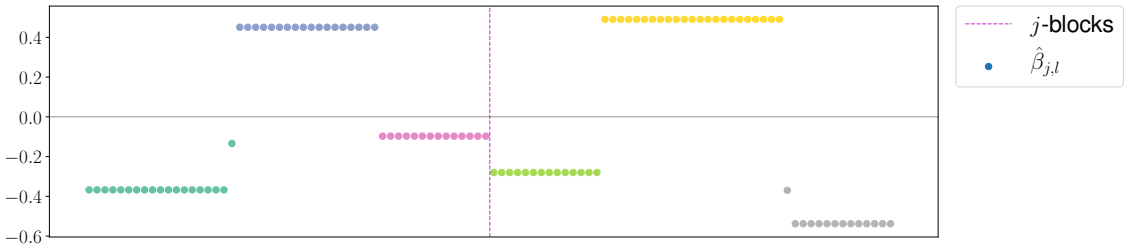


Fig. 10: Illustration of the denoising step on the cut-points detection phase. Within a block (separated with the dotted pink line), the different colors represent $\hat{\beta}_{j,l}$ with corresponding $\mu_{j,l}$ in distinct estimated $I^\star_{j,k}$. When a $\hat{\beta}_{j,l}$ is "isolated", it is assigned to its "closest" group.

A.3 TCGA genes screening

Figure 11 illustrates the screening procedure followed to reduce the high-dimension of the TCGA datasets to make the multiple testing related methods tractable. We then fit an 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 get one (or more) relevant cut-point. It appears that taking the top$-P$ features with $P = 50$ is a reasonable choice for each considered dataset.

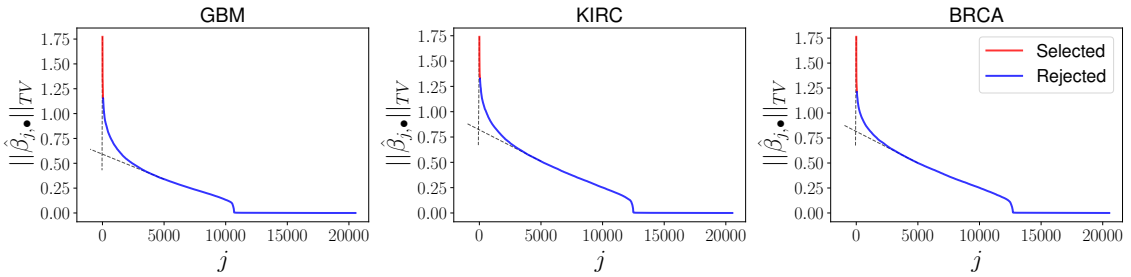


Fig. 11: $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ obtained on univariate Binacox fits for the three considered datasets. Top$-P$ selected features appear in red, and it turns out that taking $P = 50$ coincides with the elbow (represented with the dotted grey lines) in each three curves.

A.4 Results on BRCA and KIRC data

Figure 12 presents the results obtained by the considered methods on the BRCA cancer dataset for the top-10 features ordered according to the Binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values. Table 4 summarizes the detected cut-points values for each method. It turns out that the selected genes are very relevant for cancer studies (for instance, NPRL2 is a tumor suppressor gene [Huang et al., 2016]), and more particularly for breast cancer studies: for instance, HBS1L expression is known for being predictive of breast cancer survival [Antonov et al., 2014, Antonov, 2011, BioProfiling, 2009], while FOXA1 and PPFIA1 are highly related to breast cancer, see Badve et al. [2007] and Dancau et al. [2010] respectively.

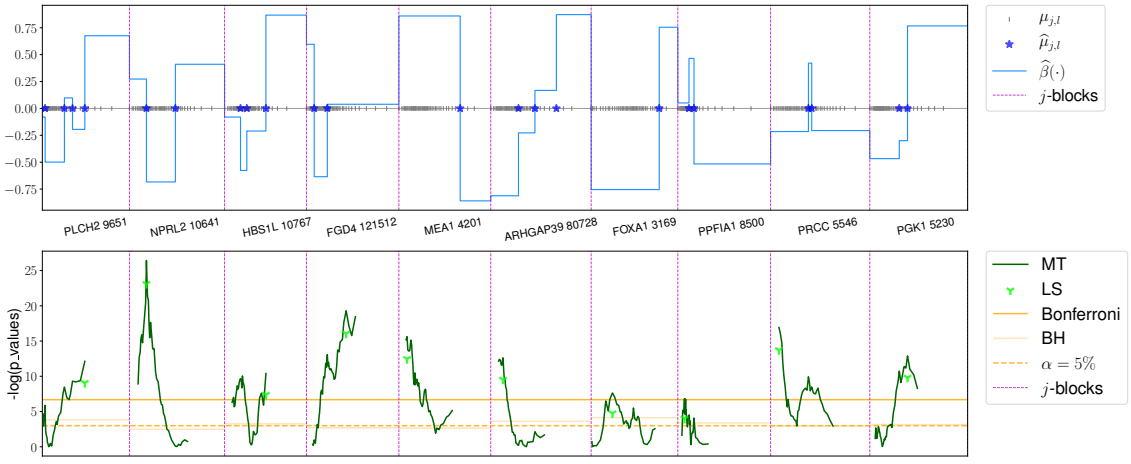


Fig. 12: Illustration of the results obtained on the top-10 features ordered according to the Binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the BRCA dataset.

Finally, Figure 13 gives the results obtained by the considered methods on the KIRC cancer dataset for the top-10 features ordered according to the Binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values and Table 5 summarizes the detected cut-points values for each method. Once again, the selected genes are relevant for cancer studies including kidney cancer. 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).

Table 4: Estimated cut-points values for each method on the top−10 genes presented in Figure 12 for the BRCA cancer.

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

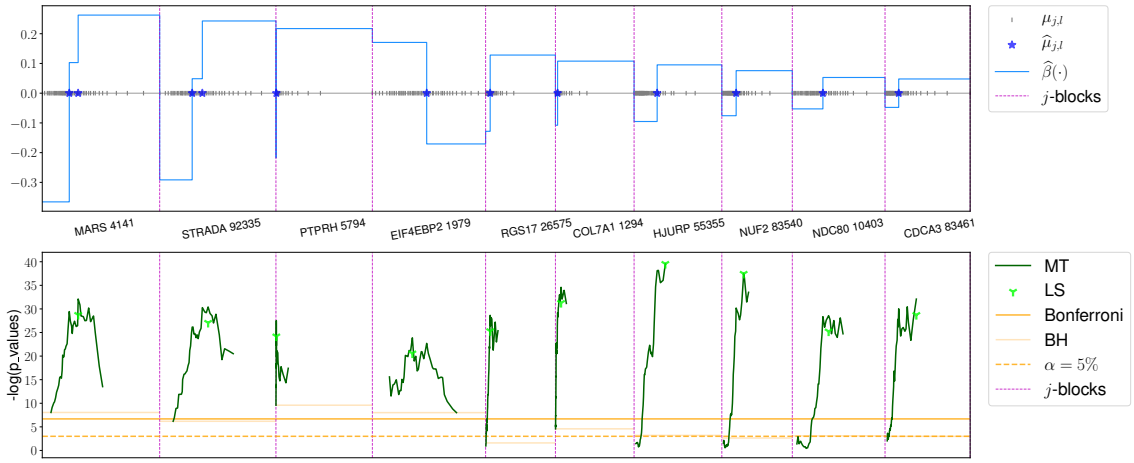


Fig. 13: Illustration of the results obtained on the top−10 features ordered according to the Binacox $\|\hat{\beta}_{j,\bullet}\|_{\mathrm{TV}}$ values on the KIRC dataset.

Table 5: Estimated cut-points values for each method on the top−10 genes presented in Figure 13 for the KIRC cancer.

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

B Proofs

In this section, we provide the proofs of the main theoretical results. Before that, we derive some preliminaries that will be used in the proofs.

B.1 Preliminaries to the proofs

Additional notations. For $u, v \in \mathbb{R}^m$, we denote by $u \odot v$ the Hadamard product $u \odot v = (u_1 v_1, \dots, u_m v_m)^\top$. We denote by $\mathrm{sign}(u)$ the subdifferential of the function $u \mapsto |u|$, that is $\mathrm{sign}(u) = \{1\}$ if $u > 0$, $\mathrm{sign}(u) = \{-1\}$ if $u < 0$, and $\mathrm{sign}(u) = [-1, 1]$ if $u = 0$. We write $\partial(\phi)$ the subdifferential mapping of a convex functional ϕ. We adopt in the proofs counting processes notations. 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 $\bar{N}(t) = n^{-1} \sum_{i=1}^n N_i(t)$. For every vector v, let $v^{\otimes 0} = 1$, $v^{\otimes 1} = v$, and $v^{\otimes 2} = vv^\top$ (outer product). Let $\tau > 0$ be the finite study duration.

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

$$\omega_{j,l} = 5.64 \sqrt{\frac{c + \log(p+d) + \mathscr{L}_{n,c}}{n}} + 18.62 \frac{(c + \log(p+d) + 1 + \mathscr{L}_{n,c})}{n} \tag{14}$$

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

Properties of binarsity penalty. We define $\omega = (\omega_{1,\bullet}, \dots, \omega_{p,\bullet})$ the weights vector, with $\omega_{j,1} = 0$ for all $j = 1, \dots, p$. Then, we rewrite the total variation part in binarsity 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 remark that for all $\beta_{j,\bullet} \in \mathbb{R}^{d_j+1}$, $\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} = \|\omega_{j,\bullet} \odot D_j \beta_{j,\bullet}\|_1$, where $\odot$ denotes the component-wise product (Hadamard product). 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

$$\mathbf{D} = \mathrm{diag}(D_1, \cdots, D_p) \text{ and } \mathbf{T} = \mathrm{diag}(T_1, \cdots, T_p). \tag{15}$$

We further prove that binarsity is a sub-additive penalty (see Kutateladze [2013] for the definition of sub-additive).

Lemma 1. *For all $\beta, \beta' \in \mathbb{R}^{p+d}$, one has*

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

Proof of Lemma 1. The hyperplane $\mathrm{span}\{u \in \mathbb{R}^{d_j+1} : \mathbf{1}_{d_j+1}{}^\top u = 0\}$ is a convex cone, then the indicator function δ_1 is sublinear (i.e., positively homogeneous + subadditive [Kutateladze, 2013]). Furthermore, the total variation penalization satisfies triangular inequality,

which gives the first statement of Lemma 1. To prove the second one, we use the fact that $\delta_1(\beta_{j,\bullet}) + \delta_1(-\beta_{j,\bullet}) \geq 0$, then we obtain

$$\text{bina}(-\beta) = \sum_{j=1}^{p} \Big(\|\beta_{j,\bullet}\|_{\text{TV},\omega_{j,\bullet}} + \delta_1(-\beta_{j,\bullet}) \Big) \leq \sum_{j=1}^{p} \Big(\|\beta_{j,\bullet}\|_{\text{TV},\omega_{j,\bullet}} + \delta_1(\beta_{j,\bullet})) \Big),$$

which concludes the proof of Lemma 1.

□

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

$$dN_i(t) = Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}dt + dM_i(t)$$

where the martingales M_i are square integrable and orthogonal.

With this notations, we define, for all $t \geq 0$ and f, the processes

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

for $r = 0, 1, 2$ and where X_i^B is the i-th row of the binarized matrix $\boldsymbol{X}^B$.

The empirical loss ℓ_n can then be rewritten as

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

Together with this loss, we introduced the loss

$$\begin{aligned}\ell(f) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big\{ f(X_i) - \log\big(S_n^{(0)}(f,t)\big)\big\} Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}dt \\ &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \log\big(e^{f(X_i)}/S_n^{(0)}(f,t)\big) Y_i(t)\lambda_0^\star(t)e^{f^\star(X_i)}dt. \end{aligned} \tag{16}$$

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

$$\begin{aligned}\nabla\ell_n(f_\beta) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{ X_i^B - \frac{S_n^{(1)}(f_\beta,t)}{S_n^{(0)}(f_\beta,t)}\Big\} dN_i(t) \\ &= \nabla\ell(f_\beta) + H_n(f_\beta) \end{aligned} \tag{17}$$

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

$$H_n(f_\beta) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \big\{ X_i^B - S_n^{(1)}(f_\beta,t)/S_n^{(0)}(f_\beta,t)\big\} dM_i(t) \tag{18}$$

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

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

with $\bar{f}(t) = \sum_{i=1}^n Y_i(t)e^{f^\star(X_i)}f(X_i)/S_n^{(0)}(f^\star,t)$. Lemma 3 below connects it to our empirical divergence.

B.2 Lemmas

Thereafter are some lemmas useful for the proof of our theorem. Their proofs are postponed to Section B.4

The following lemma is a consequence of the Karush-Kuhn-Tucker (KKT) optimality conditions [Boyd and Vandenberghe, 2004] for a convex optimization and the monotony of subdifferential mapping.

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

$$\langle \nabla \ell(f_{\hat\beta}), \hat\beta - \beta\rangle \le -\langle H_n(f_{\hat\beta}), \hat\beta - \beta\rangle - \langle h, \hat\beta - \beta\rangle.$$

The following lemma is derived from the self-concordance definition and Lemma 1 in Bach [2010].

Lemma 3. *Let $\hat\beta$ be defined by Equation (6) and $\beta \in \mathscr{B}_{p+d}(R)$, the following inequalities hold almost-surely*

$$KL_n(f^\star, f_\beta) - KL_n(f^\star, f_{\hat\beta}) + (\hat\beta - \beta)^\top \nabla \ell(f_{\hat\beta}) \ge 0 \tag{19}$$

$$\|f^\star - f_\beta\|_n^2 \frac{\psi(-\|f^\star - f_\beta\|_\infty)}{\|f^\star - f_\beta\|_\infty^2} \le KL_n(f^\star, f_\beta) \le \|f^\star - f_\beta\|_n^2 \frac{\psi(\|f^\star - f_\beta\|_\infty)}{\|f^\star - f_\beta\|_\infty^2}. \tag{20}$$

Let us define the nonnegative definite matrix

$$\widehat\Sigma_n(f^\star, \tau) = \sum_{i=1}^n \int_0^\tau \Big(X_i^B - \breve X_n(t)\Big)^{\otimes 2} \frac{Y_i(t)\exp f^\star(X_i)}{S_n^{(0)}(f^\star, t)} d\bar N(t),$$

where

$$\breve X_n(t) = \frac{\sum_{i=1}^n X_i^B Y_i(t) e^{f^\star(X_i)}}{\sum_{i=1}^n Y_i(t) e^{f^\star(X_i)}}.$$

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 our main theorem requires for the matrix $\widehat\Sigma_n(f^\star, \tau)$ to fulfill a compability condition. The following lemma shows that this is true with a large probability as long as Assumption 2 is true.

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

$$L_j = \{a_j^1, \dots a_j^{b_j}\} \subset \{1, \dots, d_j + 1\}, \tag{21}$$

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

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

where

$$\Xi_\tau(L) = 4|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\{(1+e^{2f^\star_\infty}\Lambda_0^\star(\tau))\sqrt{2/n\log(2(p+d)^2/\varepsilon)} + (2e^{2f^\star_\infty}\Lambda_0^\star(\tau)/r_\tau)t^2_{n,p,d,\varepsilon}\},$$

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

and

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

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 matrix of same size. For any $L=[L_1,\dots,L_p]$ concatenation of index subsets, then*

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

B.3 Proof of Theorem 3

Combining Lemmas 2 and 3, we get

$$KL_n(f^\star, f_{\hat\beta}) \le KL_n(f^\star,f_\beta)+(\hat\beta-\beta)^\top\nabla\ell(f_{\hat\beta}) \le KL_n(f^\star,f_\beta)-\langle H_n(f_{\hat\beta}),\hat\beta-\beta\rangle-\langle h,\hat\beta-\beta\rangle.$$

If $-\langle H_n(f_{\hat\beta}),\hat\beta-\beta\rangle-\langle h,\hat\beta-\beta\rangle<0$, the theorem holds. Let us assume for now that $-\langle H_n(f_{\hat\beta}),\hat\beta-\beta\rangle-\langle h,\hat\beta-\beta\rangle\ge 0$.

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

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

This gives

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

Using the fact that $\langle \text{sign}(u), u\rangle = \|u\|_1$, we have that

$$\begin{aligned}
-\langle h, \hat{\beta}-\beta\rangle &\leq 2\sum_{j=1}^{p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)} \odot D_j(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_1 \\
&\quad - 2\sum_{j=1}^{p} \|(\omega_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)} \odot D_j(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_1 \\
&= 2\sum_{j=1}^{p} \|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\text{TV},\omega_{j,\bullet}} - 2\sum_{j=1}^{p} \|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j^{\complement}(\beta)}\|_{\text{TV},\omega_{j,\bullet}}. \qquad (22)
\end{aligned}$$

Inequality (22) therefore becomes

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

Using the fact that $\mathbf{T}\mathbf{D} = \mathbf{I}_{p+d}$ (see their definitions in Equation (15)), we get

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

On the event

$$\mathscr{E}_n := \left\{ |\mathbf{T}^\top H_n(f_{\hat{\beta}})| \leq (\omega_{1,1}, \ldots, \omega_{p,d_p+1}) \right\} \qquad (23)$$

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

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

Hence,

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

One therefore has

$$KL_n(f^\star, f_{\hat{\beta}}) \leq KL_n(f^\star, f_\beta) + 3\sum_{j=1}^{p} \|(\hat{\beta}_{j,\bullet}-\beta_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_{\text{TV},\omega_{j,\bullet}}. \qquad (24)$$

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

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

this means that $\hat{\beta} - \beta \in \mathscr{C}_{\mathrm{TV},\omega}(\mathcal{A}(\beta))$ and $\mathbf{D}(\hat{\beta} - \beta) \in \mathscr{C}_{1,\omega}(\mathcal{A}(\beta))$. Now returning to (24), by Lemma 4 and under Assumption 2, we get

$$KL_n(f^\star, f_{\hat{\beta}}) \leq KL_n(f^\star, f_\beta) + \frac{\|f_{\hat{\beta}} - f_\beta\|_n}{\sqrt{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))}}, \tag{25}$$

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 (25) fulfills

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

By (20) in Lemma 3, we get that

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

In addition, one can easily check that

$$\max_{i=1,\dots,n} \sup_{\beta \in \mathscr{B}_{p+d}(R)} |f_\beta(X_i)| \leq \sqrt{p} R,$$

hence

$$\|f^\star - f_\beta\|_\infty \leq \max_{i=1,\dots,n} \left\{|f^\star(X_i)| + |f_\beta(X_i)|\right\} \leq c_{p,R,K^\star} \text{ and}$$
$$\|f^\star - f_{\hat{\beta}}\|_\infty \leq \max_{i=1,\dots,n} \left\{|f^\star(X_i)| + |f_\beta(X_i)|\right\} \leq c_{p,R,K^\star}.$$

Now, using the fact that the function $u \mapsto \frac{\psi(-u)}{u^2}$ is decreasing, we get

$$\|f^\star - f_\beta\|_n \leq \sqrt{\frac{c_{p,R,K^\star}^2}{\psi(-c_{p,R,K^\star})} KL_n(f^\star, f_\beta)},$$
$$\text{and } \|f^\star - f_{\hat{\beta}}\|_n \leq \sqrt{\frac{c_{p,R,K^\star}^2}{\psi(-c_{p,R,K^\star})} KL_n(f^\star, f_{\hat{\beta}})}.$$

With theses bounds inequality (25) yields

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

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

$$KL_n(f^\star, f_{\hat{\beta}}) \leq KL_n(f^\star, f_\beta) + \frac{\varrho}{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa^2_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))} + \frac{2c^2_{p,R,K^\star}}{\varrho\psi(-c_{p,R,K^\star})} KL_n(f^\star, f_\beta) + KL_n(f^\star, f_{\hat{\beta}}))$$

and

$$\Big(1 - \frac{2c^2_{p,R,K^\star}}{\varrho\psi(-c_{p,R,K^\star})}\Big) KL_n(f^\star, f_{\hat{\beta}}) \leq \Big(1 + \frac{2c^2_{p,R,K^\star}}{\varrho\psi(-c_{p,R,K^\star})}\Big) KL_n(f^\star, f_\beta) + \frac{\varrho}{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa^2_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))}.$$

By choosing $\varrho > 2c^2_{p,R,K^\star}/\psi(-c_{p,R,K^\star})$, we obtain

$$KL_n(f^\star, f_{\hat{\beta}}) \leq (1+\xi) KL_n(f^\star, f_\beta) + \frac{1}{1 - \frac{2c^2_{p,R,K^\star}}{\varrho\psi(-c_{p,R,K^\star})}} \frac{\varrho}{(\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta)))\kappa^2_{\mathbf{T},\hat{\zeta}}(\mathcal{A}(\beta))}.$$

where

$$1 + \xi = \frac{\frac{\varrho\psi(-c_{p,R,K^\star})}{2c^2_{p,R,K^\star}} + 1}{\frac{\varrho\psi(-c_{p,R,K^\star})}{2c^2_{p,R,K^\star}} - 1} = 1 + \frac{2}{\frac{\varrho\psi(-c_{p,R,K^\star})}{2c^2_{p,R,K^\star}} - 1}.$$

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

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

Finally,

$$KL_n(f^\star, f_{\hat{\beta}}) \leq (1+\xi) KL_n(f^\star, f_\beta) + \frac{512\varrho}{(1 - \frac{2c^2_{p,R,K^\star}}{\varrho\psi(-c_{p,R,K^\star})})} \frac{|\mathcal{A}(\beta)| \max_{j=1,\ldots,p} \|(\omega_{j,\bullet})_{\mathcal{A}_j(\beta)}\|^2_\infty}{\kappa_\tau^2((\mathcal{A}(\beta)) - \Xi_\tau(\mathcal{A}(\beta))}.$$

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

Computation of $\mathbb{P}[\mathscr{E}_n^\complement]$. From the definition of H_n in Equation (18), $\mathbf{T}^\top H_n(f_{\hat{\beta}})$ has the form

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

So each component of this vector has the form needed to applied 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 lengths $d_j + 1$, $j = 1, \ldots, p$. We then denote by $\big(\mathbf{T}^\top H_n\big)_{j,l}$ the l-th component of the jth block.

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

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

We now use the Theorem 3 from Gaïffas and Guilloux [2012], hence we obtain

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

Then by choosing the $\omega_{j,l}$ as in (14), we conclude that $\mathbb{P}[\mathscr{E}_n^\complement] \leq 28.55e^{-c}$ for some $c > 0$. □

B.4 Proof of the Lemmas

B.4.1 Proof of Lemma 2

To characterize the solution of the problem (6), the following result can be sraightforwardly obtained using the Karush-Kuhn-Tucker (KKT) optimality conditions [Boyd and Vandenberghe, 2004] for a convex optimization. A vector $\hat\beta \in \mathbb{R}^{p+d}$ is an optimum of the objective function in (6) if and only if there exists three sequences of subgradients $\hat h = (\hat h_{j,\bullet})_{j=1,\dots,p}$ with $\hat h_{j,\bullet} \in \partial\big(\|\hat\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}}\big)$, $\hat g = (\hat g_{j,\bullet})_{j=1,\dots,p}$ with $\hat g_{j,\bullet} \in \partial\big(\delta_1(\hat\beta_{j,\bullet})\big)$ and $\hat k \in \partial\big(\delta_{\mathscr{B}_{p+d}(R)}(\hat\beta)\big)$ such that

$$\nabla \ell_n(f_{\hat\beta}) + \hat h + \hat g + \hat k = \mathbf{0}, \tag{26}$$

where

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

and where $\mathcal{A}(\hat\beta)$ is the active set of $\hat\beta$, see (9). The subgradient $\hat g_{j,\bullet}$ belongs to

$$\partial\big(\delta_1(\hat\beta_{j,\bullet})\big) = \big\{v \in \mathbb{R}^{d_j+1} : \langle \hat\beta_{j,\bullet} - \beta_{j,\bullet}, v\rangle \geq 0, \text{ for all } \beta \text{ such that } \mathbf{1}^\top\beta_{j,\bullet} = 0\big\},$$

and $\hat k$ to

$$\partial\big(\delta_{\mathscr{B}_{p+d}(R)}(\hat\beta)\big) = \big\{v \in \mathbb{R}^{p+d} : \langle \hat\beta - \beta, v\rangle \geq 0, \text{ for all } \beta \text{ such that } \|\beta\|_2 \leq R\big\},$$

From the Equality (26), consider a $\beta \in \mathbb{R}^{p+d}$, we obtain

$$\langle \nabla\ell_n(f_{\hat\beta}), \hat\beta - \beta\rangle + \langle \hat h + \hat g + \hat k, \hat\beta - \beta\rangle = 0$$

and, with Equation (17)

$$\langle \nabla\ell(f_{\hat\beta}), \hat\beta - \beta\rangle + \langle H_n(f_{\hat\beta}), \hat\beta - \beta\rangle + \langle \hat h + \hat g + \hat k, \hat\beta - \beta\rangle = 0$$

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

□

B.4.2 Proof of Lemma 3

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

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

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

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

and

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

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

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

We observe that for all $k = 0, 1, 2\ldots$

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

Then

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

and

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

Differenciating again, we obtain

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

Therefore, we have

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

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

$$G^{''}(0)\frac{\psi(-\|f_2\|_\infty)}{\|f_2\|_\infty^2} \leq G(\eta) - G(0) - \eta G^{'}(0) \leq G^{''}(0)\frac{\psi(\|f_2\|_\infty)}{\|f_2\|_\infty^2}. \tag{27}$$

We will apply inequalities in (27) in two situations:

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

In case #1,

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

so

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

With the left bound of the self-concordance inequality (27), we get result 1 of lemma 3.

In case# 2, one gets

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

which gives result 2 of Lemma 3.

□

B.4.3 Proof of Lemma 4

For any concatenation of index sets $L = [L_1, \dots, L_p]$, 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 6, which assures that if Assumption 2 is fulfilled our random bound $\hat{\kappa}_\tau(L)$ is bounded away from 0 with large probability. It bears resemblance with Theorem 4.1 of Huang et al. [2013] apart from the fact that we work here in a fixed design setting.

Lemma 6. *Let $L = [L_1, \dots, L_p]$ be a concatenation of index sets, then the following*

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

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

Proof of Lemma 6. The proof is adapted from the proof of Theorem 4.1 in Huang et al. [2013] and it is divided in 3 steps.

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

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

The (m,m') 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

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

which, in our case,is bounded by 4. We moreover know that

$$\int_0^\tau Y_i(t)dN_i(t)\leq 1\ \text{ for all }\ i=1,\dots,n.$$

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

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

Via an union bound, we get that

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

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

$$\mathbb{P}\Big[\hat{\kappa}_\tau^2(L)\geq\bar{\kappa}_\tau^2(L)-4|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\sqrt{2/n\log(2(p+d)^2/\varepsilon)}\Big]\geq 1-\varepsilon. \tag{28}$$

Step 2. Let

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

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

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

Hence

$$\bar{\Sigma}_n(f^\star,\tau) = \widetilde{\Sigma}_n(f^\star,\tau) - \frac{1}{n}\int_0^\tau S_n^{(0)}(f^\star,s)(\breve{X}_n(s)-\bar{X}_n(s))^{\otimes 2}\lambda_0^\star(s)ds. \tag{29}$$

We first boud the second term on the right-hand side of (29). Let

$$\Delta_n(s) = \frac{1}{n}S_n^{(0)}(f^\star,s)(\breve{X}_n(s)-\bar{X}_n(s)) = \frac{1}{n}\sum_{i=1}^n Y_i(s)e^{f^\star(X_i)}(X_i^B-\bar{X}_n(s))$$

so that for each (m,m')

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

In our setting, for each i and all $t\le\tau$, $Y_i(t)\exp(f^\star(X_i)\le e^{f_\infty^\star}$. By Hoeffding inequality implies

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

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') component of $\Delta_n^{\otimes 2}(s)$ is given by

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

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

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

where $\Lambda_0^\star(\tau)=\int_0^\tau\lambda_0^\star(s)ds$.

By Lemma 4.2 in Huang et al. [2013], we obtain

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

Thanks to (29), Lemma 5, and the above two probability bounds, we known that

$$\bar{\kappa}_\tau^2(L) \ge \tilde{\kappa}_\tau^2(L) - 8e^{2f_\infty^\star}\Lambda_0^\star(\tau)|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\frac{t_{n,p,d,\varepsilon}^2}{r_\tau}, \tag{30}$$

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

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

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

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

$$\tilde{\kappa}_\tau^2(L)\ge\kappa_\tau^2(L)-4e^{2f^\star_\infty}\Lambda_0^\star(\tau)|L|\Big(\frac{8\max_j(d_j+1)\max_{j,l}\omega_{jl}}{\min_{j,l}\omega_{j,l}}\Big)^2\sqrt{2/n\log(2(p+d)^2/\varepsilon)}, \quad (31)$$

Finally, the conclusion follows from (28), (30) and (31).This finishes the proof of Lemma 6. □

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

$$\mathscr{C}_{1,\omega}(K)=\Big\{u\in\mathbb{R}^d:\sum_{j=1}^p\|(u_{j,\bullet})_{K_j^\complement}\|_{1,\omega_{j,\bullet}}\le 3\sum_{j=1}^p\|(u_{j,\bullet})_{K_j}\|_{1,\omega_{j,\bullet}}\Big\}, \quad (32)$$

the following holds

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

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

□

B.5 Proof of Lemma 5

We have that $|\beta^\top\tilde{\Sigma}\beta-\beta^\top\Sigma\beta|\le\|\beta\|_1^2\max_{j,l}|\tilde{\Sigma}_{j,l}-\Sigma_{j,l}|$. Then, we get $\beta^\top\tilde{\Sigma}\beta\ge\beta^\top\Sigma\beta-\|\beta\|_1^2\max_{j,l}|\tilde{\Sigma}_{j,l}-\Sigma_{j,l}|$. So to get the desired result, it sufficient to control $\|\beta\|_1$ using the cone $\mathscr{C}_{\mathrm{TV},\omega}$. Note that for all $j=1,\dots,p$, we have $T_jD_j=I_{d_j+1}$, where I_{d_j+1} denotes the identiy matrix in $\mathbb{R}^{d_j+1}$. Then, we have for any β

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

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

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

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

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

Besides, one has $\|\beta_{j,\bullet}\|_{\mathrm{TV},\omega_{j,\bullet}} \leq 2 \max_{j,l} \omega_{j,l} \|\beta_{j,\bullet}\|_1$. Hence, we get

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

□

References

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

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

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

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

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993.

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

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

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

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

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

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

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

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

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

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

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