Binacox: automatic cut-points detection in high-dimensional Cox model

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

[1]LPSM, UMR 8001, Sorbonne University, Paris, France [2]Modal'X, UPL, Univ Paris Nanterre, F92000 Nanterre France [3]LaMME, UEVE and UMR 8071, Paris Saclay University, Evry, France
[4]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

Objectives

1. Introduce the cut-points detection problem
2. Present the estimation procedure
3. Give theoretical guarantees
4. Illustrate the method on simulated and real high-dimensional data

Introduction

Translating significant continuous prognostic biomarkers into clinical decision often requires determining cut-points. We introduce a prognostic method called *Binacox* to deal with this problem of detecting multiple cut-points per features in a multivariate high-dimentional survival setting.

Let us denote T and C the times of the event of interest and censoring times respectively, $X \in \mathbb{R}^p$ the vector of features, $Z = T \wedge C$ the right-censored time and $\Delta = \mathbb{1}(\{T \leq C\})$ the censoring indicator. Assume that intensity of events for patient i is given by

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

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

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

with $I_{j,k}^\star = (\mu_{j,k-1}^\star, \mu_{j,k}^\star]$ for $k \in \{1, \dots, K_j^\star + 1\}$. Our goal is to estimate simultaneously $\mu^\star$ and $\beta^\star$.

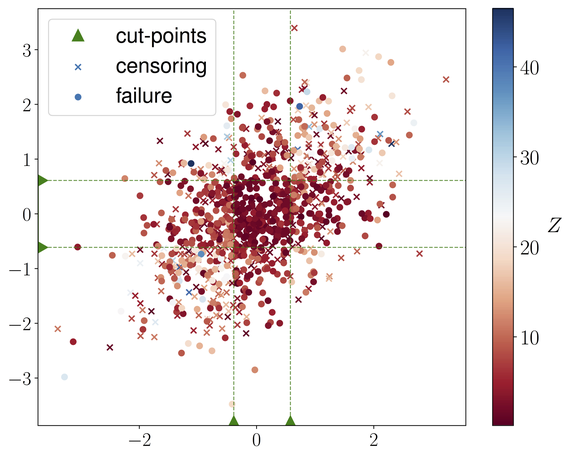


Fig.1: Illustration of data simulated according to (1).

Binarization

First, we one-hot encode all features to obtain

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

with $x_{i,j,l}^B = \begin{cases} 1 & \text{if } x_{i,j} \in I_{j,l}, \\ 0 & \text{otherwise,} \end{cases}$ and where $I_{j,l} = (\mu_{j,l-1}, \mu_{j,l}]$ with $\mu_{j,l} = l/(d_j + 1)$ for instance. Hence, we define

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

where $f_{\beta_{j,\bullet}}(x_i) = \sum_{l=1}^{d_j+1} \beta_{j,l} \mathbb{1}(x_{i,j} \in I_{j,l})$. Thus, f_β is a candidate for the estimation of $f^\star$ in (2).
We obtain the binarized partial negative log-likelihood

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

Estimation procedure

The optimization problem is

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

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

$$\mathrm{bina}(\beta) = \sum_{j=1}^{p} \Big(\sum_{l=2}^{d_j+1} \hat{\omega}_{j,l} |\beta_{j,l} - \beta_{j,l-1}| + \delta_1(\beta_{j,\bullet}) \Big),$$

with

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

Let $\mathcal{A}_j(\hat{\beta}) = \{l : \hat{\beta}_{j,l} \neq \hat{\beta}_{j,l-1}\} = \{\hat{l}_{j,1}, \dots, \hat{l}_{j,s_j}\}$, then one get

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

with $s_j = |\mathcal{A}_j(\hat{\beta})| = \widehat{K}_j$.

Fast oracle inequality in prediction

One has $KL_n(f^\star, f_{\hat{\beta}}) \leq (1 + c_1) \inf_{\substack{\beta \in \mathscr{B}_{p+d} \\ \forall j, \mathbf{1}^\top \beta_{j,\bullet} = 0}} \Big\{ KL_n(f^\star, f_\beta) + c_2 |\mathcal{A}(\beta)| \max_{j=1,\dots,p} \|(\hat{\omega}_{j,\bullet})_{\mathcal{A}_j(\beta)}\|_\infty^2 \Big\}$ with high-probability, where $\hat{\omega}_{j,l} = \mathcal{O}\big(d_j \sqrt{\pi_n \log(p+d)/n}\big)$ are data-driven weights, $\pi_n = |\{i = 1, \dots, n : \delta_i = 1\}|/n$, and c_1, c_2 are positive constants (with c_2 resulting from compatibily conditions).

Simulation study

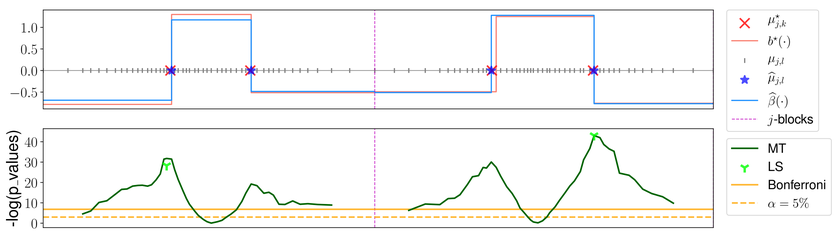


Fig.2: Results on data simulated in Figure 1.

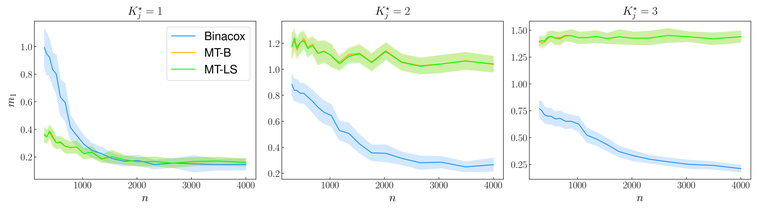
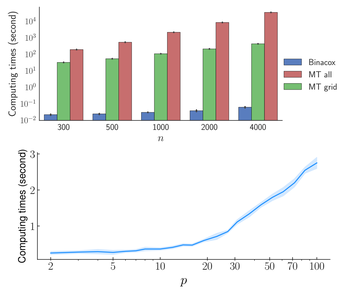


Fig.3: Comparison using metric $m_1 = |\mathcal{S}'|^{-1} \sum_{j \in \mathcal{S}'} \mathcal{H}(\mathcal{M}_j^\star, \widehat{\mathcal{M}}_j)$ where $\mathcal{M}_j^\star$ (resp. $\widehat{\mathcal{M}}_j$) is the set of true (resp. estimated) cut-points for feature j and $\mathcal{H}$ the Hausdorff distance, and $\mathcal{S}'$ the indexes of features with at least one true cut-point.



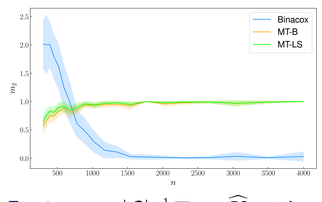


Fig.4: $m_2 = |\mathcal{S}|^{-1} \sum_{j \in \mathcal{S}} \widehat{K}_j$ with $\mathcal{S}$ the indexes of features with no cut-point.

Results on TCGA data

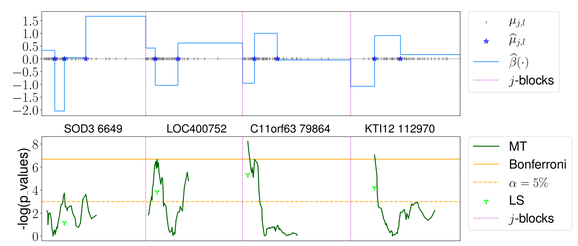


Fig.5: Illustration on the top-4 genes of GBM cancer. For instance, the first gene SOD3 is related to an antioxidant enzyme that may protect in particular the brain from oxidative stress, which is believed to play a key role in tumor formation.

Table: Comparison of risk prediction in terms of C-index on three TCGA datasets. Taking into account the detected cut-points significantly improves predictions.

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

Conclusion

Our method provides a new way to model nonlinear features associations, and powerful interpretation aspects that could be useful in both clinical research and daily practice: in addition to its raw feature selection ability, the estimated cut-points could directly be used in clinical routine.
Software: `github.com/SimonBussy/binacox`.

Contact Information

Simon Bussy, PhD Student
Email: `simon.bussy@gmail.com`
Website: `www.simonbussy.com`