56ièmes Journées de Statistique

Variable Selection With Error Control For High Dimensional Competing Risk Data

---

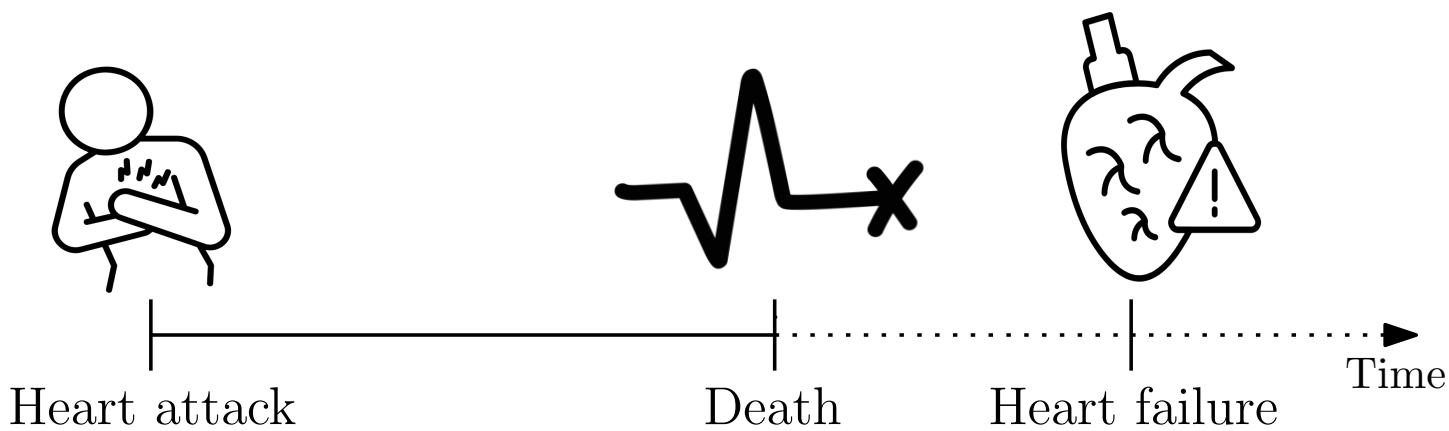
Presentation: CANNAFARINA Hugo,
1st year PhD student (Inria MODAL)

Supervisors: MAROT Guillemette (METRICS, Inria MODAL)
BABYKINA Evgeniya (METRICS)

Introduction

Competing risks:

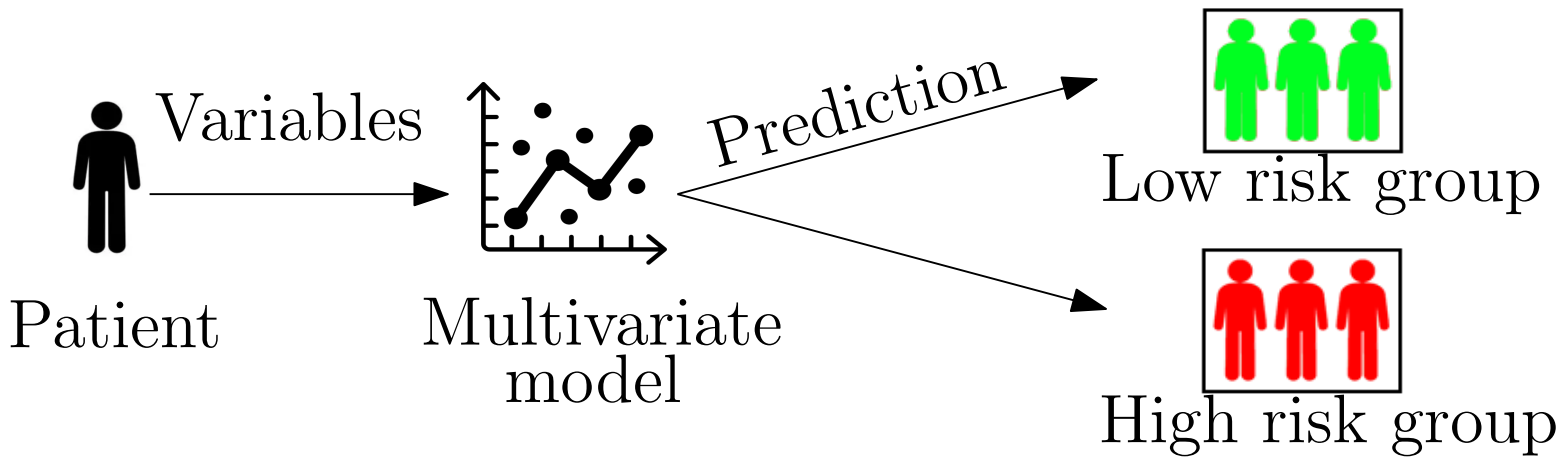
- Common in biomedical studies
- Occur when the event of interest cannot be observed because another event occurs first
- **Example** : time to heart failure



Introduction

Objective:

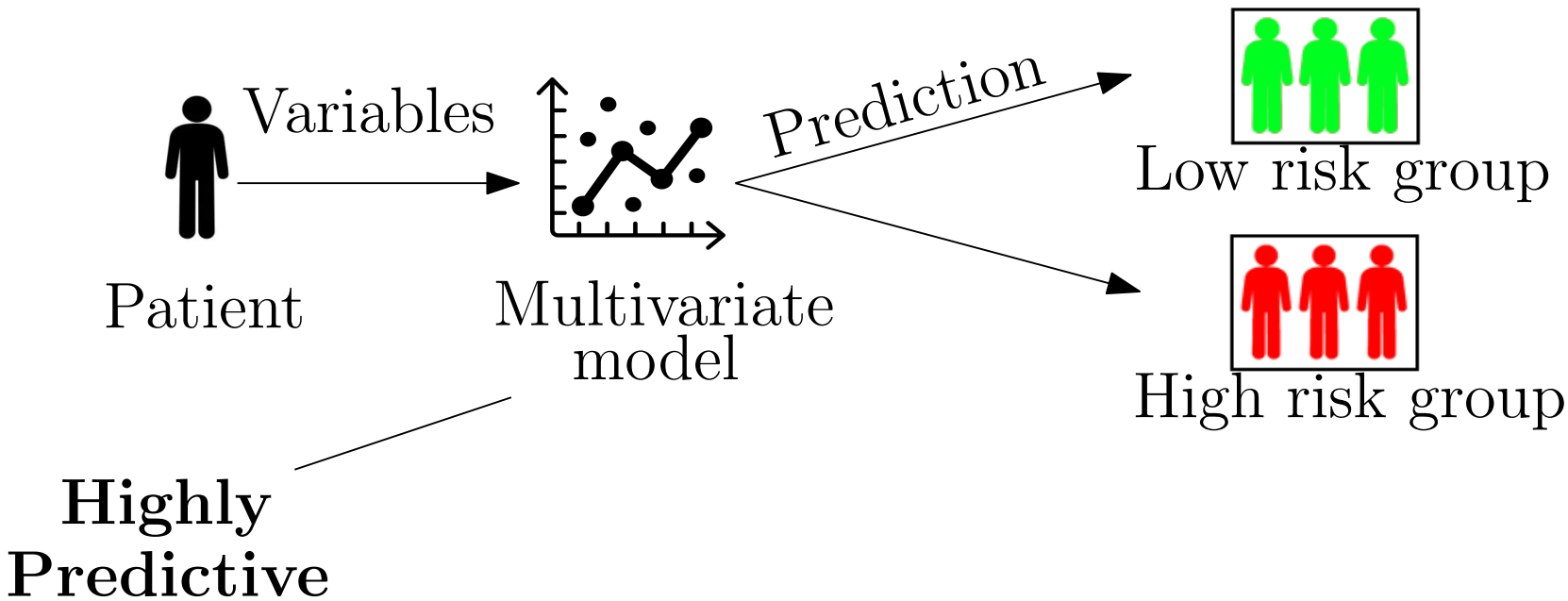
Select variables from **high dimensional data** to create multivariate models with **competing risks outcome** for patient stratification.



Introduction

Objective:

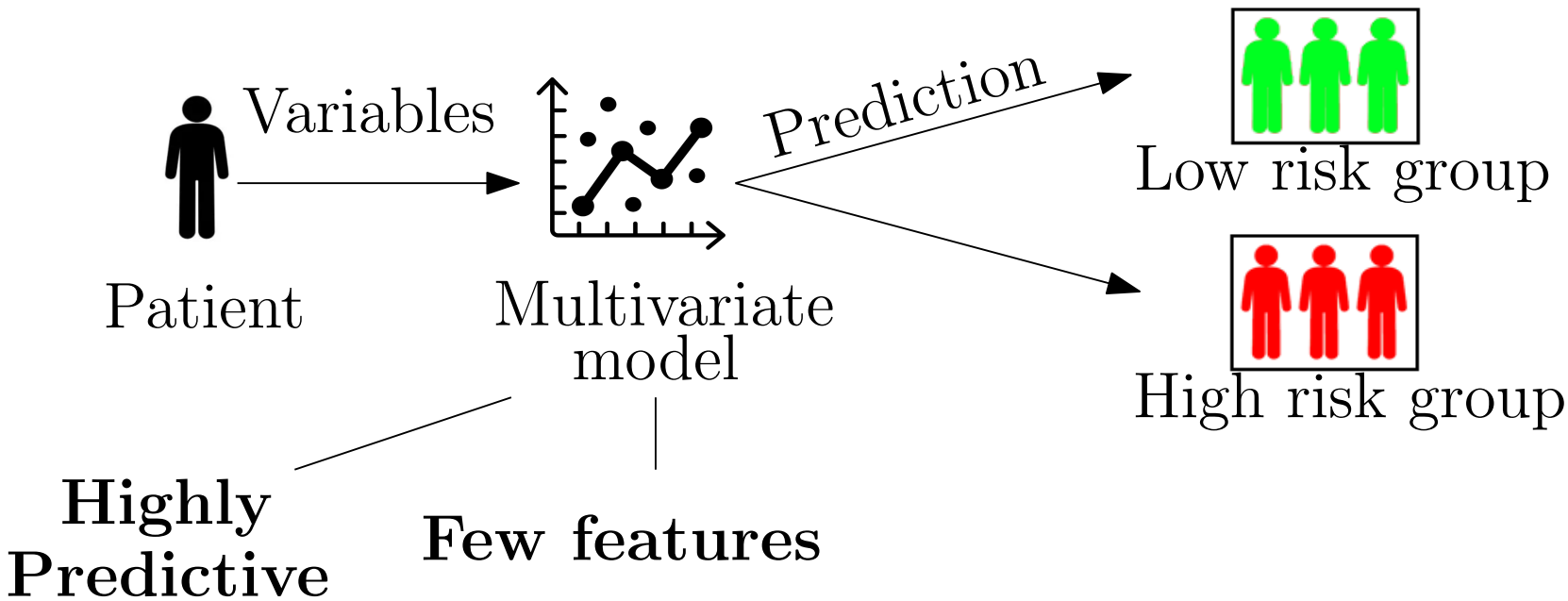
Select variables from **high dimensional data** to create multivariate models with **competing risks outcome** for patient stratification.



Introduction

Objective:

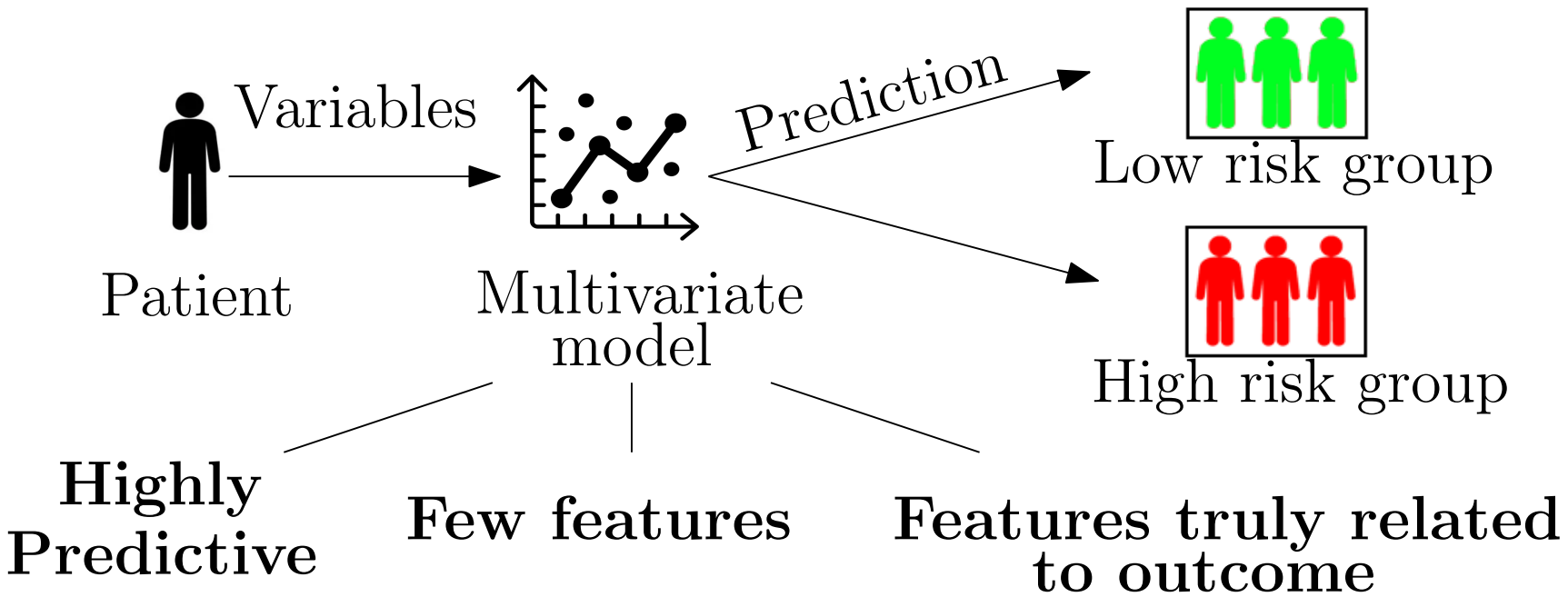
Select variables from **high dimensional data** to create multivariate models with **competing risks outcome** for patient stratification.



Introduction

Objective:

Select variables from **high dimensional data** to create multivariate models with **competing risks outcome** for patient stratification.



Introduction

Objective:

Select variables from **high dimensional data** to create multivariate models with **competing risks outcome** for patient stratification.

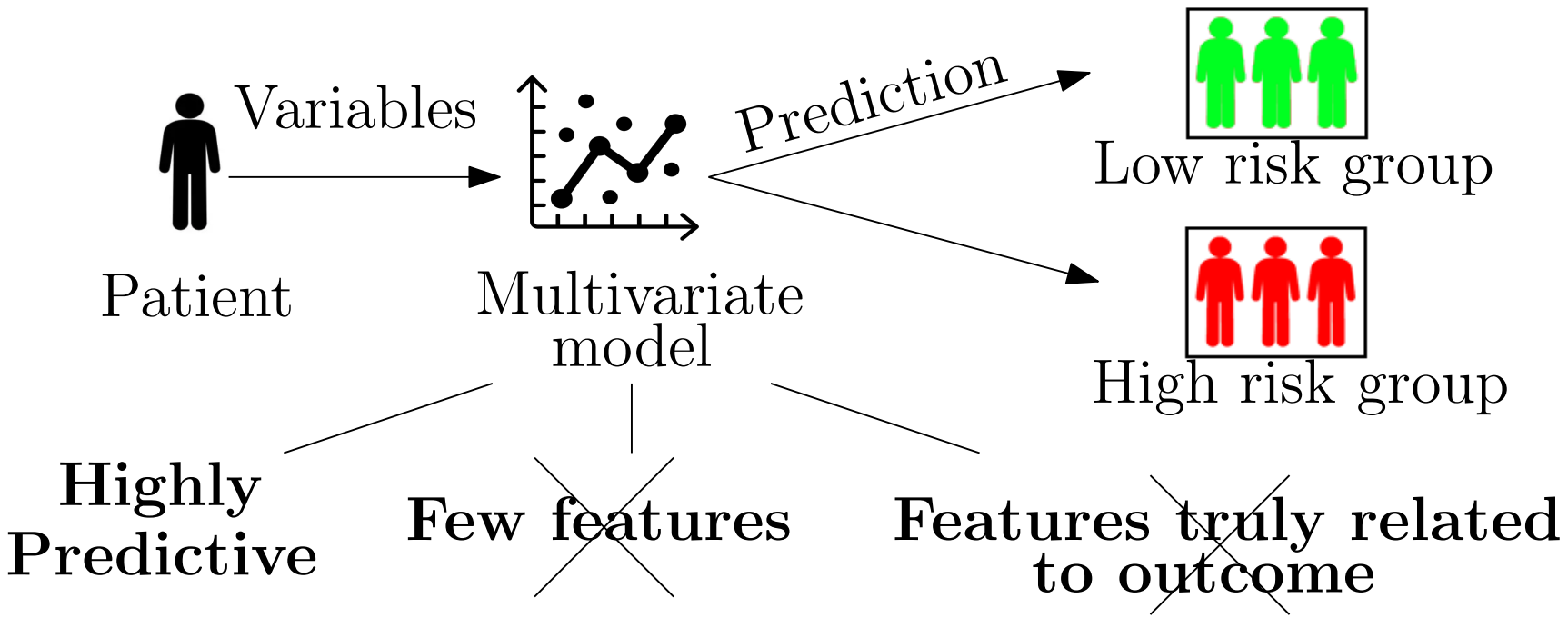
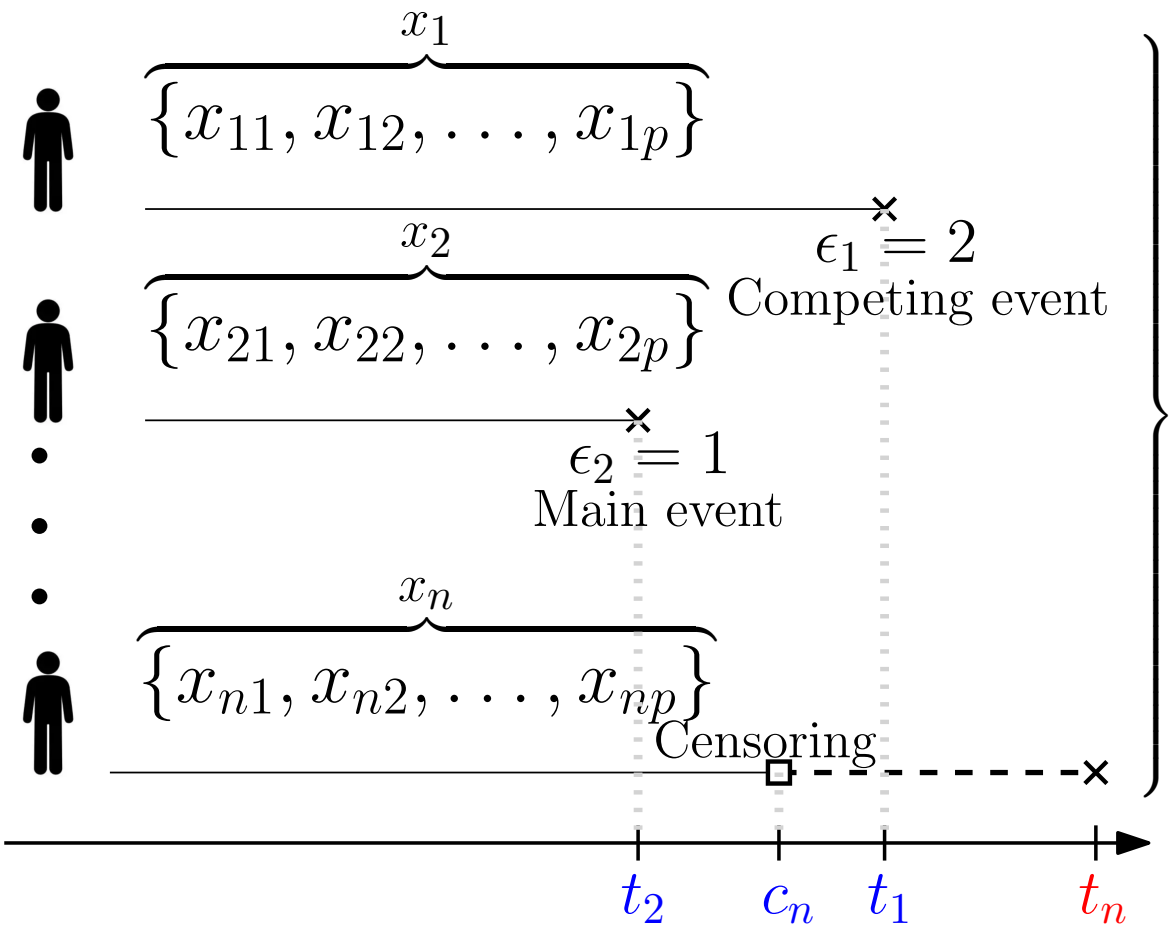


Table of Contents

Observed data



- **High dimension** : $p \gg n$
- ε indicator variable of the event (event 1 of interest)
- Observe $\min(t_i, c_i)$, δ_i, $\delta_i \varepsilon_i$, x_i with $\delta_i = I(t_i \leq c_i), i = 1 \ldots n$

CIF and subdistribution hazard

Cumulative Incidence Function (CIF) :

$$F_1(t; \mathrm{x}) = P(T \leq t, \varepsilon = 1 | \mathrm{x})$$

We want:

- High CIF in the high risk group
- Low CIF in the low risk group

$\longrightarrow$ Select variables that are related to the CIF.

CIF and subdistribution hazard

Cumulative Incidence Function (CIF) :

$$F_1(t;\mathrm{x}) = P(T \leq t, \varepsilon = 1|\mathrm{x})$$

We want:

- High CIF in the high risk group
- Low CIF in the low risk group

$\longrightarrow$ Select variables that are related to the CIF.

Subdistribution hazard

$$\begin{aligned}\gamma_1(t;\mathrm{x}) &= dF_1(t;\mathrm{x}) / \{1 - F_1(t;\mathrm{x})\} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\{t \leq T \leq t + \Delta t, \varepsilon = 1 \mid T \geq t \cup (T \leq t \cap \varepsilon \neq 1), \mathrm{x}\}\end{aligned}$$

Fine-Gray regression

Fine-Gray model (Fine and Gray 1999)

$$\gamma_1(t; x) = \gamma_{10} \exp(\beta^\top x)$$

- β vector of covariates of dimension $p \times 1$
- γ_{10} unspecified baseline hazard function

Partial log-likelihood :

$$I(\beta) = \sum_{i=1}^{n} \int_0^\infty \left[\beta^T x_i - \log \Sigma_j w_j(u) Z_j(u) \exp\left(\beta^T x_j\right)\right] \times w_i(u) dN_i(u)$$

with

- $N_i(t) = I\left(T_i \leq t, \varepsilon_i = 1\right)$
- $Z_j(t) = 1 - N_j(t-)$
- $w_i(t) = I\left(C_i \geq T_i \wedge t\right) \hat{G}(t) / \hat{G}\left(T_i \wedge t\right)$ with $G(t) = P(C \geq t)$.

Penalized Fine-Gray regression

Penalized Fine-Gray model

- Minimise $Q(\beta) = -l(\beta) + p_\lambda(\beta)$
- Selects variables in high dimension

Penalty choices :

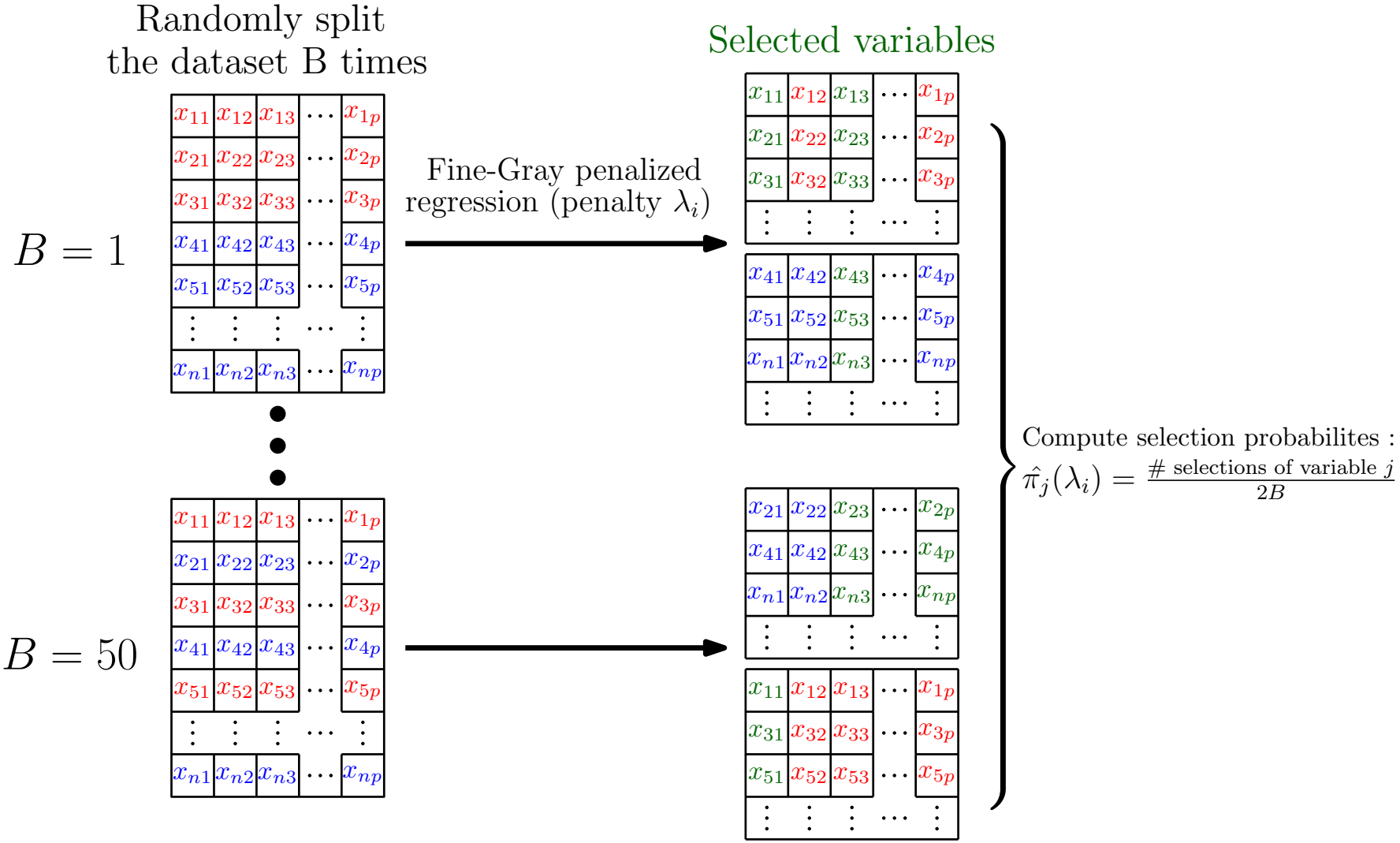
- Lasso (Tibshirani 1996): $p_\lambda(\beta_j) = \lambda|\beta_j|$
- SCAD (Fan and Li 2001) :
  $p'_\lambda(\beta_j) = \lambda I(|\beta_j| \leq \lambda) + \frac{(\alpha\lambda - |\beta_j|)_+}{\alpha - 1} I(|\beta_j| > \lambda)$
- MCP (Zhang 2010) : $p'_\lambda(\beta_j) = \left(\lambda - \frac{|\beta_j|}{\eta}\right)_+$

Optimal regularization parameter chosen by **cross-validation** or **BIC** minimization $\longrightarrow$ **many falsely selected variables**

Table of Contents

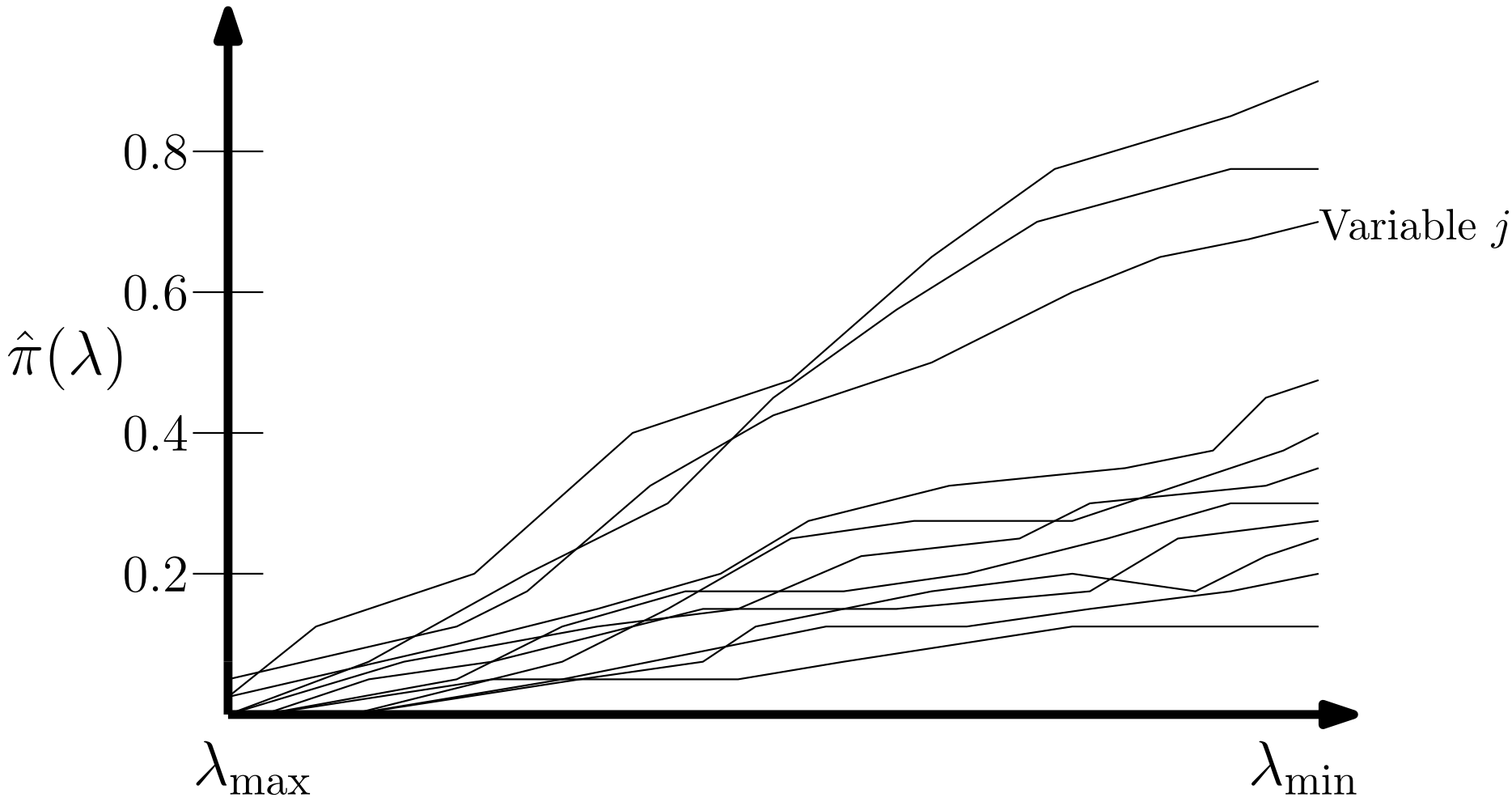
Stability Selection

For $\lambda_i \in \Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{15}\}$

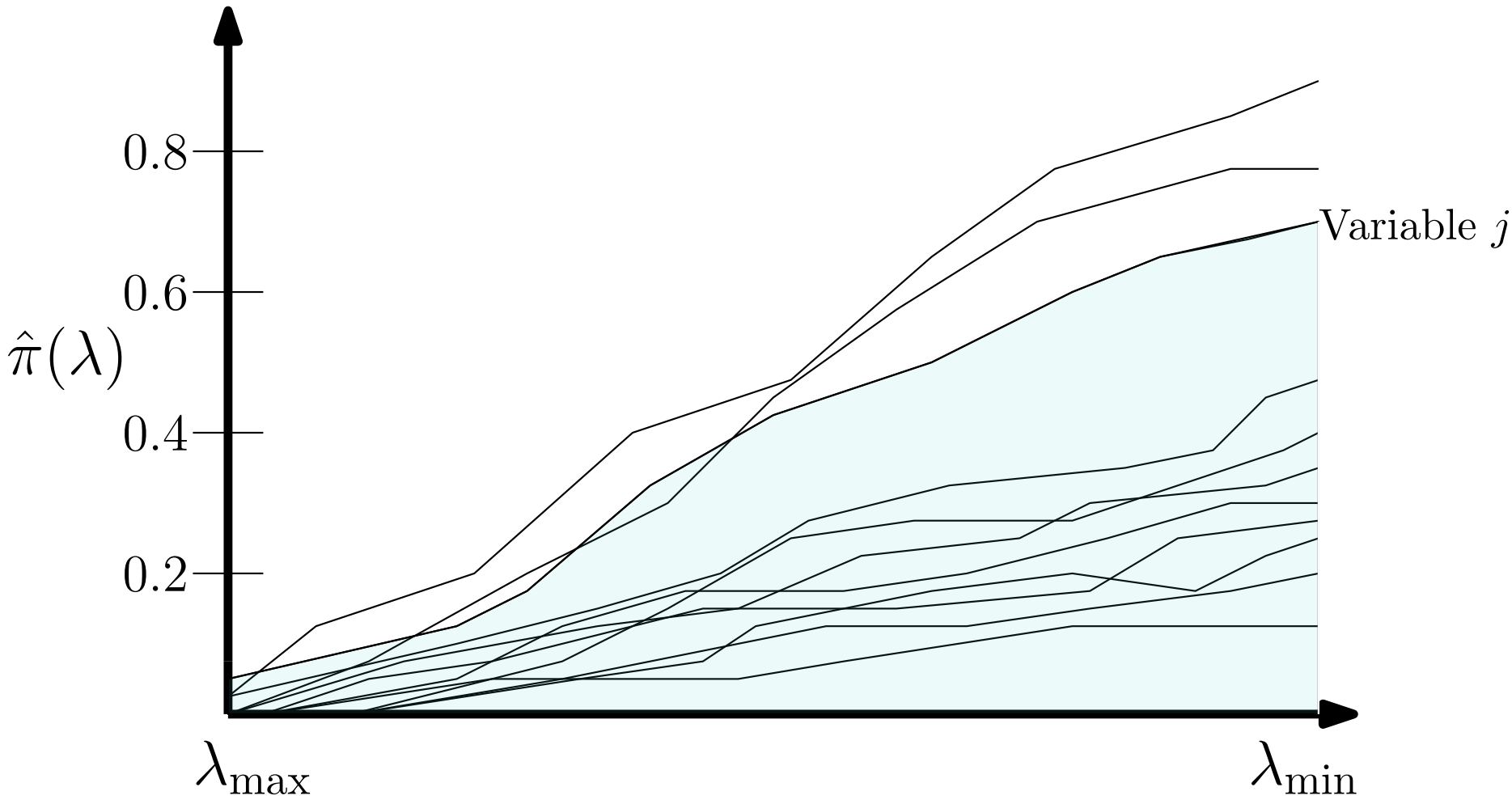


Meinshausen and Bühlmann 2010

Stability Paths



Stability Paths



$\longrightarrow$ Variables with a larger integral of the stability path are more likely to be truly related to the outcome.

Integrated path Stability Selection

Selected variables

Select the subset of variables $\hat{S}_{IPSS} \subseteq \{1, \ldots, p\}$ such that

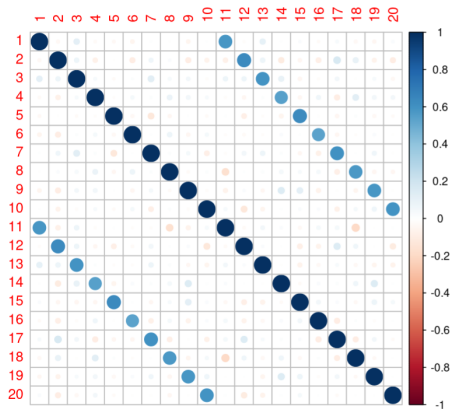
$$\hat{S}_{\text{IPSS}} = \left\{ j : \int_{\Lambda} h(\hat{\pi}_j(\lambda))\, C_{\Lambda} \lambda^{-1} d\lambda \geq \tau^* \right\}$$

with $h(x) = (2x - 1)^2 I(x \geq 0.5)$ and C_{Λ} a constant.

- τ^* depends on FP^*, the maximum desired amount of falsely selected variables.
- Under some regularity conditions, the average number of false positives in $\hat{S}_{\text{IPSS}}$ is below FP^* (Melikechi and Miller 2024)

Simulation study

- 100 datasets with 5000 variables each were generated according to Toeplitz's model : $x_i \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{jk} = \rho^{|j-k|}$ and $\rho = 0.6$, $i \in 1 \ldots n$ with $n = 300$.



- 10 out of 5000 variables, indexed by T, were set to be true predictors : $\beta_i \sim U(-1, 1), i \in T$ and $\beta_i = 0, i \notin T$
- The CIF was simulated as

$$F_1(t \mid \mathrm{x}) = 1 - \{1 - 0.75\,[1 - \exp(-t)]\}^{\exp(\mathrm{x}^T \beta)}$$

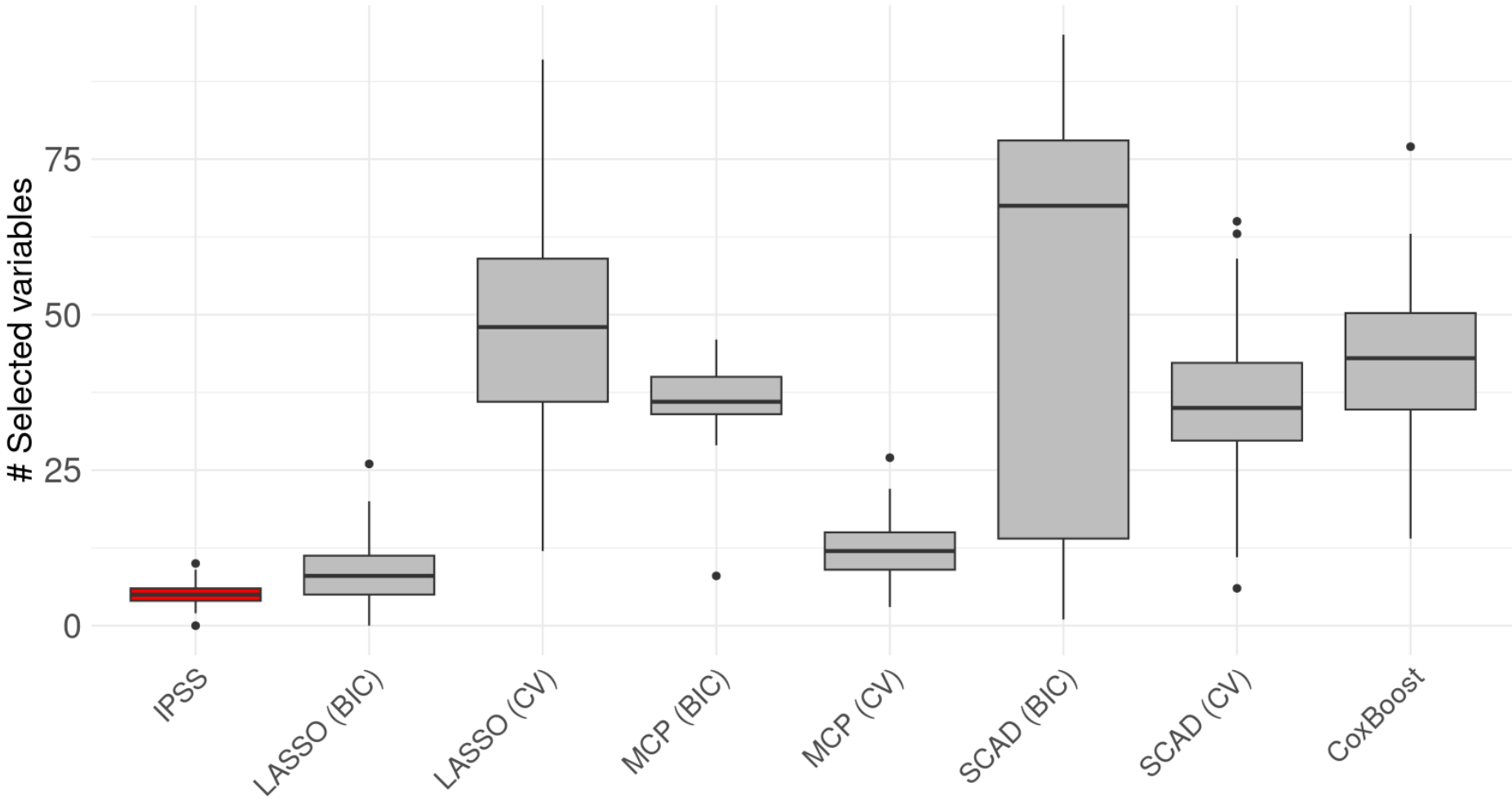
- The censoring rate was fixed at 0.3.

Evaluation criterion

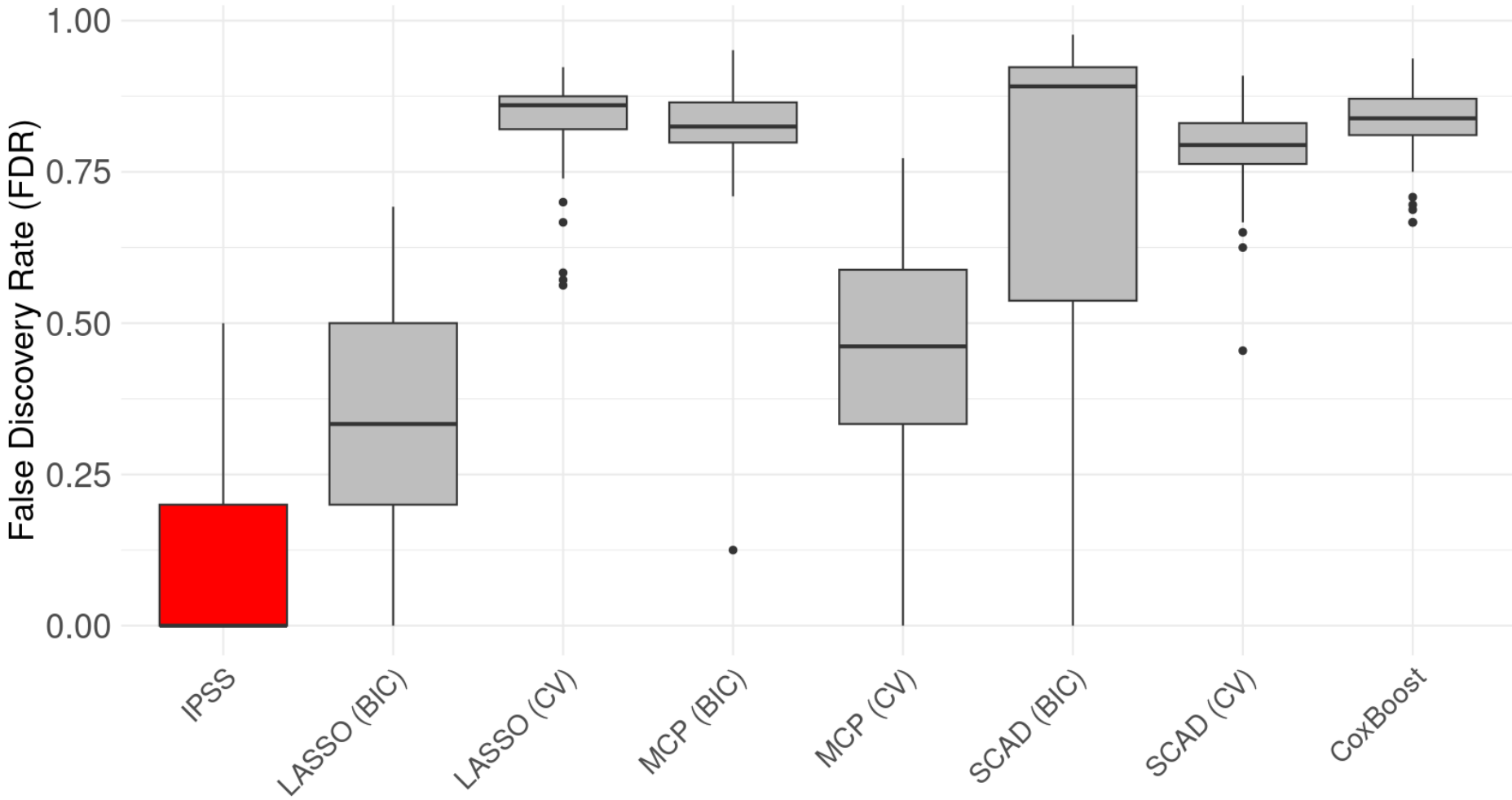
1. **Model size**: The number of variables included in the selected multivariate model.
2. **False Discovery Rate (FDR)**: The proportion of selected variables that are not associated with the outcome ($\beta = 0$).
3. **C-Index**: A value closer to 1 indicates better predictive performance.

τ^* was chosen in IPSS to control the FDR at 0.1.

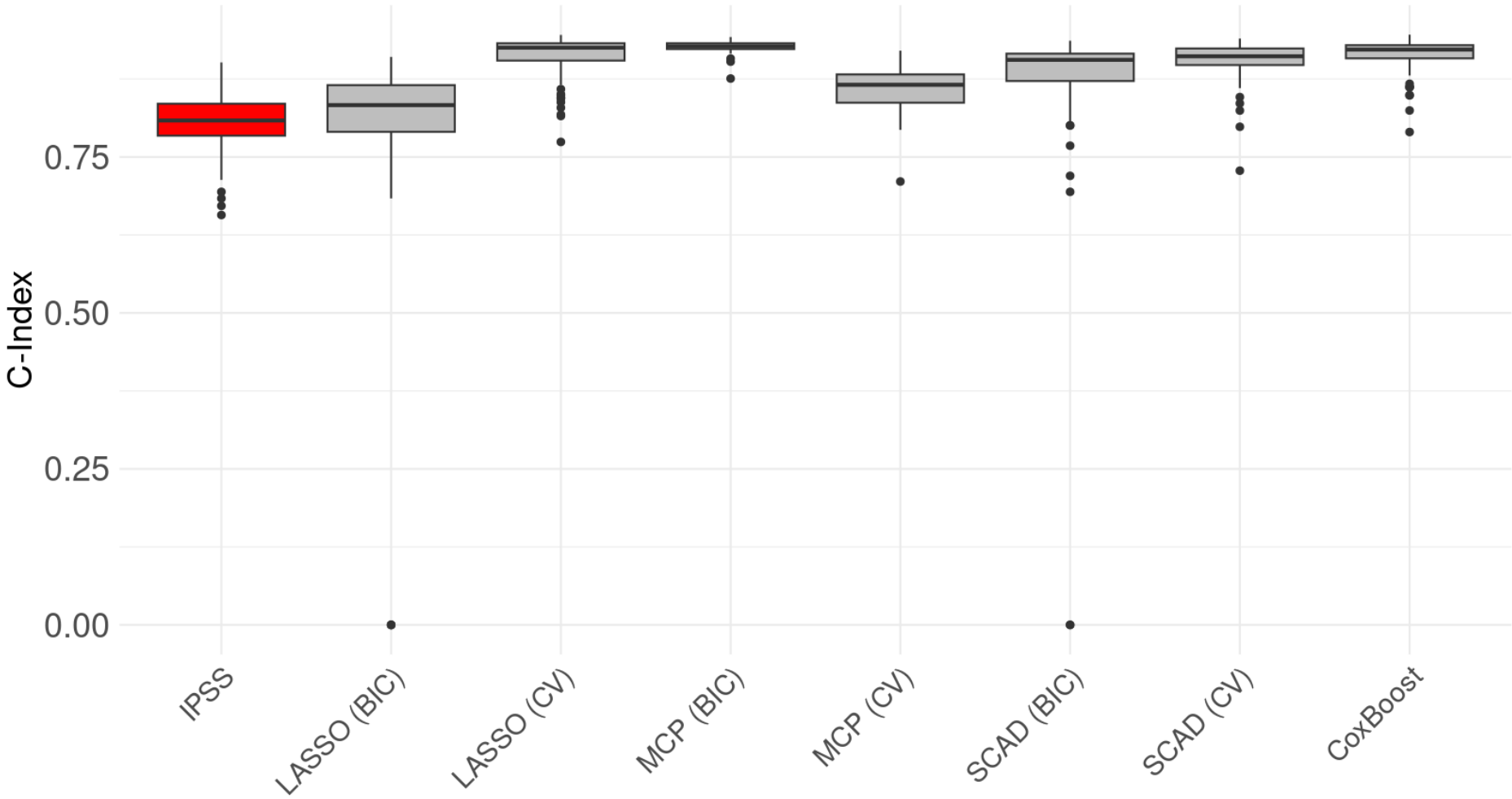
Results - Model size



Results - False Discovery Rate (FDR)



Results - C-index



Conclusion

- The proposed method enables model selection for **high-dimensional data with a competing risks outcome**.
- On simulated data, IPSS **outperforms existing methods** in terms of **model sparsity** and **false positive control**, while remaining competitive in predictive performance.
- R package in development
- The method still needs to be evaluated on real data to assess its effectiveness in selecting variables for patient stratification.

References

[1] Jianqing Fan and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties". In: *Journal of the American statistical Association* 96.456 (2001), pp. 1348–1360.

[2] Jason P Fine and Robert J Gray. "A proportional hazards model for the subdistribution of a competing risk". In: *Journal of the American statistical association* 94.446 (1999), pp. 496–509.

[3] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72.4 (2010), pp. 417–473.

[4] Omar Melikechi and Jeffrey W Miller. "Integrated path stability selection". In: *arXiv preprint arXiv:2403.15877* (2024).

[5] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.

[6] Pierre JM Verweij and Hans C Van Houwelingen. "Cross-validation in survival analysis". In: *Statistics in medicine* 12.24 (1993), pp. 2305–2314.

[7] Cun-Hui Zhang. "Nearly unbiased variable selection under minimax concave penalty". In: (2010).

Appendix

Cross-Validated partial log-likelihood (Verweij and Van Houwelingen 1993) :

$$\text{CVL}(\lambda) = \sum_{k=1}^{10} \left(l\left(\hat{\beta}_m^{(-f_k)}\right) - l^{(-f_k)}\left(\hat{\beta}_m^{(-f_k)}\right)\right),$$

where:

- $l(\cdot)$ denotes the log-likelihood of the Fine-Gray model,
- $l^{(-f_k)}(\cdot)$ indicates the log-likelihood calculated without including the observations in the k-th fold,
- $\hat{\beta}_m^{(-f_k)}$ is the vector of estimated regression coefficients obtained by fitting the model on the data set that excludes the observations in the k-th fold f_k.