Improving variable selection properties by leveraging external data

Paul Rognon-Vael [1,2] David Rossell [1] Piotr Zwiernik [1,3]

[1]UPF - U. Pompeu Fabra [2]UPC - U. Politècnica de Catalunya [3]UoT - U. Toronto

Abstract

- **Motivation**: In transfer learning or data integration settings, external information is available on the likeliness of variables to belong to the true underlying model. How much can we gain by leveraging that information?
- **Results**: We show how external information dependent ℓ_0 penalties attain model selection consistency under milder conditions than standard ℓ_0 penalties, and they also attain faster model recovery rates.

Variable selection in high dimension

Linear model:

$$y = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$$

with $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_n, \mathbb{I})$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$

We want to recover:

$$S := \{i \text{ s.t. } \beta_i^* \neq 0\}$$

with size $|S| = s$ unknown

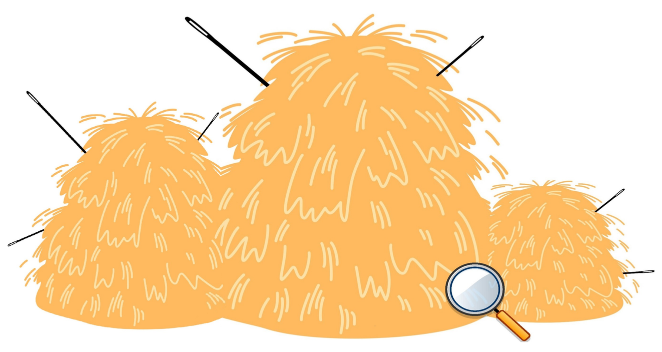
High-dimension ($n < p$): sparsity imposed with for example ℓ_0 penalization (a.k.a Bayesian variable selection, BIC...). We select a subset of variables /model:

$$\hat{S} := \underset{\text{models } M}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M\|_2^2 + \kappa|M|$$

External information

In many cases, there is external information on varying sparsity across β^*:

$$\boldsymbol{\beta}^* = (\underbrace{\beta_1^*, \ldots, \beta_{|B_1|}^*}_{\text{Block 1 less sparse}}, \underbrace{\beta_{|B1|+1}^*, \ldots, \beta_{|B_1|+|B_2|}^*}_{\text{Block 2 more sparse}}, \ldots)$$



Some cases: data integration (e.g. functional annotations on genes, past experiments); expert knowledge; meta data on variables.

Example in multi-omics:

Type of variable	Clinical	Copy-number variation	miRNA	Mutation	mRNA
Average number	9	57,927	784	15,682	22,980

Table 1. Average number of variables by block type in 15 multi-omics datasets from The Cancer Genome Atlas.

Block informed selection

Idea: Assume a given partition in blocks $B_1, \ldots, B_b$. B_j has size p_j and s_j active $\beta_i^* \neq 0$. Let penalty κ vary by block.

$$\hat{S}^b := \underset{\text{models } M}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M\|_2^2 + \sum_{j=1}^{b} \kappa_j|M_j| \quad (1)$$

where $M_j = M \cap B_j$

Many examples of improved inference in **applications**. But **theory**?

- How much can we earn?
- Can we lose?
- How to set penalties?

Milder conditions for variable selection consistency

Variable selection consistency with $\hat{S}$

Assume:

(A1) $\kappa \gtrsim \ln(p-s)$

(A2) $\sqrt{n\rho(\boldsymbol{X})}\beta_{\min}^* \gtrsim \sqrt{\kappa} + \sqrt{\ln(s)}$

then $P(\hat{S} = S) \to 1$ as $n, p \to +\infty$

Variable selection consistency with $\hat{S}^b$

Assume:

(A3) $\kappa_j \gtrsim \ln(p_j - s_j) \quad \forall j$

(A4) $\sqrt{n\rho(\boldsymbol{X})}\beta_{\min,j}^* \gtrsim \sqrt{\kappa_j} + \sqrt{\ln(s_j)} \quad \forall j$

then $P(\hat{S}^b = S) \to 1$ as $n, p \to +\infty$

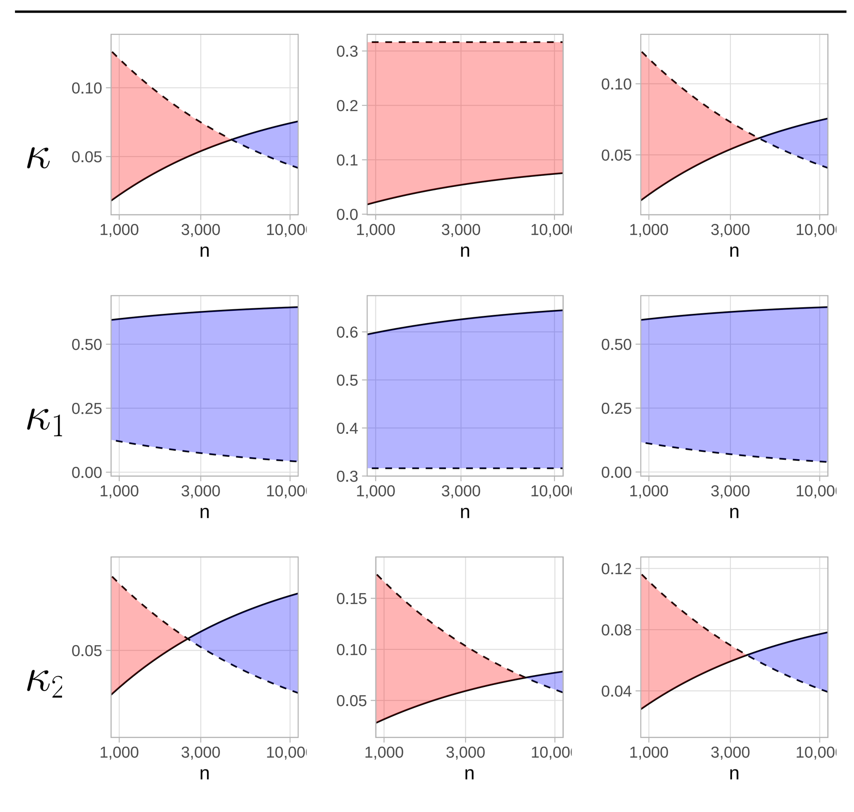
$\hat{S}^b$ can be variable selection consistent when $\hat{S}$ is not.

($\rho(\boldsymbol{X})$ depends on the correlation between variables in S and outside.)

Conditions for consistency in a two-block example

Example	1	2	3
$p-s$	$3n/2$	$e^{n/20}$	n
$p_1 - s_1$	$3n/2 - \sqrt{n}$	$e^{n/20} - n^2$	$n/2$
$p_2 - s_2$	$\sqrt{n}$	n^2	$n/2$
$s_1 = s_2$	$3\ln(n)/2$	$3\ln(n)/2$	$3\ln(n)/2$

$\beta_{\min,1}^* = 2/3$, $\beta_{\min,2}^* = \beta_{\min}^* = 1/10$, and $\rho(\mathbf{X}) = 1$



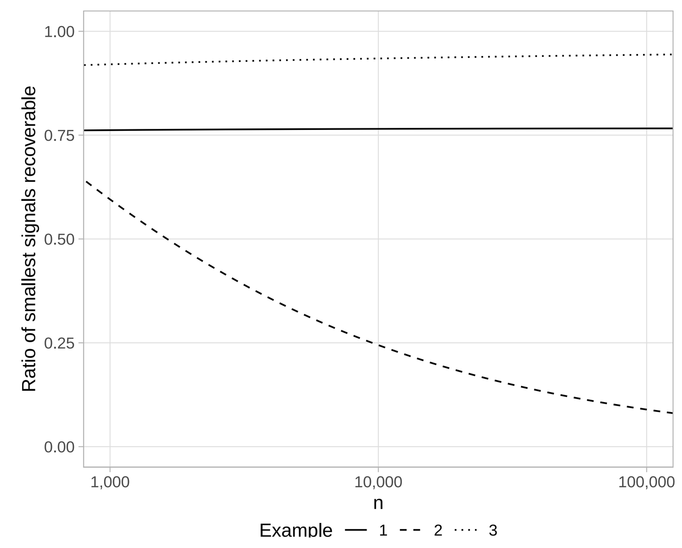
Smallest recoverable signals

Standard - $\hat{S}$:

$$\beta_{\min}^* = O\left(\sqrt{\frac{2\ln(p-s)}{n\rho(\boldsymbol{X})}} + \sqrt{\frac{2\ln(s)}{n\rho(\boldsymbol{X})}}\right)$$

Block informed - $\hat{S}^b$:

$$\beta_{\min,j}^{*,b} := O\left(\sqrt{\frac{2\ln(p_2-s_2)}{n\rho(\boldsymbol{X})}} + \sqrt{\frac{2\ln(s_2)}{n\rho(\boldsymbol{X})}}\right)$$



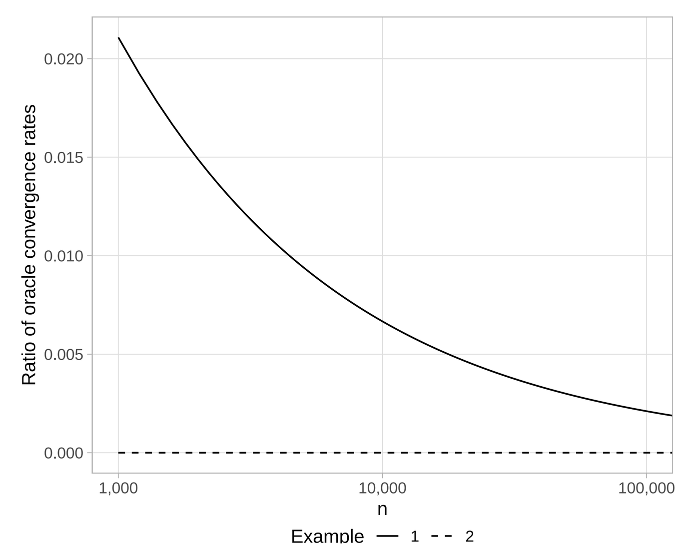
Convergence rate at the oracle (optimizing rate)

Standard oracle penalty - $\hat{S}$:

$$\kappa^* \approx \frac{\ln(p/s - 1)}{\sqrt{n\rho(\boldsymbol{X})}\beta_{\min}^*} + \sqrt{n\rho(\boldsymbol{X})}\beta_{\min}^*$$

Block informed oracle penalty - $\hat{S}^b$:

$$\forall j \ \kappa_j^* \approx \frac{\ln(p_j/s_j - 1)}{\sqrt{n\rho(\boldsymbol{X})}\beta_{\min,j}^*} + \sqrt{n\rho(\boldsymbol{X})}\beta_{\min,j}^*$$



How to set block penalties in practice?

An estimator of sparsity is:

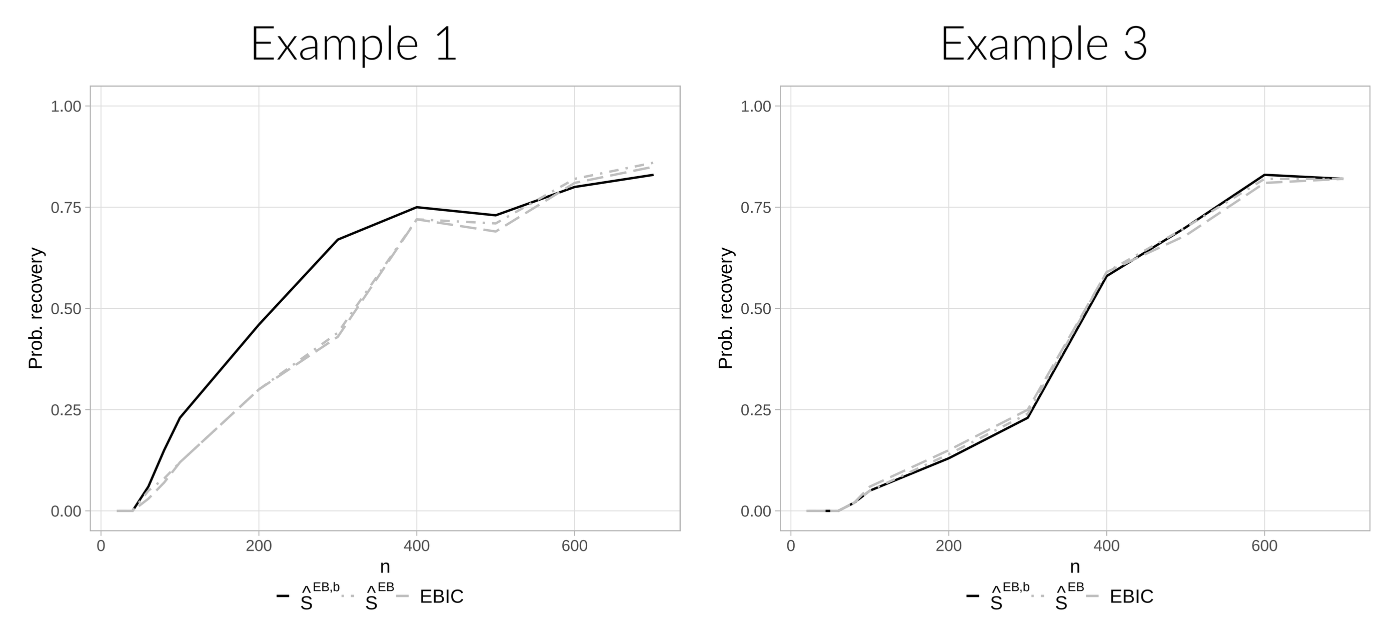
$$\hat{s}_j := \sum_{i \in B_j} \sum_M \frac{e^{-\|\boldsymbol{y}-\boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M\|_2^2 - \sum_{j=1}^b \kappa_j|M_j|}}{\sum_L e^{-\|\boldsymbol{y}-\boldsymbol{X}_L\hat{\boldsymbol{\beta}}_L\|_2^2 - \sum_{j=1}^b \kappa_j|L_j|}} \mathcal{I}(i \in M) \quad \left(\approx \sum_{i \in B_j} P(\beta_i \neq 0|\boldsymbol{y})\right)$$

If A3 and A4 hold, $\hat{s}_j/p_j$ consistent. $\hat{s}_j/p_j$ approximates an empirical Bayes estimate for the prior inclusion probability in B_j

Algorithm:

1. Set $\kappa_j = \kappa^\circ = \ln(p) + \frac{1}{2}\ln(n)$ for $j = 1, \ldots, b$. Compute $\hat{s}_j/p_j$ for $j = 1, \ldots, b$.
2. Obtain $\hat{S}^{EB,b}$ solving (1) with $\kappa_j^{EB} = \ln(p_j/\hat{s}_j - 1) + \frac{1}{2}\ln(n)$.

Simulations



Check out the arXiv!

paul.rognon@gmail.com