High-dimensional variable selection and external information

Paul Rognon-Vael
David Rossell
Piotr Zwiernik

Variable selection in high dimension

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

with $\epsilon \sim \mathrm{N}\left(\mathbf{0}_n, \sigma^2 \mathbb{I}\right), \boldsymbol{X} \in \mathbb{R}^{n\times p}$
wlog set $\sigma = 1$

Find: $\quad S := \{i \text{ s.t. } \beta_i^\star \neq 0\}$
with size $|S| = s$ unknown

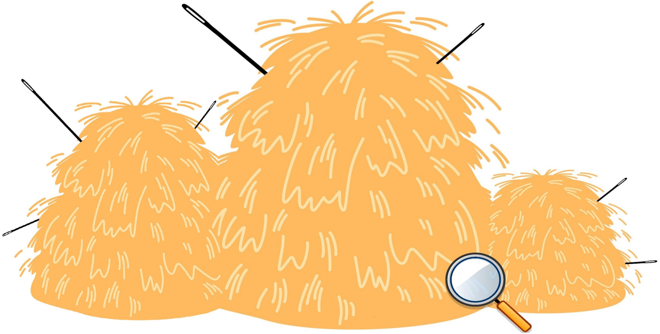
High-dimension ($n < p$): sparsity imposed with for example **L_0 penalization** (Bayesian variable selection, BIC...). We select a subset of variables /model:

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

External information

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

$$\beta^* = (\underbrace{\beta^*_1, \ \ldots \ , \beta^*_{|B_1|}}_{\text{Block 1 less sparse}}, \underbrace{\beta^*_{|B1|+1}, \ \ldots \ , \beta^*_{|B_1|+|B_2|}}_{\text{Block 2 more sparse}}, \ \ldots)$$



External information

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: Average number of variables by block type in 15 multi-omics datasets from The Cancer Genome Atlas [4] analyzed in [1]

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**.

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

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.

Smallest recoverable signals

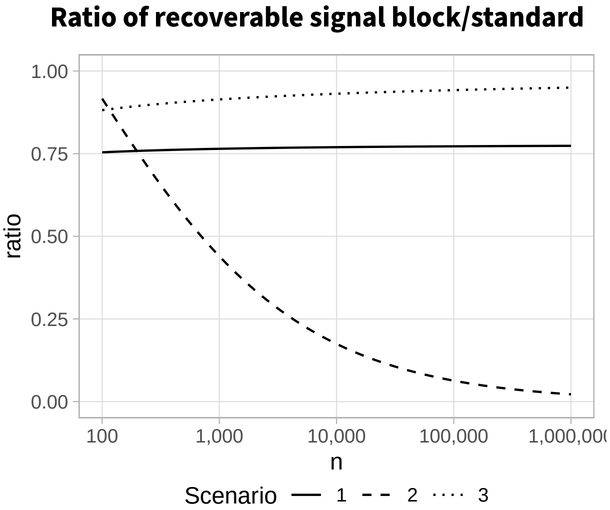
Standard - $\hat{S}$:

$$\beta^*_{\min} = O\left(\sqrt{\frac{2\ln(p-s)}{n}} + \sqrt{\frac{2\ln(s)}{n}}\right)$$

Block informed - $\hat{S}^b$:
in each block B_j,

$$\beta^*_{\min,j} := O\left(\sqrt{\frac{2\ln(p_j-s_j)}{n}} + \sqrt{\frac{2\ln(s_j)}{n}}\right)$$

Scenario	**1**	**2**	**3**
$p-s$	$\frac{3}{2}n$	$e^{n/10}$	n
$p_j - s_j$	$\sqrt{n}$	n^2	$n/2$



Convergence rate at the oracle (optimizing rate)

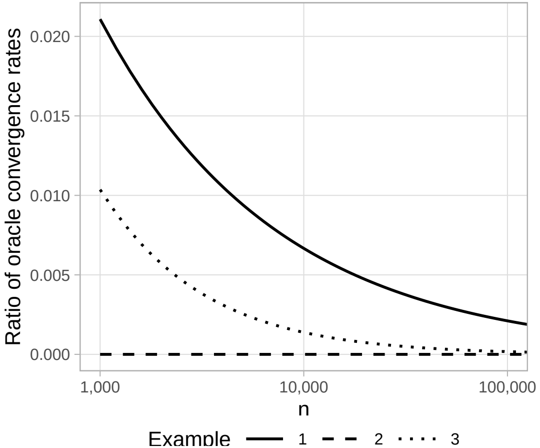
Standard oracle penalty - $\hat{S}$:

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

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

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

Ratio of prob. error block/standard



How to set block penalties? An estimator of sparsity

For any model M,

$$NC(M) = \frac{e^{-\|\boldsymbol{y}-\boldsymbol{X}_M\hat{\beta}_M\|_2^2 - \sum_{j=1}^{b} \kappa_j |M_j|}}{\sum_L e^{-\|\boldsymbol{y}-\boldsymbol{X}_L\hat{\beta}_L\|_2^2 - \sum_{j=1}^{b} \kappa_j |L_j|}}$$

is a **posterior probability** for model M from a Bayesian perspective (BIC Schwarz [3]).

An **estimator of sparsity** is:

$$\widehat{s}_j := \sum_{i \in B_j} \sum_{M} NC(M)\, \mathcal{I}(i \in M) \qquad \left(\approx \sum_{i \in B_j} P(\beta_i \neq 0 | \boldsymbol{y}) \right)$$

Two nice properties:

- If A3 and A4 hold, $\widehat{s}_j/p_j$ **consistent**. If A3 only holds, $\widehat{s}_j/p_j \leq s_j/p_j$ as $n \to +\infty$.
- $\widehat{s}_j/p_j$ matches the **empirical Bayes** estimate of the prior inclusion probability in B_j

Two-stage proposal - Empirical Bayes

Algorithm:

1. Set $\kappa_j = \kappa^\circ = \ln(p) + \frac{1}{2}\ln(n)$ for $j = 1, \ldots, b$. Compute $\widehat{s}_j/p_j$ for $j = 1, \ldots, b$.

2. Select the model:

$$\hat{S}^{EB,b} := \arg\min_{M} \|\boldsymbol{y} - \boldsymbol{X}_M \hat{\beta}_M\|_2^2 + \sum_{j=1}^{b} \kappa_j |M_j| \quad \text{where} \quad \forall j \ \kappa_j = \ln(p_j/\widehat{s}_j - 1) + \frac{1}{2}\ln(n)$$

Setting $\kappa_j = \ln(p_j/\widehat{s}_j - 1) + \frac{1}{2}\ln(n)$ is essentially setting prior inclusion probabilities for variables in B_j by empirical Bayes.

Two-stage proposal - Empirical Bayes

Properties of the algorithm:

- $P(\hat{S}^{EB,b} \subseteq S) \to 1$ as $n, p \to \infty$ (no betamin assumption)

- If the following betamin condition holds:

$$\text{(A5)} \quad \sqrt{n\rho(\boldsymbol{X})}\beta^*_{\min,j} \gtrsim \sqrt{\ln(p_j/s_j^L - 1) + \frac{1}{2}\ln(n)} + \sqrt{\ln(s_j)} \quad \forall j$$

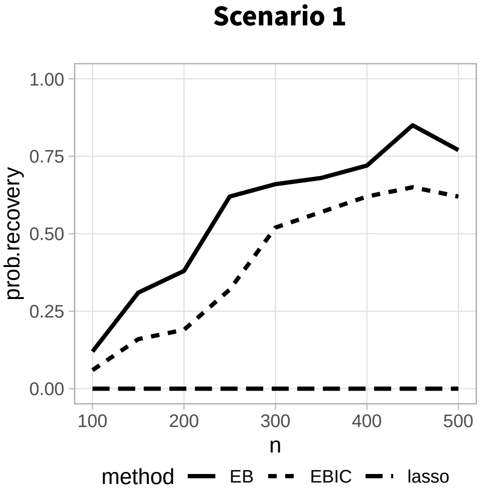
  where $s_j^L := \#\{\beta_i^* \in B_j | \sqrt{n\rho(\boldsymbol{X})}\beta_i^* \gtrsim \sqrt{\kappa^\circ} + \sqrt{\ln(s_j)}\}$

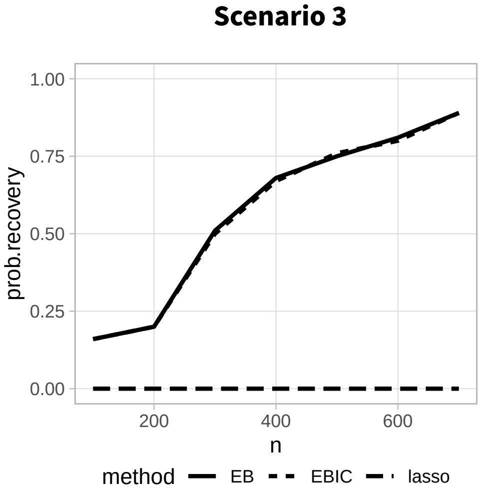
  then:

$$P(\hat{S}^{EB,b} = S) \to 1 \quad \text{as } n, p \to \infty$$

A5 is essentialy A4 for penalty $\kappa_j = \ln(p_j/s_j - 1) + \frac{1}{2}\ln(n)$, but slightly stricter. The difference depends on how many signals are missed in step 1.

Simulations





Takeaways

We provide **theoretical guarantees** and a **practical method** for the incorporation of external information / data in selection procedure. In doing so, you can:

1. tailor the procedure to varying sparsity constraints informed by external data
2. recover smaller signals
3. get better selection results without increasing the sample size (faster convergence)
4. the more discriminative is the external information / data the better.

References I

- **Yingxia Li, Ulrich Mansmann, Shangming Du, and Roman Hornung.** Benchmark study of feature selection strategies for multi-omics data. *BMC Bioinformatics*, 23(1):412, October 2022.
- **Omiros Papaspiliopoulos and David Rossell.** Bayesian block-diagonal variable selection and model averaging. *Biometrika*, 104(2):343–359, 04 2017.
- **Gideon Schwarz.** Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978.
- **Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz.** Review of the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, pages 68–77, 2015.

Convergence rates at the oracle

Assume the **penalties are set at their oracle values** κ^{OR} and κ_j^{OR}.

With **standard L_0 ($\hat{S}$)**, the **oracle convergence rate** is:

$$OR := 24\, c\, e^{-\frac{1}{2}\left[\frac{n\rho(\Sigma)}{32}{\beta^{\star}_{min}}^2 - \max\{\ln(p-s), \ln(s)\}\right]}$$

With **block informed L_0 ($\hat{S}^b$)**, the **oracle convergence rate** is:

$$OR^b := 12(2^{2b} - 2b)\, c' \sum_{j=1}^{b} e^{-\frac{1}{2}\left[\frac{n\rho(\Sigma)}{32}{\beta^{\star}_{min,j}}^2 - \max\{\ln(p_j - s_j), \ln(s_j)\}\right]}$$

Ratio of bounds on convergence rates at the oracle penalties:

$$\frac{OR^b}{OR} \sim (2^{2b-1} - b) \sum_{j=1}^{b} e^{-\frac{1}{2}\left[\frac{n\rho(\Sigma)}{24}\left({\beta^{\star}_{\min,j}}^2 - {\beta^{\star}_{\min}}^2\right)\right]} e^{-\frac{1}{2}\left[\max\{\ln(p-s), \ln(s)\} - \max\{\ln(p_j - s_j), \ln(s_j)\}\right]}$$

Necessary conditions in correlated settings

Theorem

a) If for some $j = 1, \ldots, b$, $\lim_{n\to\infty} \frac{\kappa_j}{\underline{\lambda}_j^2 \ln(p_j - s_j)} < 1$, where $\underline{\lambda}_j$ depends on the correlation of $\boldsymbol{X}_{B_j \setminus S_j}$, then $P(\hat{S}^b = S) \not\to 1$ as $n, p \to \infty$.

b) If for some $j \in \{1, \ldots, b\}$, $\sqrt{n\bar{\lambda}}\beta^\star_{\min,j} = o\left(\sqrt{\kappa_j}\right)$, where $\bar{\lambda} := \lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)$, then $P(\hat{S}^b = S) \not\to 1$ as $n, p \to \infty$.

c) If for some $j \in \{1, \ldots, b\}$, $\sqrt{n\bar{\lambda}}\beta^\star_{\min,j} = o\left(\underline{\lambda}_j \sqrt{\ln(p_j - s_j)}\right)$, then $P(\hat{S}^b = S) \not\to 1$ as $n, p \to \infty$

Properties of $\hat{s}/p$

For a fixed set of penalties κ_j, denote:

$$S_j^S := \left\{\beta_i^\star \in s_j \,\middle|\, \sqrt{n\bar{\lambda}}|\beta_i^\star| = o\left(\sqrt{\kappa_j}\right)\right\}$$

$$S_j^L := \left\{\beta_i^\star \in s_j \,\middle|\, \sqrt{\frac{n\rho(\boldsymbol{X})}{8}}|\beta_i^\star| - \sqrt{\kappa_j} = \sqrt{\ln(s_j)} + c_j\right\}$$

Assume $\kappa_j \gtrsim \ln(p_j - s_j)$ and $|S_j^S| = O(p_j - s_j)$ for every $j = 1, \ldots, b$, then :

$$\frac{|S_j^L|}{p_j} \leq \lim_{n,p\to\infty} \mathbb{E}\left(\frac{\widehat{s}_j}{p_j}\right) \leq \frac{s_j - |S_j^S|}{p_j} \qquad \text{for all } j = 1, \ldots, b$$

Bayesian interpreation of L_0 penalties

Let model M be a p-dimensional vector of variable inclusion indicators $m_i = I(\beta_i \neq 0)$. Consider the joint prior on parameters and models is:

$$p(\beta, M \mid \theta) = p(\beta \mid M) p(M \mid \theta)$$

Posterior model probabilities are:

$$p(M \mid \boldsymbol{y}, \theta) \propto p(\boldsymbol{y} \mid M) p(M \mid \theta) \quad \text{where} \quad p(\boldsymbol{y} \mid M) \approx p(\boldsymbol{y} \mid \tilde{\beta}^{(M)}) n^{-|M|/2} \text{ ([3])}$$

and:

$$\ln p(M \mid \boldsymbol{y}) \approx -\|\boldsymbol{y} - \boldsymbol{X}_M \hat{\beta}_M\|_2^2 - \tfrac{1}{2} \ln(n)|M| + \ln p(M \mid \theta) + \text{cst}$$

Assume independent inclusion variable, and inclusion probabilities constant by block:

$$p(M \mid \theta) = \textstyle\prod_{i=1}^{p} \text{Bern}\left(m_i; \theta_i\right) I(M \in \mathcal{M}) \text{ and } \forall\, i \in B_j,\ \theta_i = \theta^{(j)}$$

Then $\ln p(M \mid \theta)$ defines the block penalties

$$\kappa_j = \tfrac{1}{2} \ln(n) + \ln\left({\theta^{(j)}}^{-1} - 1\right)$$

Thresholding in orthogonal setting ($\mathbf{X}^\top \mathbf{X} = n\, I$)

Selection with **most Bayesian procedures [2], LASSO and L_0 penalty** operate by **thresholding the MLE**.

A **generic threshold estimator**:

$$\hat{S} := \left\{ i : |\hat{\beta}_i| > \tau \right\},$$

with standard L_0: $\tau = \sqrt{\frac{2\kappa}{n}}$

A **generic block informed threshold estimator**:

$$\hat{S}_j^b := \left\{ i \in B_j : |\hat{\beta}_i| > \tau_j \right\} \quad \text{and} \quad \hat{S}^b = \bigcup_j \hat{S}_j^b$$

with block informed L_0: $\tau_j = \sqrt{\frac{2\kappa_j}{n}}$

Selection consistency when $\mathbf{X}^\top \mathbf{X} = n\, I$

Theorem

Suppose that τ and $\beta^\star_{min}$ satisfy:

$$\tau \geq \sqrt{2\ln(p-s)/n} \quad \text{and} \quad \beta^\star_{min} - \tau \geq \sqrt{2\ln(s)/n}$$

then
$$P(\hat{S} = S) \to 1$$

Theorem

Suppose that the τ_j's and $\beta^\star_{min,j}$'s satisfy:

$$\tau_j \geq \sqrt{2\ln(p_j - s_j)/n} \quad \text{and} \quad \beta^\star_{min,j} - \tau_j \geq \sqrt{2\ln(s_j)/n}$$

then
$$P(\hat{S}^b = S) \to 1$$

Necessary conditions when $\mathbf{X}^\top\mathbf{X} = n\,I$

Theorem

a) If for some $j \in \{1,\ldots,b\}$, $\lim_{n\to\infty} \frac{\tau_j}{\sqrt{\frac{2\ln(p_j - s_j)}{n}}} < 1$, then $P(\widehat{S}^b \subseteq S) \nrightarrow 1$.

b) Assume for some $j \in \{1,\ldots,b\}, \forall\, i \in S_j\ \beta_i^\star = \beta_{\min,j}^\star$ and $s_j/p_j \leq c < 1$.

If $\lim_{n\to\infty} \frac{\beta_{\min,j}^\star - \tau_j}{\sqrt{\frac{\pi}{2}\frac{\ln(s_j)}{n}}} \leq 1$ then $P(S \subseteq \widehat{S}^b) \nrightarrow 1$.

c) Assume for some $j \in \{1,\ldots,b\}, \forall\, i \in S_j\ \beta_i^\star = \beta_{\min,j}^\star$ and $s_j/p_j < 1$.

If $\lim_{n\to\infty} \frac{\beta_{min,j}^*}{\sqrt{\frac{2\ln(p_j - s_j)}{n}} + \sqrt{\frac{\pi}{2}\frac{\ln(s_j)}{n}}} < 1$ then $P(\widehat{S}^b = S) \nrightarrow 1$