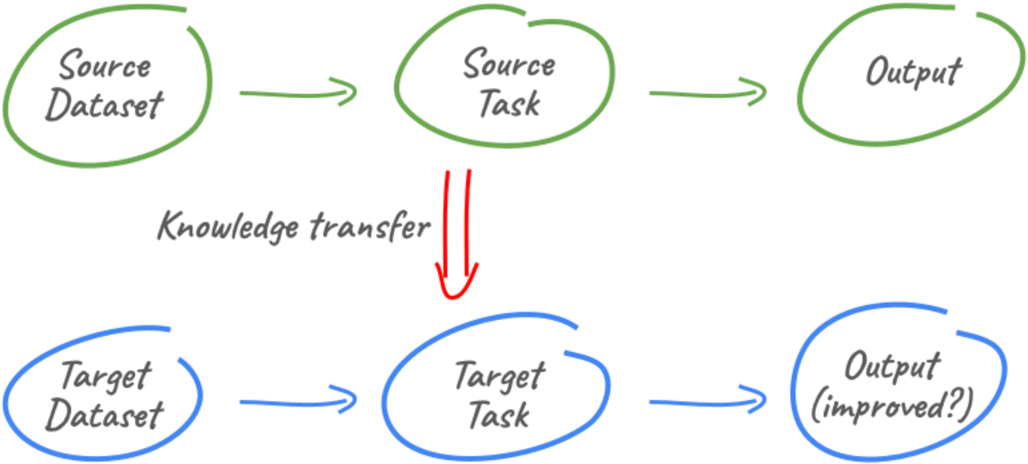
Transfer learning for variable selection: fundamental limits and a practical solution

Paul Rognon-Vael
joint work with David Rossell and Piotr Zwiernik

Transfer learning

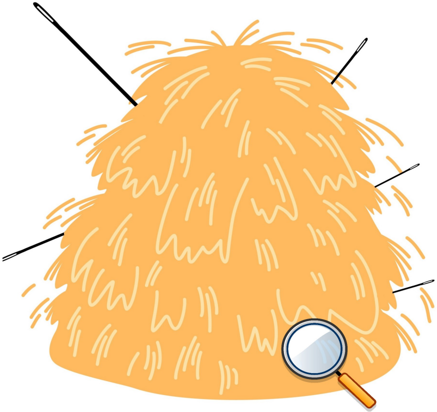


Variable selection in high dimensional linear models

Consider:
$y = \boldsymbol{X}\beta^\star + \epsilon \quad$ with $\epsilon \sim \mathrm{N}\left(\boldsymbol{0}_n, \sigma^2 \mathbb{I}\right)$

$\boldsymbol{X} \in \mathbb{R}^{n \times p}$ with p possibly larger than n,
wlog, set $\sigma = 1$

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



Variable selection with ℓ_0 penalty

For some positive function κ (e.g. $\kappa = \ln(n)$ in BIC)

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

Direct **link to Bayesian spike-and-slab regression**, set prior on subsets and coefficients: $\pi(M, \beta) = \pi(M)\pi(\beta \mid M)$; $\pi(M) = \Pi_{i=1}^{p} \text{Bern}(m_i; \underbrace{(e^{\kappa - \ln(n)/2} + 1)^{-1}}_{\textbf{prior inclusion prob.}})$ and $m_i = I(\beta_i \neq 0)$

$\hat{S}$ matches the mode of the posterior $p(M|\mathbf{y})$ (under regularity conditions).

ℓ_0 penalties:

- have **superior selection properties**,
- are **much more computationally tractable** with recent progress in discrete optimization and MCMC methods[1, 6, 7].

Transfer learning for variable selection

Let $\hat{S}(\mathcal{D}_S)$ be the subset selected in the *source* dataset $\mathcal{D}_S$. Form two blocks:

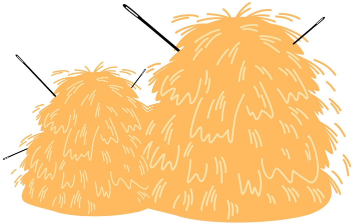
- $B_1 \ := \ \{i \in \hat{S}(\mathcal{D}_S)\}$
- $B_2 \ := \ \{i \notin \hat{S}(\mathcal{D}_S)\}$

Examples:

- In genomics, public databases register found gene-disease associations.
- In causal inference, sets of confounders may have been learnt in related problem.

If selection in the *source* is informative for selection in the *target*:

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



Transfer informed variable selection

Idea: Selection in the *source* dataset gives us prior knowledge on the likelihood of a variable to be truly associated to the outcome in the *target* dataset.

Since ℓ_0 penalties $\leftrightarrow$ prior inclusion probabilities, it's natural to let **the penalty vary by block**. We consider **transfer informed penalties**:

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

where $M_j = M \cap B_j$.

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

- How much can we earn in theory and in practice?
- Can we lose?
- How to set penalties?

Milder conditions for variable selection consistency

Variable selection consistency with $\hat{S}$

If and only if:

$$\text{(A1)} \quad \sqrt{\kappa} \gtrsim \sqrt{\ln(p-s)}$$

$$\text{(A2)} \quad \sqrt{\kappa} \lesssim \sqrt{n\rho(\boldsymbol{X})}\beta^*_{\min} - \sqrt{\ln(s)}$$

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

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

If and only if:

$$\text{(A3)} \quad \sqrt{\kappa_j} \gtrsim \sqrt{\ln(p_j - s_j)} \quad \forall j$$

$$\text{(A4)} \quad \sqrt{\kappa_j} \lesssim \sqrt{n\rho(\boldsymbol{X})}\beta^*_{\min,j} - \sqrt{\ln(s_j)} \quad \forall j$$

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

$\hat{S}^l$ is variable selection consistent in wider class of regimes $(n, \mathbf{p}, \mathbf{s}, \beta^*)$ than $\hat{S}$.

Smallest recoverable signals

For $\kappa = \ln(p-s)$ and $\kappa_j = \ln(p_j - s_j)\ \forall j$, the smallest signal recoverable is:

Standard - $\hat{S}$

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

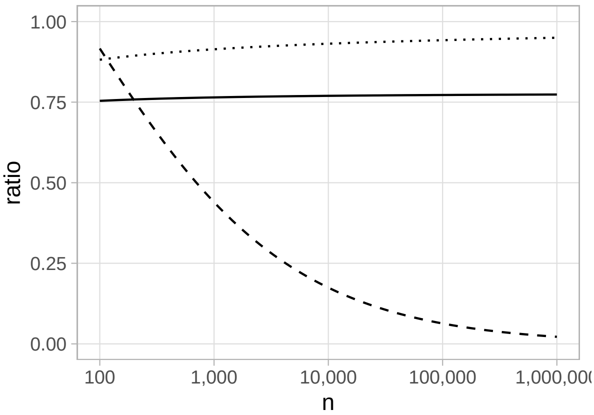
Transfer informed - $\hat{S}^I$

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
$p_1 - s_1$	$\sqrt{n}$	n^2	$n/2$

Ratio of recoverable signal informed/standard



Convergence rate for oracle penalties (min. bound)

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

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

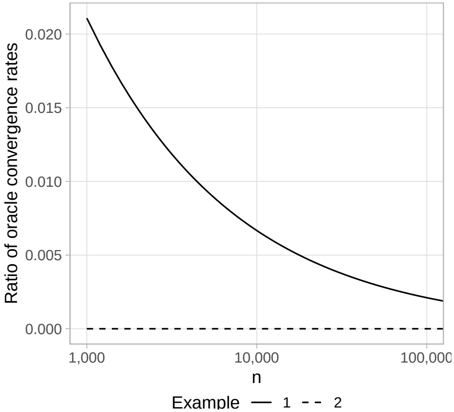
Informed oracle penalty - $\hat{S}^I$:

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

Ratio of oracle convergence rates :

$$\frac{Or.Conv.Rate^I}{Or.Conv.Rate} \sim 2^{2b}\sum_{j=1}^{b}\frac{p_j-s_j}{p-s}\, e^{-\left[n\rho(\Sigma)\left({\beta^{\star}_{\min,j}}^2-{\beta^{\star}_{\min}}^2\right)\right]}$$

Ratio of prob. error informed /standard



In the orthogonal case, we can show improvements in minimax rates.

In practice? Informed empirical Bayes

Idea 1: In less sparse blocks, we can safely lower the penalties. We estimate the sparsity in each block, thus adapting to how informative the transfer is.

Idea 2: Using the connection between ℓ_0 penalties and prior inclusion probabilities and **empirical Bayes**, an estimator of sparsity in each block s_j for any set of κ_j:

$$\hat{s}_j := \sum_{i \in B_j} \sum_{\text{subsets } M : i \in M} pmp(M) \quad \text{where} \quad pmp(M) = \frac{e^{-\|\boldsymbol{y} - \boldsymbol{X}_M \hat{\beta}_M\|_2^2 - \sum_{j=1}^{b} \kappa_j |M_j|}}{\sum_L e^{-\|\boldsymbol{y} - \boldsymbol{X}_L \hat{\beta}_L\|_2^2 - \sum_{j=1}^{b} \kappa_j |L_j|}}$$

Two-stage algorithm:

1. Compute $\hat{s}_j / p_j$ with $\kappa_j = \kappa^\circ = \ln(p) + \frac{1}{2}\ln(n)$ for $j = 1, \dots, b$.
2. Select the subset:

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

Properties of the two-stage algorithm

Standard empirical Bayes - $\hat{S}^{EB}$

- $P(\hat{S}^{EB} \subseteq S) \to 1$
- Assume condition on signals:

$$\sqrt{n\rho(\boldsymbol{X})}\beta^*_{\text{min}} \gtrsim \sqrt{\ln(p/s^L - 1) + \frac{1}{2}\ln(n)} + \sqrt{\ln(s)},$$

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

Informed empirical Bayes - $\hat{S}^{EB,I}$

- $P(\hat{S}^{EB,I} \subseteq S) \to 1$ (slightly slower rate)
- Assume **milder** condition on signals:

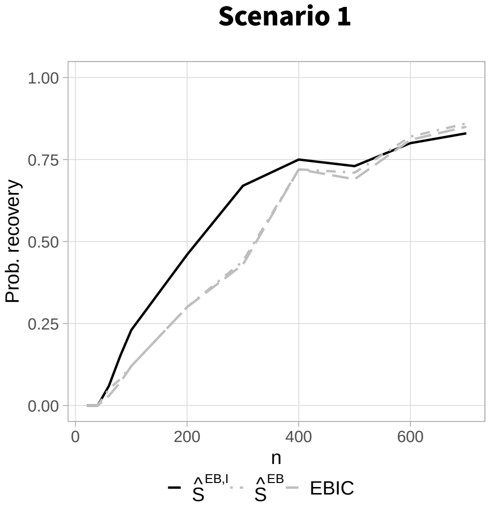
$$\sqrt{n\rho(\boldsymbol{X})}\beta^*_{\text{min},j} \gtrsim \sqrt{\ln(p_j/s_j^L - 1) + \frac{1}{2}\ln(n)} + \sqrt{\ln(s_j)} \quad \forall j,$$

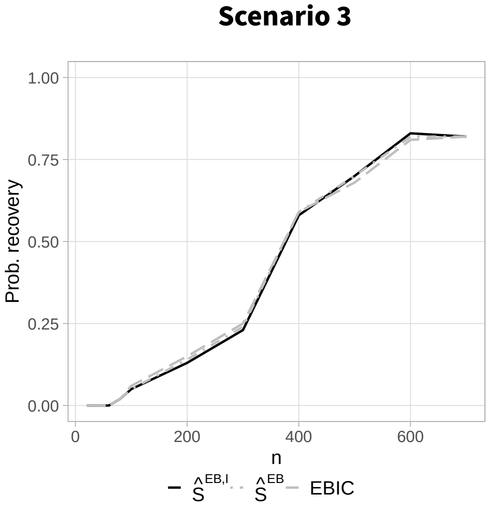
then, $\quad P(\hat{S}^{EB,I} = S) \to 1 \quad$ as $n, p \to \infty$.

Ratio of convergence rates

- If signals are weak $Conv.Rate^I / Conv.Rate \approx \sum_{j=1}^{b} \frac{p_j - s_j}{p - s} e^{-n\rho(\boldsymbol{X})({\beta^*_{\text{min},j}}^2 - {\beta^*_{\text{min}}}^2)}$
- If signals are strong $Conv.Rate^I / Conv.Rate \gtrsim 1$

Simulations





Takeaways

1. We introduce and study a class of ℓ_0 penalties for transfer learning in variable selection, grounded in Bayesian reasoning.
2. We show one can push fundamental limits on selection consistency with transfer learning.
3. We quantify how much can be earned in theory with transfer learning with oracle penalties.
4. We propose a concrete data-based approach to set penalties that realize most of the benefits of the oracle:
   - softer conditions for consistency,
   - faster convergence in hard and moderately easy settings,
   - minor loss in rate in very easy settings.

References I

- **Dimitris Bertsimas and Bart Van Parys.** Sparse high-dimensional regression. *The Annals of Statistics*, 48(1):300–323, 2020.
- **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.

References II

- **Yun Yang, Martin J. Wainwright, and Michael I. Jordan.** On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497 – 2532, 2016.

- **Quan Zhou, Jun Yang, Dootika Vats, Gareth O. Roberts, and Jeffrey S. Rosenthal.** Dimension-free mixing for high-dimensional bayesian variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1751–1784, 10 2022.

External data - meta data on variables

More generally, we may have **external data** on variables that partition the set of variables in $b > 2$ blocks:

Example - nature of variables

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 [5] analyzed in [2]

Typically, a block of genomic markers is much sparser than a block of clinical signals.

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:

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

With **block informed ℓ_0 ($\hat{S}^I$)**, the **oracle convergence rate** is:

$$OR^I := 12(2^{2b} - 2b)\, c' \sum_{j=1}^{b} e^{-\frac{1}{2}\left[\frac{n\rho(\Sigma)}{24}{\beta^{\star}_{min,\,j}}^{2} - \max\{\ln(p_j-s_j), \ln(s_j)\}\right]}$$

Ratio of bounds on convergence rates at the oracle penalties:

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

Necessary conditions in correlated settings

Theorem

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

b) If for some $j \in \{1, \dots, b\}$, $\lim_{n\to\infty} \sqrt{n\bar{\lambda}}\beta^\star_{\min,j} - \sqrt{2\kappa_j} < \infty$, where $\bar{\lambda} := \lambda_{\max}\left(\frac{1}{n}\boldsymbol{X}_S^\top \boldsymbol{X}_S\right)$, then $P(\hat{S}^I \supseteq S) \not\to 1$ as $n, p \to \infty$.

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

Bayesian interpretation of ℓ_0 penalties

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

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

Posterior subset probabilities are:

$$p(M \mid \boldsymbol{y}, \theta) \propto p(\boldsymbol{y} \mid M)p(M \mid \theta) \quad \text{where} \quad p(\boldsymbol{y} \mid M) \approx p(\boldsymbol{y} \mid \tilde{\beta}^{(M)})n^{-|M|/2} \text{ ([4])}$$

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 different prior inclusion probabilities 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

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

Empirical Bayes inspired informed penalties

The empirical Bayes estimate of the prior inclusion probability $\theta^{(j)}$ maximizes the marginal likelihood,

$$\hat{\theta}^{(j)} \;=\; \arg\max_{\theta^{(j)}} p(\boldsymbol{y} \mid \theta^{(j)}).$$

It also satisfies the fixed point equation

$$\hat{\theta}^{(j)} = \tfrac{1}{p_j} \textstyle\sum_{i \in B_j} P\left(\beta_i \neq 0 \mid \boldsymbol{y}, \hat{\theta}\right)$$

We can approximate the above equation by replacing $\hat{\theta}$ in the RHS by an initial guess θ°:

$$\hat{\theta}^{(j)} \approx \frac{1}{p_j} \sum_{i \in B_j} P\left(\beta_i \neq 0 \mid \boldsymbol{y}, \theta^{\circ}\right) = \frac{1}{p_j} \sum_{i \in B_j} \sum_{\text{subsets } M: i \in M} P\left(M \mid \boldsymbol{y}, \theta^{\circ}\right)$$

Using that $pmp(M)$ can be seen as a posterior model probability,

$$\frac{\hat{s}_j}{p_j} = \frac{1}{p_j} \sum_{i \in B_j} \sum_{\text{subsets } M: i \in M} pmp(M) \approx \frac{1}{p_j} \sum_{i \in B_j} \sum_{\text{subsets } M: i \in M} P\left(M \mid \boldsymbol{y}, \theta^{\circ}\right) \approx \hat{\theta}^{(j)}$$

Properties of $\hat{s}/p$

For a fixed set of penalties κ_j, denote:

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

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 :

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

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

Selection with **most Bayesian procedures [3], 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**:

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

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 $j \in \{1, \dots, b\}$, $\lim_{n\to\infty} \frac{\tau_j}{\sqrt{\frac{2\ln(p_j - s_j)}{n}}} < 1$, then $P(\hat{S}^b \subseteq S) \not\to 1$.

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

If $\lim_{n\to\infty} \frac{\beta^\star_{\text{min},j} - \tau_j}{\sqrt{\frac{\pi}{2}\frac{\ln(s_j)}{n}}} \leq 1$ then $P(\hat{S}^b \supseteq S) \not\to 1$.

c) Assume for some $j \in \{1, \dots, b\}, \forall\, i \in S_j\ \beta_i^\star = \beta^\star_{\text{min},j}$ 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 $P(\hat{S}^b = S) \not\to 1$