Pushing the limits of variable selection with external data

David Rossell
Universitat Pompeu Fabra & Barcelona School of Economics

Work 1 (COVID19): Jack Jewson, Li Li, Laura Battaglia, Stephen Hansen & Piotr Zwiernik
Work 2 (Theory): Paul Rognon-Vael & Piotr Zwiernik
Work 3 (Methods): Miquel Torrens and & Omiros Papaspiliopoulos

OBayes 2025, Athens

Motivation

High-dimensional variable selection (large p) requires strong assumptions if consistent recovery is desired as $n \to \infty$

- Sparsity (bound on truly active variables)
- Smallest signal (betamin conditions)
- Correlation between covariates (eigenvalue conditions)

External information often available $\Rightarrow$ variables not exchangeable a priori

- Biomedicine: gene annotations, clinical history vs genomic markers
- Transfer learning: findings in related problems/populations (e.g. cancer type)
- Causal inference: variables highly correlated with treatment are special
- ...

Abundant work showing empirical gains from data integration. Theory lacking

Graphical model application

Question: in what counties did COVID19 evolve in a coordinated manner? Is co-evolution related to social media, geographical distance and number of flights?

Data: weekly infection rates 01/2020 - 11-2023 ($n = 97$ weeks) for $p = 332$ USA meta-counties

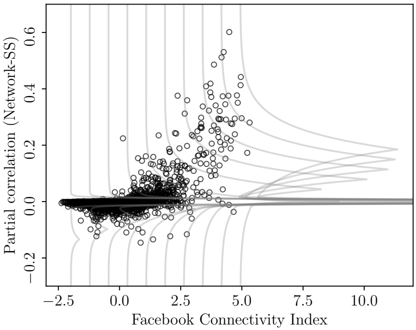
Regress log-infection rate on vaccination, containment measures, pop density, temperature, time & county fixed effects, AR1 term

- Model explains 90% of variance ($R^2 = 0.9$)
- Large residual partial correlation across some counties, i.e. COVID rates systematically higher/lower than predicted

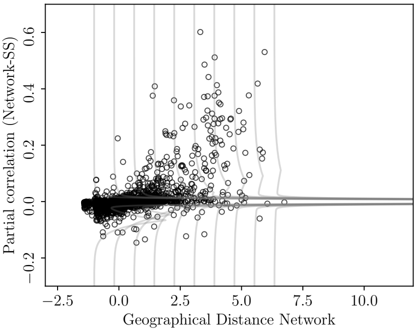
COVID. Fitted model

Regress partial correlations onto network data: Facebook, distance, flights

- Model selection: what partial correlations are 0? Prob of 0 vs. networks?
- Estimation: partial correlations? Their mean / variance vs networks?



Facebook index



1/geographical distance

Goal: study theoretical benefits of external info for **variable selection**.

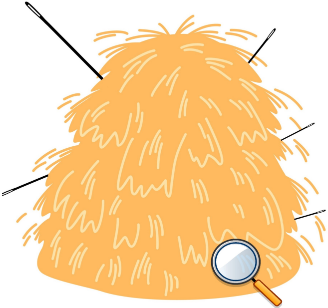
- Relax signal strength or sparsity assumptions
- Improve model selection consistency rates

Consider linear regression with many covariates

$$y = X\beta^* + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$, X is $n \times p$, and wlog $\sigma^2 = 1$

Goal: variables with non-zero effect, when $p \gg n$?



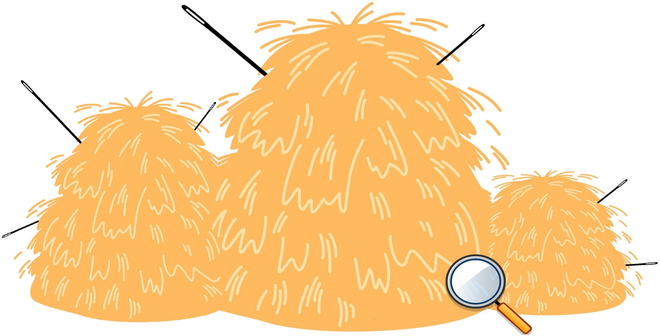
Outline

1. The theory
2. The methods
3. An application

Key idea: block-informed variable selection

- Partition variables into blocks using external data.
- Variables in less sparse or stronger signal blocks penalized less.

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



Key idea: block-informed variable selection

For simplicity we consider L_0 penalties.

$$\hat{S} = \arg\max_M \left\{ \ell(y; \hat{\beta}_M) - \sum_{j=1}^{b} \kappa_j |M_j| \right\}$$

ℓ: log-likelihood; $\hat{\beta}_M$: MLE for model M; κ_j: penalty for block $j = 1, \ldots, b$

For example, BIC corresponds to $\kappa_1 = \ldots = \kappa_b = \log(n)$

Theory applies directly to Zellner's prior on β_M and Binomial or Beta-Binomial priors on models. Therein, we'd let prior inclusion prob depend on the blocks

Intuition: conditions for consistent recovery

To attain $P(\hat{S} = S) \to 1$ as $n \to \infty$, we essentially need

$$\text{Standard L0: } \sqrt{\log(p-s)} \preceq \sqrt{\kappa} \preceq \sqrt{n\rho(X)}\beta_{\min} - \sqrt{\log(s)}$$

$$\text{Block L0:} \sqrt{\log(p_j - s_j)} \preceq \sqrt{\kappa_j} \preceq \sqrt{n\rho(X)}\beta_{\min,j} - \sqrt{\log(s_j)}$$

p_j = size of block j; s_j = number of non-zeroes; $s = \sum_{j=1}^{b} s_j$. $\beta_{\min,j}$ = smallest non-zero $|\beta_j|$ in block j; $\rho(X)$ = smallest eigenvalue related to X_S

If range of feasible κ or κ_j empty, consistent recovery is not possible

- Standard L0: narrow window when s and $p - s$ are large.
- Smaller p_j, larger $\beta_{\min,j} \Rightarrow$ easier support recovery.

Summary of oracle results

If an oracle sets optimal penalties κ and κ_j

- There are settings where consistent selection possible for block L0, but impossible for standard L0
- When both are consistent, block L0 has better rates for $P(\hat{S} = S)$
- Tight bounds shown for both sequence model and regression

Smallest recoverable signals

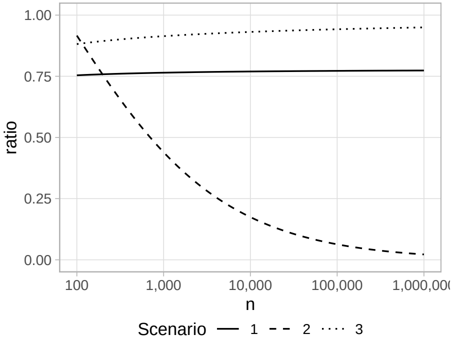
$$\text{Standard L0: } \beta_{\min} \asymp \sqrt{\frac{2\log(p-s)}{n}} + \sqrt{\frac{2\log(s)}{n}}$$

$$\text{Standard L0: } \beta_{\min,j} \asymp \sqrt{\frac{2\log(p_j-s_j)}{n}} + \sqrt{\frac{2\log(s_j)}{n}}$$

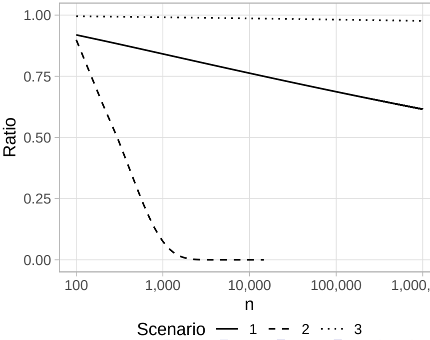
Illustration

Scenario	1 (midly informative)	2 (strongly informative)	3 (uninformative)
s_1, s_2	$1.5 \log n$	$1.5 \log n$	$1.5 \log n$
$p - s$	n	$e^{n/10}$	n
$p_1 - s_1$	$\sqrt{n}$	n^2	$n/2$

Ratio of recoverable signal
block / standard



Ratio of error prob
block/standard



Outline

Bayesian framework

Empirical Bayes method (aka non-oracle) achieves the theoretical improvements

Let's switch to a (more general) Bayesian formulation

- Variable inclusion indicators $\gamma_j = \text{I}(\beta_j \neq 0) \sim \text{Bern}(\pi_j)$
- Spike and slab $\beta \mid \gamma \sim \prod_{\gamma_j=1} N(\beta_j; 0, g)$
- Meta-covariates $w_j \in \mathbb{R}^q$ for variable j (extends the block idea)

Model prior inclusion probabilities as

$$\text{logit}(\pi_j) = w_j^T \theta$$

Idea used by many, e.g. van de Wiel, Te Beest & Münch. Scand Journ Stat 2018 & references therein

Empirical vs. full Bayes

We consider empirical Bayes. Posterior model probabilities

$$p(\gamma \mid y, \hat{\theta}) \propto p(y \mid \gamma) p(\gamma \mid \hat{\theta})$$
$$\hat{\theta} = \arg\max_{\theta} p(y \mid \theta)$$

We use data twice, but $p(\gamma \mid \hat{\theta})$ learns for data (e.g. improved consistency with external data)

Compare to full Bayes, setting a prior $p(\theta)$.

$$p(\gamma \mid y) \propto p(y \mid \gamma) p(\gamma)$$
$$p(\gamma) = \int p(\gamma \mid \theta) p(\theta) d\theta$$

Learning θ doesn't help model selection. All that matters is marginal prior $p(\gamma)$

Example: no meta-covariates. If $\pi \sim \text{Beta}(a, b)$ then $p(\gamma) = \text{Beta-Binomial}(\gamma; a, b)$ (Scott & Berger AOS 2006). Data plays no role in $p(\gamma)$

EBayes solution

Issue in obtaining $\hat{\theta}$: the marginal likelihood is a sum over 2^p models

$$\hat{\theta} = \arg\max_{\theta} p(y \mid \theta) = \sum_{\gamma} p(y \mid \gamma) p(\gamma \mid \theta)$$

Proposition

Under our specified priors, we can evaluate gradient at linear cost in p

$$\nabla_{\theta} \log p(y \mid \theta) = \sum_{i=1}^{p} w_i [P(\beta_i \neq 0 \mid y, \theta) - P(\beta_i \neq 0 \mid \theta)]$$

- Gradient evaluation requires an MCMC run at each θ
- Analogous expectation-propagation algorithm requires a single run (at $\theta = 0$)

Interpretation

Setting gradient to 0 gives a fixed-point equation.

Illustration: suppose that w_i defines b blocks

$$\sum_{i \in B_j} P(\beta_i \neq 0 \mid \theta) = \sum_{i \in B_j} P(\beta_i \neq 0 \mid y, \theta)$$

A simple algorithm

1. Set $l = 0$, $\theta^{(0)} = 0$ (uniform prior $p(\gamma)$ on models)
2. For block b, $\hat{\pi}_b = \frac{1}{p_j} \sum_{i \in B_j} P(\beta_i \neq 0 \mid \theta^{(l)})$
3. Set $\theta_b^{(l)}$ matching $\hat{\pi}_b$ (inverse logit)
4. Set $l = l + 1$, go back to 2 until convergence

Outline

An easy problem?

Goal: Effect of T treatments on outcome, adjusting for J controls.

Consider GLM $y_i \sim p(y_i \mid \eta_i, \phi)$ for $i = 1, \ldots, n$ with linear predictor

$$\eta_i = \sum_{t=1}^{T} \alpha_t d_{i,t} + \sum_{j=1}^{J} \beta_j x_{i,j}$$

- $\alpha = (\alpha_1, \ldots, \alpha_T)$: treatment effects
- $\beta = (\beta_1, \ldots, \beta_J)$: control coefficients
- ϕ dispersion parameter (e.g. σ^2 for Gaussian outcomes)

Focus is on $J \gg n$, fixed T. Issues

- Standard high-dimensional methods. Often run into under-selection
- Fixes to avoid under-selection. Often run into over-selection

Standard high-dim methods (LASSO, BMS etc) usually assume

1. Sparsity
2. All controls to be treated equally (exchangeable)

For finite n, sparsity can lead to under-selection, specially if controls correlated with treatment

Example. $T = 1$, $J = 49$, $n = 1000$, errors $\sim N(0,1)$

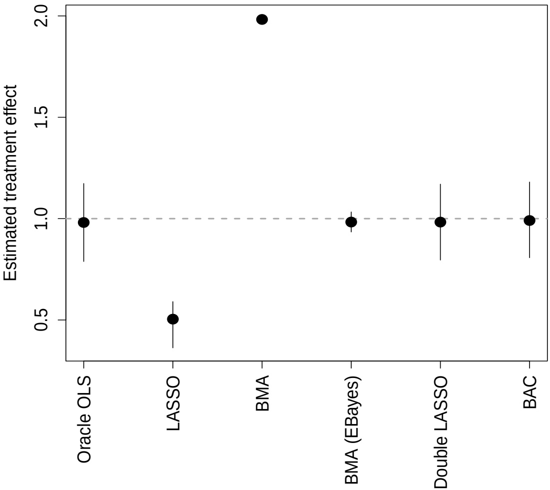
- Truly $\alpha^* = 1$, $\beta_1^* = \ldots, \beta_6^* = 1$, rest truly zero
- Treatment correlated with Controls 1-6 (full confounding)

	d	x_1	x_2	x_3	x_4	x_5	x_6	x_7	...	x_{49}
y	✓	✓	✓	✓	✓	✓	✓	-	...	-
d		✓	✓	✓	✓	✓	✓	-	...	-

BMA (default Zellner and BetaBin(1,1) priors) includes treatment but no controls

DML/BAC are causal inference methods to avoid omitted variables

BMA-EBayes (our method) learns that there's high confounding



Fix: encourage including controls that are correlated with the treatment

- Double LASSO (Belloni, Chernozhukov, Hansen, Rev Econ Stud 2014)
- Bayesian Adjustment for Confounders (Wang, Parmigiani, Dominici Biometrics 2012)
- ...

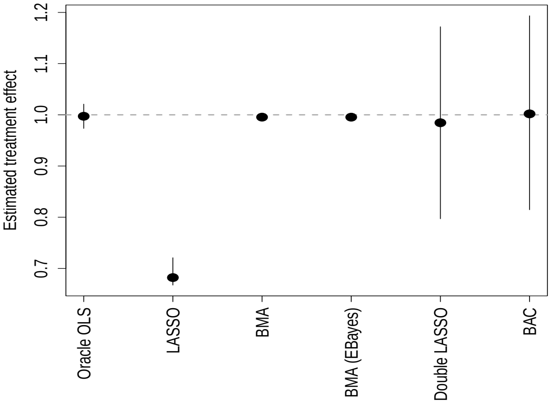
Theorem: DLASSO's $\hat{\alpha}_t$ is asymptotically Normal :-)

- Prevents omitted variable biases. Relevant under high confounding between treatment-controls
- May over-select $\Rightarrow$ variance inflation. Relevant under low confounding

Extended example

Same, but now treatment correlated with Controls 7-12 instead of 1-6

	d	x_1	...	x_6	x_7	...	x_{12}	x_{13}	...	x_{49}
y	✓	✓	...	✓	-	...	-	-	...	-
d		-		-	✓	...	✓	-	...	-



Empirical Bayes

Idea: learn from data whether there's high/low confounding (or neither)

Let $w_j \in \mathbb{R}^T$ measure association between control j and T treatments, e.g. regression coef. of treatment on controls

Method: regress inclusion prob on w_j, e.g. $\text{logit} P(\beta_j \neq 0) = w_j^T \theta$

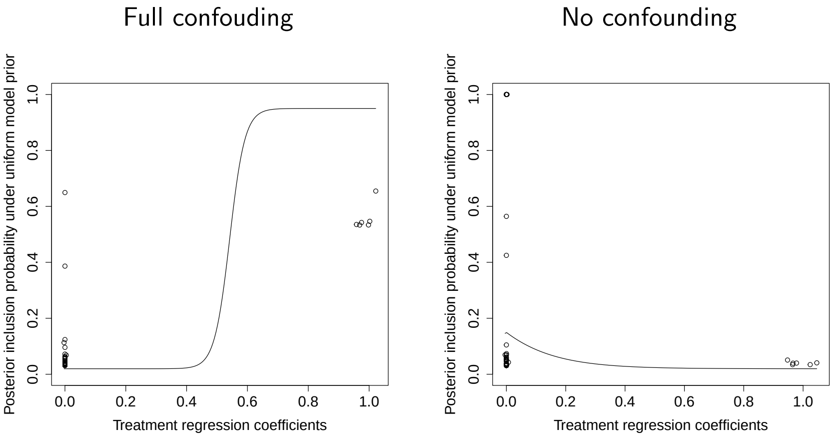
Def. *Confounding coefficient for treatment t (κ_t):* correlation between w_j's and true inclusion ($\beta_j \neq 0$)

Prop. EBayes estimate (argmax of marginal likelihood) matches the prior and posterior expectation of κ_t

$$E(\kappa_t \mid \hat{\theta}) = E(\kappa_t \mid y, \hat{\theta})$$

Example

Prior inclusion prob vs. control's association with treatment (w_j). Note that the marginal likelihood fit can be a bit "aggressive"



Final thoughts

Integrating external data is very Bayesian. One can push the mathematical limits for model recovery

- Milder sparsity/betamin conditions, faster rates of support recovery
- Empirical Bayes method achieves practical gains
- Empirical findings in many applications support the idea

Alas, full Bayes cannot attain the gains (unless prior matches the truth)

Computation is feasible

Broader implications

- Parameter estimation
- Control of false discoveries in multiple testing
- Transfer learning

Main references

· Jewson, Li, Battaglia, Hansen, Rossell, Zwiernik. Graphical model inference with external network data. Biometrics 2024

· Rognon-Vael, Rossell, Zwiernik. Improving variable selection properties by using external data. arxiv 2502.15584 (2025)

· Torrens, Papaspiliopoulos, Rossell. Confounder importance learning for treatment effect inference. Bayesian Analysis 2025

Funding

· Europa Excelencia EUR2020-112096 by the AEI / 10.13039/501100011033 and European Union NextGenerationEU/PRTR

· Consolidación Investigadora MCIN/AEI/UE - CNS2022-135963

· PID2022-138268NB-I00, financed by MCIN/AEI/10.13039/501100011033 and FEDER, UE