BIOS7240 – High Dimensional Data Analysis, Spring 2023

Spectral Deconfounding via Perturbed Sparse Linear Models

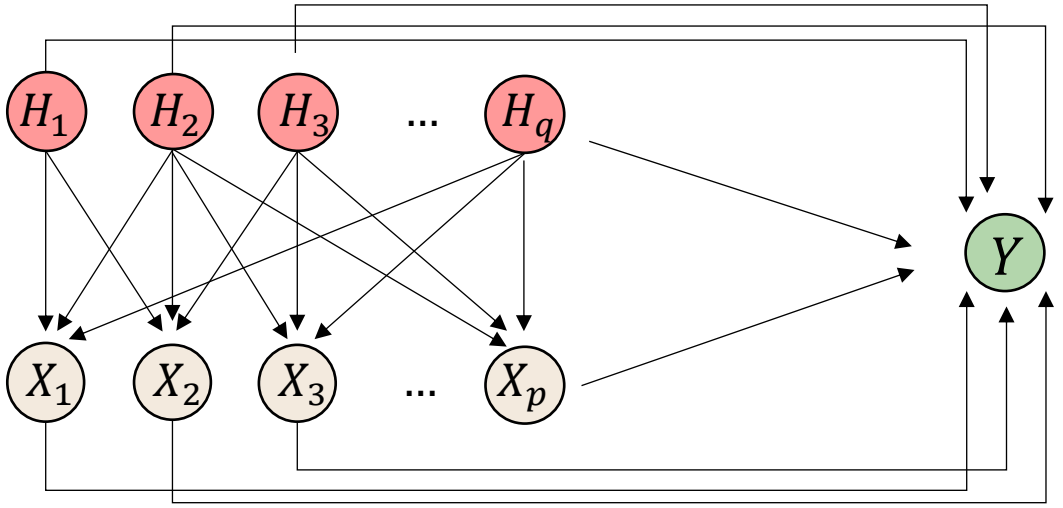
Ćevid, Bühlmann and Meinshausen (2020)

Jamie Merchant

April 24, 2023

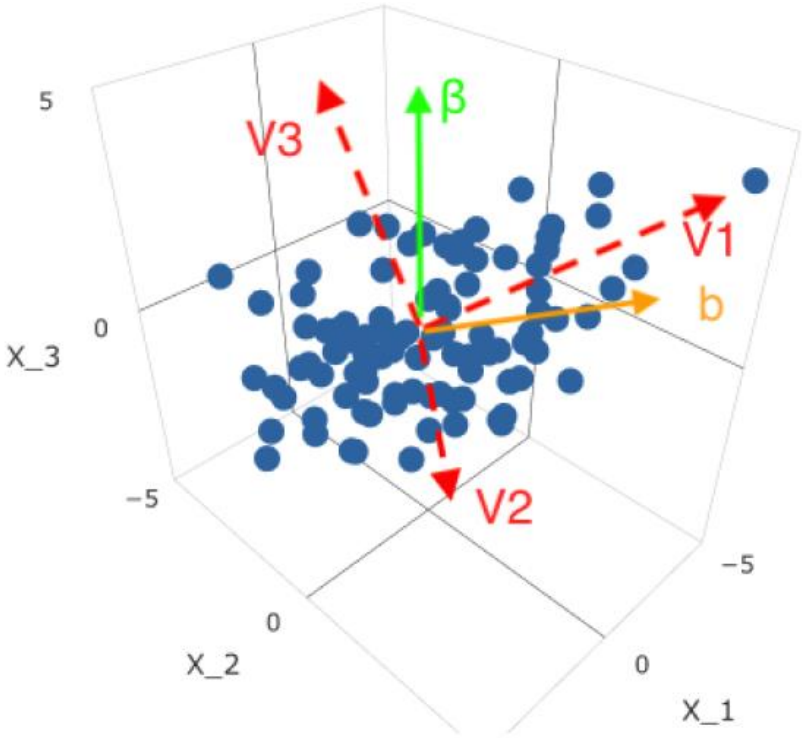
Introduction

- **Motivation:**
  - "Dense" confounding in a high-dimensional setting
  - N observations
  - p predictor variables X
  - q (unobserved) confounding variables H
  - Unobserved confounding is a problem for many high-dimensional regression methods
    - Makes interpretation of coefficients difficult
    - Particularly an issue in genetic studies



Big idea

1. Perform a spectral transformation of the observed data $\mathbf{X}$ that shrinks the largest singular values in some way and create a new $\widetilde{\mathbf{X}}$
2. The perturbation caused by confounding, $\mathbf{Xb}$, will be mostly in the direction of the first few singular values of $\mathbf{X}$, so $\widetilde{\mathbf{X}}\mathbf{b}$ will be very small
3. The deconfounded signal, $\widetilde{\mathbf{X}}\boldsymbol{\beta}$, will not be shrunk nearly as much
4. Estimates will now more closely reflect the true effect of $\mathbf{X}$ with the effect of confounding greatly reduced.



The Confounding Model

$$Y = X\beta + H\delta + \nu \tag{2.1}$$

- Where $X \in \mathbb{R}^{n \times p}$ is the matrix of predictors and $H \in R^{n \times q}$ represents the hidden confounding variables, which exhibit correlation with X.
- Assume that X and H have i.i.d. rows and are jointly Gaussian and that $\nu \in \mathbb{R}^n$ is a vector of sub-Gaussian errors with mean zero and standard deviation σ_ν, independent of X and H.
- Since the model does not change under the transformation $H \leftarrow H\mathrm{Cov}(H)^{-\frac{1}{2}}\delta$, we can assume without loss of generality that $\mathrm{Cov}(H) = I_q$

Note that by L_2 projection, X can be written as:

$$X = H\Gamma + E \tag{2.2}$$

- Where $\Gamma \in \mathbb{R}^{q \times p}$ such that $\mathrm{Cov}(H, E) = 0$
- Columns of E are allowed to be correlated with covariance matrix Σ_E

The Perturbed Linear Model

$$Y = X(\beta + b) + \epsilon \tag{2.3}$$

- Sparse coefficient vector β has been altered by the perturbation vector $b \in \mathbb{R}^p$
- Here assume that the rows of X are i.i.d. sub-Gaussian vectors with mean zero and covariance matrix $\Sigma = \mathrm{Cov}(X)$.

Relates to confounding model because we can rewrite (2.1) as:

Part of the confounding that cannot be explained by X

$$Y = X(\beta + b) + (\overbrace{H\delta - \underbrace{Xb}}) + \nu$$

Part of the confounding effect $H\delta$ that is correlated with X

- Error given by $\epsilon = (H\delta - Xb) + \nu$, which by construction of b is uncorrelated with X and thus independent of X.
- $\sigma^2 = Var(H\delta - Xb + \nu) \leq ||\delta||_2^2 + \sigma_\nu^2$

Spectral transformations

"The idea is to first transform our data by applying some specific linear transformation $F: \mathbb{R}^n \to \mathbb{R}^n$ and then perform the Lasso algorithm."

- $X \to \tilde{X} \coloneqq FX$
- $Y \to \tilde{Y} \coloneqq FY$
- $\hat{\beta} = \arg\min_{\beta}\{\frac{1}{n}\left|\left|\tilde{Y} - \tilde{X}\beta\right|\right|_2^2 + \lambda||\beta||_1$

$$F = U \underbrace{\begin{bmatrix} \tilde{d}_1/d_1 & 0 & \dots & 0 \\ 0 & \tilde{d}_2/d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{d}_r/d_r \end{bmatrix}}_{\tilde{D}^*} U^T$$

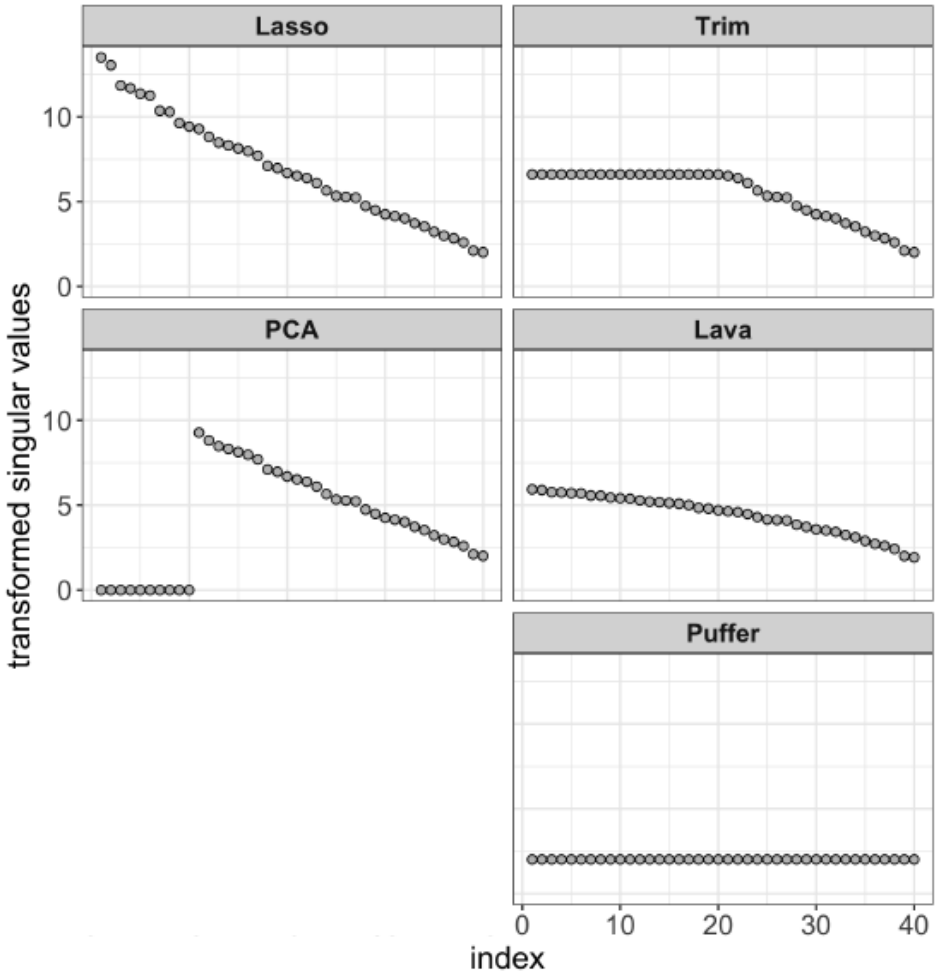
- By SVD, X can be rewritten UDV^T with U orthonormal, so:

$$\tilde{X} = FX = \left(U\tilde{D}^*U^T\right)UDV^T = U\tilde{D}V^T$$

"Lasso performs best when the predictors are uncorrelated and when the errors are independent. Therefore a good choice needs to find a good balance between a well behaved error term $\tilde{\epsilon} = F\epsilon$, well behaved design matrix $\tilde{X}$, and well behaved perturbation term $\tilde{X}b$

Spectral transformations

- The authors propose the **"trim"** transform:
  - $\tilde{d}_i = \min(d_i, \tau)$, with the median singular value representing a "good choice" for τ
- Other alternatives:
  - **PCA** – effectively shrink the first few singular values to zero
  - **Lava** - $d_i = \sqrt{\frac{n\lambda_2 d_i^2}{n\lambda_2 + d_i^2}}$ - solution to the optimization problem proposed by a linear model with a coefficient vector consisting of a dense and sparse component
  - **Puffer** – map all non-zero singular values to 1



Main result

- For the model in (2.1) with $\max_i \Sigma_{ii} = \mathcal{O}(1)$ and $cond(\Sigma_E) = \mathcal{O}(1)$ and with $\lambda_{min}(\Sigma)$ bounded away from zero;
- Under certain assumptions about the data (A1) and a spectral transformation of the data (A2, A3)

Then for penalty level $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$, the ℓ_1-estimation error of the Lasso has the following rate despite the presence of confounding:

$$\left\|\hat{\beta} - \beta\right\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right)$$

- This is the same as the ℓ_1 rate for the Lasso **without** confounding.

Assumptions

(A1) $\lambda_{min}(\Gamma) = \lambda_{min}\big(Cov(X, H)\big) = \Omega(\sqrt{p})$

Assume additionally that a spectral transformation with $\lambda_{max}(F) = 1$ satisfies

(A2) $\lambda_{max}\big(\tilde{X}\big) = \mathcal{O}_p(\sqrt{p})$

(A3) $\phi^2_{\tilde{\Sigma}} = \Omega_p\big(\lambda_{min}(\Sigma)\big)$

- $\phi_M := \inf_{||\alpha||_1 \leq 5||\alpha_S||_1} \frac{\sqrt{\alpha^T M \alpha}}{\frac{1}{\sqrt{s}}||\alpha_S||_1}$; similar to our restricted eigenvalue condition
- λ is used in this article to refer to the singular values of a matrix, not the eigenvalues.

Assumptions (in words)

$$\textbf{(A1)}\lambda_{min}(\Gamma) = \lambda_{min}\big(Cov(X,H)\big) = \Omega\big(\sqrt{p}\big)$$

"Confounding is dense in the sense that each confounding variable is correlated with many predictors… if the confounding is dense in the confounding model, then the induced coefficient perturbation in the underlying perturbed linear model is small."

- If $\frac{B}{A}$ is $\mathcal{O}_p(1)$, then A is $\Omega_p(B)$, i.e. A has asymptotically at least the same rate as B.
- So $\lambda_{min}(\Gamma) = \Omega\big(\sqrt{p}\big) \Leftrightarrow \sqrt{p} = \mathcal{O}_p\big(\lambda_{min}(\Gamma)\big) \Leftrightarrow P\left(\sqrt{p} > \lambda_{min}(\Gamma)\right) \to 0$

Assumptions (in words)

(A2) $\lambda_{max}(\tilde{X}) = \mathcal{O}_p(\sqrt{p})$

"In the confounding model we have $\Sigma = \Gamma^T\Gamma + \Sigma_E$*, i.e. the covariance matrix of X has additional low-rank component* $\Gamma^T\Gamma$*, which causes the top several singular values of X to be very large…*

[this assumption] requires the transformed singular values [of $\tilde{X}$*] to be of order* $\sqrt{p}$*."*

(A3) $\phi^2_{\tilde{\Sigma}} = \Omega_p(\lambda_{min}(\Sigma))$

"This assumption says that the compatibility constant $\phi_{\tilde{\Sigma}}$ *does not substantially decrease after applying our transformation F. We want to show that by shrinking the singular values we have not shrunk our signal X too much."*

Proof

Theorem 2

$$\left|\left|\hat{\beta} - \beta\right|\right|_1 \leq \frac{C_1 s\lambda}{\phi^2_{\tilde{\Sigma}}} + \frac{C_2 \left|\left|\tilde{X}b\right|\right|_2^2}{n\lambda}$$

$\left|\left|\tilde{X}b\right|\right|_2^2$ decreases as the largest singular values of X are decreased

Standard bound for the ℓ_1-error of the Lasso

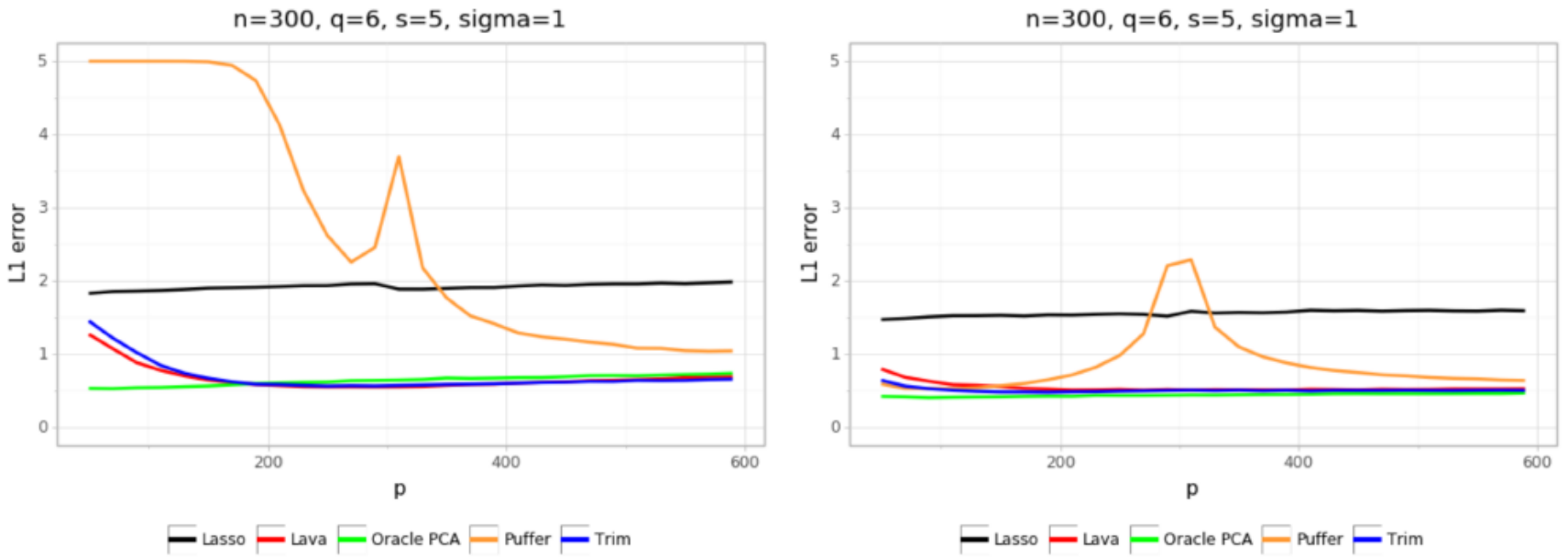
$\phi^2_{\tilde{\Sigma}}$ decreases as the largest singular values of X are decreased

The problem is shrinking the second term enough without increasing the first term.

Simulations

- Generate data from the confounding model (2.1)
  - $Y = X\beta + H\delta + \nu$
  - $X = H\Gamma + E$
- **Parameters**
  - Take $\Sigma_E = \sigma_E^2 I_p$, where $\sigma_E = 2$
  - $\beta = (1,1,1,1,1, 0, \dots, 0)$
  - For a fixed q of hidden confounders, sample Γ_{ij} and δ_i independently as standard normal random variables
  - Noise level $\sigma = 1$ as the standard deviation of ϵ
  - Same as perturbed model, drawing rows of X from $N(0, \Sigma)$ where $\Sigma = \Gamma^T \Gamma + I_p$
- **Feature size/sample size**
  - p = up to 600
  - q = up to 6
  - n=200
- 4096 independent simulations

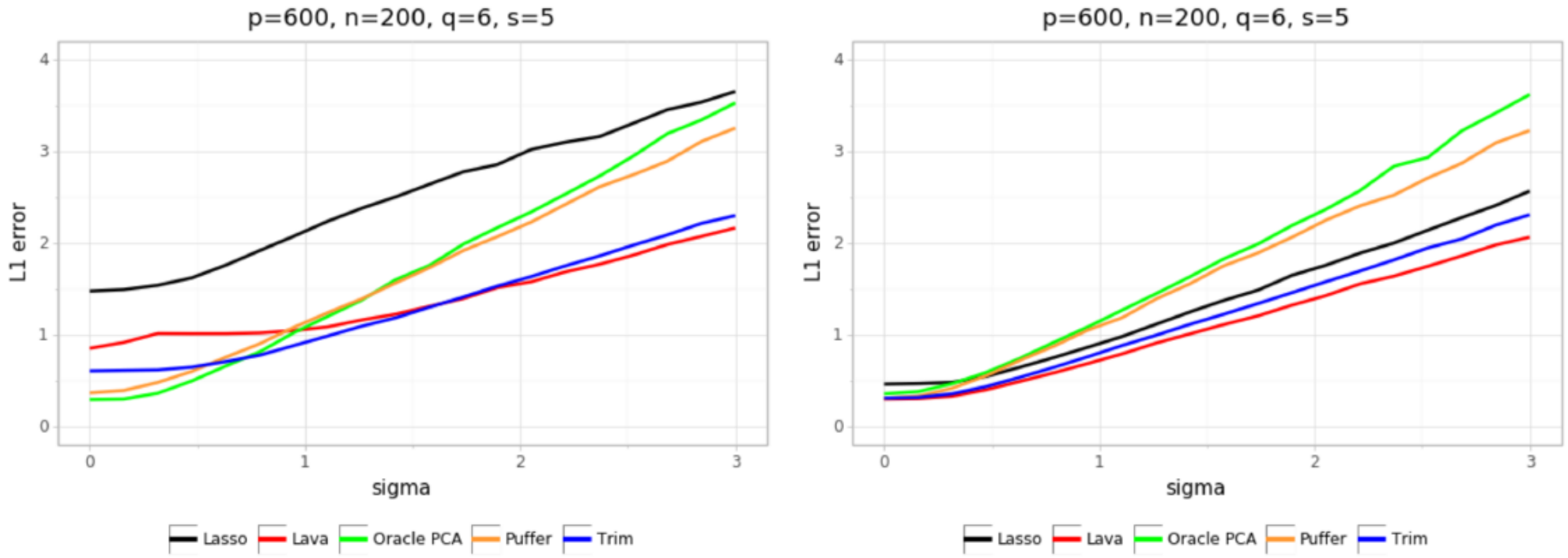
ℓ_1 error $\left|\left|\hat{\beta} - \beta\right|\right|_1$ vs. number of predictors p



Left plot: the penalty chosen by cross-validation

Right plot: penalty chosen by oracle value for which the estimation error is minimal

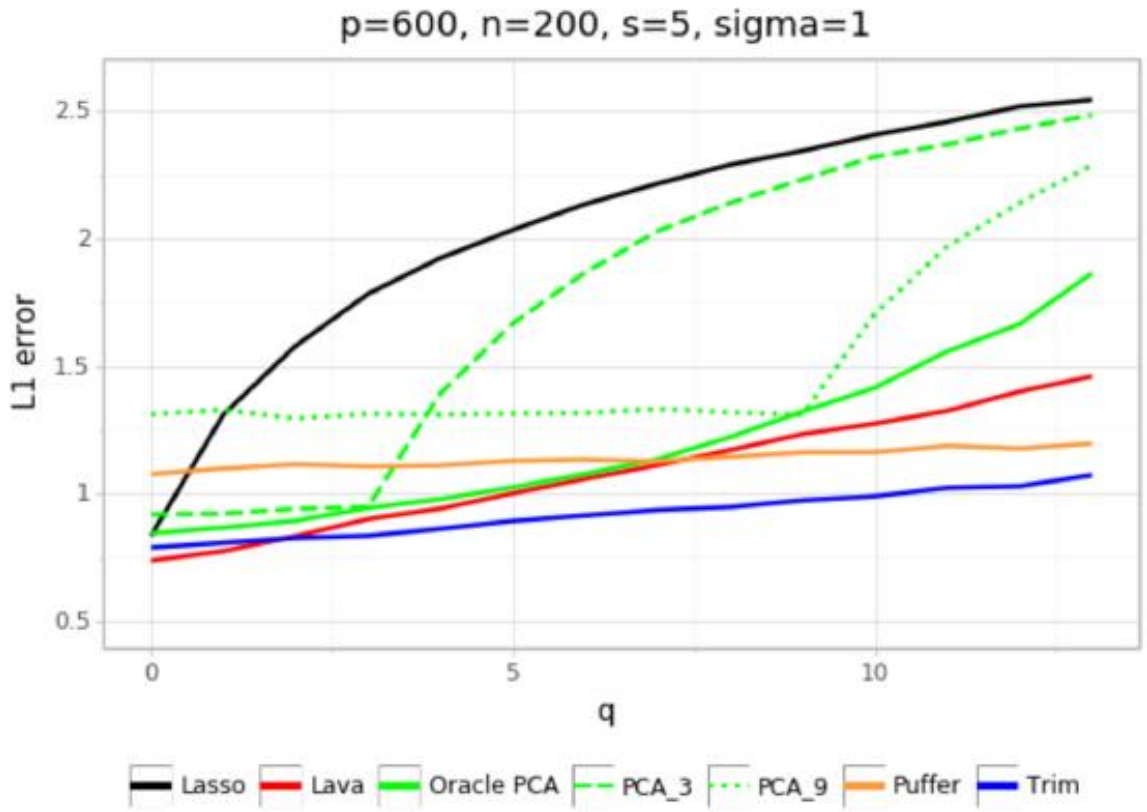
ℓ_1 error $\left|\left|\hat{\beta} - \beta\right|\right|_1$ vs. size of noise



Left plot: Confounding model

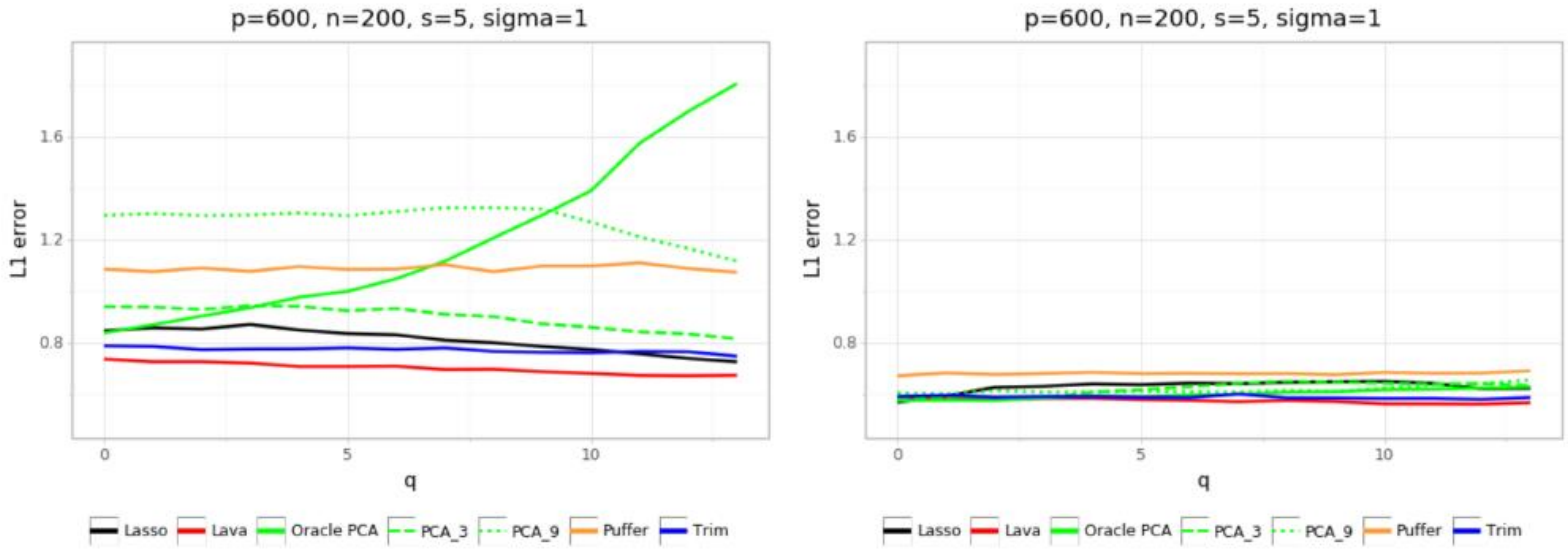
Right plot: Perturbed linear model

ℓ_1 error $\left|\left|\hat{\beta} - \beta\right|\right|_1$ vs. number of confounders q



ℓ_1 error $\left|\left|\hat{\beta} - \beta\right|\right|_1$ vs. confounders, $b = 0$

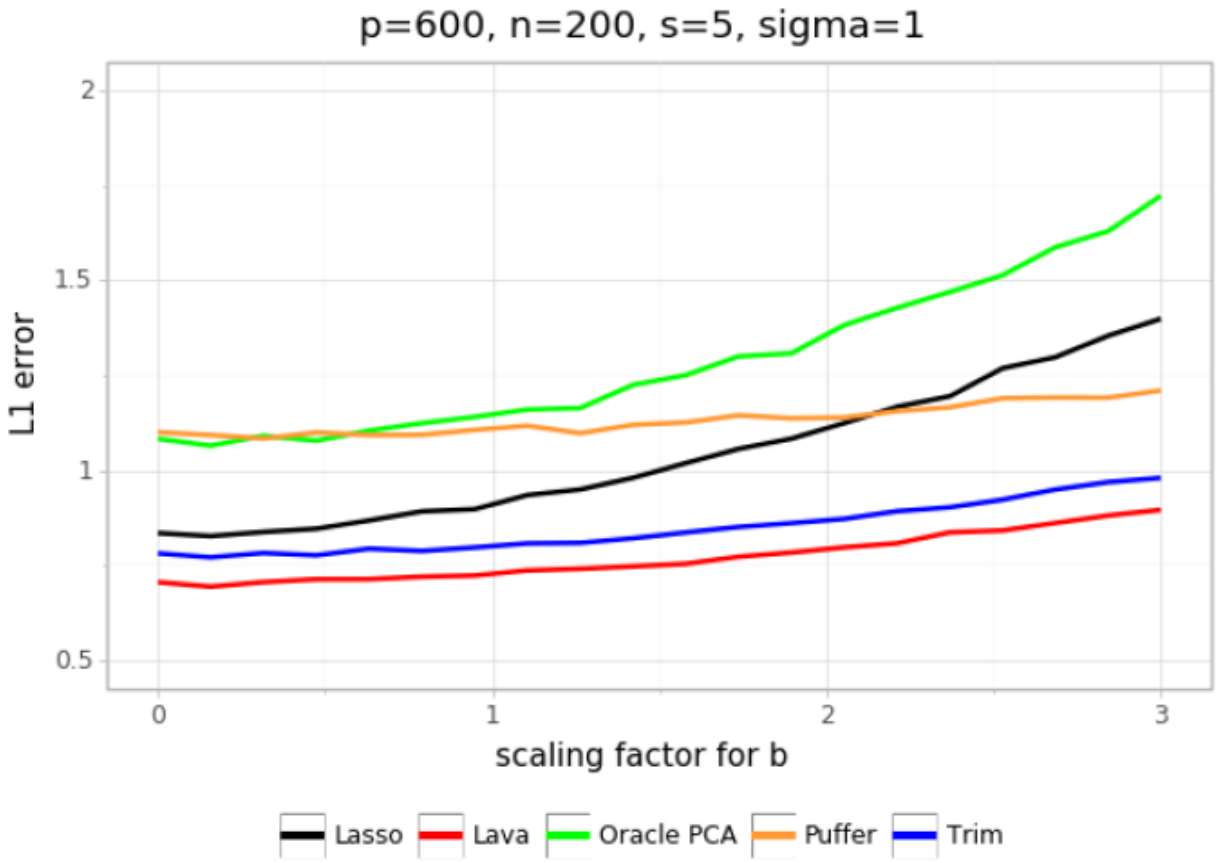
Sparse linear model where $\Sigma = \Gamma^T\Gamma + I_p$, i.e. the confounding model where the induced perturbation b is set to $b = 0$.



Left plot: Penalty chosen by CV

Right plot: Penalty chosen by the oracle value which minimizes ℓ_1-error

ℓ_1 error $\left|\left|\hat{\beta} - \beta\right|\right|_1$ vs. scale of perturbation b



Application

- Data from the GTEx Portal (http://gtexportal.org)
- p =14,713 protein-coding genes from $n = 491$ samples.
- Includes $q = 65$ proxies for "hidden confounders"
  - Genotyping principal components
  - PEER factors
- Gene expression as quantified by amount of mRNA in cell created from each gene
- Prior knowledge allows us to "regress out"* the confounders from X and create new $X^{(k)}$

* i.e. regress each column in X by q confounders, then use the residuals from these regressions in place of the original X.

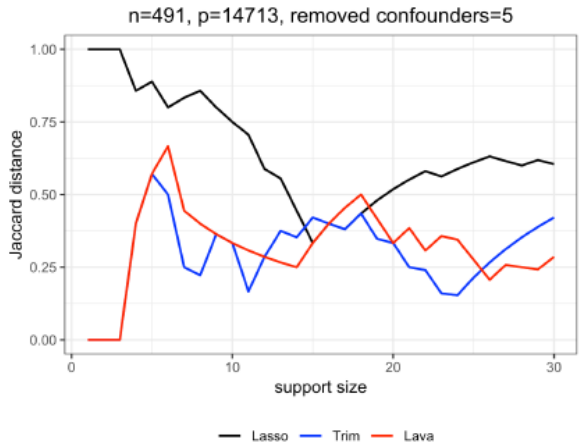
Application

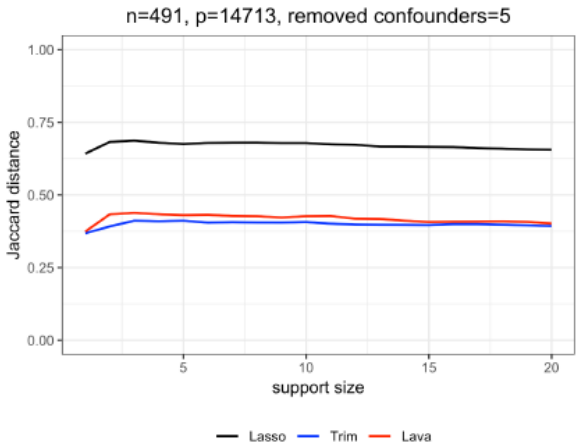
- Pick a random gene as your outcome Y. Remaining genes are X.
- For a fixed value k, regress out first k given confounder proxies from X
- For each $s = 1, \dots, 20$ apply a given method on X and $X^{(k)}$ with regularization parameter λ chosen as the largest value such that the support size of $\hat{\beta}$ equals a prespecified value s.
- Measure **Jaccard distance** $J(A, B) = \frac{A \triangle B}{A \cup B}$ on the supports of $\hat{\beta}_s$ and $\hat{\beta}_s^{(k)}$

$$\text{e.g. } \hat{\beta}_s = \begin{pmatrix} \hat{\beta}_A \\ \hat{\beta}_B \\ \hat{\beta}_C \\ \hat{\beta}_D \\ \hat{\beta}_E \end{pmatrix}, \hat{\beta}_s^{(k)} = \begin{pmatrix} \hat{\beta}_A \\ \hat{\beta}_B \\ \hat{\beta}_C \\ \hat{\beta}_D \\ \hat{\beta}_F \end{pmatrix} \Rightarrow J\left(supp\left(\hat{\beta}_s\right), supp\left(\hat{\beta}_s^{(k)}\right)\right) = \frac{2}{6}$$

* $\triangle$ = symmetric difference operator, i.e. $A \triangle B$= elements in either A or B but not both v

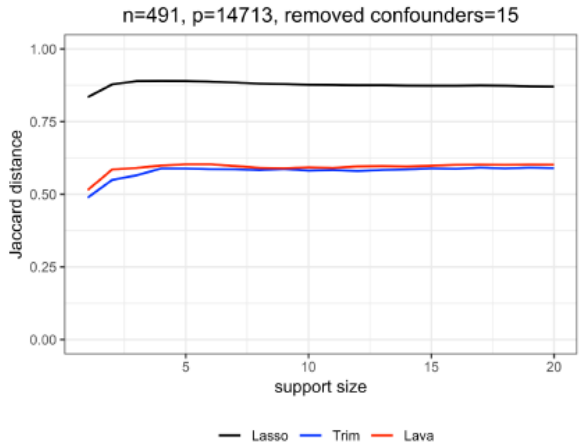
Application

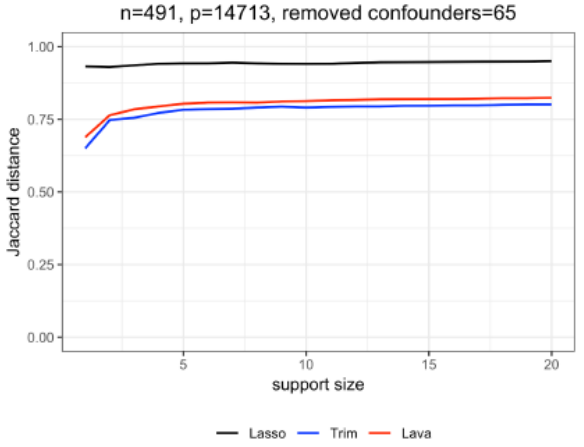




Upper left: 5 confounders removed, one randomly chosen response.

Upper right: 5 confounders removed, averaged over 500 randomly chosen responses.





Lower left: 15 confounders removed, averaged over 500 randomly chosen responses.

Lower right: 65 confounders removed, averaged over 500 randomly chosen responses.

Discussion

- How realistic are the assumptions?
- Can they be tested?
- Why use Jaccard distance and not ℓ_1 norm for application?

References

Ćevid, D., Bühlmann, P. and Meinshausen N. (2020) Spectral Deconfounding via Perturbed Sparse Linear Models. *Journal of Machine Learning Research*, 21:1-41.