Doubly Debiased Lasso:
High-Dimensional Inference under Hidden Confounding

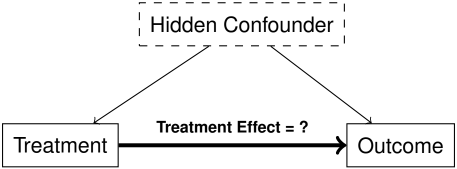
Zijian Guo

Department of Statistics, Rutgers University

Confounding and Causal Effects

“Confounding is one of the most fundamental impediments to the elucidation of causal inferences from empirical data." (Pearl, 2009)

Causal inference with hidden confounders “remains a fertile field for methodological research." (Imbens, Rubin, 2015)



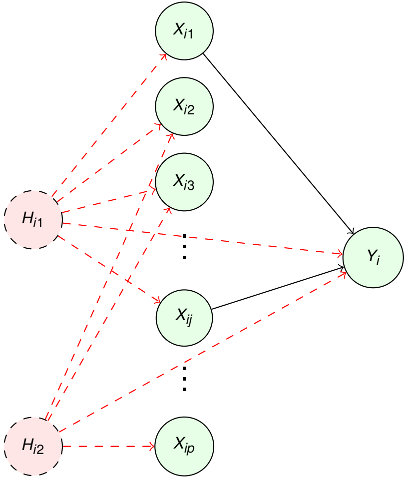
- OLS or Lasso is biased.
- Construction of instrumental variables.

Today’s Talk Focus

Domagoj Ćevid

Peter Bühlmann

Guo Z, Ćevid D, Bühlmann P. (2022) Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding. Ann. Statist. 50(3): 1320-1347.



The hidden *H* (e.g. population, environment) affects both the landmark gene (Y) and other genes (X).

Hidden Confounder Model

Linear Structural Equation Model, for $1 \leq i \leq n$,

$$Y_i \leftarrow \beta^{\intercal} X_{i\cdot} + \phi^{\intercal} H_{i\cdot} + e_i \quad X_{i\cdot} = \Psi^{\intercal} H_{i\cdot} + E_{i\cdot} \tag{1}$$

- Observed: Y_i and $X_{i\cdot} \in \mathbb{R}^p$
- High-dimension with $\|\beta\|_0 \leq k$
- Small number of hidden $H_{i\cdot} \in \mathbb{R}^q$
- Dense Hidden Confounders.

Dense Confounding

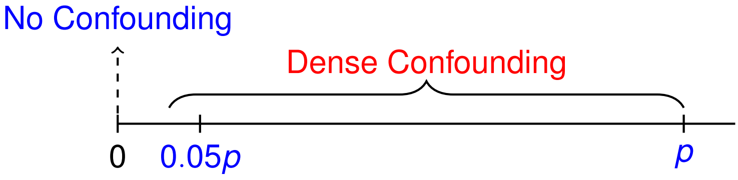


Figure: Number of confounded X_{ij}

Research goal

Inference for β_j in High-Dim and Dense Confounding.

Doubly Debiased Lasso

Perturbed Linear Model

$$Y_i \leftarrow \beta^\intercal X_{i\cdot} + \phi^\intercal H_{i\cdot} + e_i$$

Marginalizing out $H_{i\cdot}$ (Approximating $\phi^\intercal H_{i\cdot}$ by $b^\intercal X_{i\cdot}$)

$$Y_i = (\beta + b)^\intercal X_{i\cdot} + \epsilon_i \quad \text{and} \quad X_{i\cdot} = \Psi^\intercal H_{i\cdot} + E_{i\cdot}, \qquad (2)$$

$$\text{with } \epsilon_i = e_i + \phi^\intercal H_{i\cdot} - b^\intercal X_{i\cdot} \quad \text{and} \quad b = \left(\Sigma_E + \Psi\Psi^\intercal\right)^{-1}\Psi^\intercal\phi.$$

Dense Confounding setting:

- $\frac{1}{\sqrt{n}}\|Xb\|_2$ is of constant order.
- Dense $\Psi^\intercal \in \mathbb{R}^p \rightarrow \|b\|_2 \lesssim \sqrt{1/p}$.

Spectral Deconfounding Estimator of β

For a spectral transformation $\mathcal{Q} \in \mathbb{R}^{n\times n}$

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2n}\|\mathcal{Q}\left(Y - X\beta\right)\|_2^2 + \lambda_1 \sum_{j=1}^{p} \frac{\|\mathcal{Q}X_{\cdot j}\|_2}{\sqrt{n}}|\beta_j|, \quad (3)$$

where $\lambda_1 = A\sigma_\epsilon\sqrt{\log p/n}$ is a tuning parameter.

- Lasso estimator: $\mathcal{Q} = \mathrm{I}$
- Trim transformation $\mathcal{Q}$ (Ćevid, Bühlmann, Meinshausen, 2018)

Trim Transform

For $X \in \mathbb{R}^{n\times p}$, $X = U\Lambda V^{\intercal}$. (Focus on $p > n+1$)

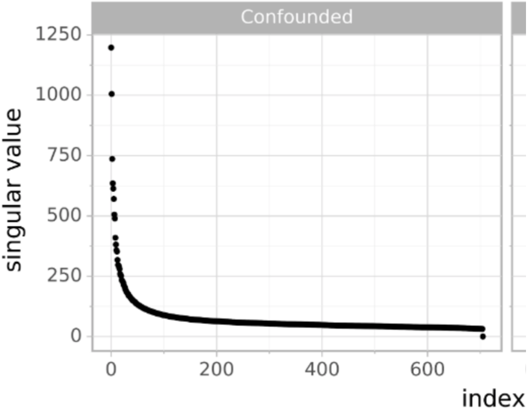
- ρ-trim $\mathcal{Q} \in \mathbb{R}^{n\times n}$ shrinks top $\lfloor \rho n \rfloor$ singular values of X.

$$\mathcal{Q}X = UH_{\rho}(\Lambda)V^{\intercal}$$

- $H_{\rho}(\Lambda)$ shrinks $\Lambda_{1,1}, \Lambda_{2,2}, \cdots, \Lambda_{\lfloor \rho n \rfloor -1, \lfloor \rho n \rfloor -1}$ to $\Lambda_{\lfloor \rho n \rfloor, \lfloor \rho n \rfloor}$.
- Robust: median trim with $\rho = \frac{1}{2}$

GTEx project (Lonsdale et al., 2013)

$p = 12,646$ gene expression levels; $n = 706$ individuals.



Two Sources of Bias

Confidence Interval Construction for β_j with $1 \leq j \leq p$?

Standard Debiased Lasso: Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014).

$$Y - X_{-1}\widehat{\beta}_{-1} = \epsilon + X_1(\beta_1 + b_1) + X_{-1}b_{-1} + X_{-1}(\beta_{-1} - \widehat{\beta}_{-1}).$$

- $X_{-1}b_{-1}$: hidden confounder bias
- $X_{-1}(\beta_{-1} - \widehat{\beta}_{-1})$: bias of estimating β

Hidden Confounding Bias

Introduce $\mathcal{P} \in \mathbb{R}^{n\times n}$ to shrinks singular values of the nuisance design X_{-1}.

$$\mathcal{P}(Y - X_{-1}\widehat{\beta}_{-1}) = \mathcal{P}[\epsilon + X_1(\beta_1 + b_1)] + \mathcal{P}X_{-1}b_{-1} + \mathcal{P}X_{-1}(\beta_{-1} - \widehat{\beta}_{-1}).$$

$\mathcal{P}$: reduce the bias $X_{-1}b_{-1}$

- $\frac{1}{\sqrt{n}}\|\mathcal{P}X_{-1}b_{-1}\|_2 \lesssim \sqrt{q/\min\{n,p\}}$
- $\frac{1}{\sqrt{n}}\|X_{-1}b_{-1}\|_2$ can be of constant order.

High-dimensional Bias

Decouple the relationship (with $\lambda \asymp \sqrt{\log p/n}$)

$$\widehat{\gamma} = \arg\min_{\gamma \in \mathbb{R}^{p-1}} \left\{ \frac{1}{2n} \|\mathcal{P}(X_1 - X_{-1}\gamma)\|_2^2 + \lambda \sum_{j=2}^{p} \frac{\|\mathcal{P}X_{\cdot,j}\|_2}{\sqrt{n}} |\gamma_j| \right\}.$$

Define the residue $Z_1 = X_1 - X_{-1}\widehat{\gamma}$ as

$$\text{KKT Condition}: \ \|\frac{1}{n}(\mathcal{P}Z_1)^{\intercal}\mathcal{P}X_{-1}\|_\infty \leq \lambda$$

Doubly Debiased Lasso

$$\widetilde{\beta}_1 = \frac{(\mathcal{P}Z_1)^{\intercal}\mathcal{P}(Y - X_{-1}\widehat{\beta}_{-1})}{(\mathcal{P}Z_1)^{\intercal}\mathcal{P}X_1}.$$

Decompose the error $\widetilde{\beta}_1 - \beta_1$

$$\frac{(\mathcal{P}Z_1)^{\intercal}\mathcal{P}\epsilon}{(\mathcal{P}Z_1)^{\intercal}\mathcal{P}X_1} + \frac{(\mathcal{P}Z_1)^{\intercal}\mathcal{P}X_{-1}b_{-1}}{(\mathcal{P}Z_1)^{\intercal}\mathcal{P}X_1} + \frac{(\mathcal{P}Z_1)^{\intercal}\mathcal{P}X_{-1}(\beta_{-1} - \widehat{\beta}_{-1})}{(\mathcal{P}Z_1)^{\intercal}\mathcal{P}X_1} + b_1.$$

- $\mathcal{P} = \mathrm{I}$: standard debiased estimator.
- $\mathcal{P}$ depends on parameter of interest.

Confidence Interval

A confidence interval (CI) with asymptotic coverage $1-\alpha$ can be obtained as

$$\mathrm{CI}(\beta_1) = \left(\widetilde{\beta}_1 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{\sigma}_\epsilon^2 \cdot Z_1^\intercal \mathcal{P}^4 Z_1}{(Z_1^\intercal \mathcal{P}^2 X_1)^2}}, \widetilde{\beta}_1 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{\sigma}_\epsilon^2 \cdot Z_1^\intercal \mathcal{P}^4 Z_1}{(Z_1^\intercal \mathcal{P}^2 X_1)^2}}\right),$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of standard normal.

We estimate the variance of $\epsilon_i = Y_i - (\beta + b)^\intercal X_{i.}$ by

$$\widehat{\sigma}_\epsilon^2 = \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}\|\mathcal{Q}(Y - X\widehat{\beta})\|_2^2.$$

Normality and Efficiency

Assumptions

$$X_{i\cdot} = \Psi^{\intercal} H_{i\cdot} + E_{i\cdot} \quad \text{with} \quad \Psi = (\Psi_1, \Psi_2, \ldots, \Psi_p)$$

Sparse Precision: $\Omega_E = (\mathbb{E} E_{i\cdot} E_{i\cdot}^{\intercal})^{-1}$ satisfies

$$c_0 \leq \lambda_{\min}(\Omega_E) \leq \lambda_{\max}(\Omega_E) \leq C_0, \quad \|(\Omega_E)_{1\cdot}\|_0 \log p/n \to 0.$$

Assumptions

$$X_{i\cdot} = \Psi^{\intercal} H_{i\cdot} + E_{i\cdot} \quad \text{with} \quad \Psi = (\Psi_1, \Psi_2, \ldots, \Psi_p)$$

Sparse Precision: $\Omega_E = (\mathbb{E} E_{i\cdot} E_{i\cdot}^{\intercal})^{-1}$ satisfies

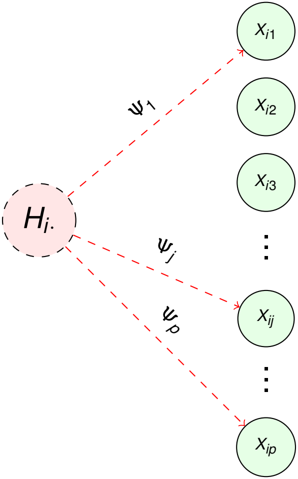
$$c_0 \leq \lambda_{\min}(\Omega_E) \leq \lambda_{\max}(\Omega_E) \leq C_0, \quad \|(\Omega_E)_{1\cdot}\|_0 \log p/n \to 0.$$

Dense Confounding: $\lambda_q(\Psi_{-1}) \gtrsim \sqrt{p}$ and
$\max\{\|\Psi(\Omega_E)_{\cdot,1}\|_2, \|\Psi_1\|_2, \|\Psi_{-1}(\Omega_E)_{-1,1}\|_2\} \lesssim \sqrt{q}$.

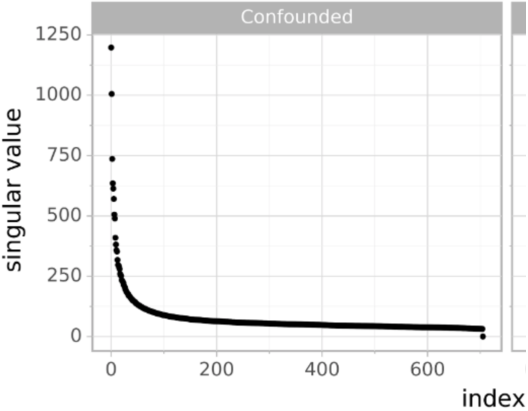
Commonly assumed in factor model.

When Dense Confounding holds?

A fixed proportion of $\{\Psi_l \in \mathbb{R}^q\}_{1\leq j\leq p}$ are i.i.d generated



$p = 12,646$ gene expression levels; $n = 706$ individuals.



Theorem 1 (G. Ćevid, Bühlmann, 2020).

Suppose $q \ll \min\{\sqrt{n}, \frac{n}{\|(\Omega_E)_{1\cdot}\|_0 \log p}\}$, $k \ll \frac{\sqrt{n}}{\log p}$ *and* $\min\{\rho, \rho_1\} \min\{n, p\} \geq 3q + 1$, *then*

$$\frac{1}{\sqrt{V}} \left(\widetilde{\beta}_1 - \beta_1\right) \xrightarrow{d} N(0, 1), \tag{4}$$

$$\mathrm{V} = \frac{\sigma_\epsilon^2 \cdot Z_1^\intercal \mathcal{P}^4 Z_1}{(Z_1^\intercal \mathcal{P}^2 X_1)^2} \quad \textit{and} \quad V^{-1} \frac{\sigma_\epsilon^2 \mathrm{Tr}(\mathcal{P}^4)}{\sigma_\eta^2 [\mathrm{Tr}(\mathcal{P}^2)]^2} \xrightarrow{p} 1. \tag{5}$$

- Sparse β and small q.

Theorem 1 (G. Ćevid, Bühlmann, 2020).

Suppose $q \ll \min\{\sqrt{n}, \frac{n}{\|(\Omega_E)_{1\cdot}\|_0 \log p}\}$, $k \ll \frac{\sqrt{n}}{\log p}$ *and* $\min\{\rho, \rho_1\}\min\{n, p\} \geq 3q + 1$, *then*

$$\frac{1}{\sqrt{V}}\left(\widetilde{\beta}_1 - \beta_1\right) \xrightarrow{d} N(0,1), \tag{4}$$

$$\mathrm{V} = \frac{\sigma_\epsilon^2 \cdot Z_1^{\intercal}\mathcal{P}^4 Z_1}{(Z_1^{\intercal}\mathcal{P}^2 X_1)^2} \quad \textit{and} \quad V^{-1}\frac{\sigma_\epsilon^2 \mathrm{Tr}(\mathcal{P}^4)}{\sigma_\eta^2[\mathrm{Tr}(\mathcal{P}^2)]^2} \xrightarrow{p} 1. \tag{5}$$

- Sparse β and small q.
- Enough trimming: $\min\{\rho, \rho_1\}\min\{n, p\} \geq 3q + 1$

Efficiency

Benchmark: Gauss-Markov $\sigma_\epsilon^2/(n\sigma_\eta^2)$
$\mathrm{ARE} = \lim_{n\to\infty} \frac{\mathrm{V}}{\sigma_\epsilon^2/(n\sigma_\eta^2)}$

$$\mathrm{ARE} \in \left[\frac{1}{\min\{c^*,1\}}, \frac{1}{(1-\rho^*)\min\{c^*,1\}}\right], \tag{6}$$

where $\rho^* = \lim_{n\to\infty} \rho_1(n)$ and $c^* = \liminf p/n \in (0,\infty]$.

1. A smaller $\rho_1(n)$ leads to a more efficient estimator.
2. For $\rho^* = 0$, $\mathrm{ARE} = \max\{1/c^*, 1\}$

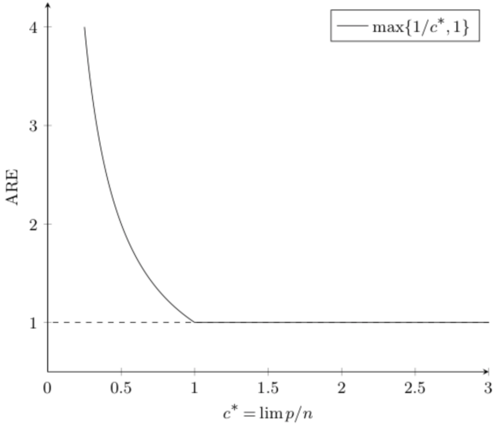
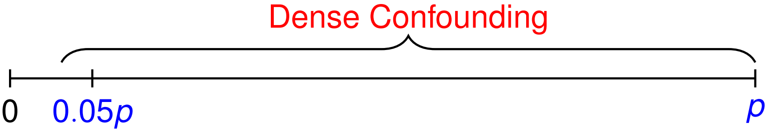


FIG 1. *The plot of* ARE *versus* $c^* = \lim p/n$, *for the setting of* $\rho^* = 0$.

Theorem 2 (G. Ćevid, Bühlmann, 2020).

If $p \geq n$ and $\rho_1(n) \to 0$, Doubly Debiased Lasso estimator achieves the Gauss-Markov efficiency bound $\frac{\sigma_{\epsilon}^2}{\sigma_{\eta}^2 n}$.

- We take $\rho_1(n) = \frac{3q+1}{n} \to 0$.
- **Efficiency: Dense Confounding=No Confounding.**



- Typical results for hidden confounder model: retain consistency but lose efficiency. (Wooldridge, 2010)

Numerical Results

Simulation Setting

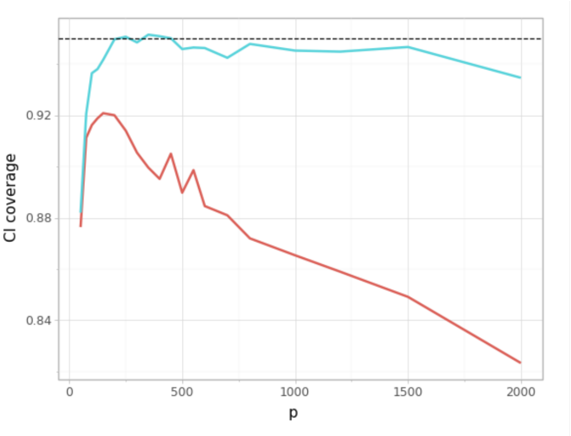
$$Y_i \leftarrow \beta^\intercal X_{i\cdot} + \phi^\intercal H_{i\cdot} + e_i$$
$$X_{i\cdot} = \Psi^\intercal H_{i\cdot} + E_{i\cdot}$$

1. $q = 3$
2. $s = 5$ and $\beta = (1, 1, 1, 1, 1, 0, \cdots 0) \in \mathbb{R}^p$.
3. $E_{i\cdot} \sim N_p(0, \mathrm{I})$, $H_{i\cdot} \sim N_q(0, \mathrm{I})$, $e_i \sim N(0, 1)$
4. $\Psi_{ij} \sim N(0, 1)$ for $1 \leq i \leq q, 1 \leq j \leq p$ and $\phi_j \sim N(0, 1)$ for $1 \leq j \leq q$
5. We repeat the simulations 5,000 times.

Dependence on p

$n = 500$ and p varies from 50 to 2,000,

- Doubly Debiased Lasso
- Standard Debiased Lasso



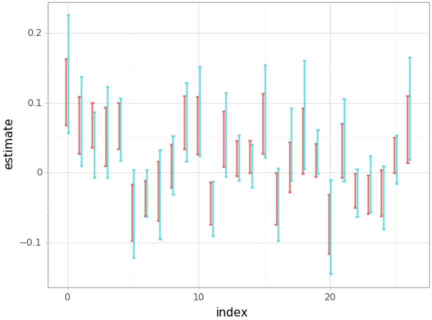
Real Data Analysis

Data: GTEx project (Lonsdale et al., 2013)

- Normalized gene expression data for the skeletal muscle tissue
- $p = 12,646$ gene expression levels over $n = 706$ individuals.
- Outcome: Landmark genes (Subramanian et al., 2017)

CI Comparison

- *x*-axis is aligned with absolute values of Lasso.
- Doubly Debiased Lasso is more conservative than Standard Debiased Lasso



Take-home Message

1. Challenges: High-dimension + hidden confounding
2. Doubly Debiased Lasso: New Trim.
3. Efficiency.

Take-home Message

1. Challenges: High-dimension + hidden confounding
2. Doubly Debiased Lasso: New Trim.
3. Efficiency.

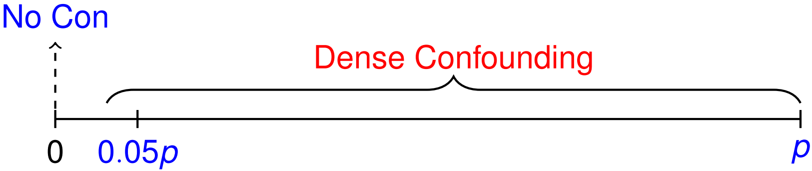


Figure: Number of confounded X_{ij}

Make good use of Dense Confounding.

Reference and Acknowledgement

Guo Z, Ćevid D, Bühlmann P. (2022) Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding. Ann. Statist. 50(3): 1320-1347.

The R package `DDL` is available at CRAN.

Acknowledgement to NSF, NIH, ETH FIM for fundings.

Thank you!

Discussion (Big Picture)

Methods to Address Confounding

	One Treatment	Many Treatments
Condition	Constructing valid IV	Dense confounding
Method	IV-based method	Doubly Debiased
Consistency	yes (valid IV)	yes
Efficiency	no	yes

Methods to Address Confounding

	One Treatment	Many Treatments
Condition	Constructing valid IV	Dense confounding
Method	IV-based method	Doubly Debiased
Consistency	yes (valid IV)	yes
Efficiency	no	yes

High-dim IVs: Gautier and Rose, 2011; Fan and Liao, 2014; Lin et al., 2015; Belloni et al., 2017; Neykov et al., 2018.

Methods to Address Confounding

	One Treatment	Many Treatments
Condition	Constructing valid IV	Dense confounding
Method	IV-based method	Doubly Debiased
Consistency	yes (valid IV)	yes
Efficiency	no	yes

High-dim IVs: Gautier and Rose, 2011; Fan and Liao, 2014; Lin et al., 2015; Belloni et al., 2017; Neykov et al., 2018.

Invalid IVs: Guo, Z., Kang, H., Cai, T. T., & Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two stage hard thresholding with voting. JRSSB, 80(4), 793-815.

Advertisement: econometrics seminar at CUHK, April 21, 2021.

Nuisance Trim

Introduce $\mathcal{P} \in \mathbb{R}^{n \times n}$ to shrinks singular values of the nuisance design X_{-1}.

- $X_{-1} = U(X_{-1})\Lambda(X_{-1})V^{\intercal}(X_{-1})$
- ρ_1-trim $\mathcal{P}$ shrinks top $\lfloor \rho_1 n \rfloor$ singular values of X_{-1} to $[\Lambda(X_{-1})]_{\lfloor \rho_1 n \rfloor, \lfloor \rho_1 n \rfloor}$

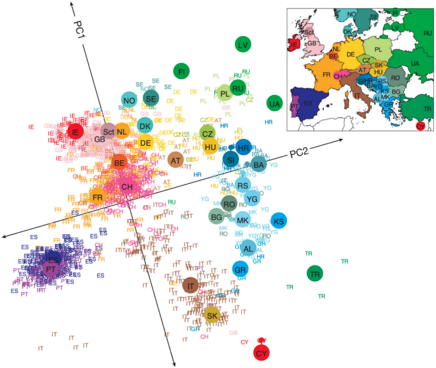
Nuisance Trim

Introduce $\mathcal{P} \in \mathbb{R}^{n \times n}$ to shrinks singular values of the nuisance design X_{-1}.

- $X_{-1} = U(X_{-1})\Lambda(X_{-1})V^{\intercal}(X_{-1})$
- ρ_1-trim $\mathcal{P}$ shrinks top $\lfloor \rho_1 n \rfloor$ singular values of X_{-1} to $[\Lambda(X_{-1})]_{\lfloor \rho_1 n \rfloor, \lfloor \rho_1 n \rfloor}$
- $\mathcal{P}$ depends on the variable of interest!
- $\mathcal{P} \neq \mathcal{Q}$.
- ρ_1 affects the efficiency.

Genes mirror geography within Europe

Data: 197,146 loci in 1,387 individuals (individual origins).



Novembre, John, et al. Nature (2008).

Measurement Error Model

SEM between outcome and covariates

$$Y_i^0 \leftarrow \beta^{\intercal} X_i^0 + \epsilon_i$$

Measurement contaminated with hidden confounders

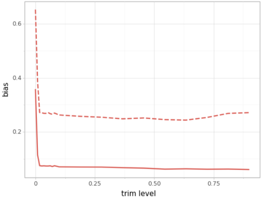
$$Y_i \leftarrow \nu^{\intercal} H_{i\cdot} + Y_i^0; \quad X_{i\cdot} \leftarrow \Psi^{\intercal} H_{i\cdot} + X_{i\cdot}^0.$$

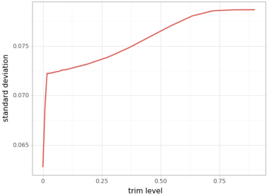
Reduced to the hidden confounding model

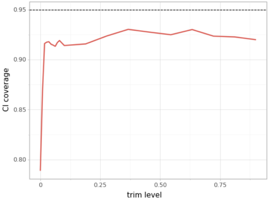
$$Y_i = \beta^{\intercal} X_{i\cdot} + (\nu - \Psi\beta)^{\intercal} H_{i\cdot} + \epsilon_i \quad \text{and} \quad X_{i\cdot} = \Psi^{\intercal} H_{i\cdot} + X_{i\cdot}^0$$

Trimming and Efficiency

- $q = 3$
- $\rho = 0.02$: shrinks 6 largest singular values
- $\rho = 0.02$: B_b and B_β are small







Dependence on n

$p = 500$ and n varies from 100 to $2,000$.