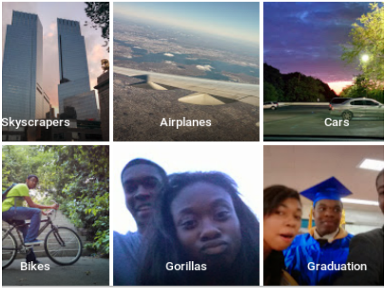
Reliable hypothesis testing paradigms in high dimensions

Yuting Wei

Carnegie Mellon University

Oct 2020

Reliable uncertainty quantification



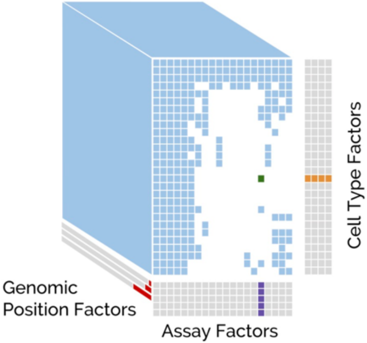
- Google photos tags two African-Americans as gorillas, 2015
- Fatal motorway collision between a Tesla and a truck, 2016

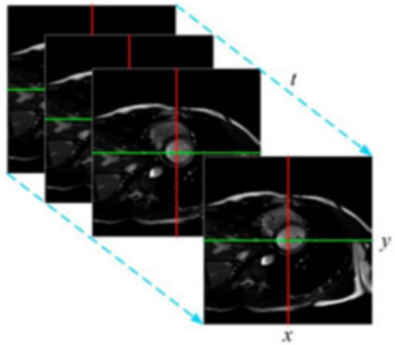
Reproducibility crisis

- Bayer Healthcare could replicate only 25% of 67 pre-clinical experiments [Prinz et al., 2011]
- Amgen could only confirm the findings in 6 out of 53 landmark cancer papers [Begley & Ellis, 2012]
- Social science papers in Science and Nature (2010 - 2015): only 13 out of 21 are consistent

https://www.bbc.com/news/science-environment-39054778

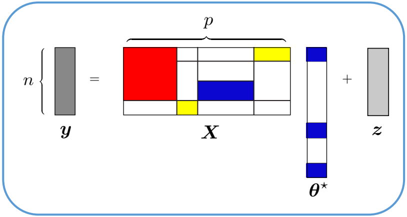
Challenges

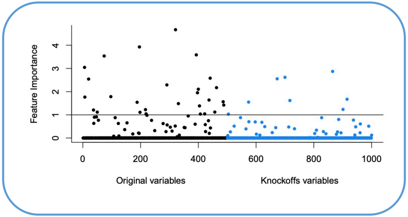




- data is of enormous dimension and dense (large n, large p)
- features can be highly correlated with each other
- signal-to-noise ratio can be small

This talk: two vignettes

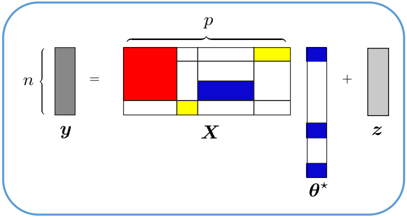


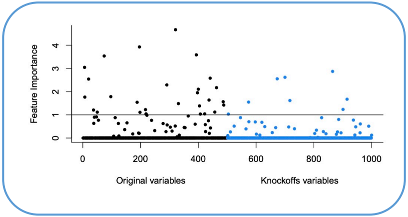


1. Lasso with general designs

   — trustworthy inference via precise distributional theory

This talk: two vignettes





1. Lasso with general designs
   — trustworthy inference via precise distributional theory
2. Derandomizing knockoffs
   — stabilizing variable selection in the knockoffs framework

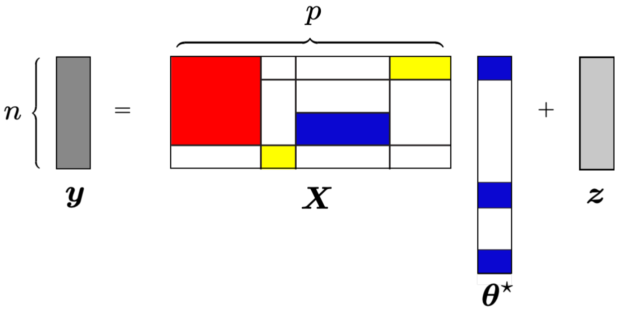
The first story: Lasso with general designs

Michael Celentano
Stanford Stat

Andrea Montanari
Stanford Stat & EE

"The Lasso with general Gaussian designs with application to hypothesis testing," M. Celentano, A. Montanari, Y. Wei, 2020. https://arxiv.org/abs/2007.13716

Lasso estimator



$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$ [Tibshirani, 1996]

Prior work: Lasso risk

Suppose $\boldsymbol{\theta}^*$ is s-sparse, $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Under restricted eigenvalue condition of design matrix $\boldsymbol{X}$,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq C\sigma\sqrt{\frac{s\log(p)}{n}}$$

[Bickel et al., 2009, Bühlmann and Van De Geer, 2011, Negahban et al., 2012, Zhao and Yu, 2006, Zhang and Zhang, 2014, Bellec et al., 2018]...

Prior work: Lasso risk

Suppose $\boldsymbol{\theta}^*$ is s-sparse, $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Under restricted eigenvalue condition of design matrix $\boldsymbol{X}$,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq C\sigma\sqrt{\frac{s\log(p)}{n}}$$

- unspecified constant

[Bickel et al., 2009, Bühlmann and Van De Geer, 2011, Negahban et al., 2012, Zhao and Yu, 2006, Zhang and Zhang, 2014, Bellec et al., 2018]...

Prior work: Lasso risk

Suppose $\boldsymbol{\theta}^*$ is s-sparse, $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Under restricted eigenvalue condition of design matrix $\boldsymbol{X}$,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq C\sigma\sqrt{\frac{s\log(p)}{n}}$$

- unspecified constant
- no distributional characterization of $\widehat{\boldsymbol{\theta}}$

[Bickel et al., 2009, Bühlmann and Van De Geer, 2011, Negahban et al., 2012, Zhao and Yu, 2006, Zhang and Zhang, 2014, Bellec et al., 2018]...

Prior work: Lasso risk

Suppose $\boldsymbol{\theta}^*$ is s-sparse, $z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. Under restricted eigenvalue condition of design matrix $\boldsymbol{X}$,

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq C\sigma\sqrt{\frac{s\log(p)}{n}}$$

- unspecified constant
- no distributional characterization of $\widehat{\boldsymbol{\theta}}$
- inadequate for statistical inference

[Bickel et al., 2009, Bühlmann and Van De Geer, 2011, Negahban et al., 2012, Zhao and Yu, 2006, Zhang and Zhang, 2014, Bellec et al., 2018]...

Exact aysmptotics under i.i.d designs

i.i.d. Gaussian design: $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$

- exact risk estimation
  [Bayati et al., 2013, Thrampoulidis et al., 2015]
- debiasing the lasso
  [Javanmard et al., 2018, Miolane and Montanari, 2018]
- precise FDP-TPP tradeoff for the Lasso
  [Su et al., 2017, Wang et al., 2020]
- exact distributional characterization
  [Miolane and Montanari, 2018]

Exact aysmptotics under i.i.d designs

i.i.d. Gaussian design: $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$

- exact risk estimation
  [Bayati et al., 2013, Thrampoulidis et al., 2015]
- debiasing the lasso
  [Javanmard et al., 2018, Miolane and Montanari, 2018]
- precise FDP-TPP tradeoff for the Lasso
  [Su et al., 2017, Wang et al., 2020]
- exact distributional characterization
  [Miolane and Montanari, 2018]

What happens with general Gaussian design $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$?

Exact aysmptotics under i.i.d designs

i.i.d. Gaussian design: $\boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{I}_p)$

- exact risk estimation
  [Bayati et al., 2013, Thrampoulidis et al., 2015]
- debiasing the lasso
  [Javanmard et al., 2018, Miolane and Montanari, 2018]
- precise FDP-TPP tradeoff for the Lasso
  [Su et al., 2017, Wang et al., 2020]
- exact distributional characterization
  [Miolane and Montanari, 2018]

What happens with general Gaussian design $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$?

— **difficulty:** non-isometry of $\|\cdot\|_1$ penalty.

This talk

$$\boldsymbol{y} = \boldsymbol{X}\theta^* + \boldsymbol{z} \ \in \mathbb{R}^n$$

This talk

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{z} \ \in \mathbb{R}^n$$

- $\boldsymbol{\theta}^* \in \mathbb{R}^p$: s-sparse

This talk

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{z} \ \in \mathbb{R}^n$$

- $\boldsymbol{\theta}^* \in \mathbb{R}^p$: s-sparse
- proportional regime: $p/n = \text{const}$, $s/p = \text{const}$

This talk

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{z} \ \in \mathbb{R}^n$$

- $\boldsymbol{\theta}^* \in \mathbb{R}^p$: s-sparse
- proportional regime: $p/n = \text{const}$, $s/p = \text{const}$
- Gaussian noise: $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$; Gaussian design: $\boldsymbol{x}_i \sim \mathcal{N}(0, \underbrace{\boldsymbol{\Sigma}}_{\text{known}})$

This talk

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{z} \ \in \mathbb{R}^n$$

- $\boldsymbol{\theta}^* \in \mathbb{R}^p$: s-sparse
- proportional regime: $p/n = \text{const}$, $s/p = \text{const}$
- Gaussian noise: $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$; Gaussian design: $\boldsymbol{x}_i \sim \mathcal{N}(0, \underbrace{\boldsymbol{\Sigma}}_{\text{known}})$

Goal: a distributional theory for general Gaussian design

Key observation

original model
$\widehat{\boldsymbol{\theta}}$

- **original model:** $\boldsymbol{y} = \boldsymbol{X}\theta + \boldsymbol{z}$

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

Key observation

original model
$\widehat{\boldsymbol{\theta}}$

fixed design model
$\widehat{\boldsymbol{\theta}}^f$

- **original model:** $\boldsymbol{y} = \boldsymbol{X}\theta + \boldsymbol{z}$

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

- **fixed design model:** $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}$, $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_p)$

$$\widehat{\boldsymbol{\theta}}^f := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\zeta^*}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

Key observation

original model
$\widehat{\boldsymbol{\theta}}$

fixed design model
$\widehat{\boldsymbol{\theta}}^f$

- **original model:** $\boldsymbol{y} = \boldsymbol{X}\theta + \boldsymbol{z}$

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

- **fixed design model:** $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}$, $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_p)$

$$\widehat{\boldsymbol{\theta}}^f := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\zeta^*}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

τ^*: effective risk level; ζ^*: effective non-sparsity

Key observation

original model $\widehat{\boldsymbol{\theta}}$ $\overset{\text{distribution}}{\approx}$ fixed design model $\widehat{\boldsymbol{\theta}}^f$

- **original model:** $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{z}$

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

- **fixed design model:** $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}$, $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_p)$

$$\widehat{\boldsymbol{\theta}}^f := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\zeta^*}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

τ^*: effective risk level; ζ^*: effective non-sparsity

Fixed point equations

$$(\tau^*, \zeta^*) \quad \xrightarrow{\text{solution}} \quad \begin{aligned} \tau^2 &= \sigma^2 + \mathsf{R}(\tau^2, \zeta) \\ \zeta &= 1 - \mathsf{df}(\tau^2, \zeta) \end{aligned}$$

Fixed point equations

$$(\tau^*, \zeta^*) \quad \xrightarrow{\text{solution}} \quad \begin{aligned} \tau^2 &= \sigma^2 + \mathsf{R}(\tau^2, \zeta) \\ \zeta &= 1 - \mathsf{df}(\tau^2, \zeta) \end{aligned}$$

$$\mathsf{R}(\tau^2, \zeta) := \underbrace{\frac{1}{n}\,\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}^f(\tau, \zeta) - \boldsymbol{\theta}^*)\right\|_2^2\right]}_{\text{in-sample prediction risk}}$$

$$\mathsf{df}(\tau^2, \zeta) := \underbrace{\frac{1}{n}\,\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}^f(\tau, \zeta)\right\|_0\right]}_{\text{degrees of freedom}}$$

Fixed point equations

$$
\color{red}{(\tau^*, \zeta^*)} \quad \xrightarrow{\text{solution}} \quad \begin{aligned} \tau^2 &= \sigma^2 + \mathsf{R}(\tau^2, \zeta) \\ \zeta &= 1 - \mathsf{df}(\tau^2, \zeta) \end{aligned}
$$

$$
\mathsf{R}(\tau^2, \zeta) := \underbrace{\frac{1}{n}\mathbb{E}\left[\left\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}}^f(\tau, \zeta) - \boldsymbol{\theta}^*)\right\|_2^2\right]}_{\text{in-sample prediction risk}}
$$

$$
\mathsf{df}(\tau^2, \zeta) := \underbrace{\frac{1}{n}\mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}}^f(\tau, \zeta)\right\|_0\right]}_{\text{degrees of freedom}}
$$

Property: solution is unique and bounded for reasonably sparse $\boldsymbol{\theta}^*$.

Main result: Lasso distribution

Theorem (Celetano, Montanari, Wei '20)

When $\boldsymbol{\theta}^*$ is sparse enough, for any 1-Lipschitz function ϕ and $\epsilon > 0$

$$\forall \lambda \in [\lambda_{\min}, \lambda_{\max}], \qquad \left| \phi\Big(\frac{\widehat{\boldsymbol{\theta}}_\lambda}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big) - \mathbb{E}\Big[\phi\Big(\frac{\widehat{\boldsymbol{\theta}}_\lambda^f}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}}\Big)\Big]\right| \le \epsilon,$$

with probability at least $1 - \frac{C}{\epsilon^4} e^{-cn\epsilon^4}$.

Main result: Lasso distribution

Theorem (Celetano, Montanari, Wei '20)

When $\boldsymbol{\theta}^*$ is sparse enough, for any 1-Lipschitz function ϕ and $\epsilon > 0$

$$\forall \lambda \in [\lambda_{\min}, \lambda_{\max}], \qquad \left| \phi\left(\frac{\widehat{\boldsymbol{\theta}}_\lambda}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}} \right) - \mathbb{E}\left[\phi\left(\frac{\widehat{\boldsymbol{\theta}}^f_\lambda}{\sqrt{p}}, \frac{\boldsymbol{\theta}^*}{\sqrt{p}} \right) \right] \right| \le \epsilon,$$

with probability at least $1 - \frac{C}{\epsilon^4} e^{-cn\epsilon^4}$.

A direct consequence:

$$\forall \lambda \in [\lambda_{\min}, \lambda_{\max}], \qquad \|\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\|_2 \approx \mathbb{E}\left[\|\widehat{\boldsymbol{\theta}}^f_\lambda - \boldsymbol{\theta}^*\|_2 \right]$$

Main result: properties for Lasso

- Lasso residual

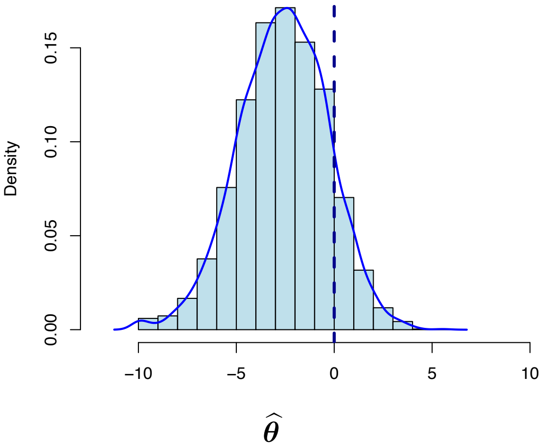
$$\mathbb{P}\left(\left|\frac{\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\theta}}\|_2}{\sqrt{n}}-\tau^*\zeta^*\right|>\epsilon\right)\leq\frac{C}{\epsilon^2}e^{-cn\epsilon^4}\,.$$

- Lasso sparsity

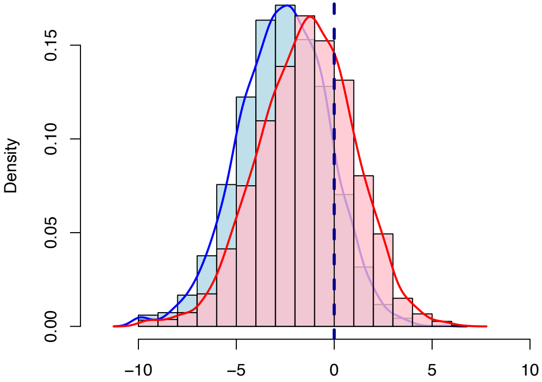
$$\mathbb{P}\left(\left|\frac{\|\widehat{\boldsymbol{\theta}}\|_0}{n}-(1-\zeta^*)\right|>\epsilon\right)\leq\frac{C}{\epsilon^3}e^{-cn\epsilon^6}\,.$$

Statistical inference: debiasing Lasso

Debiased Lasso for statistical inference



Debiased Lasso for statistical inference

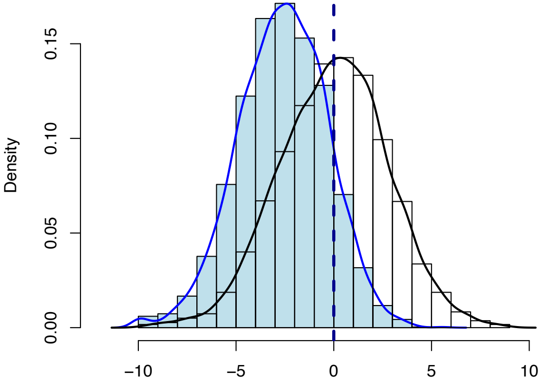


$$\widehat{\boldsymbol{\theta}}^{d} = \widehat{\boldsymbol{\theta}} + \boldsymbol{M}\boldsymbol{X}^{\top}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})$$

$\boldsymbol{M}$: surrogate for $\boldsymbol{\Sigma}^{-1} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^{\top}]^{-1}$

[Zhang and Zhang, 2014, Van de Geer et al., 2014, Javanmard and Montanari, 2014a, Javanmard and Montanari, 2014b]

Debiased Lasso for statistical inference



$$\widehat{\boldsymbol{\theta}}^d = \widehat{\boldsymbol{\theta}} + \boldsymbol{M}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})$$

$\boldsymbol{M}$: scaled version of $\boldsymbol{\Sigma}^{-1} = \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top]^{-1}$

[Javanmard et al., 2018, Miolane and Montanari, 2018, Bellec and Zhang, 2019a, Bellec and Zhang, 2019b]

Debiased Lasso

- classical debiased Lasso

$$\widehat{\boldsymbol{\theta}}_0^{\mathrm{d}} = \widehat{\boldsymbol{\theta}} + \boldsymbol{M}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}), \qquad \boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$$

Debiased Lasso

- classical debiased Lasso

$$\widehat{\boldsymbol{\theta}}_0^{\mathrm{d}} = \widehat{\boldsymbol{\theta}} + \boldsymbol{M}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}), \qquad \boldsymbol{M} = \boldsymbol{\Sigma}^{-1}$$

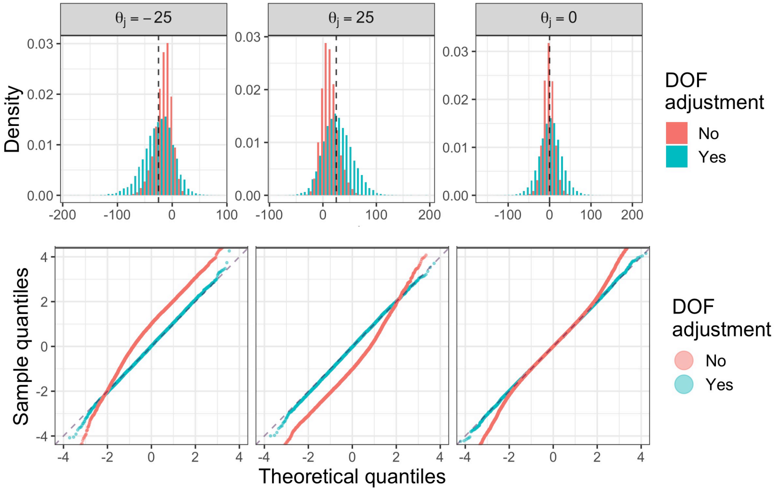
- debiased Lasso with degrees-of-freedom (DOF) adjustment

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} := \widehat{\boldsymbol{\theta}} + \boldsymbol{M}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}}), \qquad \boldsymbol{M} = \frac{\boldsymbol{\Sigma}^{-1}}{1 - \|\widehat{\boldsymbol{\theta}}\|_0 / n}$$

[Javanmard and Montanari, 2014b, Miolane and Montanari, 2018, Bellec and Zhang, 2019a, Bellec and Zhang, 2019b]

Main result: $\widehat{\boldsymbol{\theta}}^{\mathrm{d}}$ behaves like $\boldsymbol{\theta}^* + \tau^* \boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}$

Debiased Lasso with DOF adjustment



Here $p = 100$, $n = 25$, $s = 20$, $\Sigma_{ij} = 0.5^{|i-j|}$, $\sigma = 1$

DOF adjustment is successful

Theorem (Celetano, Montanari, Wei '20)

When $\boldsymbol{\theta}^*$ is sparse enough, false coverage proportion satisfies

$$\mathbb{P}\left(|\mathsf{FCP} - q| > \epsilon\right) \leq C(\epsilon) e^{-c(\epsilon) n}.$$

$$\mathsf{FCP} := \frac{1}{p} \sum_{j=1}^{p} \mathbb{1}\left\{ |\widehat{\boldsymbol{\theta}}_j^d - \boldsymbol{\theta}_j^*| > \Sigma_{j|-j}^{-1/2} \widehat{\tau} \cdot z_{1-q/2} \right\}$$

DOF adjustment is successful

Theorem (Celetano, Montanari, Wei '20)

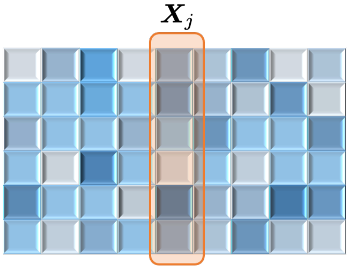
When $\boldsymbol{\theta}^*$ is sparse enough, false coverage proportion satisfies

$$\mathbb{P}\left(|\mathsf{FCP} - q| > \epsilon\right) \le C(\epsilon) e^{-c(\epsilon) n}.$$

$$\mathsf{FCP} := \frac{1}{p} \sum_{j=1}^{p} \mathbb{1}\left\{ |\widehat{\boldsymbol{\theta}}_j^d - \boldsymbol{\theta}_j^*| > \Sigma_{j|-j}^{-1/2} \widehat{\tau} \cdot z_{1-q/2} \right\}$$

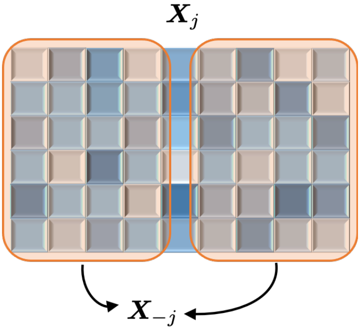
— coverage only in the average sense!

Confidence interval for a single coordinate



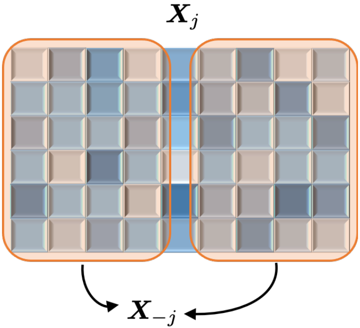
- regress $\boldsymbol{X}_j$ on $\boldsymbol{X}_{-j}$

Confidence interval for a single coordinate



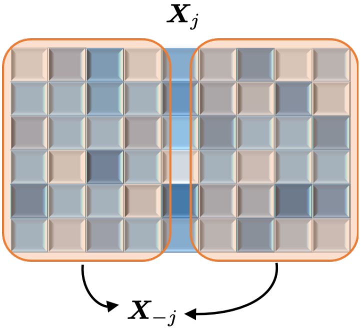
- regress $\boldsymbol{X}_j$ on $\boldsymbol{X}_{-j}$

Confidence interval for a single coordinate



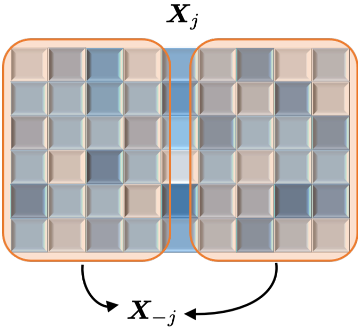
- regress $\boldsymbol{X}_j$ on $\boldsymbol{X}_{-j}$ $\quad\longrightarrow\quad$ residual $\boldsymbol{X}_j^{\perp}$

Confidence interval for a single coordinate



- regress $\boldsymbol{X}_j$ on $\boldsymbol{X}_{-j}$ $\quad\longrightarrow\quad$ residual $\boldsymbol{X}_j^{\perp}$
- obtain leave-j^{th}-coordinate-out Lasso $\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}$

Confidence interval for a single coordinate



- regress $\boldsymbol{X}_j$ on $\boldsymbol{X}_{-j}$ $\longrightarrow$ residual $\boldsymbol{X}_j^{\perp}$
- obtain leave-j^{th}-coordinate-out Lasso $\widehat{\boldsymbol{\theta}}_{\text{loo}}$
- construct confidence interval

$$\mathsf{CI}_j^{\text{loo}} := \left[\xi_j \ \pm \ \widehat{\mathsf{sd}} \cdot z_{1-\alpha/2}\right]$$

$$\xi_j = \text{correlation between } \boldsymbol{X}_j^{\perp} \text{ and } \boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\text{loo}}$$

Coverage and power

Theorem (Celetano, Montanari, Wei '20)

There exist constants $C, c, c' > 0$ *such that for all* $\epsilon < c'$,

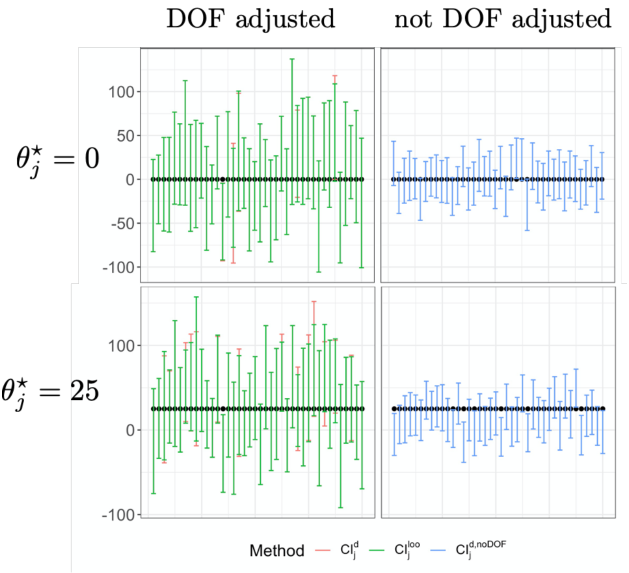
$$\left|\mathbb{P}_{\theta_j^*}\left(\theta \notin \mathrm{CI}_j^{\mathrm{loo}}\right) - \mathbb{P}_{\theta_j^*}\left(|\theta_j^* + \tau_{\mathrm{loo}}^* G - \theta| > \tau_{\mathrm{loo}}^* z_{1-\alpha/2}\right)\right| \leq C\left((1+|\theta_j^*|)\epsilon + \frac{1}{\epsilon^3} e^{-cn\epsilon^6} + \frac{1}{n\epsilon^2}\right),$$

where $G \sim \mathrm{N}(0,1)$.

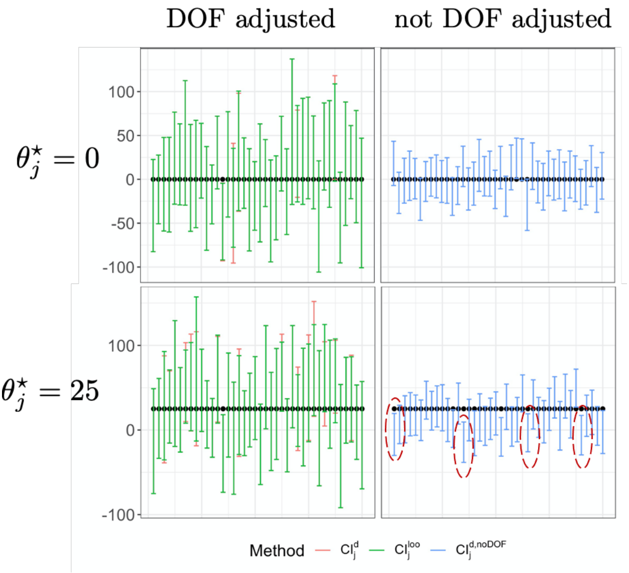
$$\mathrm{CI}_j^{\mathrm{loo}} := \left[\xi_j \pm \widehat{\mathrm{sd}} \cdot z_{1-\alpha/2}\right]$$

$$\xi_j = \text{correlation between } \boldsymbol{X}_j^{\perp} \text{ and } \boldsymbol{y} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\theta}}_{\mathrm{loo}}$$

Confidence interval for a single coordinate



Confidence interval for a single coordinate



Summary of this part

- distributional theory of Lasso/debiased Lasso for general designs
- provide confidence intervals for single coordinates with error control

"The Lasso with general Gaussian designs with application to hypothesis testing,"

M. Celentano, A. Montanari, Y. Wei, 2020. https://arxiv.org/abs/2007.13716

The second story: derandomizing knockoffs

Zhimei Ren
Stanford Stat

Emmanuel Candès
Stanford Stat & Math

"Derandomizing Knockoffs," Zhimei Ren, Yuting Wei, and Emmanuel Candès, in preparation, 2020

Stability

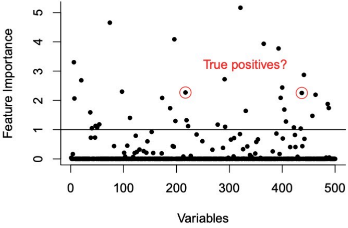
BIN YU

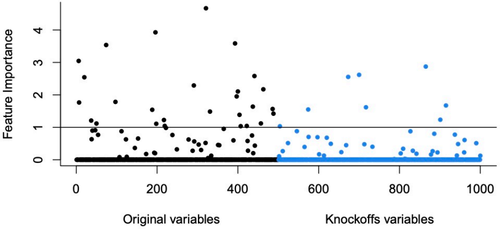
Departments of Statistics and EECS, University of California at Berkeley, Berkeley, CA 94720, USA.
E-mail: binyu@stat.berkeley.edu

Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in stability of statistical results relative to "reasonable" perturbations to data and to the model used. Jacknife, bootstrap, and cross-validation are based on perturbations to data, while robust statistics methods deal with perturbations to models.

Knockoffs framework

— [Barber et al., 2015, Candès et al., 2018]



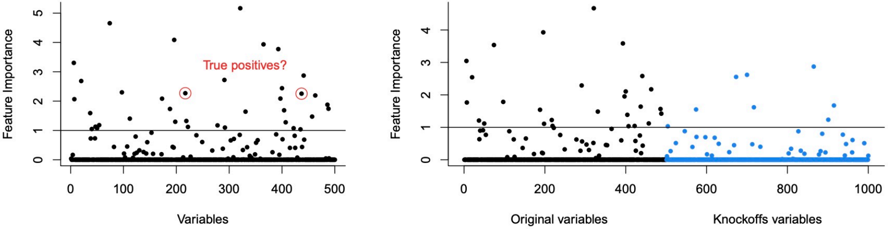


Three-step procedure:

- construct knockoff feature matrix $\widetilde{X} \in \mathbb{R}^{n \times p}$
- define feature statistics $w_j([X, \widetilde{X}, y])$ for each $j \in \{1, 2, \ldots, 2p\}$
- decide selection set $\widehat{S}$

Knockoffs framework

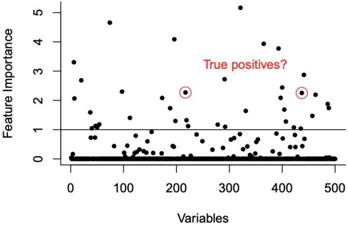
— [Barber et al., 2015, Candès et al., 2018]

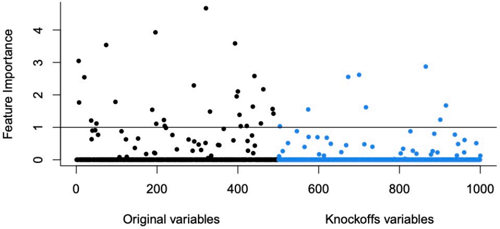


different runs ⇒ different selection sets

Knockoffs framework

— [Barber et al., 2015, Candès et al., 2018]





different runs ⇒ different selection sets

Stability

Stablity selection

Stability selection

N Meinshausen, P Bühlmann - Journal of the Royal Statistical …, 2010 - Wiley Online Library

Estimation of structure, such as in variable selection, graphical modelling or cluster analysis, is notoriously difficult, especially for high dimensional data. We introduce stability selection. It is based on subsampling in combination with (high dimensional) selection algorithms. As …

☆ 99 Cited by 2038 Related articles All 27 versions

Variable **selection** with error control: another look at **stability selection**

RD Shah, RJ Samworth - … of the Royal Statistical Society: Series …, 2013 - Wiley Online Library

Stability selection was recently introduced by Meinshausen and Bühlmann as a very general technique designed to improve the performance of a variable **selection** algorithm. It is based on aggregating the results of applying a **selection** procedure to subsamples of the data. We …

☆ 99 Cited by 246 Related articles All 20 versions

Stablity selection (original form)

1. start with the full dataset $\boldsymbol{Z}_{\text{full}} = Z_1, \ldots, Z_n$

Stablity selection (original form)

1. start with the full dataset $\boldsymbol{Z}_{\text{full}} = Z_1, \ldots, Z_n$
2. for each $m = 1, \ldots, M$
   (i) subsample without replacement to generate a smaller dataset of size $n/2$, denoted by $\boldsymbol{Z}_{(m)}$
   (ii) run the selection algorithm on $\boldsymbol{Z}_{(m)}$ to obtain a selection set $\widehat{S}^m$

Stablity selection (original form)

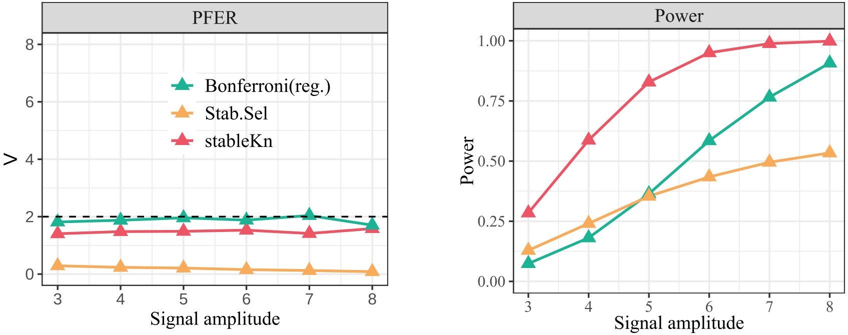
1. start with the full dataset $\boldsymbol{Z}_{\text{full}} = Z_1, \ldots, Z_n$
2. for each $m = 1, \ldots, M$
   (i) subsample without replacement to generate a smaller dataset of size $n/2$, denoted by $\boldsymbol{Z}_{(m)}$
   (ii) run the selection algorithm on $\boldsymbol{Z}_{(m)}$ to obtain a selection set $\widehat{S}^m$
3. calculate the selection fequency $\Pi_j = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{j \in \widehat{S}^m\}$

Stablity selection (original form)

1. start with the full dataset $\boldsymbol{Z}_{\text{full}} = Z_1, \ldots, Z_n$
2. for each $m = 1, \ldots, M$
   (i) subsample without replacement to generate a smaller dataset of size $n/2$, denoted by $\boldsymbol{Z}_{(m)}$
   (ii) run the selection algorithm on $\boldsymbol{Z}_{(m)}$ to obtain a selection set $\widehat{S}^m$
3. calculate the selection fequency $\Pi_j = \dfrac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{j \in \widehat{S}^m\}$
4. given a threshold $\eta > 0$, return the final selection set

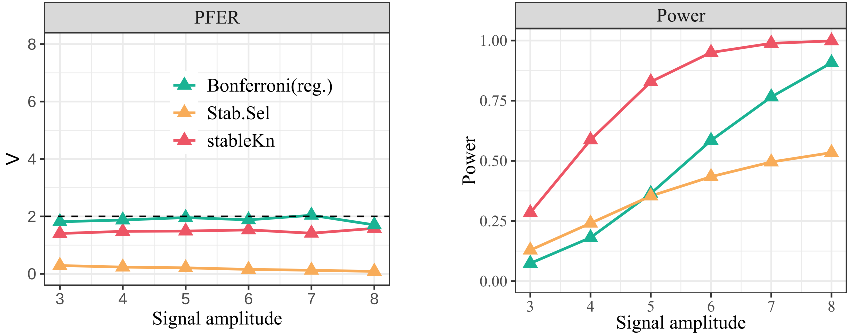
$$\widehat{S} = \{j \in [p] : \Pi_j \geq \eta\}.$$

In the large p regime?



Settings: $n = 2000$, $p = 1000$ and $\Sigma_{ij} = 0.5^{|i-j|}$. $Y \mid X \sim$ a linear model with 60 non-zero coefficients.

In the large p regime?



Settings: $n = 2000$, $p = 1000$ and $\Sigma_{ij} = 0.5^{|i-j|}$. $Y \mid X \sim$ a linear model with 60 non-zero coefficients.

subsampling leads to loss of power

stability selection

knockoffs

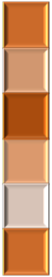
This work: derandomizing knockoffs

- Stability
- Statistical guarantees
- Improved power

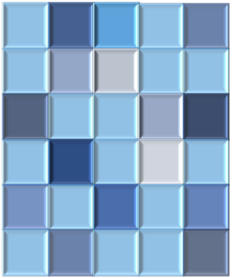
A brief review of the knockoffs framework

Step 1: construct knockoffs

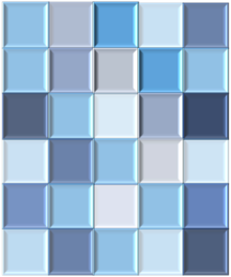
response Y



feature matrix X

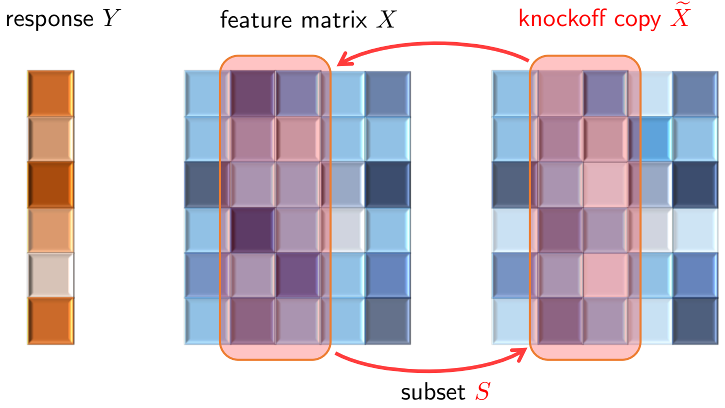


knockoff copy $\widetilde{X}$



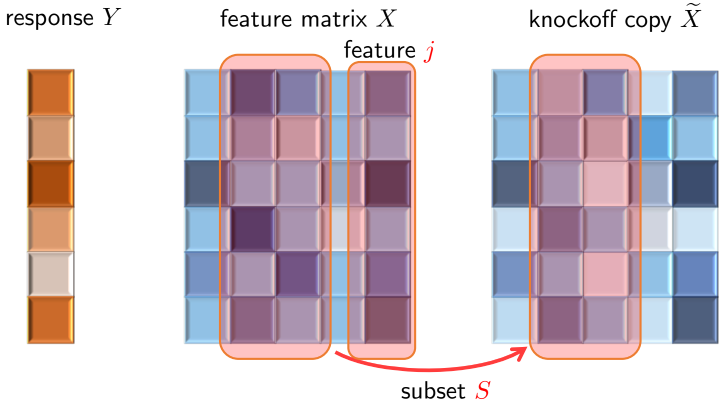
- $\widetilde{X} \perp\!\!\!\perp Y \mid X$

Step 1: construct knockoffs



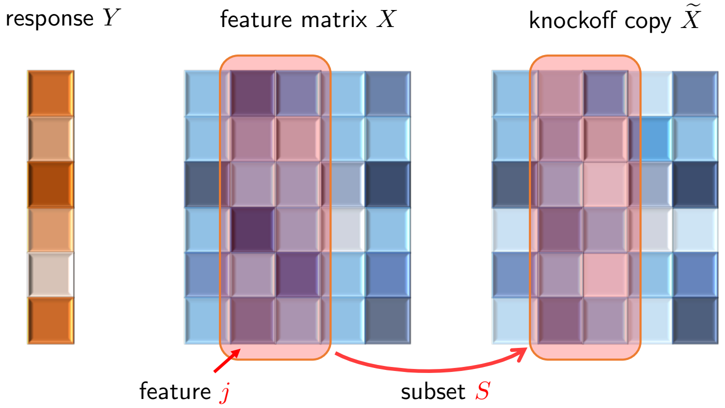
- $\widetilde{X} \perp\!\!\!\perp Y \mid X$
- for any subset $S \subset \{1, 2, \ldots p\}$: distribution $(X, \widetilde{X})_{\text{swap}(S)} \stackrel{\text{d}}{=} (X, \widetilde{X})$

Step 2: define feature statistics $w_j([X, \widetilde{X}], y)$



$$w_j([X, \widetilde{X}]_{\mathrm{swap}(S)}, y) = w_j([X, \widetilde{X}], y) \qquad j \notin S$$

Step 2: define feature statistics $w_j([X, \widetilde{X}], y)$



$$w_j([X, \widetilde{X}]_{\mathrm{swap}(S)}, y) = -w_j([X, \widetilde{X}], y) \qquad j \in S$$

Step 3: determine selection set

Model-X v-knockoff [Janson et al., 2016]

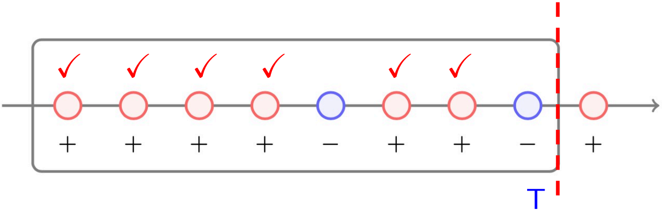
- order the features according to the magnitudes of W_j's:

$$|W_{\pi_1}| \geq |W_{\pi_2}| \geq \dots |W_{\pi_p}|$$

- reject π_j such that $j \leq T$ and $W_{\pi_j} > 0$

$$T := \inf_{k \in [p]} \left\{ \sum_{j=1}^{k} \mathbf{1}_{\{W_{\pi_j} < 0\}} \geq v \right\}$$

- if $v = 2$, stop the procedure the first time seeing 2 "$-$"s.



This work: derandomizing knockoffs

- given (X, Y), generate $m = 1, \ldots, M$ realizations of knockoffs

This work: derandomizing knockoffs

- given (X, Y), generate $m = 1, \ldots, M$ realizations of knockoffs
- for each realization of knockoff m:

$$X \longrightarrow \widetilde{X}^m \xrightarrow{Y} W^m \xrightarrow{\text{base procedure}} \hat{S}_m$$

This work: derandomizing knockoffs

- given (X, Y), generate $m = 1, \ldots, M$ realizations of knockoffs
- for each realization of knockoff m:

$$X \longrightarrow \widetilde{X}^m \xrightarrow{Y} W^m \xrightarrow{\text{base procedure}} \hat{S}_m$$

- for each feature j, define selection probability

$$\Pi_j := \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}(j \in \hat{S}_m)$$

- for a threshold η, the final selection set S is

$$\hat{S} := \{j \in [p] : \Pi_j \geq \eta\}.$$

Theoretical guarantees

Theorem (Ren, Wei, Candès 20)

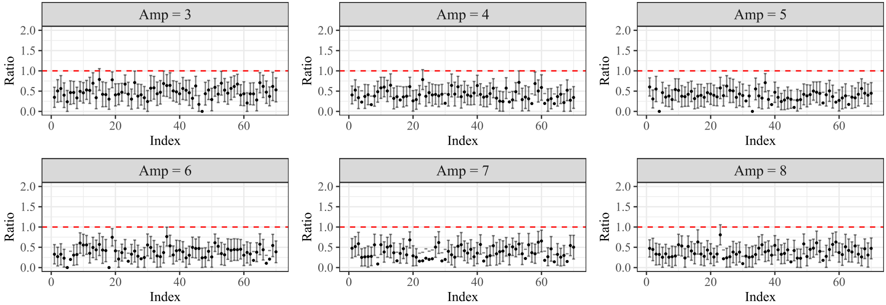
If for every $j \in \mathcal{H}_0$, the condition

$$\mathbb{P}(\Pi_j \geq 1/2) \; \leq \; \gamma \mathbb{E}[\Pi_j] \tag{1}$$

holds, then the PFER can be controlled as

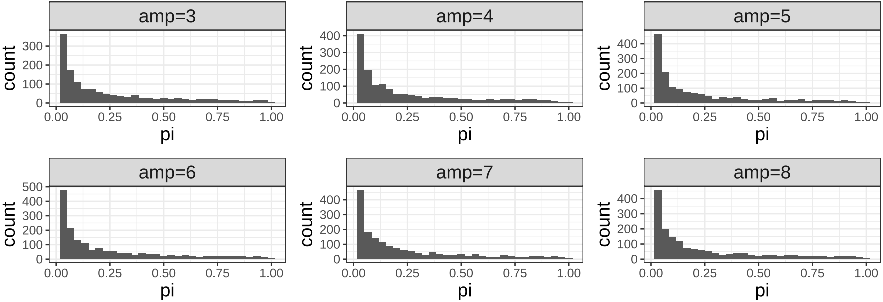
$$\mathbb{E}[V] \; \leq \; \gamma v.$$

- Per family error rate (PFER): $\mathbb{E}[V]$ (V : number of false discoveries)
- Markov's inequality gives $\gamma = 2$



Realized ratio of $\mathbb{P}(\Pi_j \geq 1/2)/\mathbb{E}[\Pi_j]$ with the 95% confidence interval, estimated from 1,000 repetitions.

How to tighten γ ? An observation...

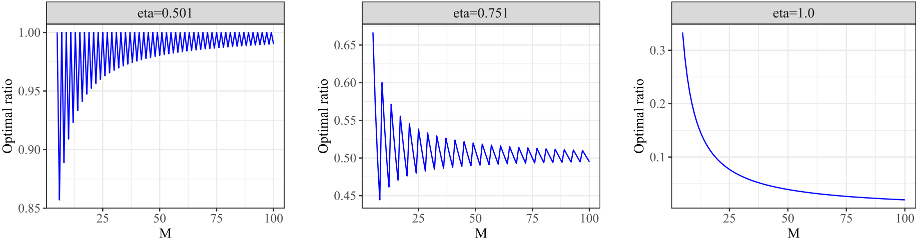


Pooled histogram of all nonzero null Π_j's.

A sharper guarantee

- If the pmf of Π_j is monotonically non-increasing for each $j \in \mathcal{H}_0$

$$\gamma = \max \sum_{m \geq M\eta} y_m,$$
$$s.t. \quad y_m \geq 0, \quad y_{m-1} \geq y_m, \; m \in [M],$$
$$\sum_{m=0}^{M} y_m \cdot \frac{m}{M} = 1.$$



Theoretical guarantees

Theorem (Ren, Wei, Candès 20)

Suppose condition (1) *holds with* $\gamma = 1$ *and the pmf of* V *is monotonically non-increasing, then the* k-FWER *can be controlled as*

$$\mathbb{P}(V \geq k) \;\leq\; \min\left\{\frac{v}{2k}, \frac{\mathbb{E}[(2Z)^{\alpha}]}{2k^{\alpha}}, \frac{\mathbb{E}[\exp(\lambda(2Z))]}{2\exp(\lambda k)}\right\}$$

- k family-wise error rate (k-FWER): $\mathbb{P}(V \geq k)$
- $Z \sim \mathrm{NB}(m, q)$ negative binomial random variable

Theoretical guarantees

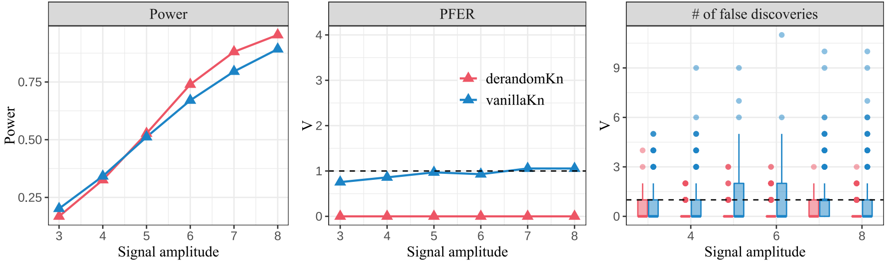
Theorem (Ren, Wei, Candès 20)

Suppose condition (1) *holds with* $\gamma = 1$ *and the pmf of* V *is monotonically non-increasing, then the* k-FWER *can be controlled as*

$$\mathbb{P}(V \geq k) \;\leq\; \min\left\{\frac{v}{2k},\; \frac{\mathbb{E}[(2Z)^{\alpha}]}{2k^{\alpha}},\; \frac{\mathbb{E}[\exp(\lambda(2Z))]}{2\exp(\lambda k)}\right\}$$

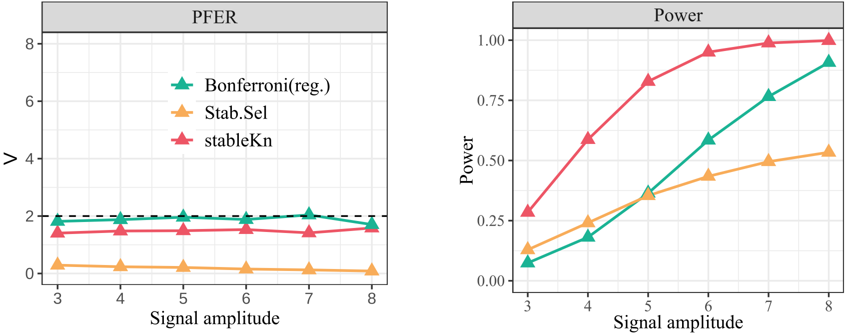
- k family-wise error rate (k-FWER): $\mathbb{P}(V \geq k)$
- $Z \sim \mathrm{NB}(m, q)$ negative binomial random variable
- minimum is also taken over α, λ
- "monotonically non-increasing" condition can be relaxed

Simulation studies: PFER control



Settings: $n = 200$, $p = 100$, $X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\Sigma_{ij} = 0.2^{|i-j|}$, and $Y \mid X \sim$ a linear model with 30 non-zero coefficients. Each nonzero coefficient β_j takes value $A/\sqrt{n}$ where A ranges in $\{3, 4, \ldots, 8\}$ and the sign is determined by i.i.d. coin flips. The locations of the non-zero signal are randomly chosen from $[p]$. We show the averaged results over 200 trials.

Simulation studies: more comparisons

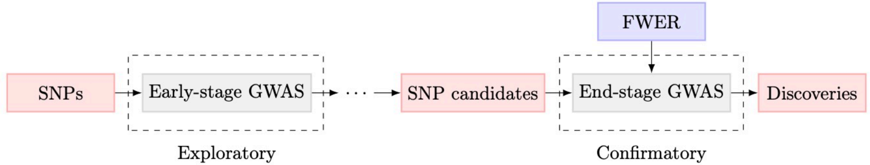


Settings: $n = 2000$, $p = 1000$ and $\Sigma_{ij} = 0.5^{|i-j|}$. $Y \mid X \sim$ a linear model with 60 non-zero coefficients.

A real data example

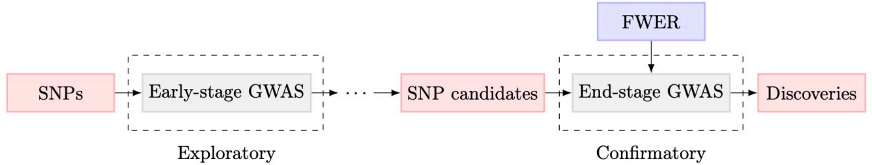
Genome-Wide Association Study (GWAS)

A typical workflow of multi-stage GWAS:



Genome-Wide Association Study (GWAS)

A typical workflow of multi-stage GWAS:



Conditional knockoffs:

- suppose a subset of candidate SNPs $\mathcal{C}$ is selected in stage one
- construct a conditional knockoff copy *only* for $X_{\mathcal{C}}$

$$(X_{\mathcal{C}}, \widetilde{X}_{\mathcal{C}})_{\mathrm{swap}(g)} \mid X_{-\mathcal{C}} \overset{\mathrm{d}}{=} (X_{\mathcal{C}}, \widetilde{X}_{\mathcal{C}}) \mid X_{-\mathcal{C}}$$

Procedures

- data: The UK biobank dataset 161k unrelated British male individuals and their disease status (prostate cancer)

Procedures

- data: The UK biobank dataset 161k unrelated British male individuals and their disease status (prostate cancer)
- early-stage: selecting p-values from [Schumacher et al., 2018] below 10^{-3} gives 4072 pre-selected SNPs

Procedures

- data: The UK biobank dataset 161k unrelated British male individuals and their disease status (prostate cancer)
- early-stage: selecting p-values from [Schumacher et al., 2018] below 10^{-3} gives 4072 pre-selected SNPs
- partition the SNPs into clusters at a level of resolution 2% and the resulting average length of the clusters is 0.226 Mb.

Procedures

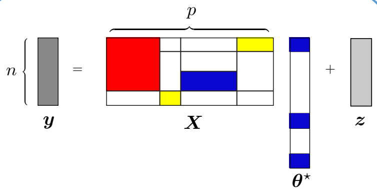
- data: The UK biobank dataset 161k unrelated British male individuals and their disease status (prostate cancer)
- early-stage: selecting p-values from [Schumacher et al., 2018] below 10^{-3} gives 4072 pre-selected SNPs
- partition the SNPs into clusters at a level of resolution 2% and the resulting average length of the clusters is 0.226 Mb.
- apply derandomized knockoffs with target FWER level 0.1 (ten runs of conditional group HMM knockoffs)

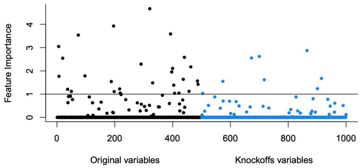
Results

Lead SNP	Chromosome	Position range (Mb)	Size	Confirmed by?
rs12621278	2	173.28-173.58	68	[Wang et al., 2015]
rs1512268	8	23.39-23.55	48	[Wang et al., 2015]
rs1016343	8	128.07-128.24	45	[Hui et al., 2014]
rs6983267	8	128.40-128.47	37	[Wang et al., 2015]
rs7121039	11	2.18-2.31	40	[Wang et al., 2015]*
rs10896449	11	68.80-69.02	62	[Wang et al., 2015]
rs7501939	17	36.05-36.18	55	[Elliott et al., 2010]
rs1859962	17	69.07-69.24	40	[Wang et al., 2015]

Discoveries at 2% resolution and the target FWER level set to 0.1 and $\eta = 1$ and $M = 10$.

Concluding remarks

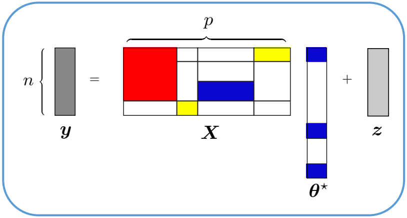


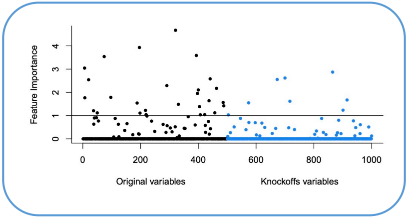


Future directions:

- unknown covariance structure
- distributional theory beyond Gaussian design
- power analysis
- more liberal criteria: FDR, FDX

Concluding remarks





Future directions:

- unknown covariance structure
- distributional theory beyond Gaussian design
- power analysis
- more liberal criteria: FDR, FDX

Thanks for your attention!

Other technical details

Intuition for DOF adjustment

- **original model:** $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{z}$

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

- **fixed design model:** $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}$

$$\widehat{\boldsymbol{\theta}}^f := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\zeta^*}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

Intuition for DOF adjustment

- **original model:** $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{z}$

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

- **fixed design model:** $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}$

$$\widehat{\boldsymbol{\theta}}^f := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\zeta^*}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} := \widehat{\boldsymbol{\theta}} + \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})}{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n}$$

$\widehat{\boldsymbol{\theta}}^f$

Intuition for DOF adjustment

- **original model:** $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{z}$

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

- **fixed design model:** $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}$

$$\widehat{\boldsymbol{\theta}}^f := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\zeta^*}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} := \underbrace{\widehat{\boldsymbol{\theta}}}_{\widehat{\boldsymbol{\theta}}^f} + \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})}{\underbrace{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n}_{\zeta^*}}$$

Intuition for DOF adjustment

- **original model:** $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{z}$

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

- **fixed design model:** $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}$

$$\widehat{\boldsymbol{\theta}}^f := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\zeta^*}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

$$\boldsymbol{\Sigma}^{-1} \cdot \zeta^*\boldsymbol{\Sigma}^{1/2}(\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}^f) = \zeta^*(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}^f - \widehat{\boldsymbol{\theta}}^f)$$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} := \widehat{\boldsymbol{\theta}} + \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})}{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n}$$

$\widehat{\boldsymbol{\theta}}^f$ $\qquad$ ζ^*

Intuition for DOF adjustment

- **original model:** $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{z}$

$$\widehat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

- **fixed design model:** $\boldsymbol{y}^f = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}^* + \tau^*\boldsymbol{g}$

$$\widehat{\boldsymbol{\theta}}^f := \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{\zeta^*}{2}\|\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1 \right\}$$

$$\boldsymbol{\Sigma}^{-1}\cdot\zeta^*\boldsymbol{\Sigma}^{1/2}(\boldsymbol{y}^f - \boldsymbol{\Sigma}^{1/2}\widehat{\boldsymbol{\theta}}^f) = \zeta^*(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}^f - \widehat{\boldsymbol{\theta}}^f)$$

$$\widehat{\boldsymbol{\theta}}^{\mathrm{d}} := \underbrace{\widehat{\boldsymbol{\theta}}}_{\widehat{\boldsymbol{\theta}}^f} + \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}})}{\underbrace{1 - \|\widehat{\boldsymbol{\theta}}\|_0/n}_{\zeta^*}} \approx \boldsymbol{\theta}^* + \tau^*\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g}$$

A simple example

Suppose $X \sim \mathcal{N}(0, \Sigma)$, how to construct $\widetilde{X}$?

$$(X, \widetilde{X}) \sim \mathcal{N}(0, G) \quad \text{where} \quad G = \begin{bmatrix} \Sigma & \Sigma - \operatorname{diag}(s) \\ \Sigma - \operatorname{diag}(s) & \Sigma \end{bmatrix}.$$

$$\widetilde{X} \mid X \sim \mathcal{N}(\mu, V)$$

where

$$\begin{aligned} \mu &= X - X\Sigma^{-1}\operatorname{diag}(s) \\ V &= 2\operatorname{diag}(s) - \operatorname{diag}(s)\Sigma^{-1}\operatorname{diag}(s) \end{aligned}$$

A simple example: Lasso coefficient difference

Run Lasso

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{2p}} \quad \frac{1}{2}\|y - [X, \widetilde{X}]\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Lasso coefficient difference statistics (LCD):

$$W_j = |\hat{\boldsymbol{\beta}}_j(\lambda)| - |\hat{\boldsymbol{\beta}}_{j+p}(\lambda)|$$

- null W_j's are symmetrically distributed
- conditional on $|W_j|$, signs of null W_j's are i.i.d. coin flips