# A practical evaluation of recent methods in high-dimensional inference

Charles Zheng

Stanford University

May 3, 2015

# Theory and Practice

| Theory | Practice |
|---|---|
| $Y \sim N(\beta' X, \sigma^2 I)$ | $Y, X_1, \ldots, X_p$ *unknown* relationship |

| Theory | Practice |
| --- | --- |
| $Y \sim \overline{N(\beta' X, \sigma^2 I)}$ | $Y, X_1, \ldots, X_p$ *unknown* relationship |
| $X_i = \begin{cases} \text{non-null} & \beta_i \neq 0 \\ \text{null} & \beta_i = 0 \end{cases}$ | $X_i = \begin{cases} \text{interesting} \\ \text{uninteresting} \\ \text{???} \end{cases}$ |

# Methods

| | Control | $p \leq n$ | $p > n$ |
|---|---|---|---|
| Classical inference (Pearson 1930) | Marginal | Yes | |
| Covariance test (Lockhart et al. 2014) | ? | Yes | Yes |
| Debiased lasso (Javanmard et al. 2014) | Marginal | | Yes |
| Knockoffs (Barber et al. 2014) | FDR | Yes | ? |
| | | | |

# Methods

But what's actually used in practice?

| | Control | $p \leq n$ | $p > n$ |
|---|---|---|---|
| Classical inference (Pearson 1930) | Marginal | Yes | |
| Covariance test (Lockhart et al. 2014) | ? | Yes | Yes |
| Debiased lasso (Javanmard et al. 2014) | Marginal | | Yes |
| Knockoffs (Barber et al. 2014) | FDR | Yes | ? |
| **Marginal screening** | ? | Yes | Yes |

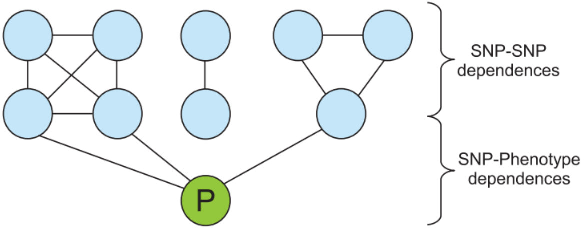
# Regression vs Marginal Screening

- Why not just test $H_i : \text{Cov}(X_i, Y) \neq 0$?
- *Even if* $Y = X\beta + \epsilon$, most non-null $X_i$ are probably also correlated

# Regression vs Marginal Screening

- Why not just test $H_i : \mathrm{Cov}(X_i, Y) \neq 0$?
- *Even if* $Y = X\beta + \epsilon$, most non-null $X_i$ are probably also correlated
- In "big data" *many* $X_i$ are correlated to $Y$, but *redundant*

Genome-wide association study




$\beta_i = 0$

$\beta_i \neq 0$

(Adapted from *Mourad 2012*)

# From theory to practice

*Theory*

- Theory of inference in linear model

*Practice*

# From theory to practice

*Theory*

- Theory of inference in linear model

- Validation of given procedure in real data with ground truth

*Practice*

# From theory to practice

*Theory*

- Theory of inference in linear model
- Theory of robust inference

- Validation of given procedure in real data with ground truth

*Practice*

# From theory to practice

*Theory*

- Theory of inference in linear model
- Theory of robust inference
- Simulation studies

- Validation of given procedure in real data with ground truth

*Practice*

# From theory to practice

*Theory*

- Theory of inference in linear model
- Theory of robust inference
- Simulation studies
- Validation on real data + **synthetic negative controls**
- Validation of given procedure in real data with ground truth

*Practice*

# Practical Validation

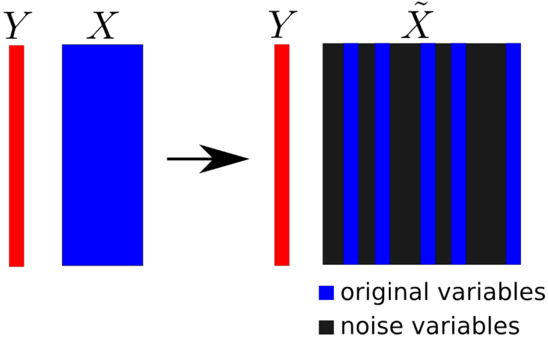
- Difficult to validate inference procedures, because we would need to know the '*true*' $\beta$
- What is the 'true' $\beta$ when the linear model is incorrect? We take

$$\beta = \mathbf{E}[xx^T]^{-1}\mathbf{E}[yx]$$

  (the 'superpopulation' model)
- We don't know the ground truth in real data... what's the next best thing?

# Idea

I give you real data *mixed in* with noise variables




- Can you identify the original columns from the noise columns?
- I can test your procedure this way, because I know the ground truth!

# Synthetic Negative Controls

- Given random vector $x \in \mathbb{R}^p$, *define* $\tilde{x} \in \mathbb{R}^{p+q}$ by by

$$\tilde{x} = \begin{pmatrix} I \\ \Gamma \end{pmatrix} x + e$$

where $\Gamma$ is a fixed matrix and $e \perp x, y$.

# Synthetic Negative Controls

- Given random vector $x \in \mathbb{R}^p$, *define* $\tilde{x} \in \mathbb{R}^{p+q}$ by by

$$\tilde{x} = \begin{pmatrix} I \\ \Gamma \end{pmatrix} x + e$$

  where $\Gamma$ is a fixed matrix and $e \perp x, y$.

- Let

$$\beta = \mathbf{E}[xx^T]^{-1}\mathbf{E}[yx], \quad \tilde{\beta} = \mathbf{E}[\tilde{x}\tilde{x}^T]^{-1}\mathbf{E}[y\tilde{x}]$$

# Synthetic Negative Controls

- Given random vector $x \in \mathbb{R}^p$, *define* $\tilde{x} \in \mathbb{R}^{p+q}$ by by

$$\tilde{x} = \begin{pmatrix} I \\ \Gamma \end{pmatrix} x + e$$

where $\Gamma$ is a fixed matrix and $e \perp x, y$.

- Let

$$\beta = \mathbf{E}[xx^T]^{-1}\mathbf{E}[yx], \quad \tilde{\beta} = \mathbf{E}[\tilde{x}\tilde{x}^T]^{-1}\mathbf{E}[y\tilde{x}]$$

- Then

$$\forall i \in \{1, \ldots, p\} : \beta_i = \tilde{\beta}_i$$

$$\forall i \in \{p+1, \ldots, p+q\} : \tilde{\beta}_i = 0$$

# Synthetic Negative Controls

- Given random vector $x \in \mathbb{R}^p$, *define* $\tilde{x} \in \mathbb{R}^{p+q}$ by by

$$\tilde{x} = \begin{pmatrix} I \\ \Gamma \end{pmatrix} x + e$$

where $\Gamma$ is a fixed matrix and $e \perp x, y$.

- Let

$$\beta = \mathbf{E}[xx^T]^{-1}\mathbf{E}[yx], \quad \tilde{\beta} = \mathbf{E}[\tilde{x}\tilde{x}^T]^{-1}\mathbf{E}[y\tilde{x}]$$

- Then

$$\forall i \in \{1, \ldots, p\} : \beta_i = \tilde{\beta}_i$$

$$\forall i \in \{p+1, \ldots, p+q\} : \tilde{\beta}_i = 0$$

- *Special case.* $X_{p+1}, \ldots, X_{p+q}$ are pure noise: this is when $\Gamma = 0$

# Using SNCs to investigate robustness

- All methods considered depend on strong assumptions (e.g. linearity, Gaussian iid errors, sparsity)
- How well do these methods work on real data where assumptions are most likely violated?

# Using SNCs to investigate robustness

- All methods considered depend on strong assumptions (e.g. linearity, Gaussian iid errors, sparsity)
- How well do these methods work on real data where assumptions are most likely violated?
- Take low-dimensional real data mixed with SNCs (synthetic negative controls): can we identify the real data while controlling Type I error (measured by rejections of SNCs)?

# What can we conclude?

- Experiments using SNCs shows that we can do well on the *hypothesis testing problem* in realistic settings, where assumptions are violated
- However, these experiments cannot tell us if we are solving the right problem!
- Is the *hypothesis testing problem* even relevant for the application? The only way to tell is validation on real, high-dimensional data with application-specific ground truth.

# Closing thoughts

*"Statistics is a science in my opinion... for if its methods fail the test of experience – not the test of logic – they are discarded."*

*" Both the statistician and the client must learn to confront the uncertainties of the world more explicitly, ... never to avoid responsibility for an ever-present understanding that all assumptions underlying data analysis are always approximations. Above all, they must base their thinking on a recognition that their assumptions will always require review and reappraisal... "*

– John Tukey

# References


- Barber, R., and Candes, E. (2014). Controlling the False Discovery Rate via Knockoffs. arXiv Preprint arXiv:1404.5609, 127. Retrieved from http://arxiv.org/abs/1404.5609
- Javanmard, A., and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. The Journal of Machine Learning Research, 15, 28692909. Retrieved from http://dl.acm.org/citation.cfm?id=2697057
- Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). a Significance Test for the Lasso. Annals of Statistics, 42(2), 413468. doi:10.1214/13-AOS1175

# Acknowledgements

Thanks to Will Fithian for useful discussions.