Confounder Adjustment in Multiple Hypothesis Testing

Qingyuan Zhao

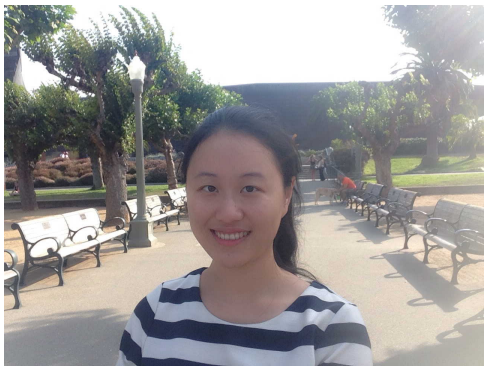
Department of Statistics, Stanford University

January 28, 2016

Slides are available at `http://web.stanford.edu/~qyzhao/`.

Collaborators

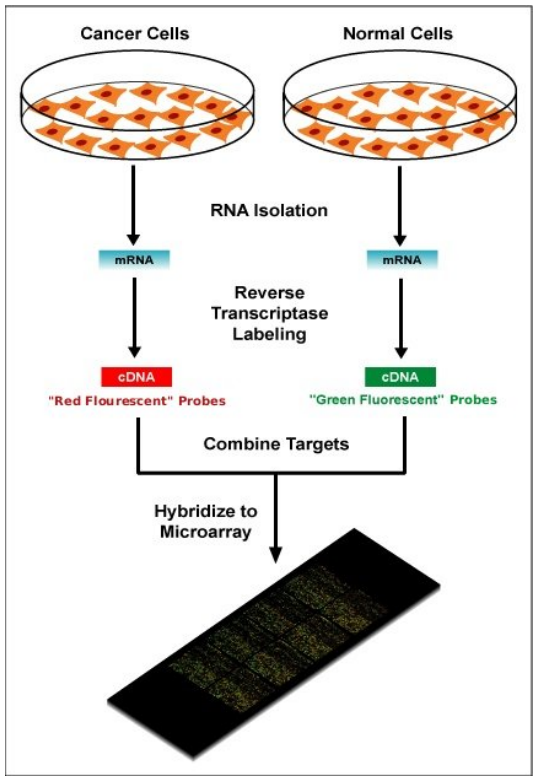
Jingshu Wang



Trevor Hastie

Art Owen

Microarray experiments



- Responses: normalized gene expression level.
- Primary variables (variables of interest): treatment, disease status, etc.
- Control covariates: age, gender, batch, date, etc.

Microarray data analysis

Biologist: "Which genes are (causally) related to this disease?"

Statistician: "Let me run some analysis."

Two common practices

1. **Sparse regression**: regress the primary variable on the genes. More common for SNP data and predictive tasks.
2. **Association tests/screening (this talk)**: for each gene, perform a significance test of correlation with the primary variable.

Statistician: "Here a short list of candidate genes with **false discovery rate (FDR)** $\leq 20\%$."

Biologist: "Good, let me validate these discoveries."

Concerns

J. P. Ioannidis. Why most published research findings are false. *Chance*, 18(4):40–47, 2005

Two major challenges to reproducibility in genetic screening:

1. **Correlated tests**: Is the FDR still controlled? If not, can we correct the statistical analysis?
   - Well studied in the last 15 years [Benjamini and Yekutieli, 2001, Storey et al., 2004, Efron, 2007, Fan et al., 2012].
2. **Confounded tests (this talk)**: the individual association tests are biased in presence of unobserved confounders. Can we still provide a good candidate list?
   - Equally long history [e.g. Alter et al., 2000, Price et al., 2006]. Still many open questions.

Confounding

Brief history

- Fisher [1935] first uses the term in experiment designs.
- Kish [1959] first uses its modern meaning:
  A mixing of effects of unobserved extraneous factors (called confounders) with the effect of interest.
- Huge literature, but mostly in causal inference.

Aliases for confounders in genetic screening:

- "systematic ancestry differences" [Price et al., 2006].
- "batch effects" (widely used by biologists).
- "surrogate variables" [Leek and Storey, 2007, 2008].
- "unwanted variation" [Gagnon-Bartsch and Speed, 2012].
- "latent effects" [Sun et al., 2012].

Example 1: gender study

Which genes are more expressed in male/female?

A microarray experiment by Vawter et al. [2004]:

- Postmortem samples from the brains of 10 individuals.
- For each individual, 3 samples from different cortices.
- Each sample is sent to 3 different labs for analysis.
- Two different microarray platforms are used by the labs.

In total, $10 \times 3 \times 3 = 90$ samples.

This example was first used by Gagnon-Bartsch and Speed [2012] to demonstrate the importance to "remove unwanted variation".

Screening

Notation

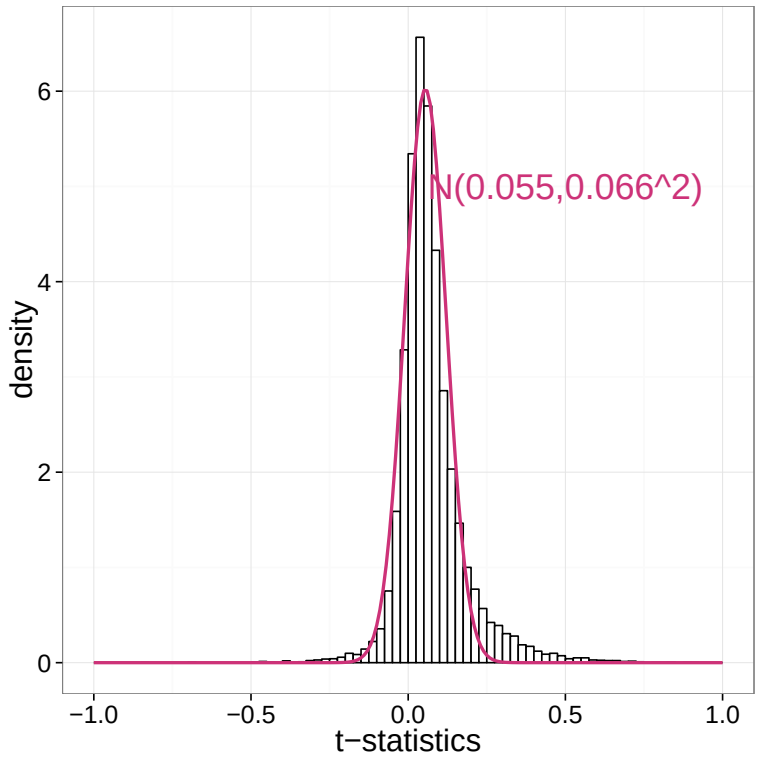
- $\mathbf{Y}$: $n \times p$ matrix of gene expression.
- $\mathbf{X}$: $n \times 1$ vector of gender.

- Simplest association test:

  Regress each column of $\mathbf{Y}$ (gene) on $\mathbf{X}$.

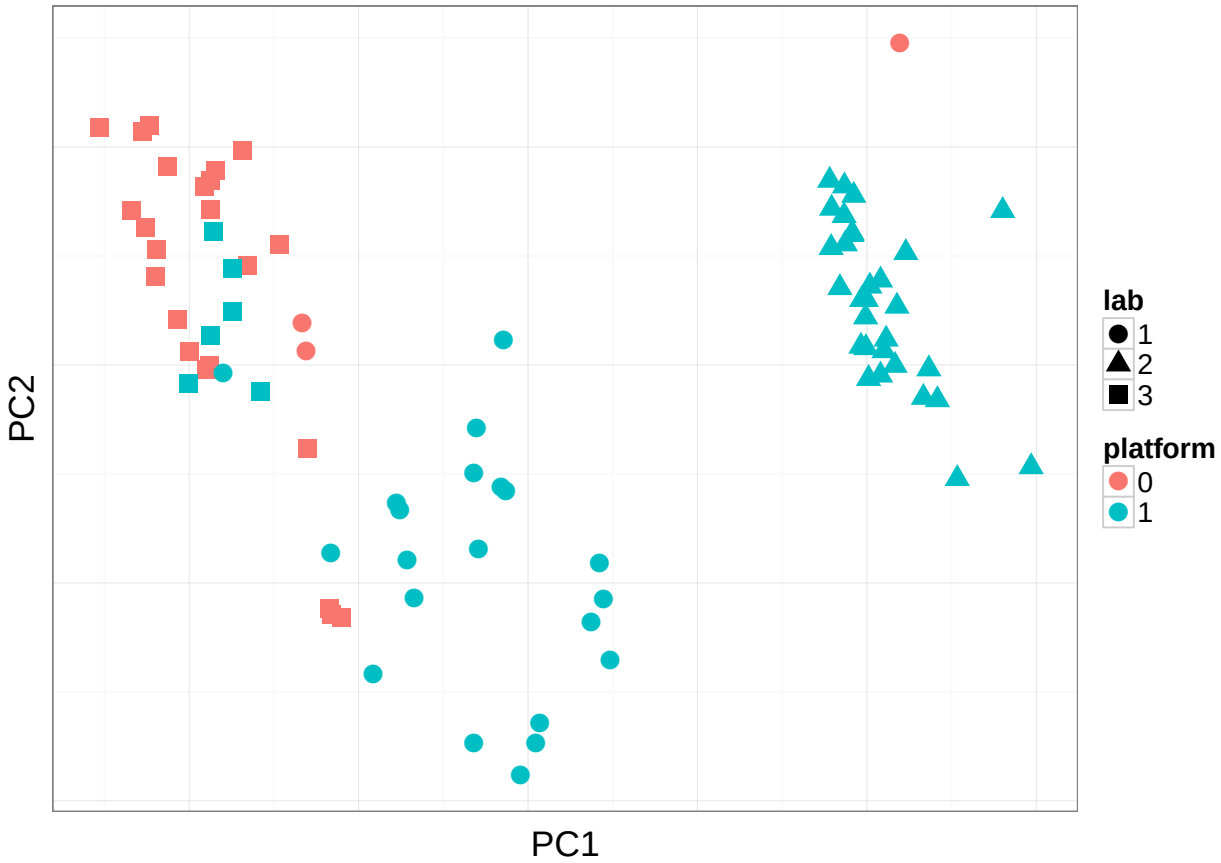
- In R, run `summary(lm(Y~X))`.
- Equivalent to a two-sample t-test with equal variance.

Histogram of t-statistics



Skewed and very underdispersed.

What happened?



Association test

Notation

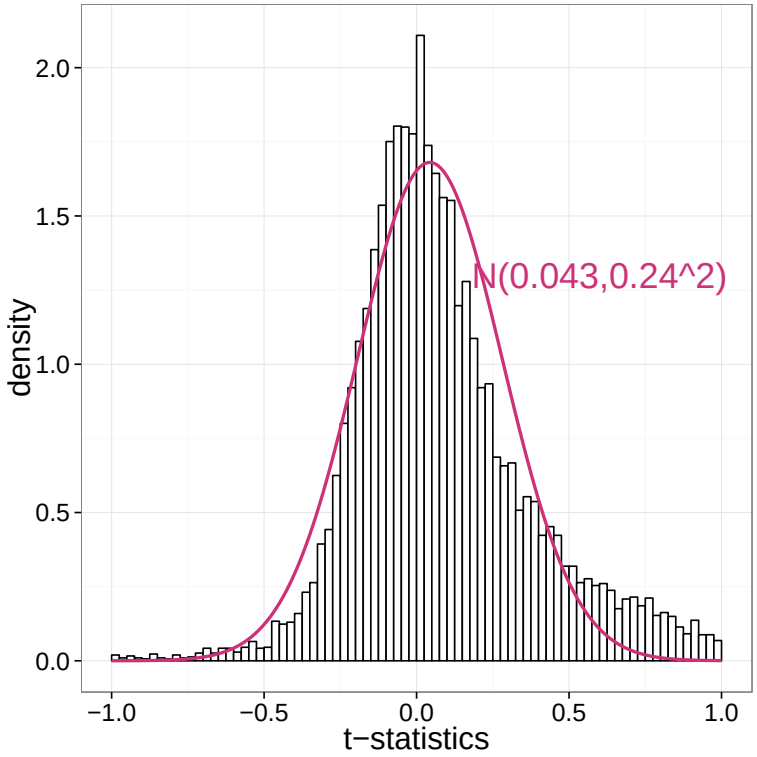
- $\mathbf{Y}$: $n \times p$ matrix of gene expression.
- $\mathbf{X}$: $n \times 1$ vector of gender.
- $\mathbf{Z}$: $n \times d$ matrix of control covariates (lab and platform).

- Modified association test:

  Regress each column of $\mathbf{Y}$ (gene) on $\mathbf{X}$ and $\mathbf{Z}$.

- In R, run `summary(lm(Y~X+Z))`.
- Report the significance of the coefficients of $\mathbf{X}$.

Histogram of t-statistics

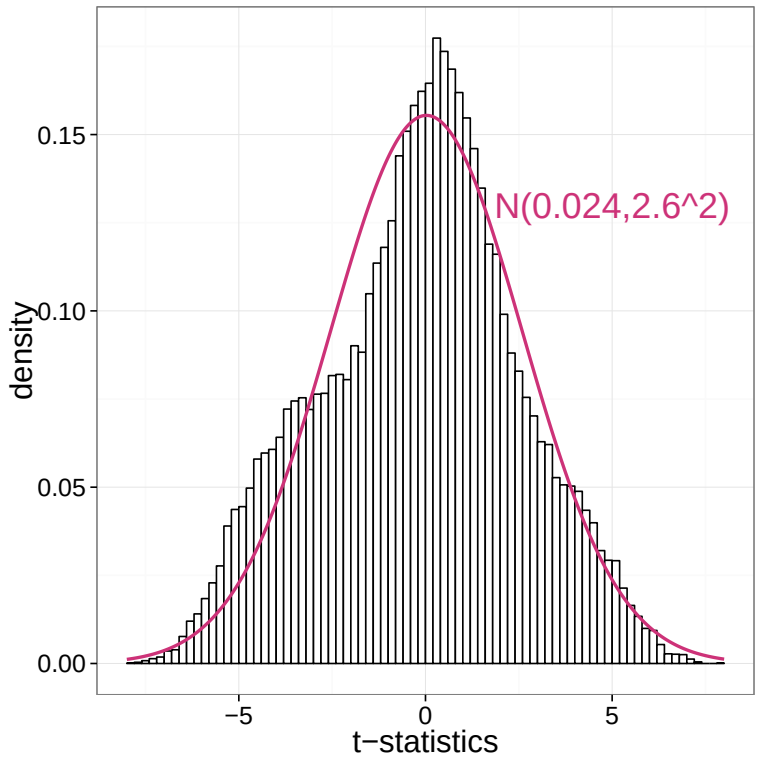


Better, but still problematic.

Reasonable guess: there are more unobserved confounders!

Example 2: COPD study

- COPD = chronic obstructive pulmonary disease.
- Singh et al. [2011] tried to find genes associated with the severity of COPD (moderate or severe).



Overdispersed and skewed.

Example 3: Mutual fund selection

Barras et al. [2010] used the following model to select mutual funds:

$$Y_{it} = \alpha_i + \gamma_i^T \mathbf{Z}_t + e_{it}, \ \ i = 1, \ldots, n, t = 1, \ldots, p.$$

- Y_{it}: observed log-return of fund i at time t.
- α_i: risk-adjusted return (Goal: find funds with positive α).
- $\mathbf{Z}_t$: systematic risk factors.

They assumed:

- $\boldsymbol{\alpha}$ is sparse (Berk and Green equilibrium);
- No unobserved risk factors (is that possible/necessary?).

Idea 0: Remove the largest principal component(s)

EIGENSTRAT [Price et al., 2006]

Regression model:

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times 1}\beta^T_{p\times 1} + \mathbf{Z}_{n\times r}\mathbf{\Gamma}^T_{p\times r} + \mathbf{E}_{n\times p},$$

where $\mathbf{Z}$ is the first r PC(s) of $\mathbf{Y}$.

- Motivation: in SNP, the largest PC(s) usually correspond to ancestry difference.
- Weakness: can easily remove true signals.

Idea 1: Use control genes

Same regression model:

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times 1}\beta^T_{p\times 1} + \mathbf{Z}_{n\times r}\mathbf{\Gamma}^T_{p\times r} + \mathbf{E}_{n\times p},$$

RUV2 [Gagnon-Bartsch and Speed, 2012]

If we know $\beta_{\mathcal{C}} = \mathbf{0}$ (negative controls),

1. Run PCA on $col_{\mathcal{C}}(\mathbf{Y})$ to obtain $\mathbf{Z}$.
2. Run the regression for $col_{\text{-}\mathcal{C}}(\mathbf{Y})$.

- Example: bacterial RNAs (spike-in controls).
- Limited to the availability and number of negative controls.

Idea 2: Sparsity

Same regression model:

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times 1}\beta^T_{p\times 1} + \mathbf{Z}_{n\times r}\mathbf{\Gamma}^T_{p\times r} + \mathbf{E}_{n\times p},$$

Idea: If β contains actual effects, it should be a sparse vector.

SVA [Leek and Storey, 2008]

Iterate between

1. Weighted PCA on $\mathbf{Y}$ (based on how likely $\beta = \mathbf{0}$).
2. Regress $\mathbf{Y}$ on $\mathbf{X}$ and the estimated PCs.

- Does not always converge.

Idea 2: Sparsity

Same regression model:

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times 1}\beta^T_{p\times 1} + \mathbf{Z}_{n\times r}\mathbf{\Gamma}^T_{p\times r} + \mathbf{E}_{n\times p},$$

Idea: If β contains actual effects, it should be a sparse vector.

LEAPP [Sun, Zhang, and Owen, 2012]

1. Run PCA on the residuals of $\mathbf{Y} \sim \mathbf{X}$.
2. Run a sparse regression.

Our contributions: a unifying framework

Missing in previous methods:

- Explicit assumptions on the latent variables.
- Model identification conditions.
- Theoretical guarantees.
- Multiple primary and secondary covariates.
- Practical guidelines: when is confounder adjustment necessary/useful?

Statistical model for confounding

- Linear model for the responses (e.g. gene expression)

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times 1}\,\boldsymbol{\beta}^T_{p\times 1} + \mathbf{Z}_{n\times r}\,\boldsymbol{\Gamma}^T_{p\times r} + \mathbf{E}_{n\times p},$$

  - $\mathbf{X}$: primary variable (disease, treatment, gender, etc.);
  - $\mathbf{Z}$: unobserved confounders;
  - $\boldsymbol{\beta}$: primary effects that we are interested in.
- Missing in the literature: dependence of $\mathbf{Z}$ and $\mathbf{X}$

$$\mathbf{Z}_{n\times r} = \mathbf{X}_{n\times 1}\boldsymbol{\alpha}^T_{r\times 1} + \mathbf{W}_{n\times r},$$

- Additional distributional assumptions:

$$X_i \overset{\text{i.i.d.}}{\sim} \text{mean } 0,\ \text{variance } 1,\ i = 1,\ldots,n,$$

$$\mathbf{E} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}),\ \mathbf{E} \perp\!\!\!\perp (\mathbf{X}, \mathbf{Z}),\ \boldsymbol{\Sigma} = \mathrm{diag}(\{\sigma_j^2\}_{j=1}^p),$$

$$\mathbf{W} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{I}_r),\ \mathbf{W} \perp\!\!\!\perp \mathbf{X}.$$

Marginal effects and direct effects

The model can be rewritten as

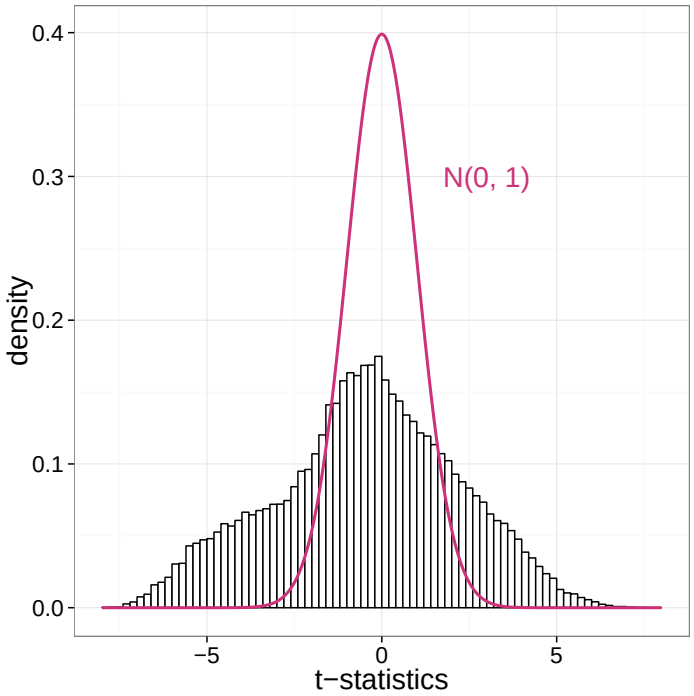
$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times 1}\left(\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\Gamma}_{p\times r}\boldsymbol{\alpha}_{r\times 1}\right)^T + (\mathbf{W}\boldsymbol{\Gamma} + \mathbf{E}),$$

which gives the population identity

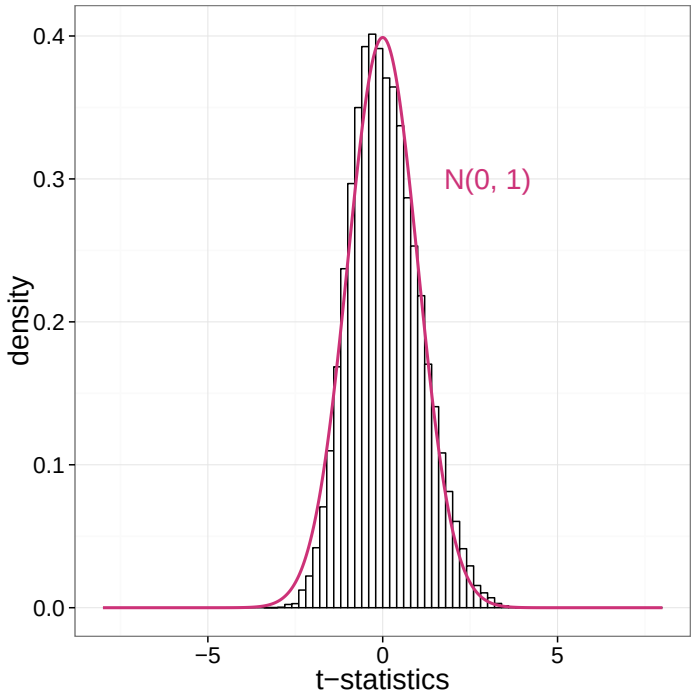
$$\boldsymbol{\tau}_{p\times 1} = \boldsymbol{\beta} + \boldsymbol{\Gamma}\boldsymbol{\alpha}.$$

- $\boldsymbol{\tau}$: marginal effects.
- $\boldsymbol{\beta}$: direct effects (more meaningful).

COPD data: marginal effects vs. direct effects



(a) Before adjustment (t-statistics for $\tau_j = 0$).



(b) After adjustment (t-statistics for $\beta_j = 0$).

Identifiability of $\boldsymbol{\beta}$

To identify $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from

$$\boldsymbol{\tau}_{p\times 1} = \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\Gamma}\boldsymbol{\alpha}_{r\times 1},$$

there are p equations but $p + r$ parameters.

Proposition [Wang, Z., Hastie, and Owen, 2015]

Suppose $\boldsymbol{\Gamma}$ can be identified. $\boldsymbol{\beta}$ is identifiable under either of the two following conditions:

1. Negative control: for a known negative control set $\mathcal{C}$,

$$\boldsymbol{\beta}_{\mathcal{C}} = \mathbf{0},\ |\mathcal{C}| \geq r,\ \operatorname{rank}(\boldsymbol{\Gamma}_{\mathcal{C}}) = r.$$

2. Sparsity: $\|\boldsymbol{\beta}\|_0 \leq \lfloor (p-r)/2 \rfloor$ (the maximum breakdown point),

$$\operatorname{rank}(\boldsymbol{\Gamma}_{\mathcal{C}}) = r,\ \forall \mathcal{C} \subset \{1,\ldots,p\} \text{ such that } |\mathcal{C}| = r.$$

Rotation

Householder transformation

$$\mathbf{X}_{n\times 1} = \mathbf{Q}\mathbf{R}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is orthogonal with $\mathbf{R} = (\|\mathbf{X}\|_2, 0, \cdots, 0)^T$.

- For simplicity, assume $\|\mathbf{X}\|_2 = \sqrt{n}$.
- Can be easily extended to multiple variables $\mathbf{X}$.

Rotation (LEAPP)

Left-Multiply $\mathbf{Q}^T$ to $\mathbf{Y} = \mathbf{X}\beta^T + \mathbf{Z}\mathbf{\Gamma}^T + \mathbf{E}$, we get

$$\mathrm{row}_1(\mathbf{Q}^T\mathbf{Y}) \sim \mathrm{N}(\sqrt{n}(\beta + \mathbf{\Gamma}\alpha), \mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Sigma}),$$

$$\mathrm{row}_{\text{-}1}(\mathbf{Q}^T\mathbf{Y}) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Sigma}).$$

Two-step estimation

1. Run factor analysis for

$$\text{row}_{-1}(\mathbf{Q}^T\mathbf{Y}) \overset{\text{i.i.d.}}{\sim} \text{N}(\mathbf{0}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \boldsymbol{\Sigma})$$

   to obtain $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Sigma}}$. Identifiability follows from classical results in factor analysis [e.g. Anderson and Rubin, 1956].

2. Run linear regression for the marginal effects

$$\frac{\text{row}_1(\mathbf{Q}^T\mathbf{Y})_{p\times 1}}{\sqrt{n}} = \hat{\boldsymbol{\Gamma}}_{p\times r}\,\boldsymbol{\alpha}_{r\times 1} + \boldsymbol{\beta}_{p\times 1} + \tilde{\mathbf{E}}_1/\sqrt{n}$$

response — design matrix — coefficients

How accurate is $\hat{\mathbf{\Gamma}}$?

Assumptions

- High-dimensional data: $n \to \infty$, $p \to \infty$.
- Assume that the factors are strong enough: $\lim_{p\to\infty} \frac{1}{p}\mathbf{\Gamma}^T\mathbf{\Sigma}^{-1}\mathbf{\Gamma}$ exists and is positive definite.
- Consistent estimate of r [Bai and Ng, 2002].

Theoretical Results for MLE

- Consistent estimate of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ [Bai and Li, 2012] and

$$\sqrt{n}(\hat{\mathbf{\Gamma}}_j - \mathbf{\Gamma}_j) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma_j^2 \mathbf{I}_r), \ \sqrt{n}(\hat{\sigma}_j - \sigma_j) \xrightarrow{d} \mathrm{N}(0, 2\sigma_j^4),$$

- Uniform consistency if $n^k/p \to \infty$ for some $k > 0$ [Wang, Z., Hastie, and Owen, 2015].

Strategy 1: Estimate β via negative controls

Recall the marginal effects are

$$\frac{\tilde{\mathbf{Y}}^T_{p\times 1}}{\sqrt{n}} = \boldsymbol{\Gamma}_{p\times r}\,\boldsymbol{\alpha}_{r\times 1} + \boldsymbol{\beta}_{p\times 1} + \tilde{\mathbf{E}}_1/\sqrt{n}$$

response — design matrix — coefficients

In the negative control scenario, we know $\boldsymbol{\beta}_{\mathcal{C}} = \mathbf{0}$.

Generalized Least Squares (GLS) estimator

$$\hat{\boldsymbol{\alpha}}^{\mathrm{NC}} = (\hat{\boldsymbol{\Gamma}}_{\mathcal{C}}^T \hat{\boldsymbol{\Sigma}}_{\mathcal{C}}^{-1} \hat{\boldsymbol{\Gamma}}_{\mathcal{C}})^{-1} \hat{\boldsymbol{\Gamma}}_{\mathcal{C}}^T \hat{\boldsymbol{\Sigma}}_{\mathcal{C}}^{-1} \tilde{\mathbf{Y}}_{1,\mathcal{C}}^T / \|\mathbf{X}\|_2$$

$$\hat{\boldsymbol{\beta}}_{\text{-}\mathcal{C}}^{\mathrm{NC}} = \tilde{\mathbf{Y}}_{1,\text{-}\mathcal{C}}^T / \|\mathbf{X}\|_2 - \hat{\boldsymbol{\Gamma}}_{\text{-}\mathcal{C}} \hat{\boldsymbol{\alpha}}^{\mathrm{NC}}$$

Note: RUV4 [Gagnon-Bartsch et al., 2013] = Ordinary Least Squares (OLS).

Asymptotic distribution of $\hat{\boldsymbol{\beta}}^{\mathrm{NC}}$

Theorem (Wang, Z., Hastie, and Owen [2015])

Under the assumptions of uniform convergence of $\hat{\boldsymbol{\Sigma}}$ *and* $\hat{\boldsymbol{\Gamma}}$ *and* $\lim_{p\to\infty} \frac{1}{|\mathcal{C}|}\boldsymbol{\Gamma}_{\mathcal{C}}^T\boldsymbol{\Sigma}_{\mathcal{C}}^{-1}\boldsymbol{\Gamma}_{\mathcal{C}} \succ \mathbf{0}$, *then for any finite index set* $\mathcal{S}$ *such that* $\mathcal{S}\cap\mathcal{C}=\emptyset$*:*

1. *If the number of negative controls* $|\mathcal{C}|\to\infty$,
$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{NC}} - \boldsymbol{\beta}_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (1+\|\boldsymbol{\alpha}\|_2^2)\boldsymbol{\Sigma}_{\mathcal{S}})$$
2. *If* $\lim_{p\to\infty}|\mathcal{C}| < \infty$,
$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{NC}} - \boldsymbol{\beta}_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (1+\|\boldsymbol{\alpha}\|_2^2)(\boldsymbol{\Sigma}_{\mathcal{S}}+\boldsymbol{\Delta}_{\mathcal{S}}))$$
where $\boldsymbol{\Delta}_{\mathcal{S}} = \lim_{p\to\infty}\boldsymbol{\Gamma}_{\mathcal{S}}(\boldsymbol{\Gamma}_{\mathcal{C}}^T\boldsymbol{\Sigma}_{\mathcal{C}}^{-1}\boldsymbol{\Gamma}_{\mathcal{C}})^{-1}\boldsymbol{\Gamma}_{\mathcal{S}}^T$.

Strategy 2: Estimate β via sparsity

Recall

$$\underbrace{\frac{\tilde{\mathbf{Y}}^T_{p\times 1}}{\sqrt{n}}}_{\text{response}} = \underbrace{\mathbf{\Gamma}_{p\times r}}_{\text{design matrix}} \underbrace{\boldsymbol{\alpha}_{r\times 1}}_{\text{coefficients}} + \boldsymbol{\beta}_{p\times 1} + \tilde{\mathbf{E}}_1/\sqrt{n}$$

Idea: if $\|\boldsymbol{\beta}\|_0 \ll p$, $\beta_j \neq 0$ is an outlier in this regression.

Robust regression estimator (simplification of LEAPP)

$$\hat{\boldsymbol{\alpha}}^{\mathrm{RR}} = \arg\min \sum_{j=1}^{p} \rho\left(\frac{\tilde{Y}_{1j}/\sqrt{n} - \hat{\mathbf{\Gamma}}_j^T \boldsymbol{\alpha}}{\hat{\sigma}_j}\right)$$

$$\hat{\boldsymbol{\beta}}^{\mathrm{RR}} = \tilde{\mathbf{Y}}_1^T/\sqrt{n} - \hat{\mathbf{\Gamma}}\hat{\boldsymbol{\alpha}}^{\mathrm{RR}}$$

Asymptotic distribution of $\hat{\boldsymbol{\beta}}^{\mathrm{RR}}$

Assumptions on the loss function $\rho(x)$

The derivatives ρ', ρ'' and ρ''' exist and are bounded.
$\rho(0) = \rho'(0) = 0$, $\rho''(0) > 0$ and $\rho'(x) \cdot x \geq 0$.
(e.g. Tukey's bisquare)

Theorem (Wang, Z., Hastie, and Owen [2015])

Under the assumptions of uniform convergence of $\hat{\boldsymbol{\Sigma}}$ *and* $\hat{\boldsymbol{\Gamma}}$ *and the above assumption of the loss function, if* $\min(\|\boldsymbol{\beta}\|_0, \|\boldsymbol{\beta}\|_1)\sqrt{n}/p \to 0$, *then for any finite index set* $\mathcal{S}$*:*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{\mathrm{RR}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (1 + \|\boldsymbol{\alpha}\|_2^2)\boldsymbol{\Sigma}_{\mathcal{S}}).$$

Oracle efficiency

In either the sparsity or negative control scenario ($|\mathcal{C}| \to \infty$):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, (1 + \|\boldsymbol{\alpha}\|_2^2)\boldsymbol{\Sigma}_{\mathcal{S}})$$

Oracle estimator

Consider the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^T + \mathbf{Z}\boldsymbol{\Gamma}^T + \mathbf{E}.$$

If $\mathbf{Z}$ were observed, the oracle OLS estimator would be

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{OLS}} - \boldsymbol{\beta}_{\mathcal{S}}) \sim \mathrm{N}(\mathbf{0}, (1 + \|\boldsymbol{\alpha}\|_2^2)\boldsymbol{\Sigma}_{\mathcal{S}}).$$

$\hat{\boldsymbol{\beta}}_{\mathcal{S}}$ is as efficient asymptotically as the oracle estimator!

Significance test for confounding

Theorem (Wang, Z., Hastie, and Owen [2015])

Under the above assumptions for oracle efficiency and the null hypothesis that $H_{0,\boldsymbol{\alpha}} : \boldsymbol{\alpha} = \mathbf{0}$*, we have*

$$n \cdot \hat{\boldsymbol{\alpha}}^T \hat{\boldsymbol{\alpha}} \xrightarrow{d} \chi^2_r$$

where χ^2_r *is the chi-square distribution with* r *degree of freedom.*

Recipes

1. Graphical diagnostics: the histogram of test statistics.
2. Positive controls: e.g. X/Y genes for gender.
3. Asymptotic χ^2 test. If significant, check $\hat{\boldsymbol{\Gamma}}$.

Multiple hypothesis testing

Two-sided asymptotic z-tests

Test $H_{j0} : \beta_j = 0$ vs. $H_{j1} : \beta_j \neq 0$ for $j = 1, \ldots, p$.

$$t_j = \frac{\sqrt{n}\hat{\beta}_j}{\hat{\sigma}_j \sqrt{1 + \|\hat{\boldsymbol{\alpha}}\|^2}}, \quad P_j = 2(1 - \Phi(|t_j|)).$$

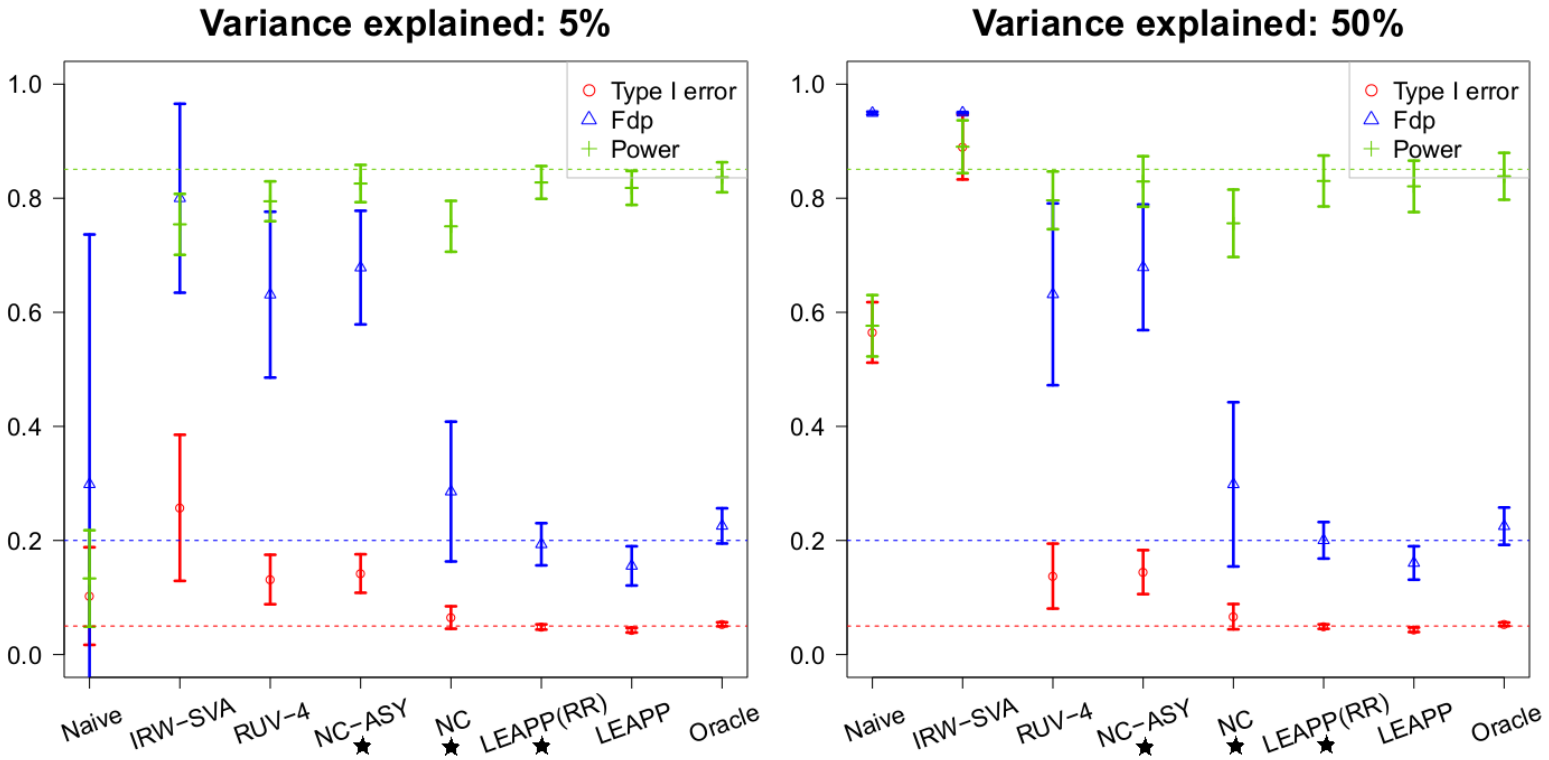
Theorem (Wang, Z., Hastie, and Owen [2015])

Under the assumptions for oracle efficiency, the overall type I error and the familywise error rate (FWER) can be asymptotically controlled.

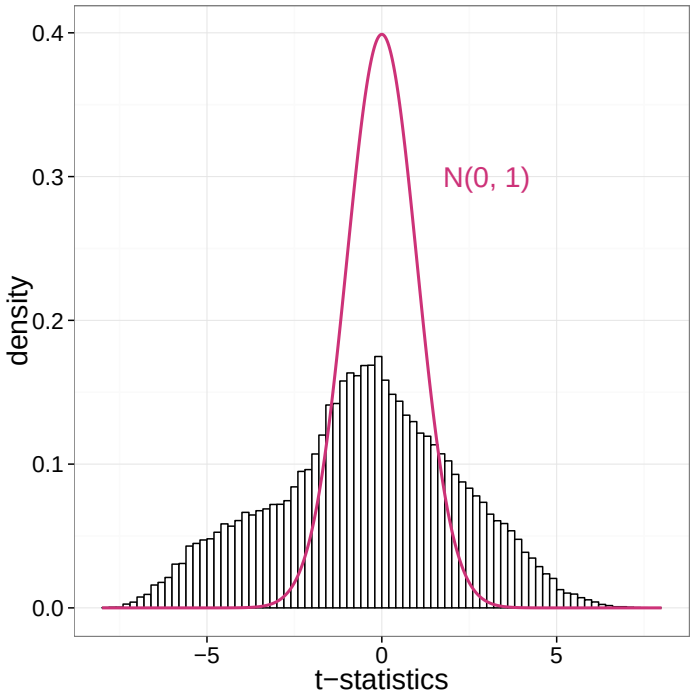
FDR control: ongoing work.

Simulation: $n = 100$, $p = 5000$ and $r = 10$

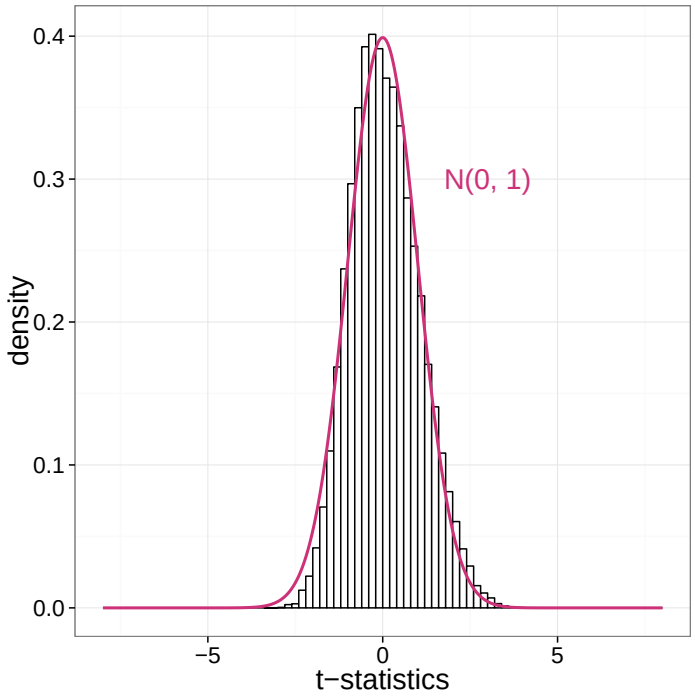
- Sparsity: $\|\beta\|_0/p = 0.05$; NC: $|\mathcal{C}| = 30$.
- $\mathbf{\Gamma}$ uniform from orthogonal matrices; $\sigma_i^2 \overset{i.i.d.}{\sim} \mathrm{InvGamma}(3, 2)$.
- Variance of $\mathbf{X}$ explained by $\mathbf{Z}$: $\max_{\boldsymbol{\rho}} \mathrm{corr}(X_i, \boldsymbol{\rho}^T \mathbf{Z}_i) = \dfrac{\|\boldsymbol{\alpha}\|_2}{1 + \|\boldsymbol{\alpha}\|_2}$.



COPD data: severity as primary variable



(a) Naive linear regression.

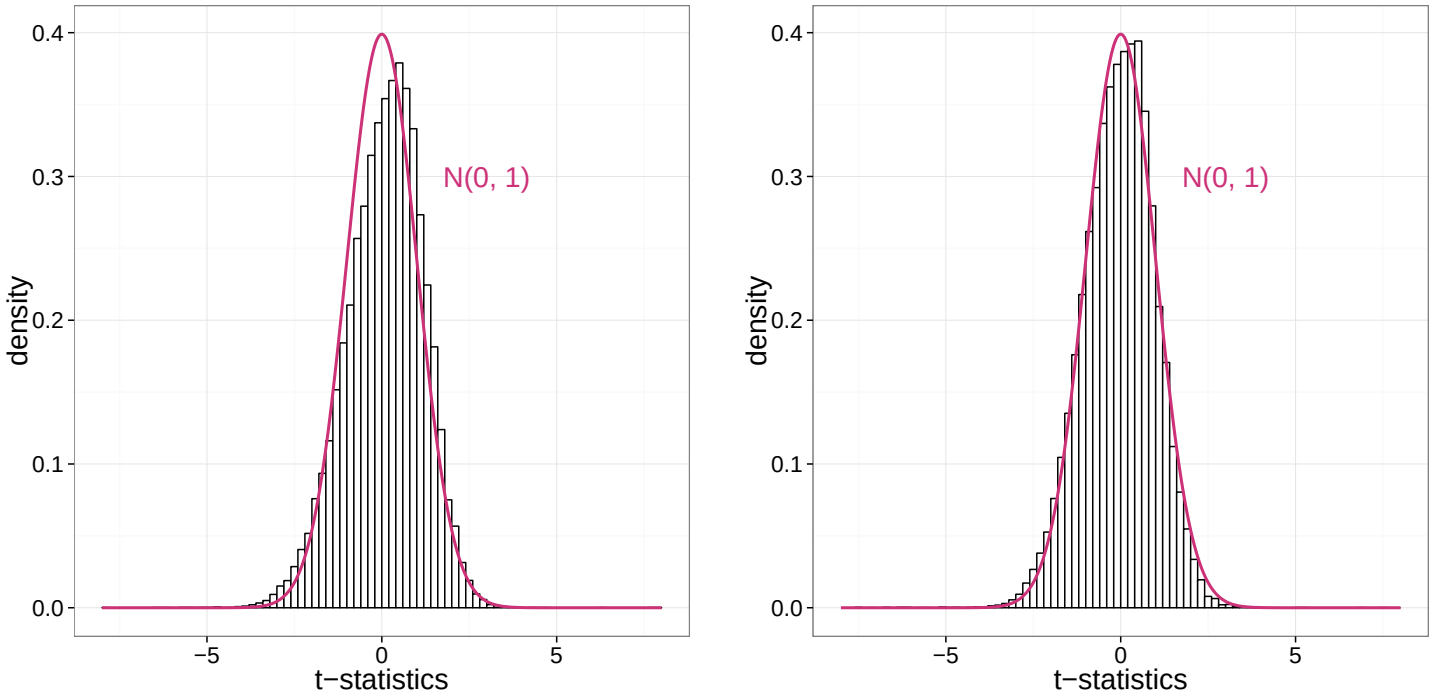


(b) After adjustment.

- $\hat{r} = 1$ [Onatski, 2010].
- $\hat{\alpha} \approx 0.98$, variance explained is approximately 22%.
- Test of confounding: p-value ≈ 0.

COPD data: gender as primary variable

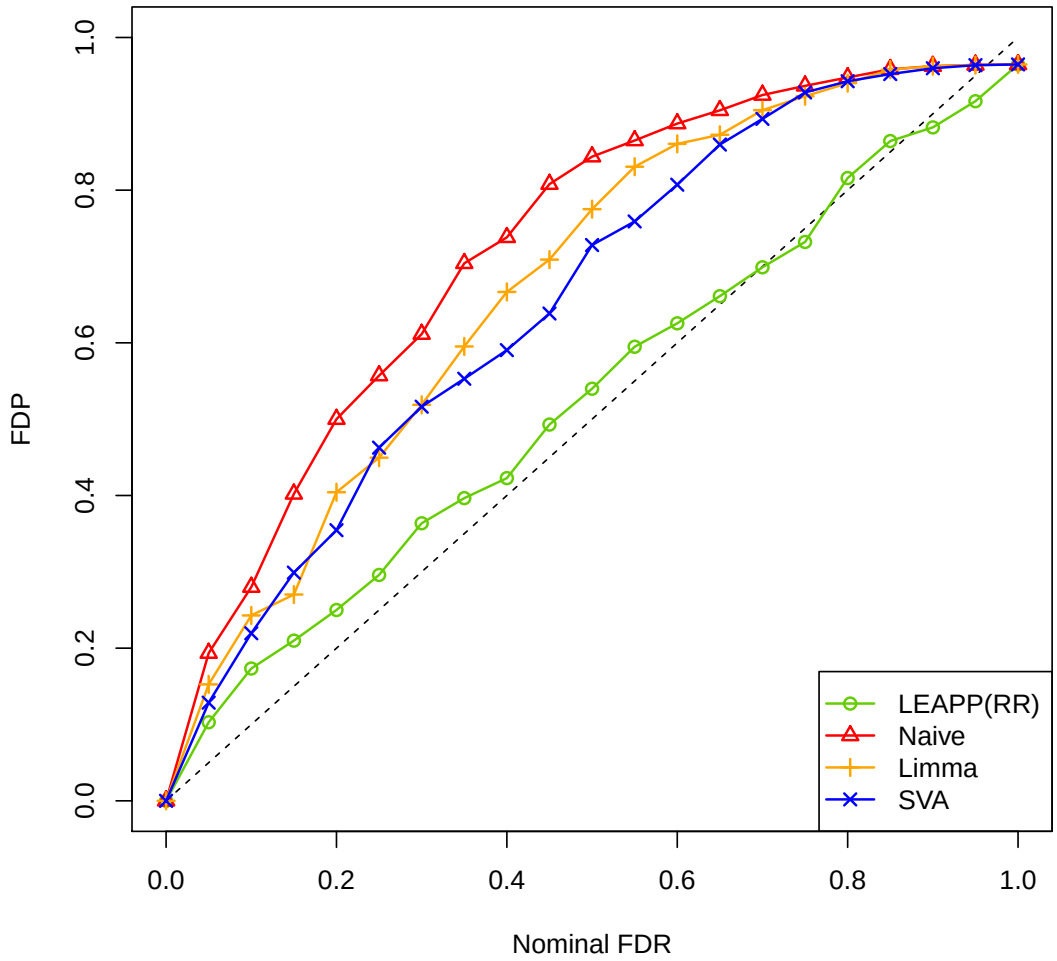
Genes associated with gender should come from X/Y chromosomes (positive controls).



(a) Naive linear regression. (b) After adjustment.

- $\hat{\alpha} \approx -0.27$, variance explained is approximately 3%.
- Test of confounding: p-value $\approx 1.2 \times 10^{-3}$.

COPD data: gender as primary variable

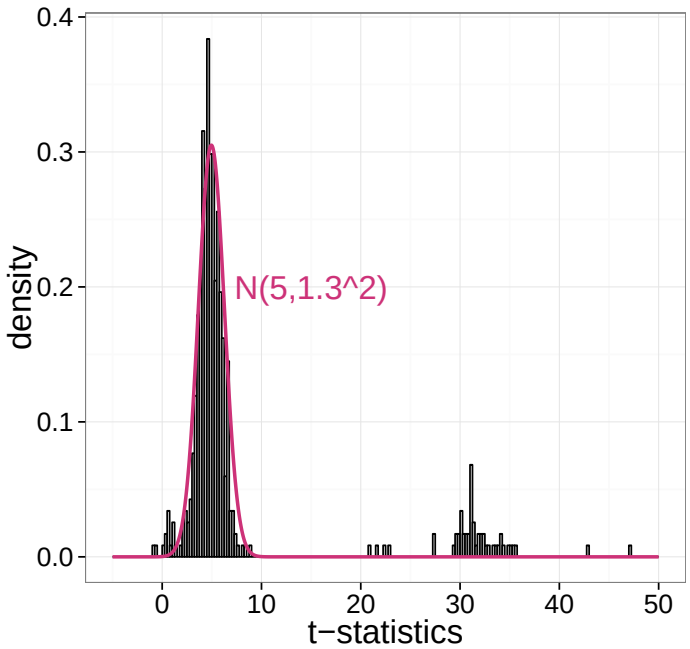


COPD data: gender as primary variable

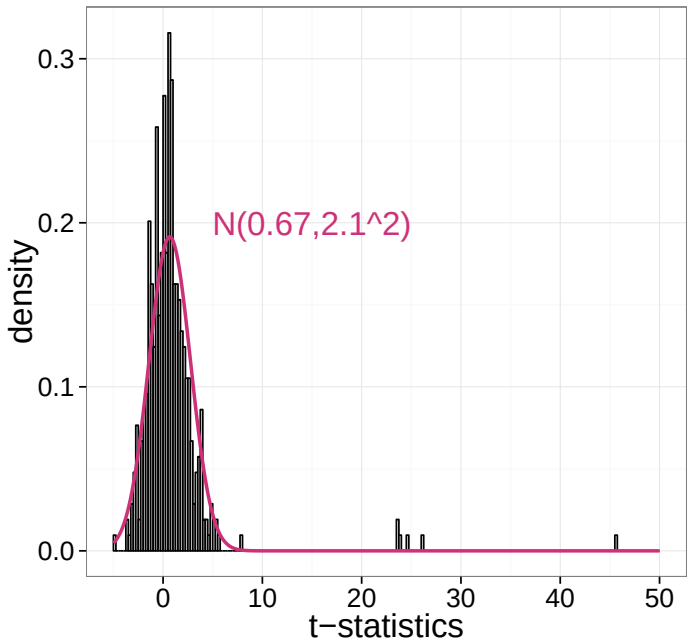
Method	X/Y Genes in Top 100
LEAPP(RR)	69
Naive	58
Limma	58
SVA	68

Mutual fund selection (preliminary results)

- $p = 469$ mutual funds with monthly returns available in CRSP database in Jan. 1980 – Dec. 2000 ($n = 240$).
- Apply the RR procedure with $r = 6$ without adjusting for any observed systematic risk factor.



(a) Naive linear regression.



(b) After adjustment.

Summary

Recap

- Linear model with unobserved confounding factors.
- Identification conditions: negative control and sparsity.
- Two-step estimation of the primary effects.
- Asymptotic distributions and oracle efficiency.
- Hypothesis tests for confounding and the primary effects.

Open problems

- Correlated noise: approximate factor models.
- Weak factors: random matrix theory.
- Non-Gaussian data: RNA-seq, GWAS.
- Beyond linearity?

Resources

- J. Wang, Z., T. Hastie, and A. B. Owen. Confounder adjustment in multiple hypothesis testing. *under revision for Annals of Statistics*, 2015.
  - Available on arXiv.
- Software: `cate` on CRAN.
  (`https://cran.r-project.org/web/packages/cate/index.html`)
  - Package vignette available online.
  - Unified interface for existing packages `sva`, `ruv`, `leapp`.
  - We also support formula:

```
results <- cate(~ gender | . - gender - 1, data, ...)
```