

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

Charles Zheng

Stanford University

May 7, 2015

# Problem and motivation

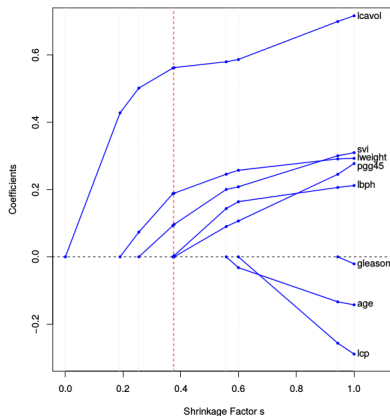
- $x \in \mathbb{R}^p, y \in \mathbb{R}$  have a joint distribution  $P$  where  $y|x \sim N(x^T \beta, \sigma^2)$
- Observe  $X = (x_1, \dots, x_n)^T$ ,  $Y = (y_1, \dots, y_n)$  iid
- Problem: test  $H_i : \beta_0 = i$  for  $i = 1, \dots, p$
- Motivation:  $x$  are SNPs (mutations),  $y$  is phenotype

|  | Control  | $p > n$ |
|--|----------|---------|
| Classical inference (Pearson 1930)     | Marginal | No      |
| Debiased lasso (Javanmard et al. 2014) | Marginal | Yes     |
| Knockoffs (Barber et al. 2014)         | FDR      | ?       |
| Covariance test (Lockhart et al. 2014) | ??       | Yes     |
| ... + FDR control (G'Sell et al. 2013) | FDR      | Yes     |
|  |          |         |

# The LASSO path

All three methods share an association with LASSO:

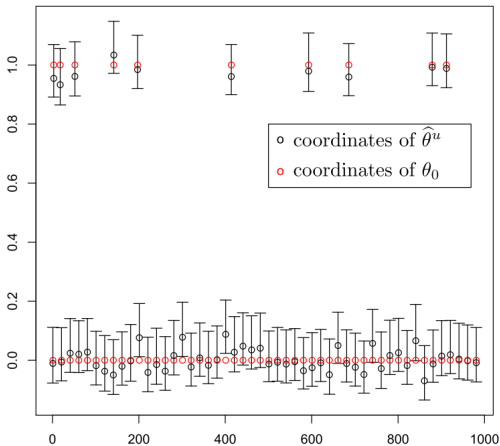
$$\hat{\beta}_{\lambda} = \operatorname{argmin}_{\beta} \frac{1}{2} \|X\beta - Y\|^2 + \lambda \|\beta\|_1$$



(Image credit: ??)

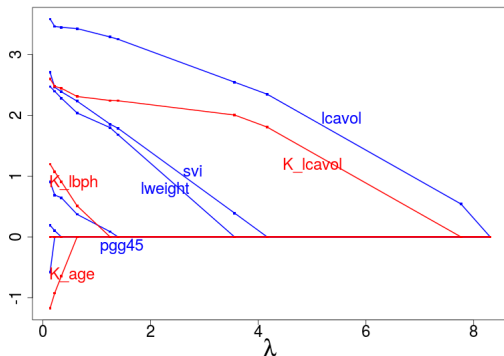
# Debiased regularized M-estimators

- (2014) Javanmard and Montanari
- Standard assumptions + sparsity condition on  $\beta$  + large  $n$  and  $p$  asymptotics



# Knockoff filter

- (2014) Barber and Candés
- *Finite sample*  $Y \sim N(X\beta, \sigma^2 I)$ ,  $n \leq p$ , control FDR
- Extension to  $p > n$ , FWER control, etc. forthcoming...



|         |         |
|---------|---------|
| lweight | 22.5652 |
| lcavol  | 20.5199 |
| svi     | 4.4871  |
| lbph    | 1.1865  |
| age     | 0.0829  |
| gleason | 0.0387  |
| lcp     | -0.2359 |
| pgg45   | -3.3742 |

# Covariance test

- (2014) Lockhart, Taylor, Tibshirani ( $\times 2$ )
- Standard assumptions  $Y \sim N(X\beta, \sigma^2 I) +$  large  $p$  asymptotics
- See *also* non-asymptotic exact test (Lee, Sun  $\times 2$ , Taylor 2015)
- What kind of Type I error does it control?

| Step | Predictor entered | Forward stepwise | Lasso |
|------|-------------------|------------------|-------|
| 1    | lcavol            | 0.000            | 0.000 |
| 2    | lweight           | 0.000            | 0.052 |
| 3    | svi               | 0.041            | 0.174 |
| 4    | lbph              | 0.045            | 0.929 |
| 5    | pgg45             | 0.226            | 0.353 |
| 6    | age               | 0.191            | 0.650 |
| 7    | lcp               | 0.065            | 0.051 |
| 8    | gleason           | 0.883            | 0.978 |

# FDR control for covariance test

- G'Sell, Wager, Chouldechova, Tibshirani (2013)
- Two methods to control FDR for covariance test... but under a *different* definition of Type I error
- Type I error is defined according to the *incremental null* (next slide)



# Formulations for hypothesis testing

- For a subset  $E$  of the variables, define  $\beta^E = (X_E^T X_E)^{-1} X_E^T X \beta$

# Formulations for hypothesis testing

- For a subset  $E$  of the variables, define  $\beta^E = (X_E^T X_E)^{-1} X_E^T X \beta$
- **The full model null.** Test multiple hypotheses  $H_i : \beta_i = 0$

# Formulations for hypothesis testing

- For a subset  $E$  of the variables, define  $\beta^E = (X_E^T X_E)^{-1} X_E^T X \beta$
- **The full model null.** Test multiple hypotheses  $H_i : \beta_i = 0$
- **Selective inference.** Condition on a randomly selected subset  $E$ , test hypotheses  $H_i : \beta_i^E = 0$  for all  $i \in E$

# Formulations for hypothesis testing

- For a subset  $E$  of the variables, define  $\beta^E = (X_E^T X_E)^{-1} X_E^T X \beta$
- **The full model null.** Test multiple hypotheses  $H_i : \beta_i = 0$
- **Incremental null.** (Informal explanation).
  - Test variables as they are added to the model
  - I make a mistake whenever I add a variable which *doesn't* improve the fit of the model
  - If a variable I added previously *becomes* redundant later, it doesn't count against me

# Formulations for hypothesis testing

- For a subset  $E$  of the variables, define  $\beta^E = (X_E^T X_E)^{-1} X_E^T X \beta$
- **The full model null.** Test multiple hypotheses  $H_i : \beta_i = 0$
- In this talk, we define type I errors according to *full model null*...

But what's actually used in practice?

|  | Control  | $p > n$ |
|--|----------|---------|
| Classical inference (Pearson 1930)     | Marginal | No      |
| Debiased lasso (Javanmard et al. 2014) | Marginal | Yes     |
| Knockoffs (Barber et al. 2014)         | FDR      | ?       |
| Covariance test (Lockhart et al. 2014) | ??       | Yes     |
| ... + FDR control (G'Sell et al. 2013) | FDR      | Yes     |
| <b>Marginal screening</b>              | ???      | Yes     |

# Statistical Validation

- These procedures are derived under strong assumptions (linearity, gaussianity, homoscedasticity)
- How well do they work in real data where these assumptions are violated?
- We could validate inference procedures in real data if only we knew the 'true'  $\beta$ , (re)defined as

$$\beta = \mathbf{E}[\mathbf{x}\mathbf{x}^T]^{-1}\mathbf{E}[\mathbf{y}\mathbf{x}]$$

# Statistical Validation

- These procedures are derived under strong assumptions (linearity, gaussianity, homoscedasticity)
- How well do they work in real data where these assumptions are violated?
- We could validate inference procedures in real data if only we knew the 'true'  $\beta$ , (re)defined as

$$\beta = \mathbf{E}[\mathbf{x}\mathbf{x}^T]^{-1}\mathbf{E}[\mathbf{y}\mathbf{x}]$$

- Possibility: take a dataset with large  $p$  and *humongous*  $n$ , so we can get an extremely precise estimate of  $\beta$  using OLS. Then test the high-dimensional inference procedures on subsamples of size  $n_0 \leq p < n$  of the data

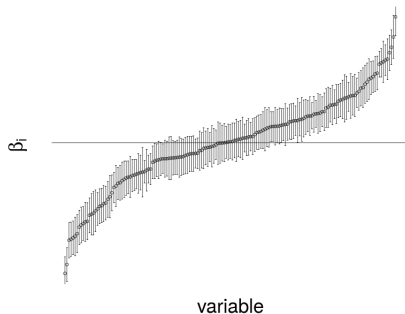


## Example: personality data

- Data with  $p = 163$  survey questions from an online personality test,  $n = 49086$  (after processing)
- Predict self-reported age of respondent,  $y$ , from their responses
- Is  $n$  large enough for us to confidently say which  $\beta_i = 0$  (for use as ground truth?)

# Example: personality data

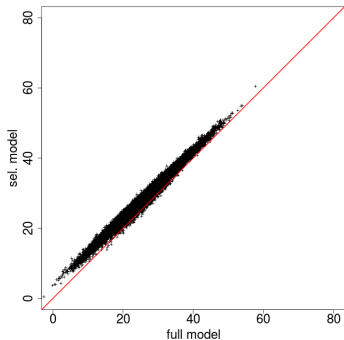
Coefficient estimates  $\pm 3$  sd



Consider declaring all variables whose intervals cross 0 to be null. Then  $p_1 = 105$  (out of 163)

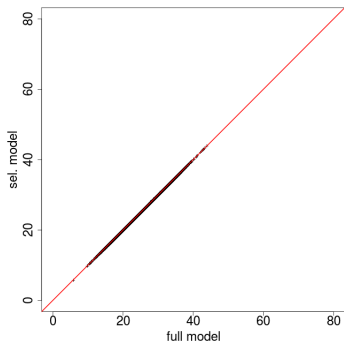
# Example: personality data

- If  $n$  were large enough, then for the selected model  $S$  we should have  $\hat{y} = \sum_{i=1}^p X_i \hat{\beta}_i$  close to  $\hat{y}_S = \sum_{i \in S} X_i \hat{\beta}_i$
- But...



# Example: personality data

- Here  $n$  is not large enough for  $p = 163$
- If we reduce the dimensionality to 15 by subsampling columns, it looks more convincing that we selected the correct 10 variables



- It is by no means *impossible* to get large enough data to estimate high-dimensional  $\beta$ , with say,  $p > 100$
- But if were *easy* to get such large  $n$  data... we wouldn't need these new inference techniques in the first place!

# Why not use simulations?

- Simulations can be used to test robustness of the procedure
- In simulations, we can add all the nonlinearities, nongaussianity, etc. that we want

# Why not use simulations?

- Advantage: In simulations, we not only know  $\beta$ , but exactly how the data is generated
- Advantage: We can vary simulation parameters and get a lot of insight about the procedure being tested

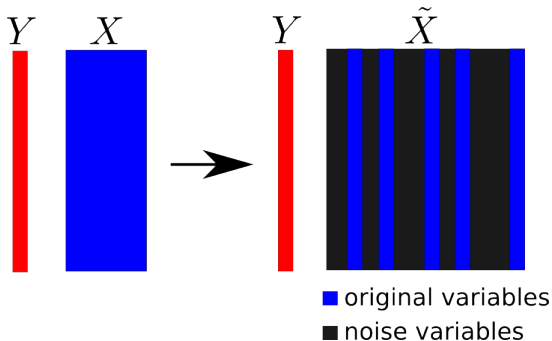
# Why not use simulations?

- Advantage: In simulations, we not only know  $\beta$ , but exactly how the data is generated
- Advantage: We can vary simulation parameters and get a lot of insight about the procedure being tested
- **Disadvantage:** Are these simulations relevant? How can we tell the simulated models are realistic?



# Idea

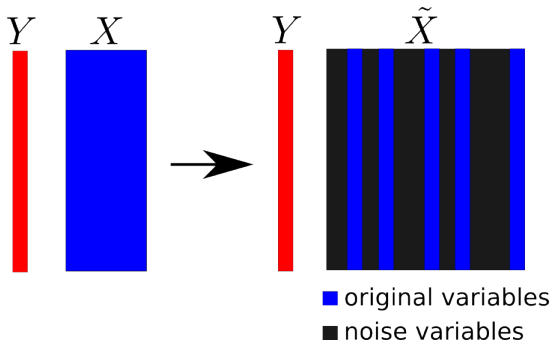
I give you real data *mixed in* with noise variables



- Can you identify the original columns from the noise columns?

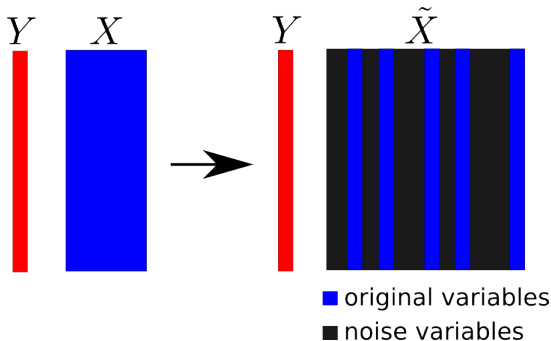
# Idea

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!

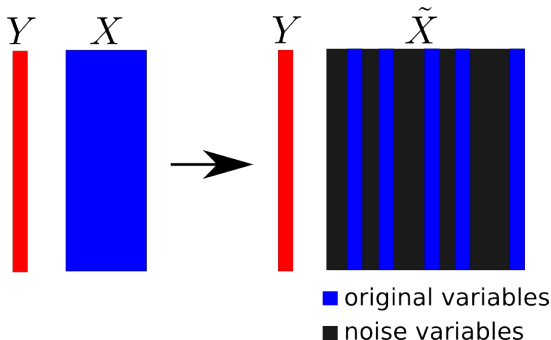
I give you real data *mixed in* with noise variables



- Can you identify the original columns from the noise columns?
- I can test your procedure this way, because I know the ground truth!
- **Caveat:** this test is unrealistically 'easy' (due to lack of correlations)

# Synthetic Negative Controls

- Synthetic negative controls (SNCs) are artificial columns *which are correlated* to  $X$ , yet still have zero (population) regression coefficients
- Suppose I give you real data + SNCs, then you apply high-dimensional inference. If you reject any SNCs, we know these are errors!
- This gives us some measure of performance on “real” data (maybe?)



# Synthetic Negative Controls

- Given random vector  $x \in \mathbb{R}^p$ , let  $e$  be noise in  $\mathbb{R}^p$  independent of  $x$ .
- Let  $\Gamma$  be a fixed  $p \times q$  matrix. *Define* synthetic negative controls  $z \in \mathbb{R}^q$  by

$$z = x'\Gamma + e$$

and let  $\tilde{x} = (x, z)$ , so that

$$\tilde{x}_1 = x_1, \dots, \tilde{x}_p = x_p$$

$$\tilde{x}_{p+1} = z_1, \dots, \tilde{x}_{p+q} = z_q$$

# Synthetic Negative Controls

- Given random vector  $x \in \mathbb{R}^p$ , let  $e$  be noise in  $\mathbb{R}^p$  independent of  $x$ .
- Let  $\Gamma$  be a fixed  $p \times q$  matrix. Define synthetic negative controls  $z \in \mathbb{R}^q$  by

$$z = x'\Gamma + e$$

and let  $\tilde{x} = (x, z)$ , so that

$$\tilde{x}_1 = x_1, \dots, \tilde{x}_p = x_p$$

$$\tilde{x}_{p+1} = z_1, \dots, \tilde{x}_{p+q} = z_q$$

- 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  $x \in \mathbb{R}^p$ , let  $e$  be noise in  $\mathbb{R}^p$  independent of  $x$ .
- Let  $\Gamma$  be a fixed  $p \times q$  matrix. Define synthetic negative controls  $z \in \mathbb{R}^q$  by

$$z = x'\Gamma + e$$

and let  $\tilde{x} = (x, z)$ , so that

$$\tilde{x}_1 = x_1, \dots, \tilde{x}_p = x_p$$

$$\tilde{x}_{p+1} = z_1, \dots, \tilde{x}_{p+q} = z_q$$

- 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

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

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

# Why is this...?

- Recall that  $\hat{\beta}_i$  is the *univariate regression* coefficient of  $Y$  on  $X_{i|-i}$ , where  $X_{i|-i}$  is the *residual* of  $X_i$  after  $X_i$  is regressed on the other columns..
- Population version:  $\beta_i = 0$  if the projection of  $X_i$  on the null space of the other covariates is uncorrelated with  $Y$



# Why is this...?

- Population version:  $\beta_i = 0$  if the projection of  $X_i$  on the null space of the other covariates is uncorrelated with  $Y$
- For  $i = 1, \dots, q$ , we have

$$\tilde{X}_{p+i} = x' \Gamma_i + E_i$$

where here  $\tilde{X}_{p+1}$  denotes the random variable (not the column of the design matrix)

# Why is this...?

- Population version:  $\beta_i = 0$  if the projection of  $X_i$  on the null space of the other covariates is uncorrelated with  $Y$
- For  $i = 1, \dots, q$ , we have

$$\tilde{X}_{p+i} = x' \Gamma_i + E_i$$

where here  $\tilde{X}_{p+1}$  denotes the random variable (not the column of the design matrix)

- The orthogonal projection  $P_X^\perp$  of  $\tilde{X}_{p+1}$  is

$$P_X^\perp \tilde{X} = P_X^\perp X \Gamma_i + P_X^\perp E_i = 0 + E_i$$

since  $P_X^\perp X = 0$ ; meanwhile since  $E_i \perp X$ ,  $P_X^\perp E_i = E_i$ .

# Why is this...?

- Population version:  $\beta_i = 0$  if the projection of  $X_i$  on the null space of the other covariates is uncorrelated with  $Y$

- The orthogonal projection  $P_X^\perp$  of  $\tilde{X}_{p+1}$  is

$$P_X^\perp \tilde{X} = P_X^\perp X \Gamma_i + P_X^\perp E_i = 0 + E_i$$

since  $P_X^\perp X = 0$ ; meanwhile since  $E_i \perp X$ ,  $P_X^\perp E_i = E_i$ .

- Since  $E_i \perp y$ , we have  $\text{Cor}(P_X^\perp \tilde{X}_{p+i}, y) = 0$ , hence  $\tilde{\beta}_{p+i} = 0$

# Why is this...?

- Population version:  $\beta_i = 0$  if the projection of  $X_i$  on the null space of the other covariates is uncorrelated with  $Y$
- Since  $E_i \perp y$ , we have  $\text{Cor}(P_X^\perp \tilde{X}_{p+i}, y) = 0$ , hence  $\tilde{\beta}_{p+i} = 0$
- And since  $\tilde{\beta}_j = 0$  for all the added variables  $j = p+1, \dots, p+q$ , it follows that  $\tilde{\beta}_i$  is unchanged for  $i = 1, \dots, p$ .

# Using SNCs to evaluate procedures

- Take low-dimensional real data mixed with SNCs (synthetic negative controls), apply inference procedure
- *Proxy for Type I error*: Rejected SNCs
- *Proxy for Power*: Rejected original variables

# A step-by-step tutorial (in R)

## 1. Take the prostate data

```
> data(prostate)
> x <- prostate[, 1:8]
> y <- prostate[, 9]
> colnames(x)
[1] "lcavol" "lweight" "age"      "lbph"      "svi"
     "lcp"    "gleason" "pgg45"
> dim(x)
[1] 97 8
```

# A step-by-step tutorial

## 2. Construct 20 synthetic negative controls

```
> GAMMA <- matrix(rnorm(8 * 20), 8, 20)
> E <- matrix(rnorm(97 * 20), 97, 20)
> sncs <- as.matrix(x) %*% GAMMA + 2 * E
> sncs <- data.frame(sncs)
> colnames(sncs)
[1] "X1"  "X2"  "X3"  "X4"  "X5"  "X6"  ...
[19] "X19" "X20"
```

## 3. Create combined design matrix

```
> x2 <- cbind(x, sncs)
```

# A step-by-step tutorial

## 4. Try marginal screening

```
> cors <- cor(x2, y)
> cors[order(-abs(cors)), , drop = F]
      [,1]
lcavol  0.7344603
svi      0.5662182
lcp      0.5488132
X6       -0.4591506
X16      0.4482263
lweight  0.4333194
X4       -0.4326898
```



# A step-by-step tutorial

## 5. Try covariance test

```
> library(covTest)
> covTest(lars(as.matrix(x2), y), as.matrix(x2), y)
$results
```

| Predictor_Number | Drop_in_covariance | P-value |
|------------------|--------------------|---------|
| 1                | 69.0292            | 0.0000  |
| 5                | 1.5390             | 0.2219  |
| 2                | 6.8094             | 0.0020  |
| 11               | 0.8559             | 0.4294  |

(Numbers 1, 5, 2 are original, 11 is a SNC)

# A step-by-step tutorial

6. Try debiased lasso (code at <http://web.stanford.edu/~montanar/sslasso/>)

```
> res <- SSLasso(as.matrix(x2), y)
[1] "10% done"
...
[1] "90% done"
> rej <- (res$up < 0) | (res$low > 0)
> names(x2)[rej]
[1] "lcavol" "lweight" "svi"
```

# A step-by-step tutorial

## 7. Try knockoffs

```
> library(knockoff)
```

```
> knockoff.filter(x2, y)
```

Call:

```
knockoff.filter(X = x2, y = y)
```

Selected variables:

|         |    |
|---------|----|
| lweight | X7 |
| 2       | 15 |

# Disclaimer!

- I am *not* proposing SNCs as a methodology for *inference*
- There is a danger of inferring *that Type I error has been controlled* from lack of rejection of SNCs. There are no formal guarantees of this!
- One should interpret results from experiments with SNCs in the same way one interprets simulation results with purely synthetic data

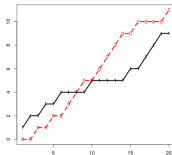
# More Experiments!

| Data        | $n$  | $p_1$ | Linear? | Gaussian? | Constant $\sigma^2$ ? |
|-------------|------|-------|---------|-----------|-----------------------|
| Personality | 3000 | 163   | No      | No        | No                    |
| fMRI        | 1750 | 53    | No      | OK        | No                    |
| HIV         | 842  | 207   | No      | Yes?      | OK?                   |
| Galaxy      | 323  | 4     | No      | OK        | No                    |

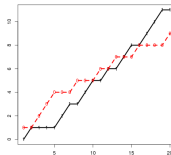
- We add  $n/2 - p_1$  synthetic negative controls
- $X$  is scaled,  $\Gamma$  is a gaussian matrix,  $\text{Var}(E)$  is chosen to yield 'interesting' results
- Personality data is subsampled

# Marginal Screening

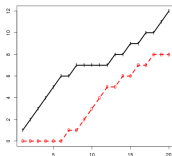
Personality



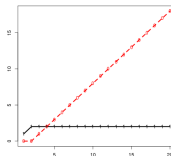
fMRI



HIV



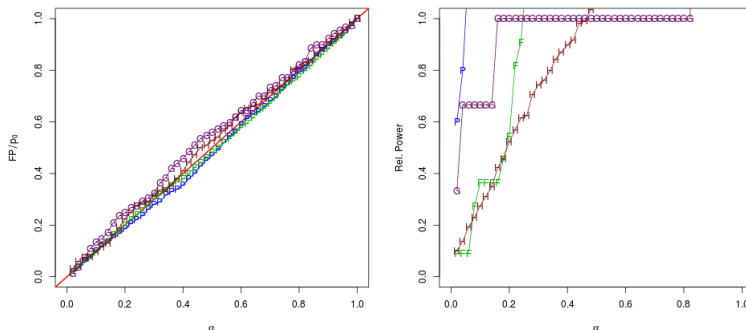
Galaxy



Legend: 0 = False positives, 1 = True positives

# Ordinary Least Squares

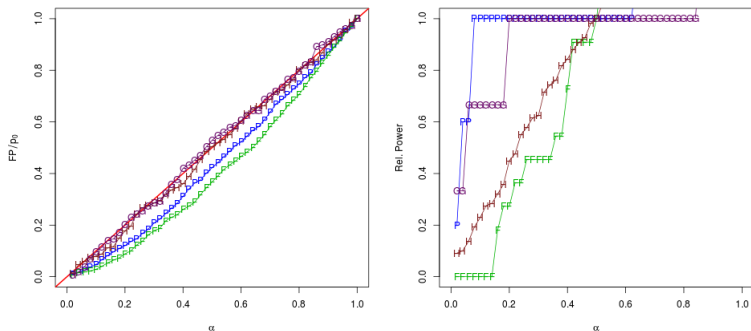
“Rel. power” =  $TP / (\text{max number of TPs at } \alpha = 0.5 \text{ for any method})$



Legend: **P** = Personality, **F** = fMRI, **H** = HIV, **G** = Galaxy

# Debiased Lasso

Can you spot the difference from the previous slide?

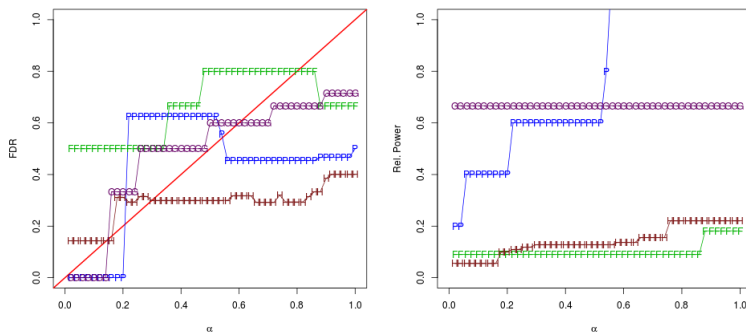


Legend: **P** = Personality, **F** = fMRI, **H** = HIV, **G** = Galaxy



# Covariance Test

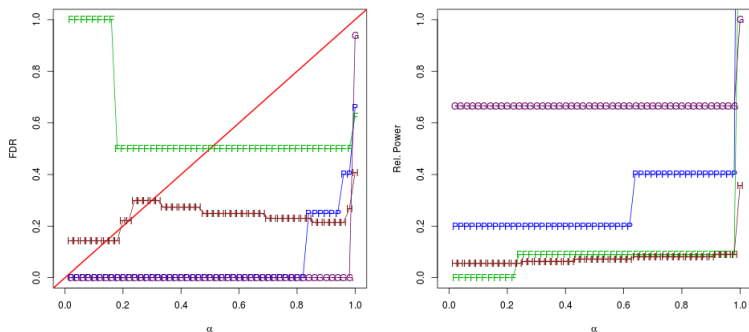
*Forward Stop:* reject first  $\hat{k}$ , where  $-\frac{1}{\hat{k}} \sum_{i=1}^{\hat{k}} \log(1 - p_i) \leq \alpha$



Legend: **P** = Personality, **F** = fMRI, **H** = HIV, **G** = Galaxy

# Covariance Test

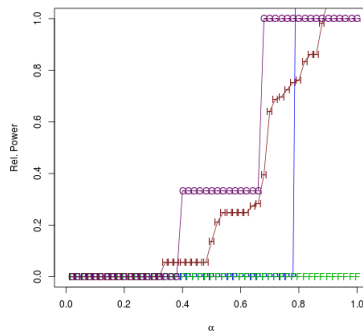
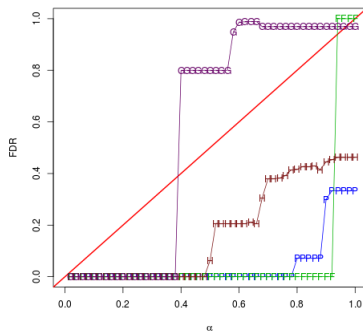
*Strong Stop*: reject first  $\hat{k}$ , where  $\frac{m}{\hat{k}} e^{\sum_{j=\hat{k}}^p \log(p_j)/j} \leq \alpha$



Legend: P = Personality, F = fMRI, H = HIV, G = Galaxy

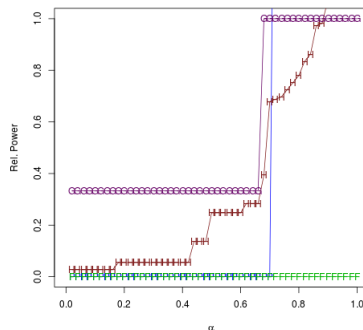
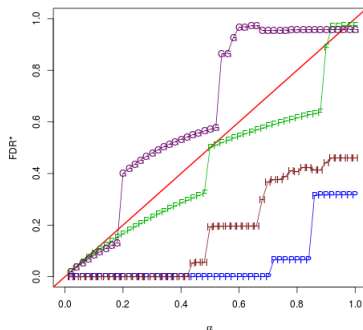
# Knockoffs

## Using Knockoff+ threshold



Legend: **P** = Personality, **F** = fMRI, **H** = HIV, **G** = Galaxy

Note:  $FDR^* = \mathbf{E}[FP/(FP + TP + 1/\alpha)]$

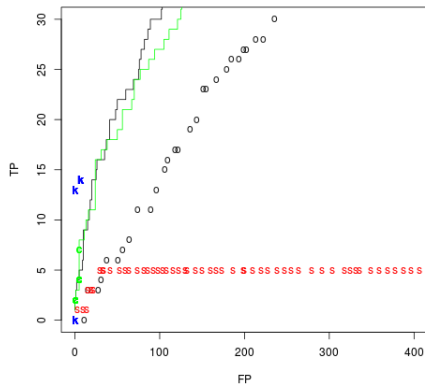


Legend: **P** = Personality, **F** = fMRI, **H** = HIV, **G** = Galaxy

# Variable Ranking Criteria

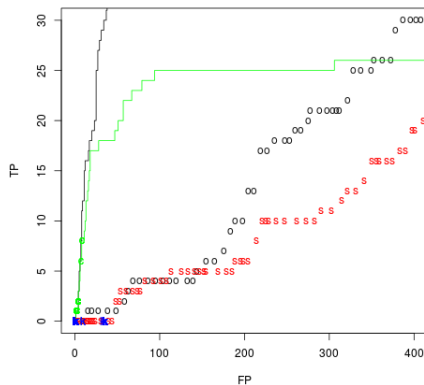
- Forget about Type I error for a second...
- Use procedures to *rank* variables by p-value
- Easy to compare procedures with different Type I criteria and also non-inference variable selection
- (Optional) score by Area Under Curve (AUC), etc.

# Variable Ranking: Personality



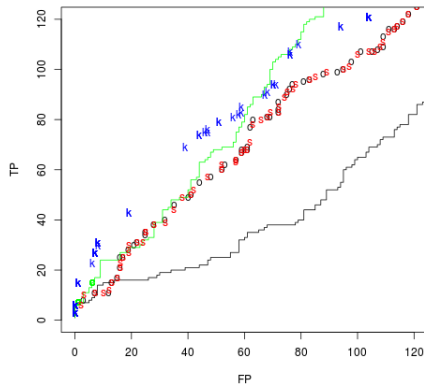
Legend: o = OLS, c = covariance test, k = knockoff, s = debiased lasso, (line) = marginal screening, (line) = lasso path

# Variable Ranking: fMRI



Legend: o = OLS, c = covariance test, k = knockoff, s = debiased lasso, (line) = marginal screening, (line) = lasso path

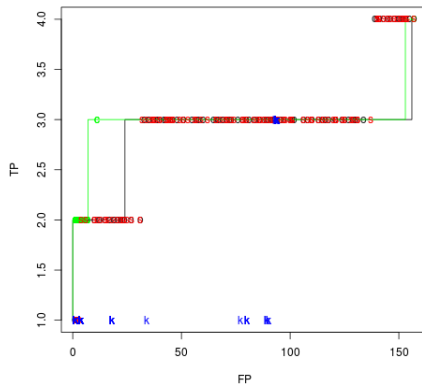
# Variable Ranking: HIV



Legend: o = OLS, c = covariance test, k = knockoff, s = debiased lasso, (line) = marginal screening, (line) = lasso path



# Variable Ranking: Galaxy

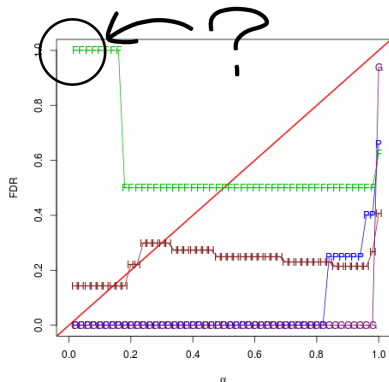


Legend: o = OLS, c = covariance test, k = knockoff, s = debiased lasso, (line) = marginal screening, (line) = lasso path

- We should not conclude too much from four experiments with rather arbitrary generation parameters...
- Debiased lasso similar to OLS but more conservative, less powerful
- Knockoffs vs covariance test:
  - Knockoffs may control FDR more robustly than Covariance test (especially at small  $\alpha$ )
  - Knockoffs and covariance are similar in power overall but have different case-by-case behavior
- Knockoffs tend to be conservative, but have good variable ranking in some cases (Personality, fMRI)
- Marginal screening remains annoyingly effective...

# Another look at covTest

*Strong stop type I error*

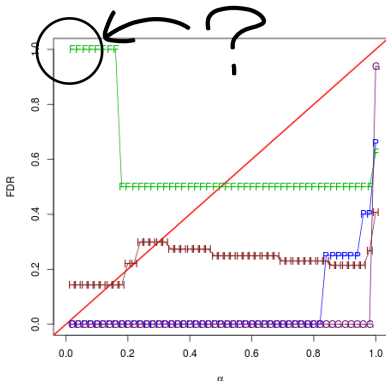


- Why did this negative control get rejected in the fMRI data at such a low  $\alpha$ ?

(thanks to Stefan!)

# Another look at covTest

*Strong stop type I error*

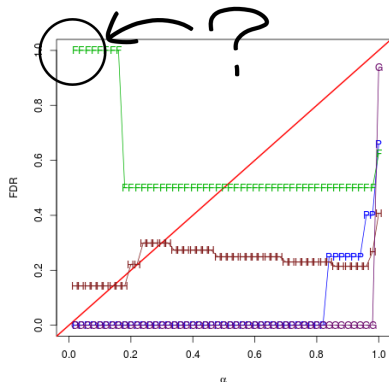


- According to the *incremental null*, the negative control was not a mistake... it is the best single predictor by far!
- The particular negative control took an *average* of the original columns

(thanks to Stefan!)

# Another look at covTest

*Strong stop type I error*



- It is true that  $\beta_i = 0$  for the rejected variable... but should we really disregard such strong "proxy" variables?

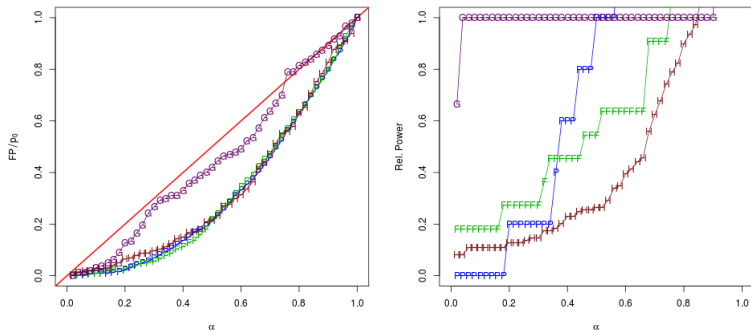
(thanks to Stefan!)

# Low sample size

| Data        | $n$ | $p$  | $p_1$ | Linear? | Gaussian? | Constant $\sigma^2$ ? |
|-------------|-----|------|-------|---------|-----------|-----------------------|
| Personality | 100 | 1500 | 163   | No      | No        | No                    |
| fMRI        | 100 | 875  | 53    | No      | OK        | No                    |
| HIV         | 100 | 421  | 207   | No      | Yes?      | OK?                   |
| Galaxy      | 100 | 161  | 4     | No      | OK        | No                    |

- Reduce the sample size to 100, so that  $p \gg n$
- Same number of negative controls, but larger added noise (easier)
- Covariance test requires estimate  $\hat{\sigma}$ : “cheat” by using OLS estimate from *original* data

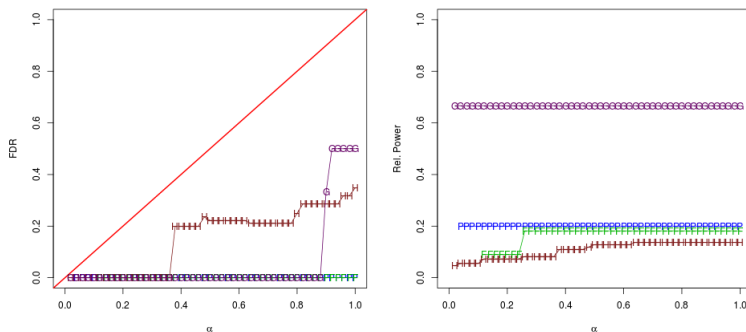
# Debiased Lasso



Legend: **P** = Personality, **F** = fMRI, **H** = HIV, **G** = Galaxy

# Covariance Test

*Forward Stop:* reject first  $\hat{k}$ , where  $-\frac{1}{\hat{k}} \sum_{i=1}^{\hat{k}} \log(1 - p_i) \leq \alpha$

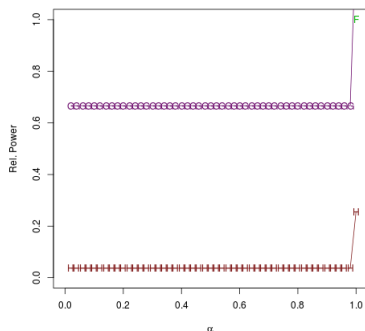
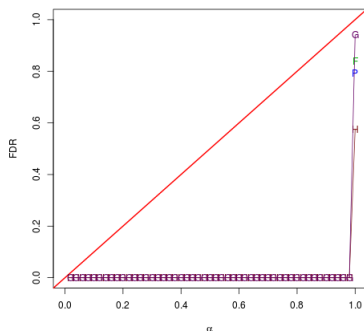


Legend: **P** = Personality, **F** = fMRI, **H** = HIV, **G** = Galaxy



# Covariance Test

*Strong Stop*: reject first  $\hat{k}$ , where  $\frac{m}{\hat{k}} e^{\sum_{j=\hat{k}}^p \log(p_j)/j} \leq \alpha$



Legend: **P** = Personality, **F** = fMRI, **H** = HIV, **G** = Galaxy

- Debiased lasso and Covariance test + forward stop continue to control Type I error while finding true positives
- Covariance test + strong stop appears too conservative

# Are these experiments actually that realistic!?

- Here we are implicitly assuming that real data always consists of a few “active variables” and many null variables
- If that’s true, it seems reasonable to model the distribution of the inactive variables conditional on knowing a superset of the active variables

# Are these experiments actually that realistic!?

- Here we are implicitly assuming that real data always consists of a few “active variables” and many null variables
- If that’s true, it seems reasonable to model the distribution of the inactive variables conditional on knowing a superset of the active variables
- But how do we know that  $\beta$  is really sparse?
- Even more importantly, how do we know that inferring coefficients of the *full model*  $\beta$  are a meaningful objective? E.g. should we consider the covariance test to have made a mistake in the fMRI data?

# Are these experiments actually that realistic!?

- Here we are implicitly assuming that real data always consists of a few “active variables” and many null variables
- If that’s true, it seems reasonable to model the distribution of the inactive variables conditional on knowing a superset of the active variables
- But how do we know that  $\beta$  is really sparse?
- Even more importantly, how do we know that inferring coefficients of the *full model*  $\beta$  are a meaningful objective? E.g. should we consider the covariance test to have made a mistake in the fMRI data?
- How can we decide between the *full model null*, the *incremental null*, or an entirely different framework altogether?
- *Feedback from the practitioner* is the only way we can tell if we have the right formulation for any particular application

# Questions to consider

- Why is OLS more powerful than lasso in some of these experiments even when  $\beta$  is sparse? Look at covariance conditions in the theory of LASSO...
- Why do knockoffs or lasso beat marginal screening/OLS in the HIV data? Was it due to how we generated the SNCs or is due to something special about the data itself?
- Suppose we wanted to validate selective inference or the incremental null. How can we do this with synthetic negative controls (other than pure noise?)

*“ Both the client and the statistician... must base their thinking on a recognition that their assumptions will always require review and reappraisal... ”*

– John Tukey

# Acknowledgements

Thanks to Will Fithian and Stefan Wager for useful discussions.



- 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>
- Gsell, Max Grazier. Wager, Stefan Chouldechova, Alexandra. Tibshirani, Robert. Sequential Selection Procedures and False Discovery Rate Control. (2013): 31. Web. 7 May 2015.
- 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