A practical evaluation of recent methods in high-dimensional inference

Charles Zheng

Stanford University

May 5, 2015

Problem and motivation

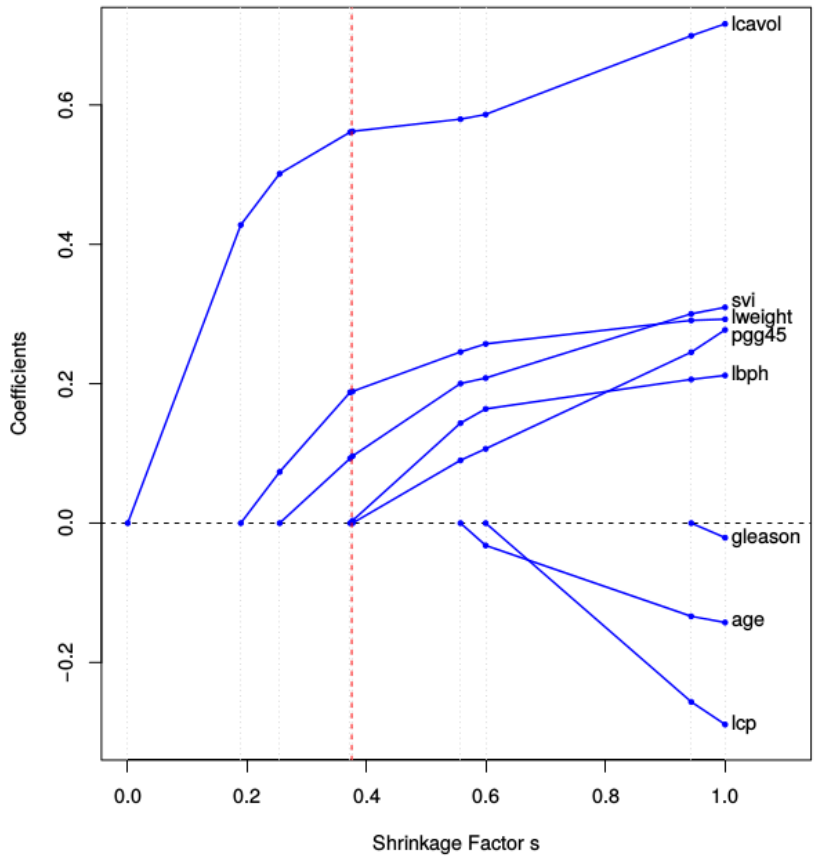
- $x \in \mathbb{R}^p, y \in \mathbb{R}$ have a joint distribution P where $y|x \sim N(x^T\beta, \sigma^2)$
- Observe $X = (x_1, \ldots, x_n)^T$, $Y = (y_1, \ldots, y_n)$ iid
- Problem: test $H_i : \beta_0 = i$ for $i = 1, \ldots, p$
- Motivation: x are SNPs (mutations), y is phenotype

Methods

	Control	$p > n$
Classical inference (Pearson 1930)	Marginal	No
Covariance test (Lockhart et al. 2014)	FWER?	Yes
Debiased lasso (Javanmard et al. 2014)	Marginal	Yes
Knockoffs (Barber et al. 2014)	FDR	?

The LASSO path

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



(Image credit: ??)

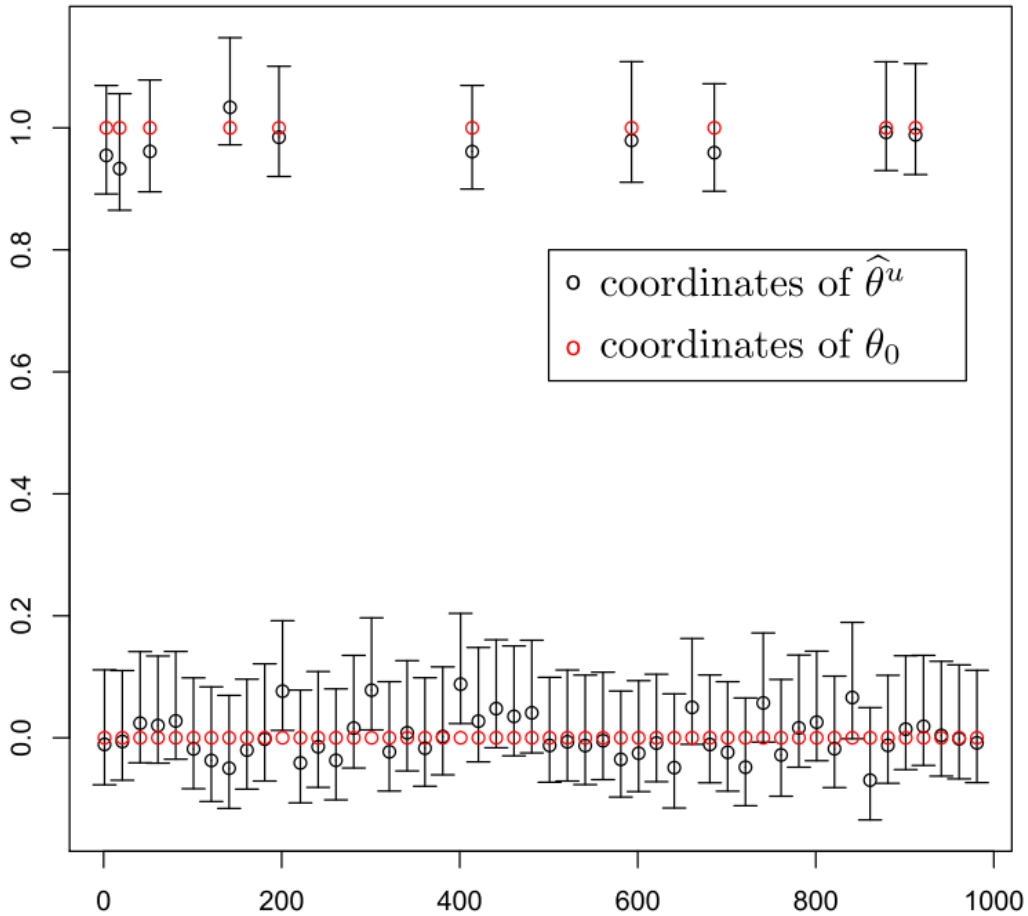
Covariance test

- (2014) Lockhart, Taylor, Tibshirani (x 2)
- Standard assumptions $Y \sim N(X\beta, \sigma^2 I)$ + large p asymptotics
- *See also* non-asymptotic exact test (Lee, Sun x 2, Taylor 2015)

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

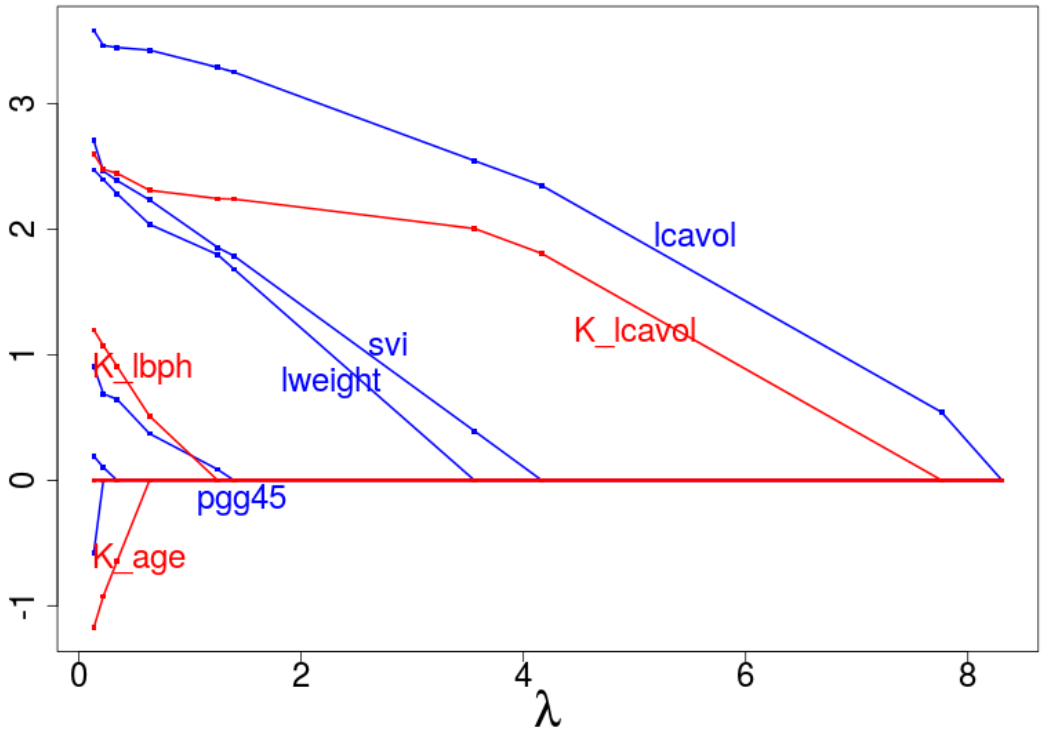
Debiased regularized M-estimators

- (2014) Javanmard and Montanari
- Standard assumptions + sparsity condition on β + 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
```

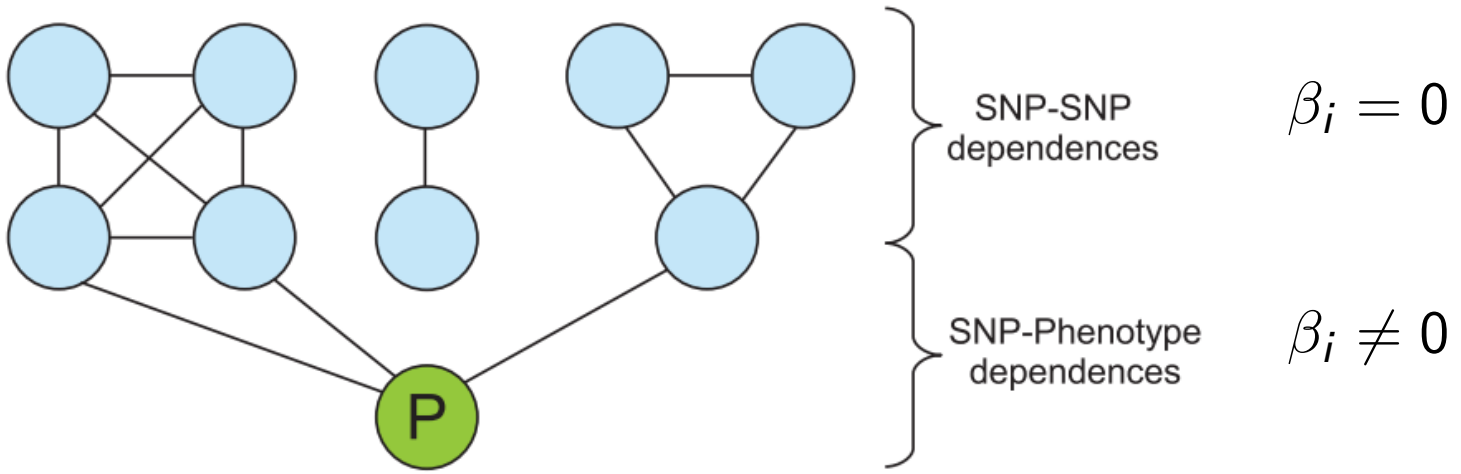
Methods

But what's actually used in practice?

	Control	$p > n$
Classical inference (Pearson 1930)	Marginal	No
Covariance test (Lockhart et al. 2014)	FWER?	Yes
Debiased lasso (Javanmard et al. 2014)	Marginal	Yes
Knockoffs (Barber et al. 2014)	FDR	?
Marginal screening	???	Yes

Regression vs Marginal Screening

Testing $H_i : \beta_i = 0$ is better than testing $H_i : \text{Cov}(X_i, Y) = 0$ when you are looking for X_i *directly* linked to Y



(Adapted from *Mourad 2012*)

Practical Validation

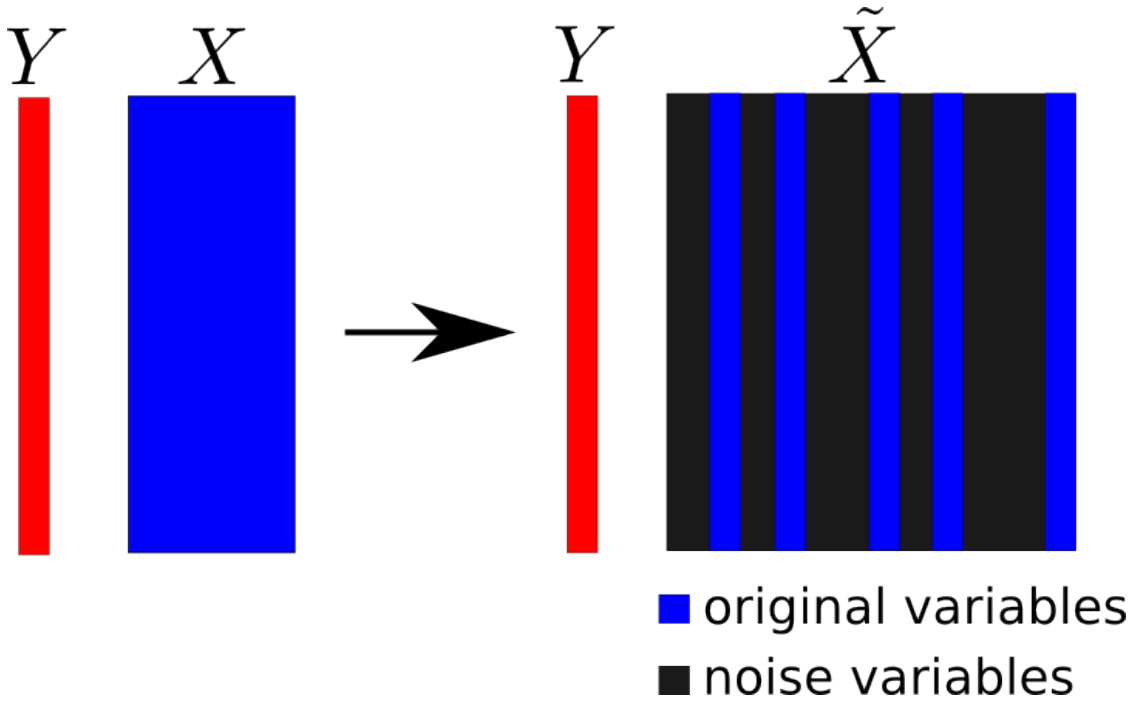
- These procedures are derived under strong assumptions (and slightly different model, fixed X), *how well do they work in real data*?
- We could validate inference procedures in real data if only we knew the '*true*' β, defined as

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

- Possibility: take a dataset with large n *and* large p (so we can estimate β easily using OLS) and test procedure on a subset $n_0 << n$ of the data
- Or...

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 Γ 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 Γ 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 Γ 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 Γ 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 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
- If your original data is high-dimensional, apply variable selection to make it low dimensional before conducting this experiment

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
```

Experiments, part 1

Data	n	p_1	Linear?	Gaussian?	Constant σ^2?
Personality	49k	163	No	No	No
fMRI	1750	44	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, Γ is a gaussian matrix, $Var(E) = Var(X\Gamma)$
- Multiple trials averaging over the randomness of generating SNCs

How could this be useful?

- Poor performance on benchmarks would tell us where our methods need improvement
  - Failure to control Type I error on benchmarks indicates a need for methods derived under weaker assumptions
  - Overly conservative Type I error control indicates a need for methods which are more adaptive to 'easy' cases
- Possible to run a Kaggle-style competition for *inference* rather than prediction
- Recognizing that different procedures can have differing strengths creates room for a diversity of approaches

Do we still need to validate on real data?

- SNCs can be used to get an idea of worst-case performance on the *hypothesis testing problem* in realistic settings
- However, how can we tell if the regression framework itself is appropriatefor the real-world problem we are trying to solve?
- Validation on real data with *scientific* ground truth is still needed

Closing thoughts

" Both the client and the statistician... 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.