False inclusion rates

Patrick Breheny

April 11

Where we're at and where we're going

- At this point, we've covered the most widely used approaches to fitting penalized regression models in the standard setting
- The remainder of the course will focus on:
  - Inference for $\boldsymbol{\beta}$
  - Other models, such as logistic regression and Cox regression
  - Other covariate structures, such as grouping and fusion
- We'll begin with inference

Inferential questions

- Up until this point, our inference has been restricted to the predictive ability of the model (which we can obtain via cross-validation)
- This is useful, of course, but we would also like to be able to ask the questions:
  - How reliable are the selections made by the model? What is its false discovery rate?
  - How accurate are the estimates yielded by the model? Can we obtain confidence intervals for β? Can we obtain confidence intervals for only the selected elements of β?

Overview

- As I've remarked previously, little progress was made on these questions until relatively recently, and the field is still very much unsettled as far as a consensus on how to proceed with inference
- Broadly speaking, I would classify the proposed approaches into four major categories:
  - Debiasing
  - False inclusion rates
  - Sample splitting/resampling
  - Selective inference

Debiasing

- We have already seen the idea of *debiasing* in the semi-penalized approach Jian discussed last week
- The basic idea behind debiasing is that frequentist inference tends to work well if $\widehat{\beta}_j \dot{\sim} \mathrm{N}(\beta_j, \mathrm{SE}^2)$
- Penalized regression estimates obviously do not have this property (with the possible exception of MCP/SCAD), so debiasing approaches attempt to construct an estimate $\tilde{\beta}_j$, based on $\widehat{\boldsymbol{\beta}}$ in some way, for which approximate unbiased normality holds

Implementations

- We have already seen one way to accomplish this: simply set $\lambda_j = 0$ for β_j (SPIDR)
- Many other approaches along these lines have been proposed, instead using analytical means to develop a bias correction term:
  - Zhang and Zhang (2014)
  - Bühlmann (2013)
  - van de Geer et al. (2013)
  - Javanmard and Montanari (2014)
- It is worth noting that these ideas are not exactly inferential approaches for penalized regression estimates, but rather ways of using penalized regression estimates as starting points for high-dimensional inference

False inclusion rates: KKT conditions

- In contrast, false inclusion rates attempt to directly estimate the error rates for coefficients selected by penalized regression estimates
- Recall the KKT conditions for the lasso:

$$\frac{1}{n}\mathbf{x}_j'\mathbf{r} = \lambda \operatorname{sign}(\widehat{\beta}_j) \qquad \text{for all } \widehat{\beta}_j \neq 0$$

$$\frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}\right| \leq \lambda \qquad \text{for all } \widehat{\beta}_j = 0$$

KKT conditions in terms of partial residuals

- Letting $\mathbf{X}_{-j}$ and $\boldsymbol{\beta}_{-j}$ denote the portions of the design matrix and coefficient vector that remain after removing the jth feature, let $\mathbf{r}_j = \mathbf{y} - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_{-j}$ denote the partial residuals with respect to feature j
- The KKT conditions thus imply that

$$\frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}_j\right| > \lambda \qquad \text{for all } \widehat{\beta}_j \neq 0$$
$$\frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}_j\right| \leq \lambda \qquad \text{for all } \widehat{\beta}_j = 0$$

and therefore that the probability that variable j is selected is

$$\mathbb{P}\left(\frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}_j\right| > \lambda\right)$$

Orthogonal case

- This suggests that if we are able to characterize the distribution of $\frac{1}{n}\mathbf{x}_j'\mathbf{r}_j$ under the null, we can estimate the number of false selections in the model
- Indeed, this is easy to do in the case of orthonormal design ($\frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{I}$)
- **Theorem:** Suppose $\frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{I}$. Then for any value of λ,

$$\mathbb{E}\,|\mathcal{S} \cap \mathcal{N}| = 2\,|\mathcal{N}|\,\Phi(-\lambda\sqrt{n}/\sigma),$$

where $\mathcal{S}$ is the set of selected variables and $\mathcal{N}$ is the set of null variables (i.e., $\{j : \beta_j = 0\}$)

Estimation

- To use this as an estimate, two unknown quantities must be estimated
  - $|\mathcal{N}|$ can be replaced by p, using the total number of variables as a bound for the null variables
  - σ^2 can be estimated by $\mathbf{r}'\mathbf{r}/(n - |\mathcal{S}|)$ (this is the simplest approach, but other possibilities exist)
- This implies the following estimate for the expected number of false discoveries:

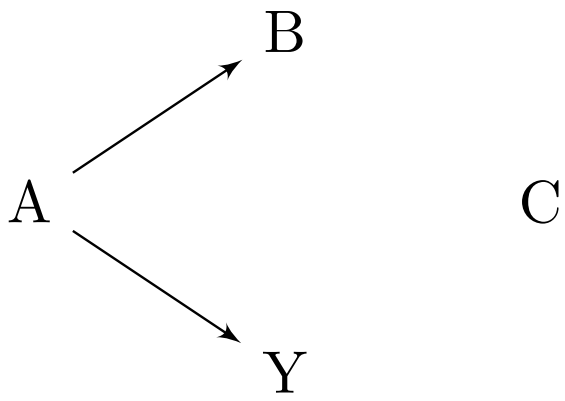
$$\widehat{\mathrm{FD}} = 2p\Phi(-\sqrt{n}\lambda/\hat{\sigma})$$

and, as an estimate of the false discovery rate:

$$\widehat{\mathrm{FDR}} = \frac{\widehat{\mathrm{FD}}}{S}$$

Causal diagram

- The case of correlated variables, however, is considerably more complex
- Consider the following causal diagram:



- Estimating the number of false selections arising from variables like B is challenging; however, simple approaches still work well for estimating false selections arising from variables like C

Advantages of this definition

- In this talk, I will define a *false inclusion* as a variable like C, that has no path (direct or indirect) between it and the outcome; this is in contrast to most other work, which consider any variable with $\beta_j = 0$ to be a false discovery
- This definition has several advantages:
  - When two variables (like A and B) are correlated, it is very challenging to distinguish between which of them (if either, or both) is driving changes in Y and which is merely correlated with Y
  - In many applications, discovering variables like B is not problematic
  - Whether or not a variable is a false inclusion is not conditional, and does not depend on λ

Independence among predictors

- The design matrix does not have to be strictly orthogonal in order for the proposed estimator to work
- **Theorem:** Suppose $\frac{1}{n}\mathbf{X}'\mathbf{X} \to \mathbf{I}$. Then for any $j : \beta_j = 0$ and any value of λ,

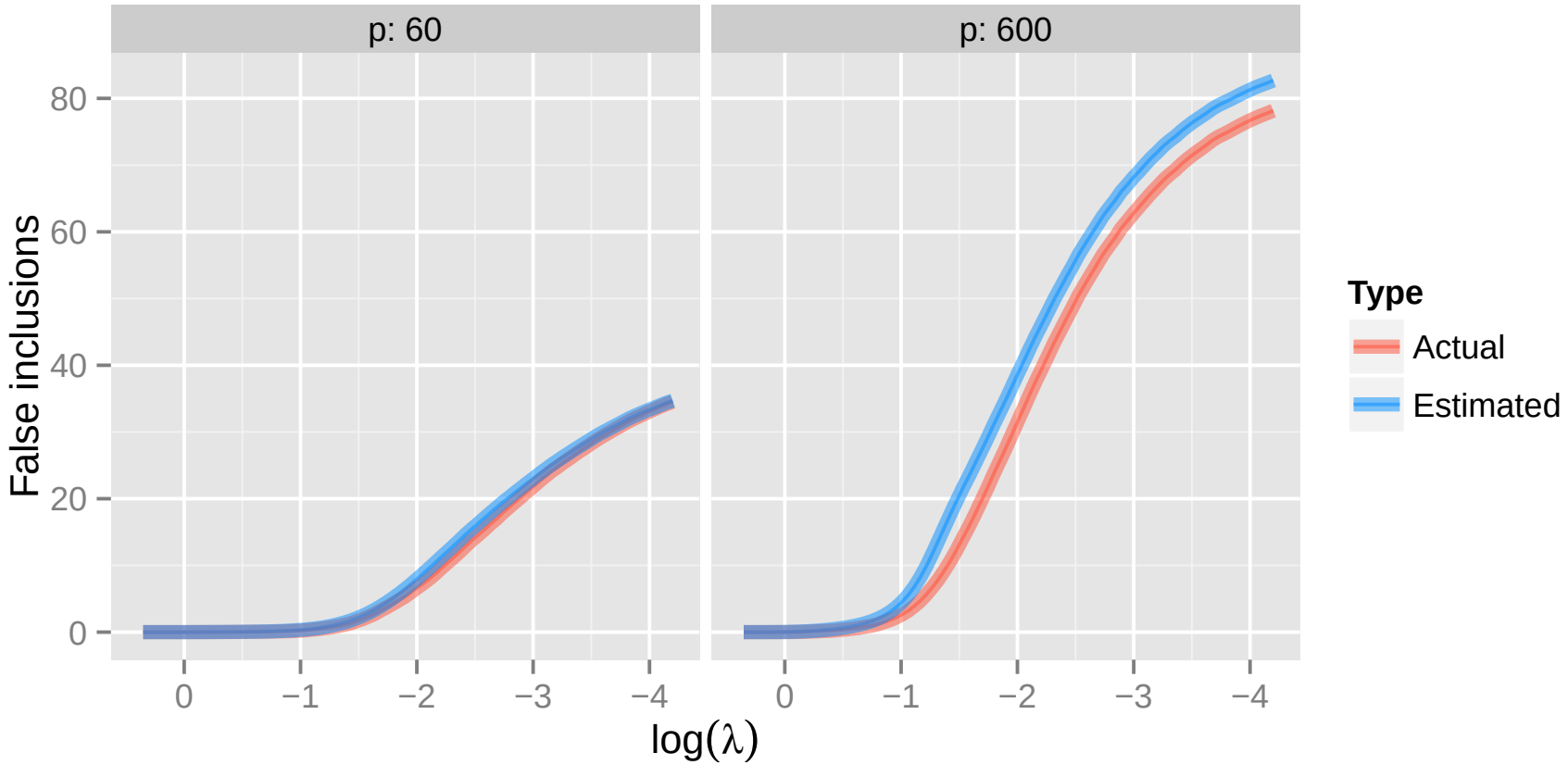
$$\frac{1}{\sqrt{n}}\mathbf{x}_j'\mathbf{r}_j \xrightarrow{\text{d}} N(0, \sigma^2)$$

- This can be relaxed to $\frac{1}{n}\mathbf{X}_{\mathcal{C}}'\mathbf{X}_{\mathcal{C}} \to \mathbf{I}$; i.e., only the variables in $\mathcal{C}$ need to be uncorrelated

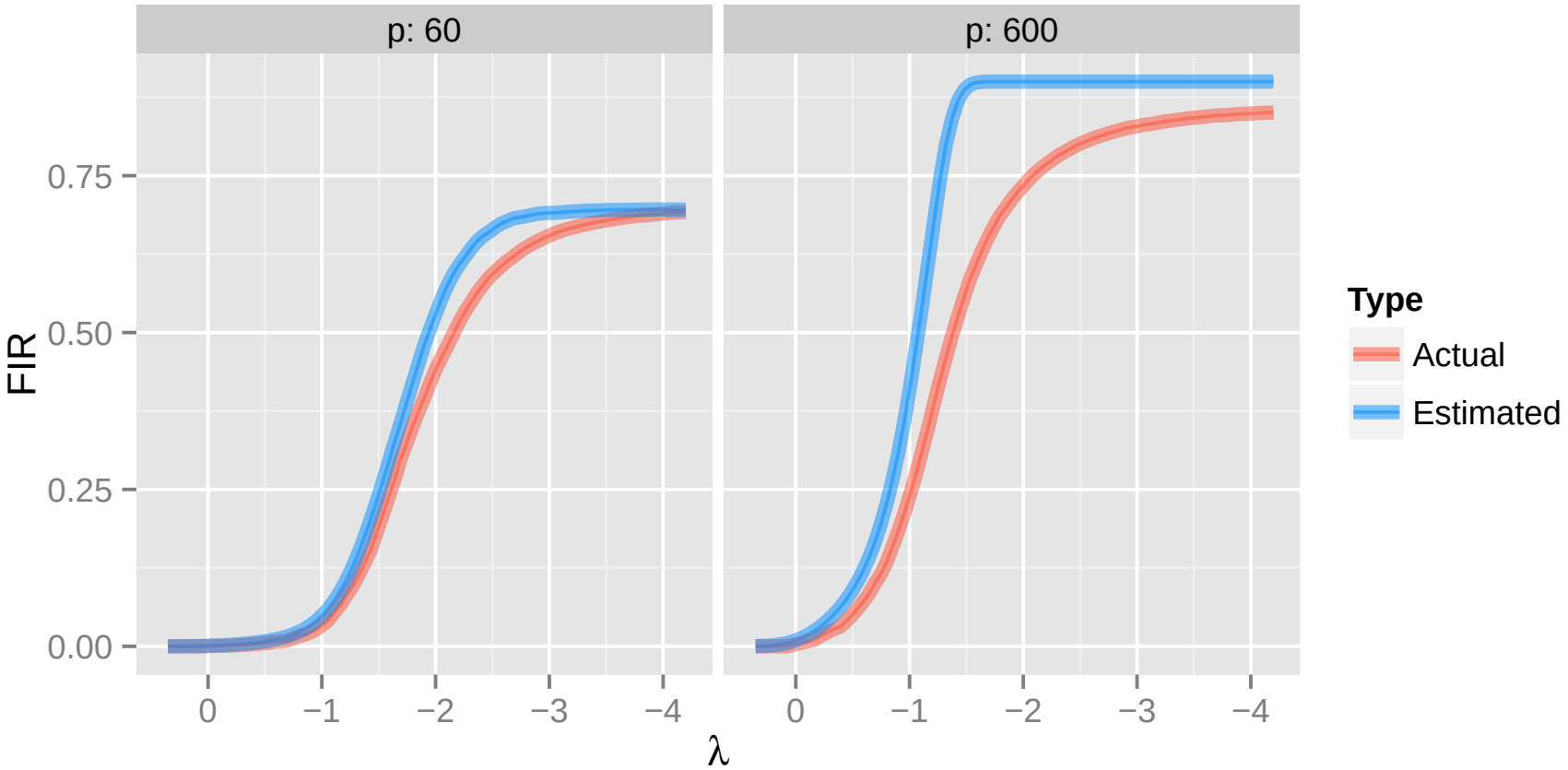
Design

- I conducted both "low-dimensional" ($n > p$) and "high-dimensional" ($n < p$) simulation studies, organized along the lines of the earlier causal diagram:
  - Six variables had $\beta_j \neq 0$ ("causative")
  - Each causative feature was correlated ($\rho = 0.5$) with m other features ("correlated"; $m = 2$ for low-dimension and 9 for high-dimensional)
  - Independent noise features ("spurious") were added to bring the total number of variables up to 60 in the low-dimensional case and 600 in the high-dimensional case
- In each setting, the sample size was $n = 100$, and the nonzero β values were set to ± 1
- Causative/Correlated/Spurious: 6/12/42 for low-dimensional, 6/54/540 for high-dimensional

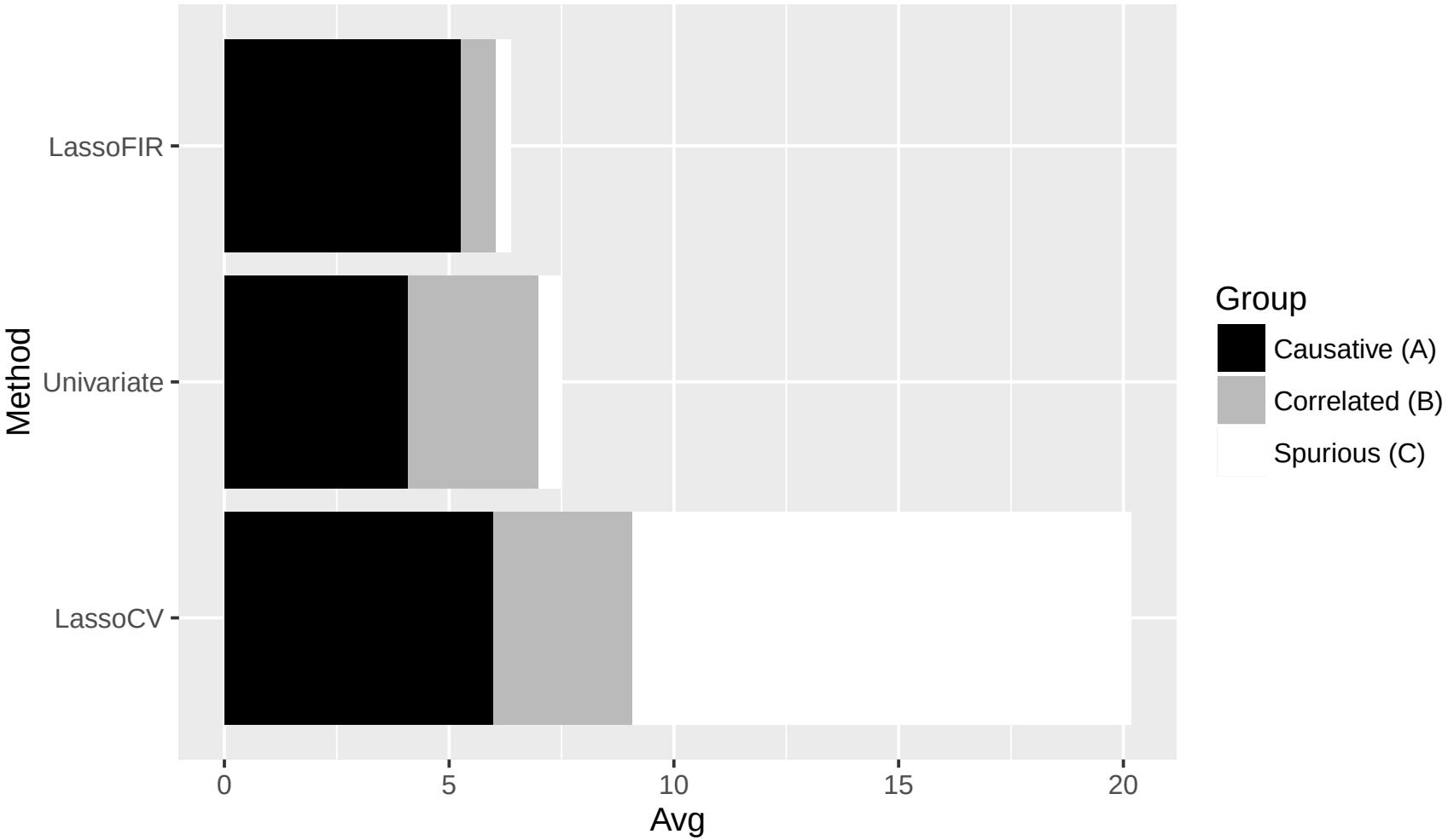
Accuracy of false inclusion estimates



Accuracy of false inclusion estimates



Comparison: High-dimensional (nominal FIR/FDR=0.1)



Correlated case

- The preceding results are something of a "best case scenario" for the proposed method, since the variables in $\mathcal{C}$ were independent
- When the null variables are dependent, the estimator becomes conservative
- **Conjecture:** Suppose $\frac{1}{n}\sum_i \mathbf{X}'\mathbf{X} = \Sigma$. Then for any λ,

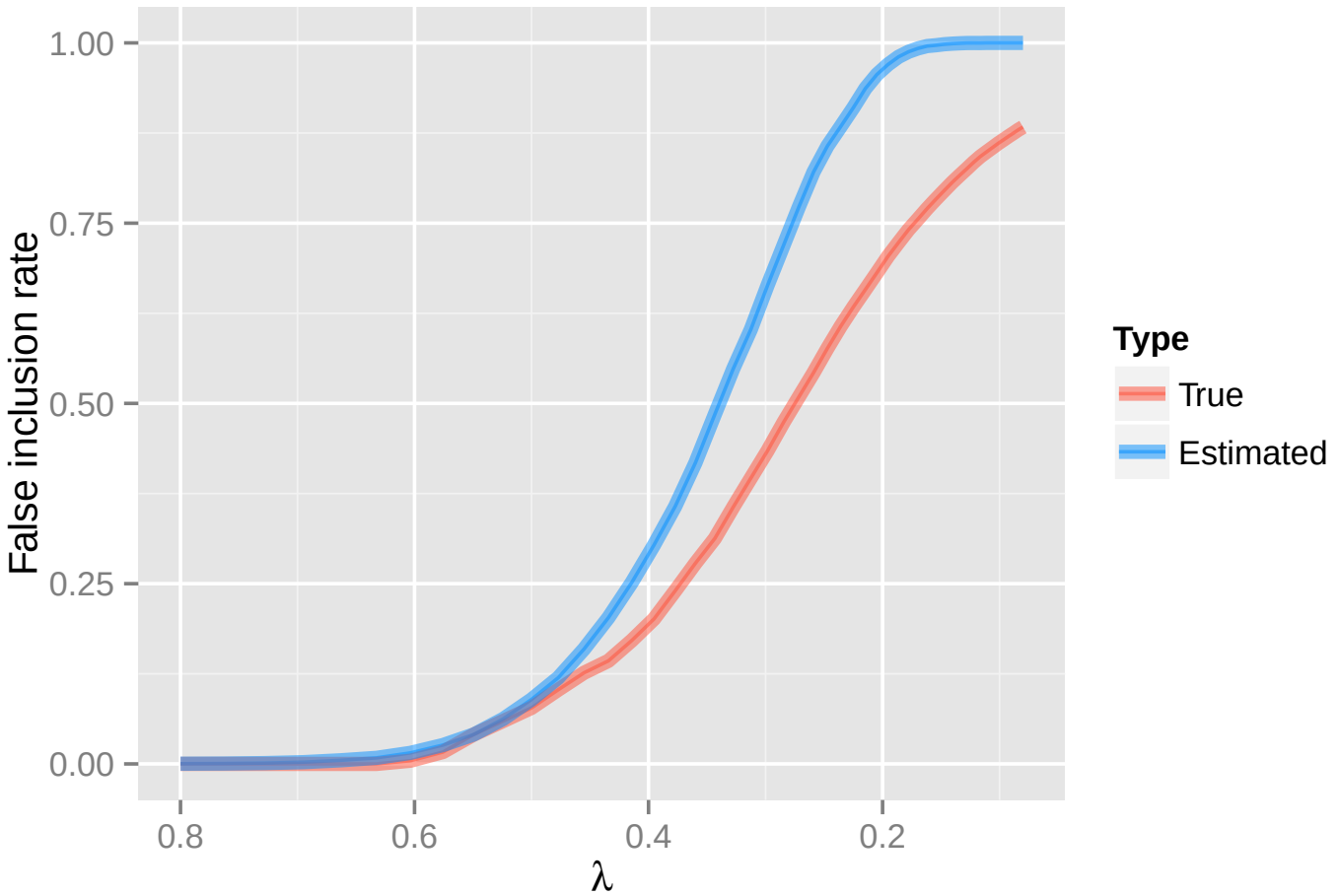
$$\mathbb{E}\,|\mathcal{S} \cap \mathcal{N}| \leq 2\,|\mathcal{N}|\,\Phi(-\lambda\sqrt{n}/\sigma)$$

- This can be shown for the case when $p = 2$, and is (my belief) likely to be true for any p and Σ

Autoregressive correlation

- I carried out the following simulation to investigate the robustness of the proposed FIR estimator in the presence of moderate correlation
- The generating model contains 6 causative features, and 494 correlated spurious features ($n = 100$, $p = 500$, $R^2 = 0.5$)
- The correlation structure on the spurious features was set to be $\mathrm{Cor}(X_i, X_j) = 0.8^{|i-j|}$

Results: Autoregressive correlation among $\mathcal{C}$

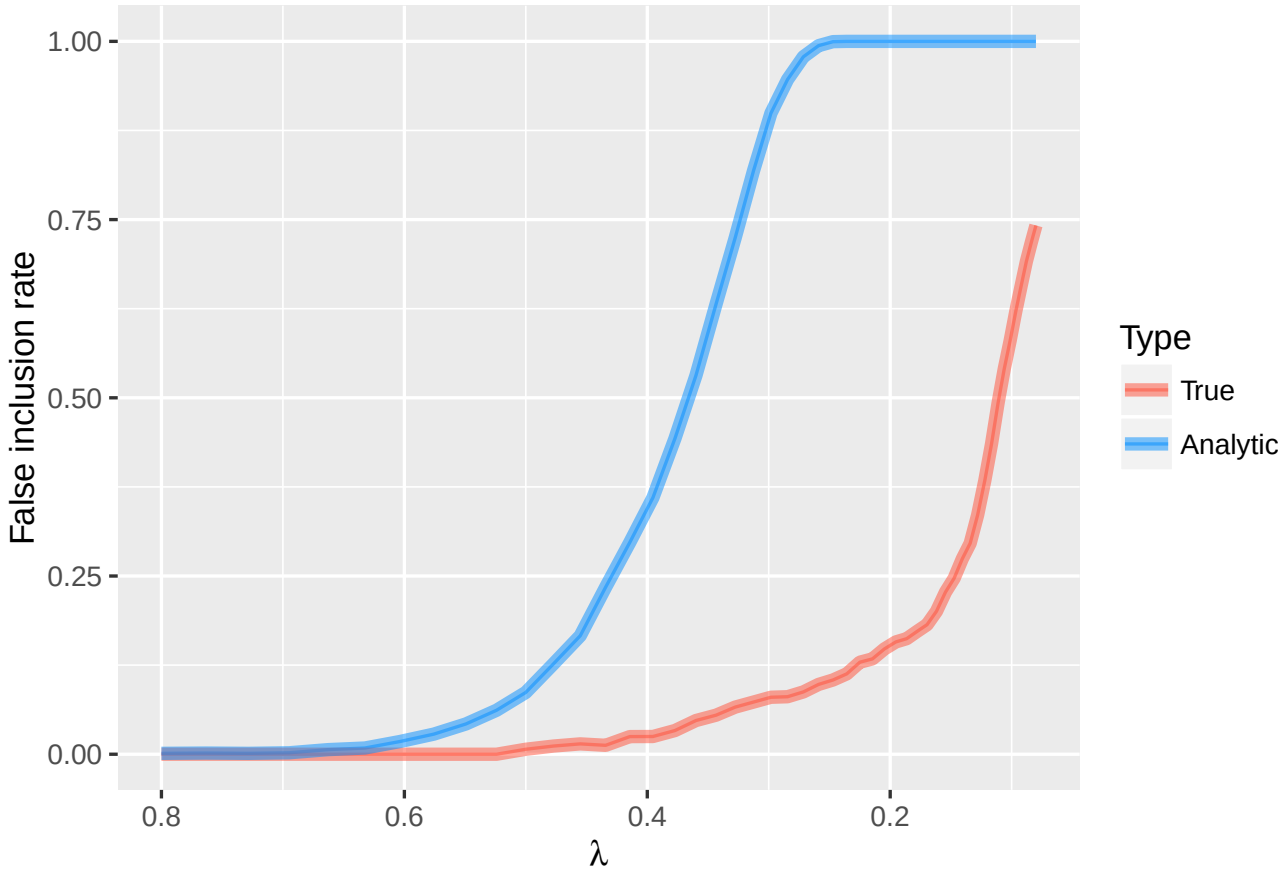


Still quite accurate, but slightly conservative

Exchangeable correlation

- It is worth mentioning that this conservatism becomes more extreme as the correlation becomes heavier
- Let us consider an extreme case: all variables in $\mathcal{C}$ have a pairwise correlation of $\rho = 0.8$ (otherwise, all settings are the same as before)

Results: Exchangeable correlation among $\mathcal{C}$



Much less accurate, although once again conservative

Breast cancer data

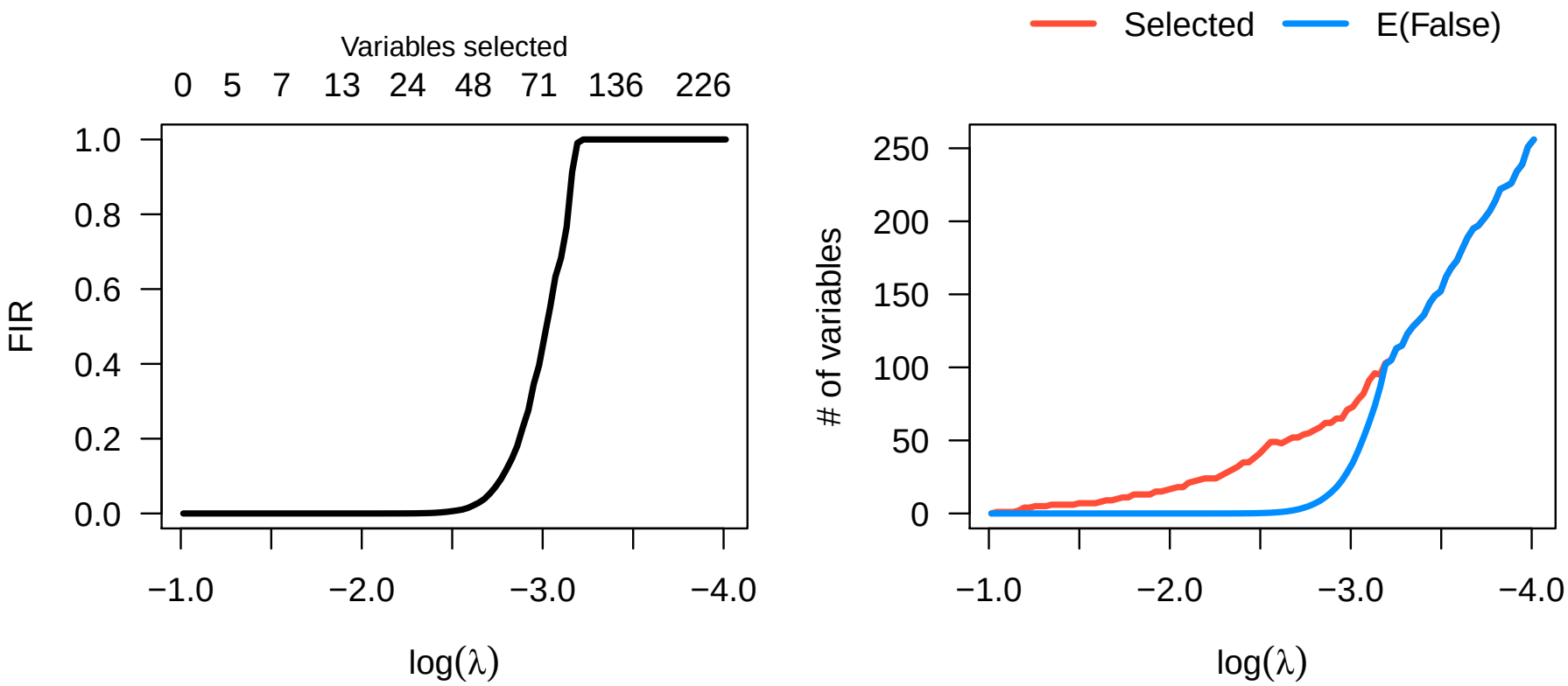
- To see how this works with real data, let's take a look at the breast cancer TCGA data ($n = 536$, $p = 17,322$)
- We can fit a lasso model with

```
fit <- ncvreg(X, y, penalty="lasso")
```

FIR plots

We can then calculate and plot false inclusion rates with

```
fir(fit)
plot(fir(fit))
```

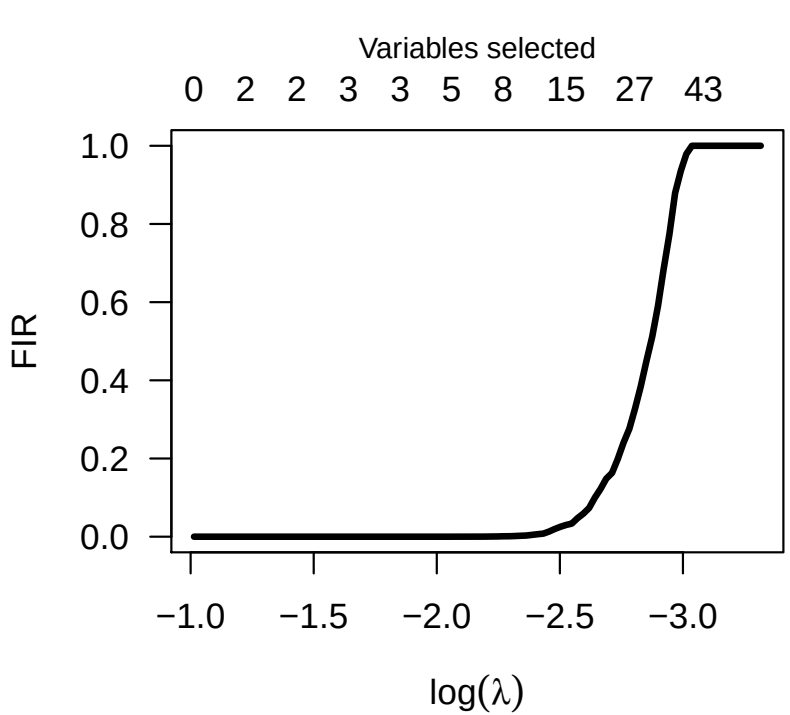


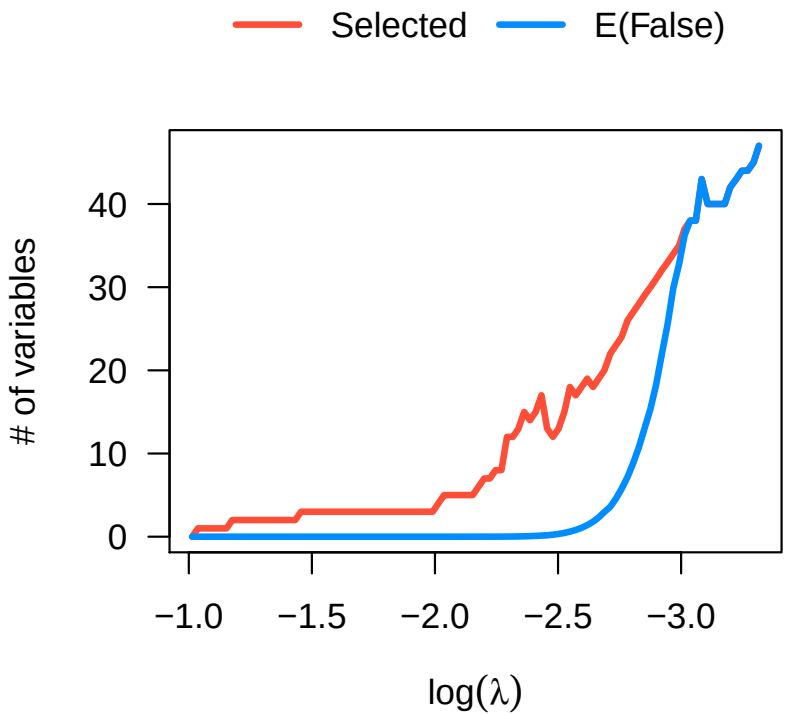
Remarks

- The lasso FIR approach discussed here selects 52 features at a FIR of 5% ($\lambda = 0.0687$)
- In contrast, using cross-validation to select λ ($\lambda = 0.0450$) we have 91 features, but a false inclusion rate of 68%
- This is in line with earlier remarks we have made: if we select λ for lasso to achieve the best prediction/estimation accuracy, we allow many false variables to enter the model

TCGA: MCP

One nice aspect of the approach is that it can be readily extended to MCP/SCAD:



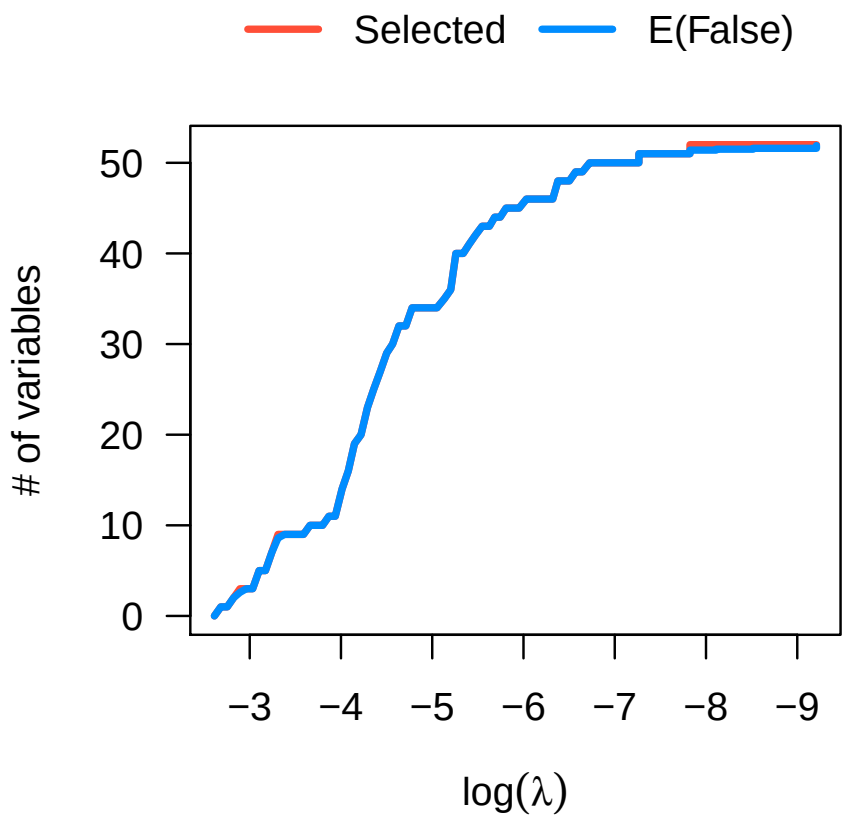


MCP Remarks

- With MCP, FIR selects 18 features at a FIR of 10% ($\lambda = 0.0946$)
- Meanwhile, using cross-validation to select λ ($\lambda = 0.0687$) we get 20 features and a false inclusion rate of 13%
- Note that the discrepancy between the two is far less severe with MCP than with lasso

SOPHIA

A GWAS example



No features can be selected with any confidence that they are not false inclusions

Conclusions

- False inclusion rates are a useful tool for assessing the reliability of variable selection in penalized regression models
- The estimator is conservative when variables are highly correlated, although this is not a fatal flaw, and the conservatism is usually mild
- The simplicity of the estimator makes it (a) available at minimal added computational cost and (b) very easy to generalize to new methods