Marginal false discovery rates

Patrick Breheny

April 3

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

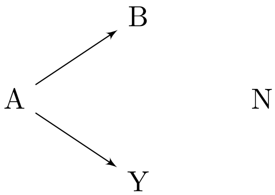
- Up until this point, our inference has been restricted to the predictive ability of the model (which we can obtain via cross-validation)
- This is useful, of course, but we would also like to be able to ask the questions:
    - How reliable are the selections made by the model? What is its false discovery rate?
    - How accurate are the estimates yielded by the model? Can we obtain confidence intervals for $\boldsymbol{\beta}$? Even for β_j not selected by the model?

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 five major categories:
    - Marginal approaches
    - Debiasing
    - Sample splitting/resampling
    - Selective inference
    - Knockoff filter

Setup

- For all of these methods, we will describe the idea behind how they work and then analyze the same set of simulated data for the sake of comparison
- Simulation setup:



- The `hdrm` package has a function called `genDataABN()` to simulate data of this type

Example data

Our example data set for the next several lectures:

- $n = 100$, $p = 60$, $\sigma^2 = 1$
- Six variables with $\beta_j \neq 0$ (category "A"):
    - Two variables with $\beta_j = \pm 1$:
    - Four variables with $\beta_j = \pm 0.5$:
- Each of the six variables with $\beta_j \neq 0$ is correlated ($\rho = 0.5$) with two other variables; i.e., there are 12 "Type B" features
- The remaining 42 variables are pure noise, $\beta_j = 0$ and independent of all other variables ("Type N")

```
genDataABN(n=100, p=60, a=6, b=2, rho=0.5,
           beta=c(1,-1,0.5,-0.5,0.5,-0.5))
```

KKT conditions

- Recall the KKT conditions for the lasso:

$$\begin{aligned} \frac{1}{n}\mathbf{x}_j'\mathbf{r} &= \lambda\,\mathrm{sign}(\widehat{\beta}_j) && \text{for all } \widehat{\beta}_j \neq 0 \\ \frac{1}{n}\left|\mathbf{x}_j'\mathbf{r}\right| &\leq \lambda && \text{for all } \widehat{\beta}_j = 0 \end{aligned}$$

- Letting $\mathbf{r}_j = \mathbf{y} - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_{-j}$ denote the partial residual with respect to feature j, this implies that

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

similar equations apply for MCP, SCAD, elastic net, etc.

Selection probabilities

- Therefore, the probability that variable j is selected is

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

- 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:

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

where $\hat{\mathcal{S}}$ is the set of selected variables and $\mathcal{N}$ is the set of null variables

Estimation

- To use this as an estimate, two unknown quantities must be estimated (this should seem familiar):
  - $|\mathcal{N}|$ can be replaced by p, using the total number of variables as an upper bound for the null variables
  - σ^2 can be estimated by $\mathbf{r}^T\mathbf{r}/(n - \left|\hat{\mathcal{S}}\right|)$
- This implies the following estimate for the expected number of false discoveries:

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

and this to estimate of the false discovery rate:

$$\widehat{\mathrm{FDR}} = \frac{\widehat{\mathrm{FD}}}{\left|\hat{\mathcal{S}}\right|}$$

Local false discovery rates

- Letting

$$z_j = \frac{\frac{1}{n}\mathbf{x}_j^T\mathbf{r}_j}{\hat{\sigma}\sqrt{n}},$$

  we therefore have $z_j \dot\sim \mathrm{N}(0,1)$
- We could therefore use this set of z-statistics to estimate feature-specific local false discovery rates as well
- Note that in this approach, we are not restricted to variables in the model; z_j can be calculated for all p features
- This is all assuming an orthonormal design; what about in the general case?

General case

- In the non-orthogonal case,

$$\frac{1}{n}\mathbf{x}_j^T\mathbf{r}_j = \beta_j^* + \frac{1}{n}\mathbf{x}_j^T\boldsymbol{\varepsilon} + \frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j}(\boldsymbol{\beta}_{-j}^* - \widehat{\boldsymbol{\beta}}_{-j})$$

- Broadly speaking, the general idea here is that:
    - For variables like B, the remainder term is not negligible
    - For variables like N, however, the remainder term *is* negligible, at least under certain conditions
- For this reason, I named these *marginal false discovery rates*, as it only establishes FDR control for variables marginally independent of the outcome ($X_j \perp\!\!\!\perp Y$), as opposed to conditional approaches that are concerned with conditional independence: $X_j \perp\!\!\!\perp Y | \{X_k\}_{k \neq j}$

Remarks

Focusing on marginal false discoveries has a few advantages:

- Allows straightforward, efficient estimation of the marginal false discovery rate (mFdr)
- Much more powerful: When two variables are correlated, distinguishing between which of them (or none, or both) is driving changes in Y and which is merely correlated with Y is challenging – even more so in high dimensions
- In many applications, discovering variables like B is not problematic

Theoretical support

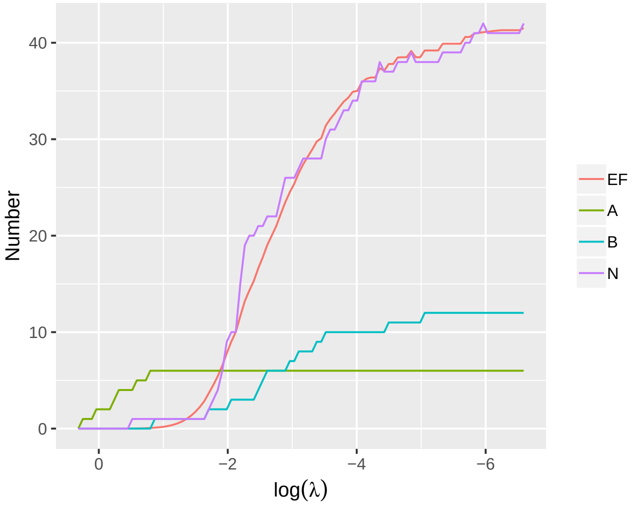
- The design matrix does not have to be strictly orthogonal in order for the proposed estimator to work; let $\mathcal{A}, \mathcal{N}$ partition $\{1, 2, \ldots, p\}$ such that $\beta_j = 0$ for all $j \in \mathcal{N}$ and the following condition holds:

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{X}'\mathbf{X} = \begin{bmatrix} \Sigma_{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\mathcal{N}} \end{bmatrix}$$

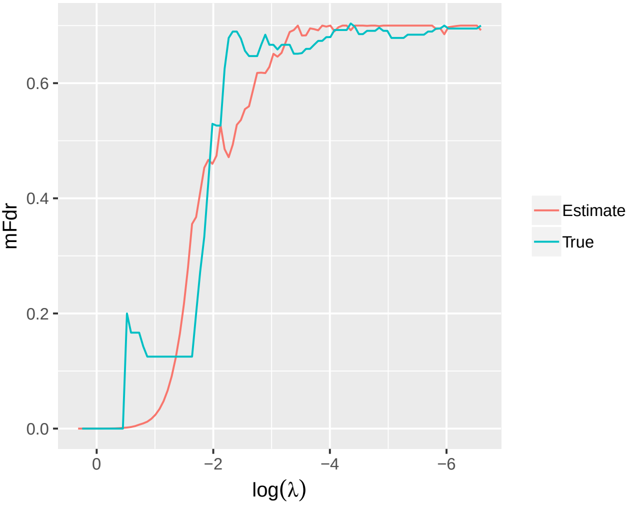
- **Theorem:** Suppose $\frac{1}{n}\mathbf{X}_{\mathcal{N}}^T\mathbf{X}_{\mathcal{N}} \to \Sigma_{\mathcal{N}} = \mathbf{I}$. Then for any $j \in \mathcal{N}$ and for λ_n such that the sequence $\sqrt{n}\lambda_n$ is bounded,

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

mFdr accuracy



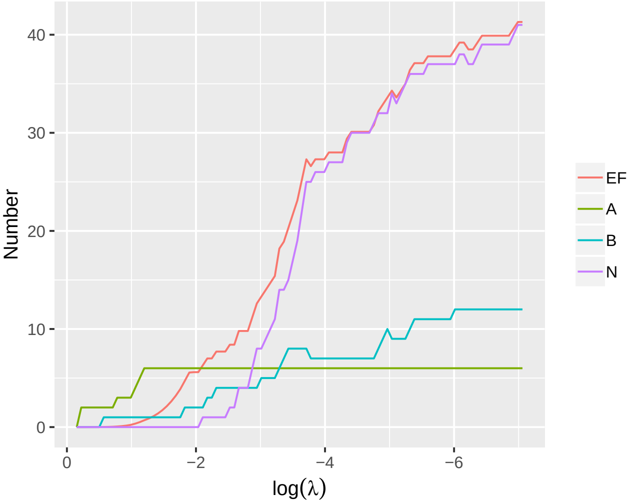
mFdr accuracy (cont'd)



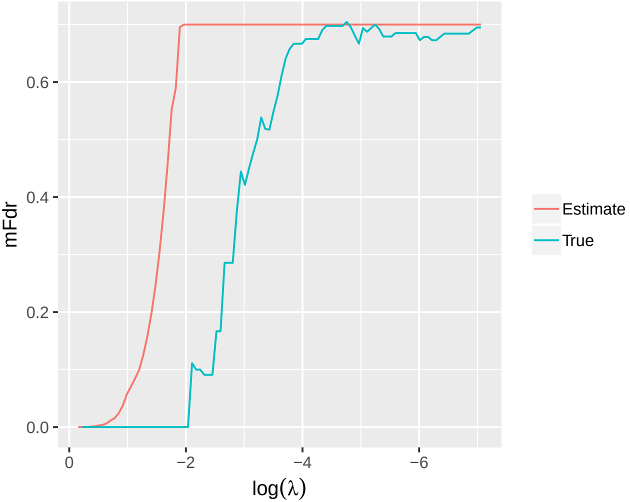
Correlated noise

- The preceding results are something of a "best case scenario" for the proposed method, since the variables in $\mathcal{N}$ were independent
- When the null variables are dependent, the estimator becomes conservative
- The reason for this is that if features are correlated, regression methods such as the lasso will tend to select a single feature and then become less likely to select other correlated features; our calculations do not account for this phenomenon

mFdr accuracy: Highly correlated noise



mFdr accuracy: Highly correlated noise (cont'd)



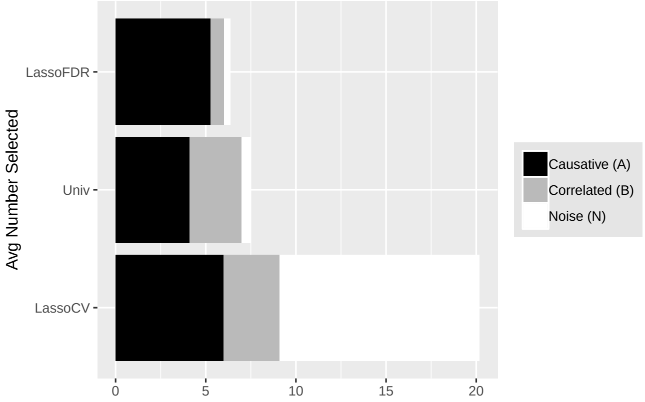
Comparison

- Being able to estimate mFdr gives us another way of choosing λ: we can choose the smallest value of λ such that $\mathrm{mFdr}(\lambda) < \alpha$
- For our example data set (uncorrelated noise; FDR methods with a nominal FDR of 10%):

	# Selected		
	A	B	N
Lasso (mFDR)	6	1	1
Univariate	6	5	1
Lasso (CV)	6	2	3

Comparison (simulation)

A more extensive comparison based on averaging across many simulated data sets:



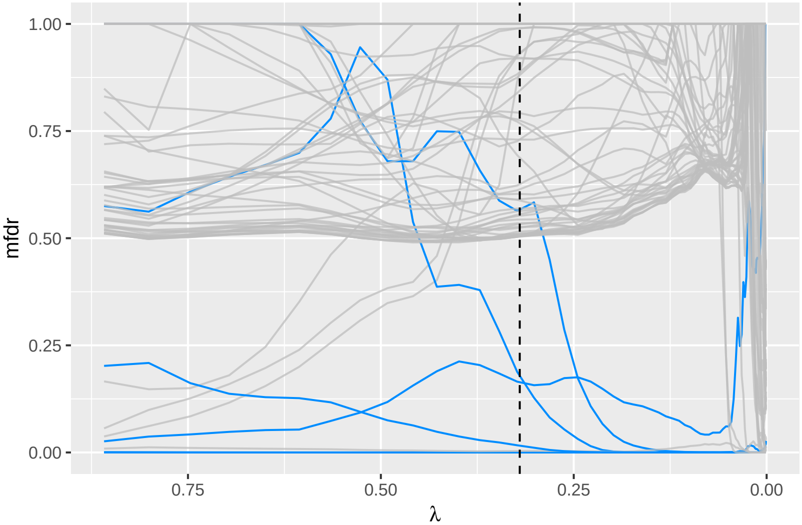
Remarks

- Cross-validation gives no control over the number of noise variables selected (and indeed, tends to select a lot of them)
- Univariate approaches give no control over the number of "Type B" variables selected (and also, tend to select a lot of them)
- Using lasso with mFdr control
  - Controls the number of noise variables selected
  - Doesn't necessarily control the number of "Type B" variables selected, but tends not to select many of them (because it's fundamentally a regression-based approach)

Tension between selection and prediction

- As we saw in our theory lectures, there tends to be a tension between variable selection and prediction, at least for the lasso: values of λ that are optimal for prediction let in too many false positives
- Conversely, if we select λ so as to limit the number of false positives, the resulting model has quite a bit of bias – prediction and estimation suffer
- By providing feature-specific inference, local false discovery rates alleviate this tension: we can select the optimal predictive model, but still have a way of quantifying which features are likely to be false discoveries

Local mfdr



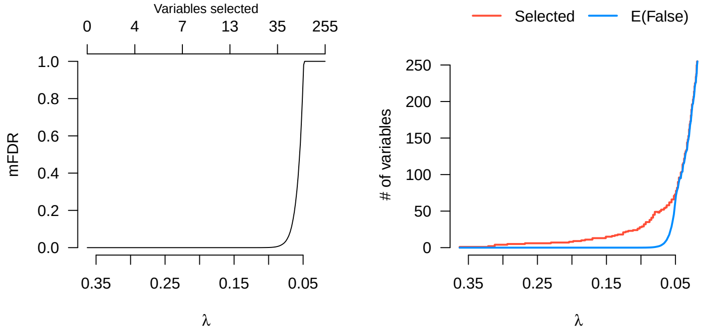
summary

```
> summary(fit, lambda=cvfit$lambda.min)
----------------------------------------------------
  Expected nonzero coefficients: 1.13
  Average mfdr (8 features)     : 0.142

    Estimate      z        mfdr
A2   -0.7167 -9.320    < 1e-04
A1    0.7228  8.970    < 1e-04
A6   -0.3045 -4.925    < 1e-04
A3    0.2730  4.357 0.00077842
B10   0.2406  4.022 0.00318218
A4   -0.2216 -3.693 0.01532035
A5    0.1378  3.102 0.11479490
B2    0.0012  1.661 1.00000000
```

Breast cancer data ($n = 536$, $p = 17,322$)

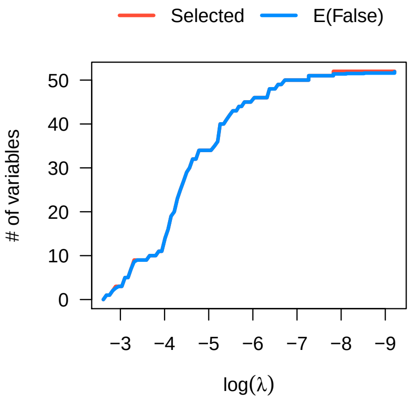
```
mfdr(fit)
plot(mfdr(fit))
```



We can select quite a few variables (≈ 50) with a low mFdr

SOPHIA ($n = 292$, $p = 705,969$)

A GWAS example



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

Conclusions

- Marginal false discovery rates are a useful tool for assessing the reliability of variable selection in penalized regression models
- The simplicity of the estimator makes it (a) available at minimal added computational cost and (b) very easy to generalize to new methods
- Some issues to be aware of, though:
  - Only controls FDR in the marginal sense (i.e., not for all $\beta_j = 0$)
  - Becomes conservative when noise features are highly correlated
- Local false discovery rates provide a way to select prediction-optimal models without worrying about the number of false selections