Marginal false discovery rates

Patrick Breheny

April 3, 2025

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 β
    - 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
  - Synthetic variable approaches (knockoff filter, Gaussian mirror)

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:

$$\begin{tikzcd} & B & \\ A \arrow[ur] \arrow[dr] & & N \\ & Y & \end{tikzcd}$$

- The `hdrm` package has a function called `gen_data_abn()` 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")

```
gen_data_abn(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:

$$\frac{1}{n}\mathbf{x}_j^\top \mathbf{r} = \lambda \operatorname{sign}(\widehat{\beta}_j) \qquad \text{for all } \widehat{\beta}_j \neq 0$$
$$\frac{1}{n}\left|\mathbf{x}_j^\top \mathbf{r}\right| \leq \lambda \qquad \text{for all } \widehat{\beta}_j = 0$$

- 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

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

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

Selection probabilities

- Therefore, the probability that variable j is selected is

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

- This suggests that if we are able to characterize the distribution of $\frac{1}{n}\mathbf{x}_j^\top \mathbf{r}_j$ under the null, we can estimate the number of false selections in the model
- A simple approximation (we'll come back to this shortly) is:

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

where $\hat{\mathcal{S}}$ is the set of selected variables and $\mathcal{N}$ is the set of "N" variables (note that $\mathcal{N}$ differs here from other lectures)

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 $\mathbf{r}^\top\mathbf{r}/(n - |\hat{\mathcal{S}}|)$
- 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:

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

Local false discovery rates

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

Remainder term

- Expanding $\mathbf{x}_j^\top \mathbf{r}_j/n$, we have

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

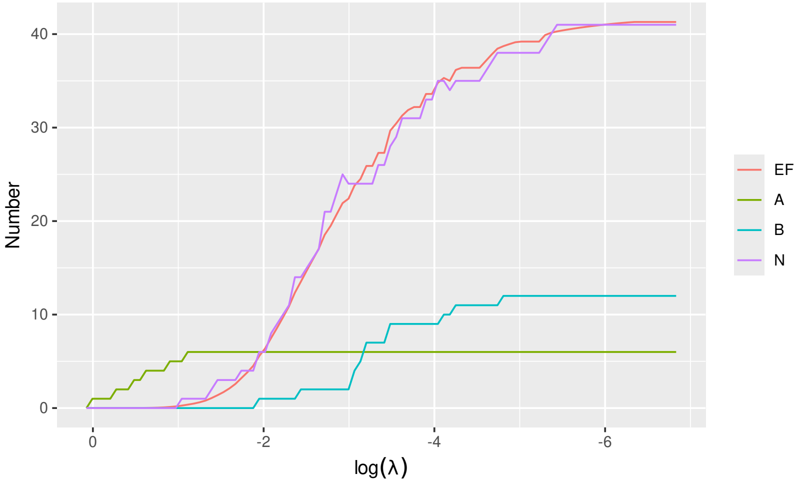
- Broadly speaking,
  - For variables like B, this remainder term is not negligible
  - For variables like N, however, this remainder term *is* negligible

Remarks

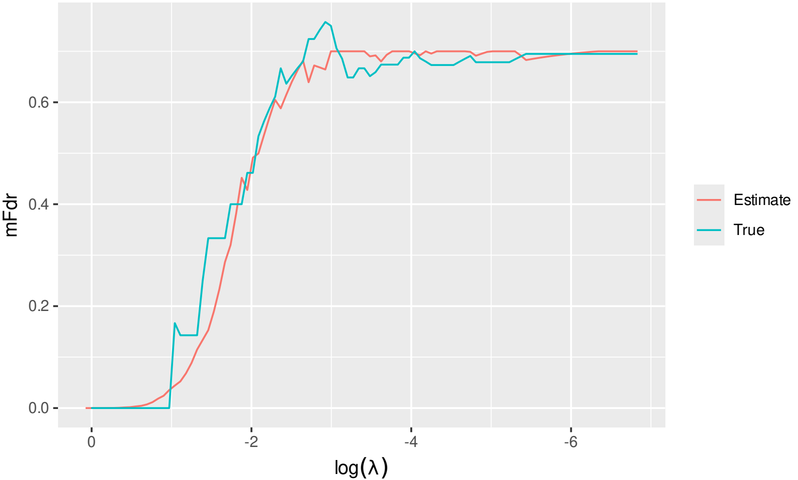
Focusing on marginal false discoveries ($X_j \perp\!\!\!\perp Y$) as opposed to conditional independence ($X_j \perp\!\!\!\perp Y | \{X_k\}_{k \neq j}$) has several advantages:

- 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

mFdr accuracy



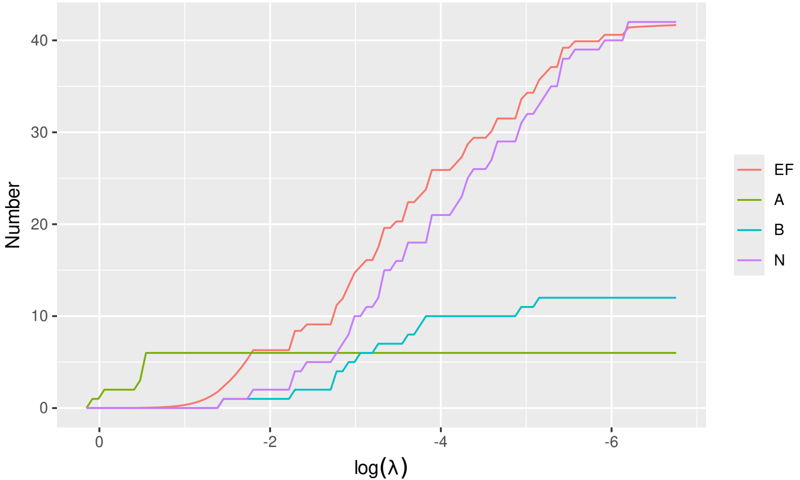
mFdr accuracy (cont'd)



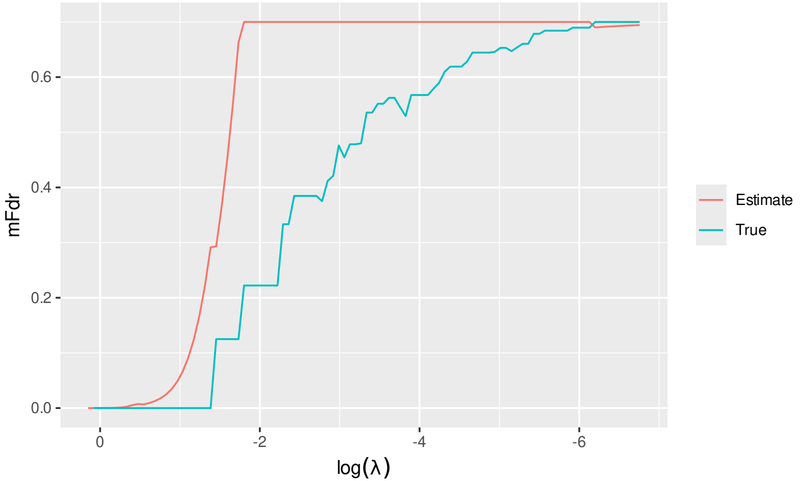
Correlated noise

- 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: $\rho_{jk} = 0.5$



mFdr accuracy, highly correlated noise (cont'd)



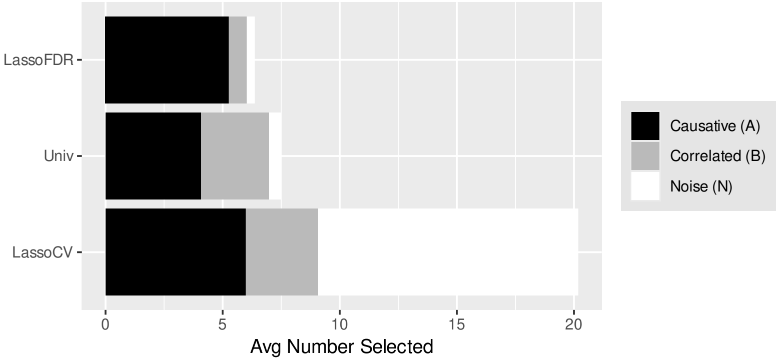
Comparison

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

	# Selected		
method	A	B	N
Lasso (mFDR)	6	0	1
Univariate	4	5	2
Lasso (CV)	6	2	21

Comparison (simulation)

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



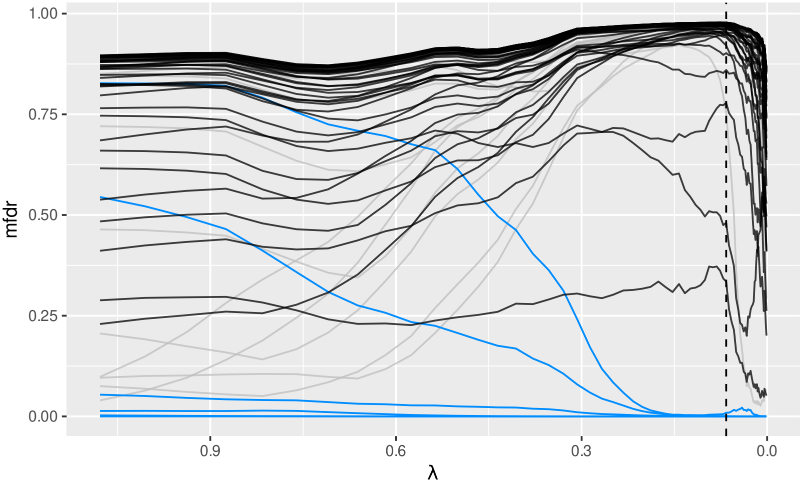
Remarks

- 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

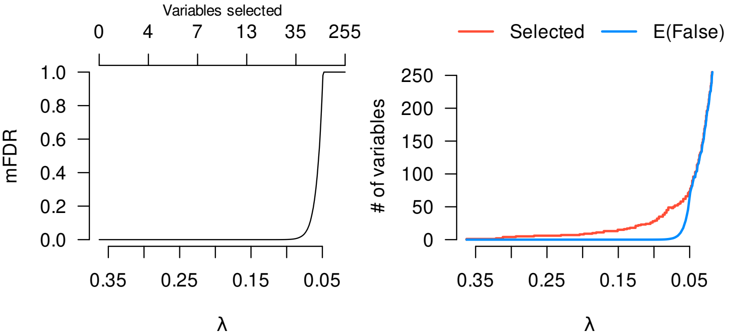
```
summary(fit, lambda=cvfit$lambda.min)
#   Nonzero coefficients         :  29
#   Expected nonzero coefficients:  20.37
#   Average mfdr (29 features)   :   0.702
#
#      Estimate        z       mfdr Selected
# A1   0.874102  11.2647    < 1e-04        *
# A2  -0.774583 -10.9076    < 1e-04        *
# A4  -0.502917  -7.1268    < 1e-04        *
# A3   0.422238   5.4744    < 1e-04        *
# A6  -0.351849  -4.7564 0.00059017        *
# A5   0.309722   4.1233 0.00915535        *
# N39 -0.200926  -2.9913 0.33482886        *
```

summary (cont'd)

```
...
# N42  0.037062   1.2128 0.95479163        *
# N26 -0.024974  -1.0778 0.96101303        *
# N24 -0.020723  -1.0491 0.96213966        *
# N2   0.018021   0.9914 0.96422830        *
# N17 -0.009270  -0.8914 0.96733745        *
# N37 -0.008625  -0.8807 0.96763605        *
# N11 -0.004770  -0.8405 0.96870108        *
# N41 -0.004774  -0.8346 0.96885141        *
# N34  0.003134   0.8183 0.96925552        *
```

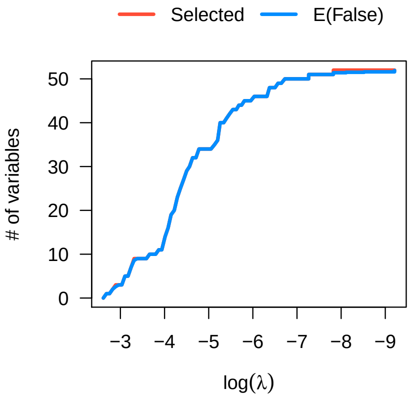
Breast cancer data ($n = 536$, $p = 17,322$)

```r
plot(mfdr(fit))
plot(mfdr(fit), type = 'EF')
```



We can select 16 genes with mfdr $< 20\%$

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



A GWAS example: No features can be selected with confidence that they are not false discoveries

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