Feature-specific Inference for Penalized Regression Models using Local False Discovery Rates

Ryan Miller

Introduction

- Data containing large numbers of variables are becoming increasingly more common
  - Comprehensive electronic databases containing hundreds of attributes
  - Technologies capable of measuring thousands of genetic expression levels

Introduction

- Data containing large numbers of variables are becoming increasingly more common
  - Comprehensive electronic databases containing hundreds of attributes
  - Technologies capable of measuring thousands of genetic expression levels
- Many traditional statistical methods struggle with these modern datasets
  - Heightened potential for false discoveries
  - Barriers to traditional estimation approaches
- These challenges motivate the need for new methods of inference

Two Approaches

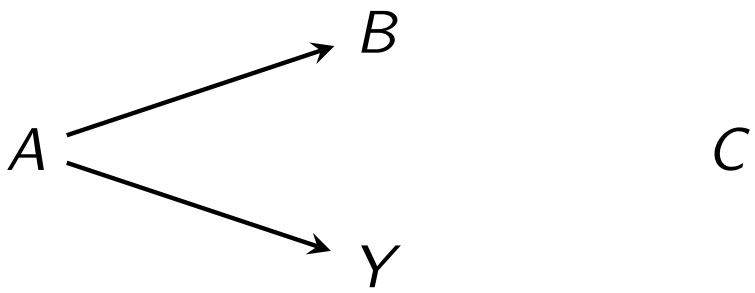
1. *Large-scale testing*
   - Test each feature's relationship with the outcome variable individually (one-at-a-time)
   - Aggregate the results in a way that controls the false discovery rate
2. *Regression modeling*
   - Model the outcome variable using all explanatory features
   - Apply traditional methods of inference (t-tests on model coefficients, ANOVA, etc.)

Outline

1. False discovery perspectives and definitions
2. The challenges "high-dimensional" data
3. An introduction to penalized regression
4. Development of feature-specific local false discovery rates for the penalized regression models
5. Simulation studies
6. BCRA1 gene expression application

False Discovery Perspectives

Consider an outcome variable Y, and set of explantory variables, X_j for $j \in \{1, \ldots, p\}$



False Discovery Perspectives:

- *marginal*: if $X_j \perp\!\!\!\perp Y$
  - B is considered a valid discovery
- *fully conditional*: if $X_j \perp\!\!\!\perp Y | X_{k \neq j}$
  - B is considered a false discovery
- *pathwise conditional*: if $X_j \perp\!\!\!\perp Y | X_k$ for $k \in M_j$
  - The status of B depends on the model

False Discovery Perspectives

- The marginal definition is used in *large-scale testing* approaches
- Regardless of the perspective taken, identifiability is an inherent issue in distinguishing 'A' variables from 'B' variables
- Regression-based approaches tend to naturally downplay the importance of variables like 'B'

Challenges to Traditional Methods

Consider the familiar regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Challenges to Traditional Methods

Consider the familiar regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

This model can be expressed more compactly:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{y}$ is a vector of outcomes for $i \in \{1 \ldots n\}$ observations
- $\mathbf{X}$ is an n by $p + 1$ matrix of explanatory variables
- β is a vector of regression coefficients
- ϵ is a vector of indepedent errors, with $\epsilon_i \sim N(0, \sigma^2)$

Challenges to Traditional Methods

- To utilize the model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ for statistical inference, β must be estimated

Challenges to Traditional Methods

- To utilize the model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ for statistical inference, β must be estimated
- *Ordinary least squares* estimates β by minimizing the residual sum of squares, $||\mathbf{y} - \mathbf{X}\beta||_2^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$, yielding:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Challenges to Traditional Methods

- To utilize the model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ for statistical inference, β must be estimated
- *Ordinary least squares* estimates β by minimizing the residual sum of squares, $||\mathbf{y} - \mathbf{X}\beta||_2^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$, yielding:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- However, when $p \geq n$ the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible
  - Further, when p approaches n these estimates become unstable (highly variable)

Penalized Regression

One method of attaining a unique solution is to impose a penalty on the size of the coefficient vector, β, then solving for estimates that minimize:

$$||\mathbf{y} - \mathbf{X}\beta||_2^2 + P_\lambda(\beta)$$

Penalized Regression

One method of attaining a unique solution is to impose a penalty on the size of the coefficient vector, $\boldsymbol{\beta}$, then solving for estimates that minimize:

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + P_\lambda(\boldsymbol{\beta})$$

The *least absolute shrinkage and selection operator*, or LASSO, penalizes the L1 norm of $\boldsymbol{\beta}$ via:

$$P_\lambda(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1 = \lambda \sum_j |\beta_j|$$

- λ is a *tuning parameter* controlling the degree of penalization

Penalized Regression - Technical Notes

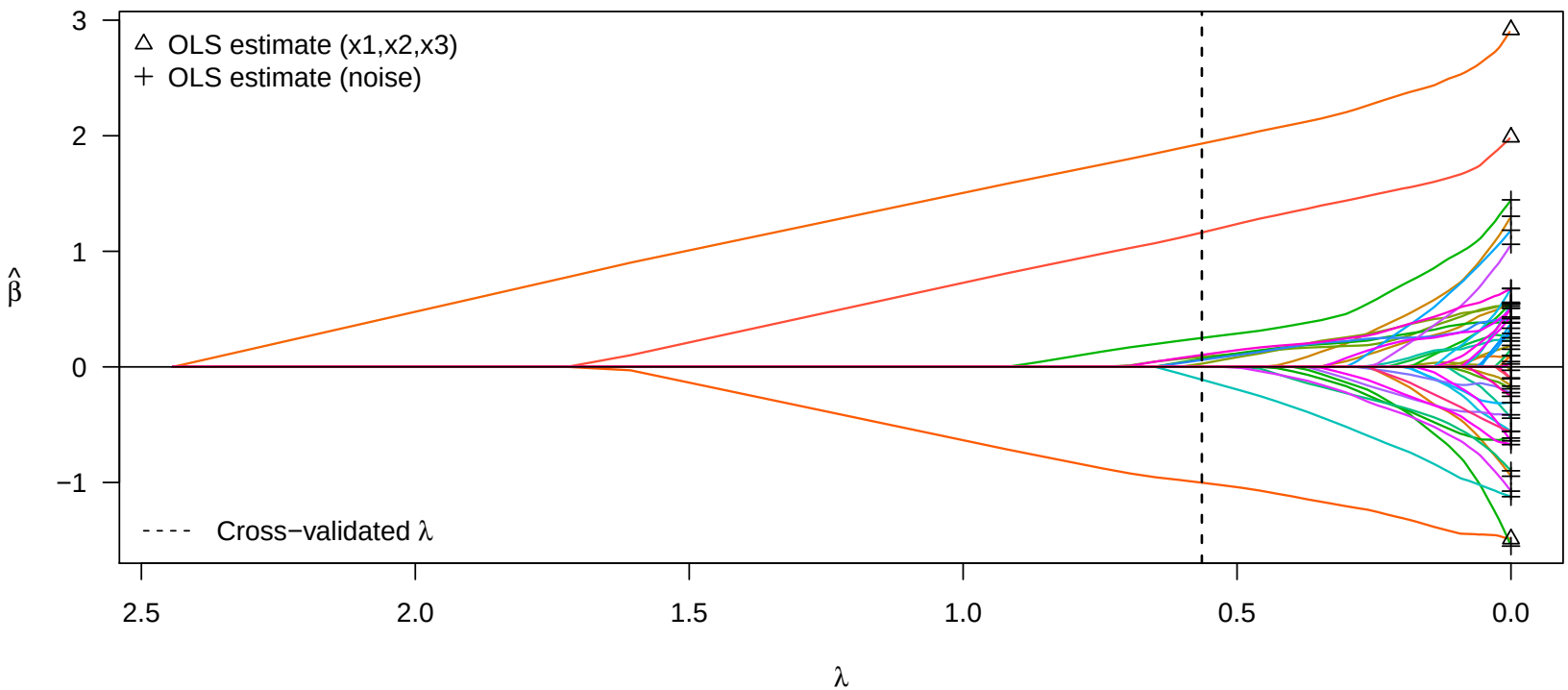
- To ensure the LASSO penalty is applied fairly to variables on different scales, the columns of $\mathbf{X}$ must be *standardized*:
  - Center each variable to have a mean of 0
  - Scale each variable so that $\sum_i \mathbf{x}_{ij}^2 = n$
- After estimation is done on the standardized scale, these estimates can be transformed to reflect the variable's original scale

Penalized Regression - A Simple Example

- $n = 100$ outcomes simulated under the model:

$$y_i = 2x_1 - 2x_2 + 3x_3 + \epsilon_i$$

- Here $\epsilon_i \sim N(0, 4)$, and we include 47 predictors unrelated to Y as "noise"



Penalized Regression - A Simple Example

- The LASSO penalty can lead to tension between models with *optimal prediction* and models that *eliminate noise*
  - Using cross-validation to choose λ favors a model containing 7 noise variables!

Penalized Regression - A Simple Example

- The LASSO penalty can lead to tension between models with *optimal prediction* and models that *eliminate noise*
  - Using cross-validation to choose λ favors a model containing 7 noise variables!
- In the remainder of this talk, we'll explore *feature selection reliability* for penalized regression models
  - I will propose a feature-specific method for quantifying the likelihood that a selected variable is a "noise variable"

Local mfdr - Introduction

- The LASSO solution is mathematically characterized by a set of equations known as the KKT conditions:

$$\frac{1}{n}\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \lambda \ \text{sign}(\hat{\beta}_j) \quad \text{if } \hat{\beta}_j \neq 0$$
$$\frac{1}{n}\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \in [-\lambda, \lambda] \quad \text{if } \hat{\beta}_j = 0$$

- From the KKT conditions, it is clear that selection by the LASSO is based upon the variable's correlation with the model's residuals

Local mfdr - Intuition

- The KKT conditions describe the conditions necessary for a variable is selected by the LASSO, we'll use them to construct a "test statistic" for the selection event

Local mfdr - Intuition

- The KKT conditions describe the conditions necessary for a variable is selected by the LASSO, we'll use them to construct a "test statistic" for the selection event
- These yields a collection of p different statistics for *any* given LASSO model
  - We can apply local false discovery rate methods to this collection as a means of feature-specific inference

Local mfdr - Intuition

- The KKT conditions describe the conditions necessary for a variable is selected by the LASSO, we'll use them to construct a "test statistic" for the selection event
- These yields a collection of p different statistics for *any* given LASSO model
  - We can apply local false discovery rate methods to this collection as a means of feature-specific inference
- In what follows we'll develop this "test statistic", establishing its suitability for local false discovery rate methods

Local mfdr - Development

- Define $\mathbf{r}_j = \mathbf{y} - \mathbf{X}_{-j}\hat{\beta}_{-j}$, where $-j$ indicates removal of the j^{th} variable
- Then, the KKT conditions imply:

$$\frac{1}{n}|\mathbf{x}_j^T \mathbf{r}_j| > \lambda \qquad \text{for all } \hat{\beta}_j \neq 0$$
$$\frac{1}{n}|\mathbf{x}_j^T \mathbf{r}_j| \leq \lambda \qquad \text{for all } \hat{\beta}_j = 0$$

- Notice that $\frac{1}{n}\mathbf{x}_j^T \mathbf{r}_j$ governs the selection of the j^{th} variable: if it is large enough in absolute value (relative to λ), feature j is selected
  - This quantity will form the basis of our selection "test statistic"

Local mfdr - Development

- In the special case of orthonormal design, where $\frac{1}{n}\mathbf{X}^T\mathbf{X} = \mathbf{I}$, it is straightforward to show:

$$\frac{1}{n}\mathbf{x}_j^T\mathbf{r}_j \sim N(\beta_j, \sigma^2/n)$$

- This suggests the normalized statistic:

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

- Under the hypothesis that $\beta_j = 0$, we expect $z_j \sim N(0, 1)$

Local mfdr - Development

- In practice $\frac{1}{n}\mathbf{X}^T\mathbf{X} \neq \mathbf{I}$, but:

$$\begin{aligned}\frac{1}{n}\mathbf{x}_j^T\mathbf{r}_j &= \frac{1}{n}\mathbf{x}_j^T(\mathbf{X}\beta + \epsilon - \mathbf{X}_{-j}\hat{\beta}_{-j}) \\ &= \frac{1}{n}\mathbf{x}_j^T\epsilon + \beta_j + \frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j}(\beta_{-j} - \hat{\beta}_{-j}).\end{aligned}$$

- Notice $\frac{1}{n}\mathbf{x}_j^T\epsilon + \beta_j$ is unaffected by the structure of $\frac{1}{n}\mathbf{X}^T\mathbf{X}$
- Thus $z_j \sim N(0,1)$ when $\frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j}(\beta_{-j} - \hat{\beta}_{-j})$ is negligible

Local mfdr - Development

- In practice $\frac{1}{n}\mathbf{X}^T\mathbf{X} \neq \mathbf{I}$, but:

$$\begin{aligned}\frac{1}{n}\mathbf{x}_j^T\mathbf{r}_j &= \frac{1}{n}\mathbf{x}_j^T(\mathbf{X}\beta + \epsilon - \mathbf{X}_{-j}\hat{\beta}_{-j}) \\ &= \frac{1}{n}\mathbf{x}_j^T\epsilon + \beta_j + \frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j}(\beta_{-j} - \hat{\beta}_{-j}).\end{aligned}$$

- Notice $\frac{1}{n}\mathbf{x}_j^T\epsilon + \beta_j$ is unaffected by the structure of $\frac{1}{n}\mathbf{X}^T\mathbf{X}$
- Thus $z_j \sim N(0,1)$ when $\frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j}(\beta_{-j} - \hat{\beta}_{-j})$ is negligible
  - If feature j is independent of other features, $\frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j} \to 0$ as n increases
  - So this term becomes asymptotically negligible provided $\sqrt{n}(\beta_{-j} - \hat{\beta}_{-j})$ is bounded in probability

Local mfdr - Development

- In practice $\frac{1}{n}\mathbf{X}^T\mathbf{X} \neq \mathbf{I}$, but:

$$\begin{aligned}\frac{1}{n}\mathbf{x}_j^T\mathbf{r}_j &= \frac{1}{n}\mathbf{x}_j^T(\mathbf{X}\beta + \epsilon - \mathbf{X}_{-j}\hat{\beta}_{-j}) \\ &= \frac{1}{n}\mathbf{x}_j^T\epsilon + \beta_j + \frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j}(\beta_{-j} - \hat{\beta}_{-j}).\end{aligned}$$

- Notice $\frac{1}{n}\mathbf{x}_j^T\epsilon + \beta_j$ is unaffected by the structure of $\frac{1}{n}\mathbf{X}^T\mathbf{X}$
- Thus $z_j \sim N(0,1)$ when $\frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j}(\beta_{-j} - \hat{\beta}_{-j})$ is negligible
  - If feature j is independent of other features, $\frac{1}{n}\mathbf{x}_j^T\mathbf{X}_{-j} \to 0$ as n increases
  - So this term becomes asymptotically negligible provided $\sqrt{n}(\beta_{-j} - \hat{\beta}_{-j})$ is bounded in probability
- If feature j is not independent of other features, z_j follows a distribution with thinner tails (we'll later explore how to accomodate this)

Local mfdr - Development

To recap:

1. The LASSO solution is characterized by the KKT conditions
2. The KKT conditions suggest $\frac{1}{n}|\mathbf{x}_j^T \mathbf{r}_j| > \lambda$ for selected variables
3. The statistic $z_j = \frac{\frac{1}{n}\mathbf{x}_j^T \mathbf{r}_j}{\hat{\sigma}/\sqrt{n}}$ can be expected to follow a $N(0,1)$ distribution when $\beta_j = 0$ and feature j is independent of the other predictors (we'll explore this soon)

Local mfdr - Development

To recap:

1. The LASSO solution is characterized by the KKT conditions
2. The KKT conditions suggest $\frac{1}{n}|\mathbf{x}_j^T \mathbf{r}_j| > \lambda$ for selected variables
3. The statistic $z_j = \frac{\frac{1}{n}\mathbf{x}_j^T \mathbf{r}_j}{\hat{\sigma}/\sqrt{n}}$ can be expected to follow a $N(0,1)$ distribution when $\beta_j = 0$ and feature j is independent of the other predictors (we'll explore this soon)

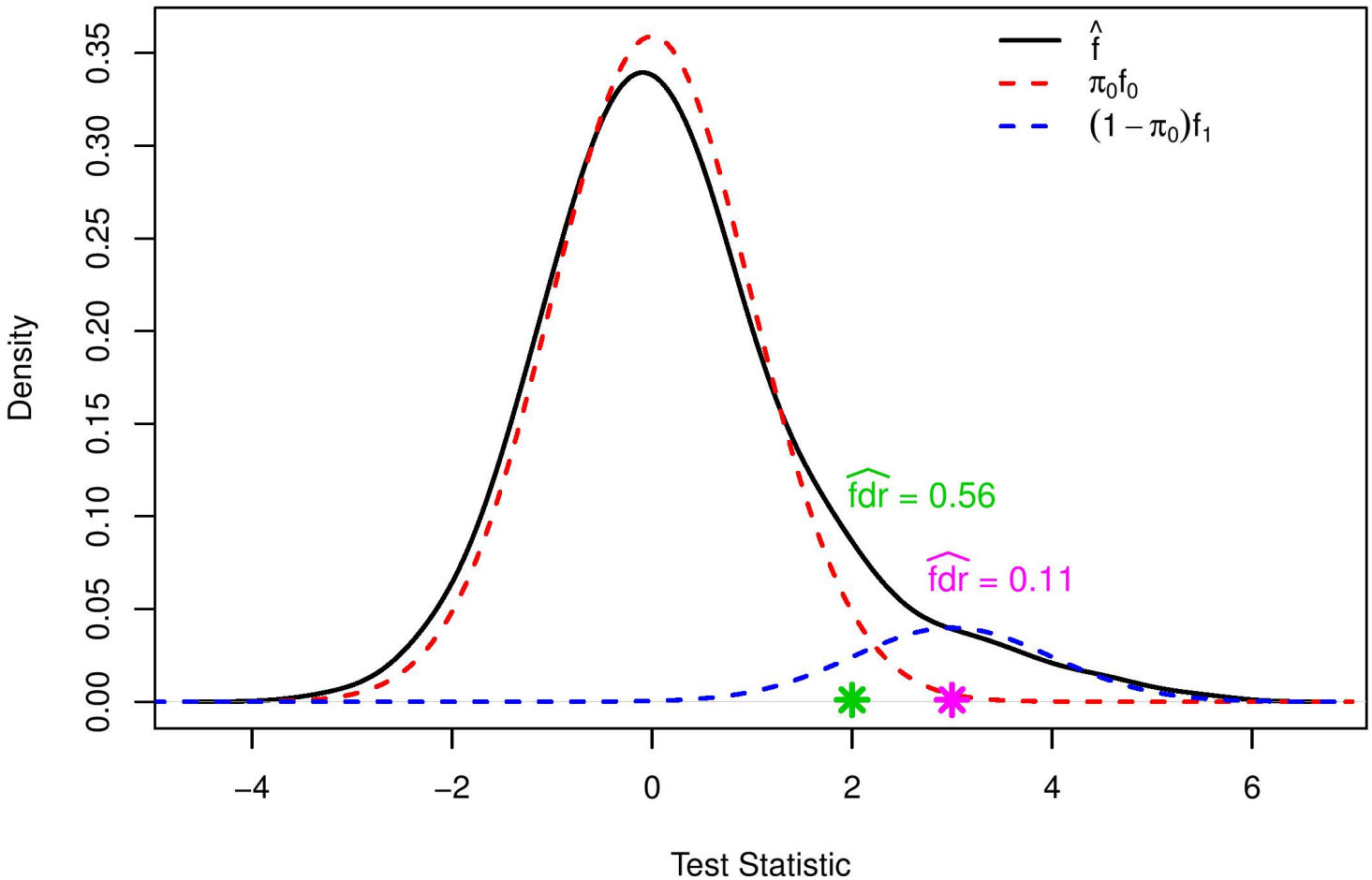
Up next:

- How can we apply *local false discovery rate* methods to the collection of z_j for $j \in \{1, \ldots, p\}$?

Local False Discovery Rates

- Suppose that feature j belongs to one of two classes, "null" or "non-null"
- The corresponding local false discovery rate is defined: $\Pr(\text{Null}|z = z_j) = \frac{\pi_0 f_0(z_j)}{f(z_j)} = fdr(z_j)$

Local False Discovery Rates



Local False Discovery Rate Estimation

There are two main approaches to local false discovery rate estimation:

1. Construct z_j such that f_0 is the $N(0,1)$ density, then estimate f using any method of density estimation: $\widehat{fdr}(z_j) = \frac{\pi_0 f_0(z_j)}{\hat{f}(z_j)}$ (Efron 2012)
2. Explicity model f via a mixture of null and non-null components: $\widehat{fdr}(z_j) = \frac{\hat{\pi}_0 \hat{f}_0(z_j)}{\sum_{k=0}^{K} \hat{\pi}_k \hat{f}_k(z_j)}$ (Stephens 2016)

Both approaches have strengths and weaknesses, the simulation results we'll present use the approach of Stephens (2016)

Simulation Studies

In our adaptation of local false discovery rates to the reliability of LASSO selections as measured using $z_j = \frac{\frac{1}{n}\mathbf{x}_j^T \mathbf{r}_j}{\hat{\sigma}/\sqrt{n}}$, we'll focus on two main questions:

1. Is the resulting inference *valid*? (ie: does it accurately measure false discoveries)
2. Is the resulting inference *powerful*? (ie: does it lead to more true positives than other existing approaches)

Simulation Studies - Setup

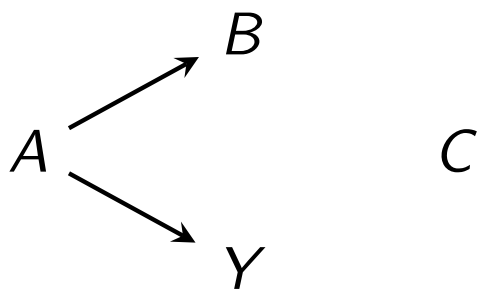
"Assumptions Met" Scenario:

- $n = 1000$ and $p = 600$
- 60 'A' variables
- 0 'B' variables
- 540 'C' variables, independent

"Assumptions Violated" Scenario:

- $n = 200$ and $p = 600$
- 6 'A' variables
- 9 'B' variables per 'A' variable ($\rho = 0.6$)
- 540 'C' variables, correlated: $cor(\mathbf{x}_j, \mathbf{x}_k) = 0.8^{|j-k|}$

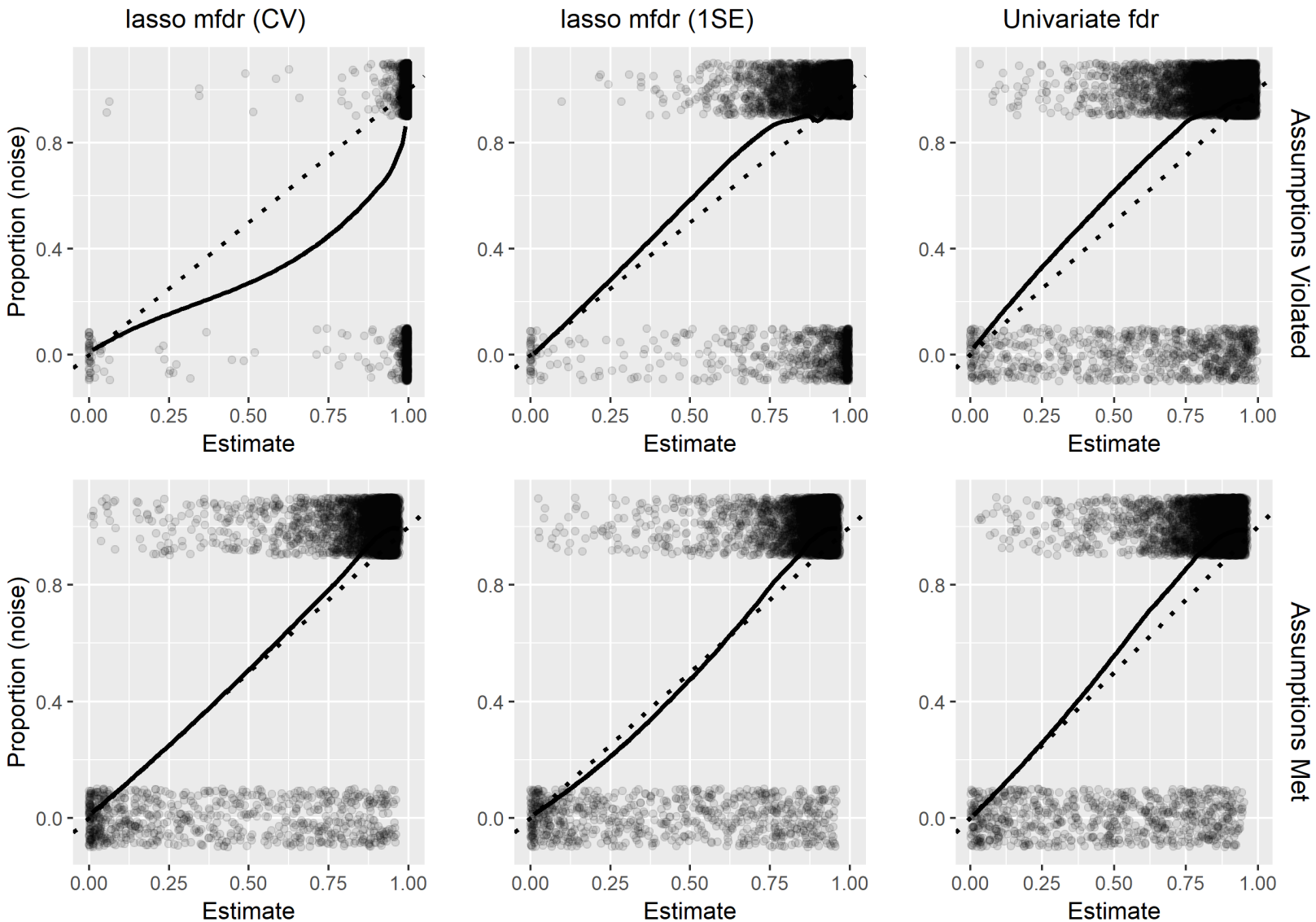
Figure 1: Causal diagram describing variable relationships



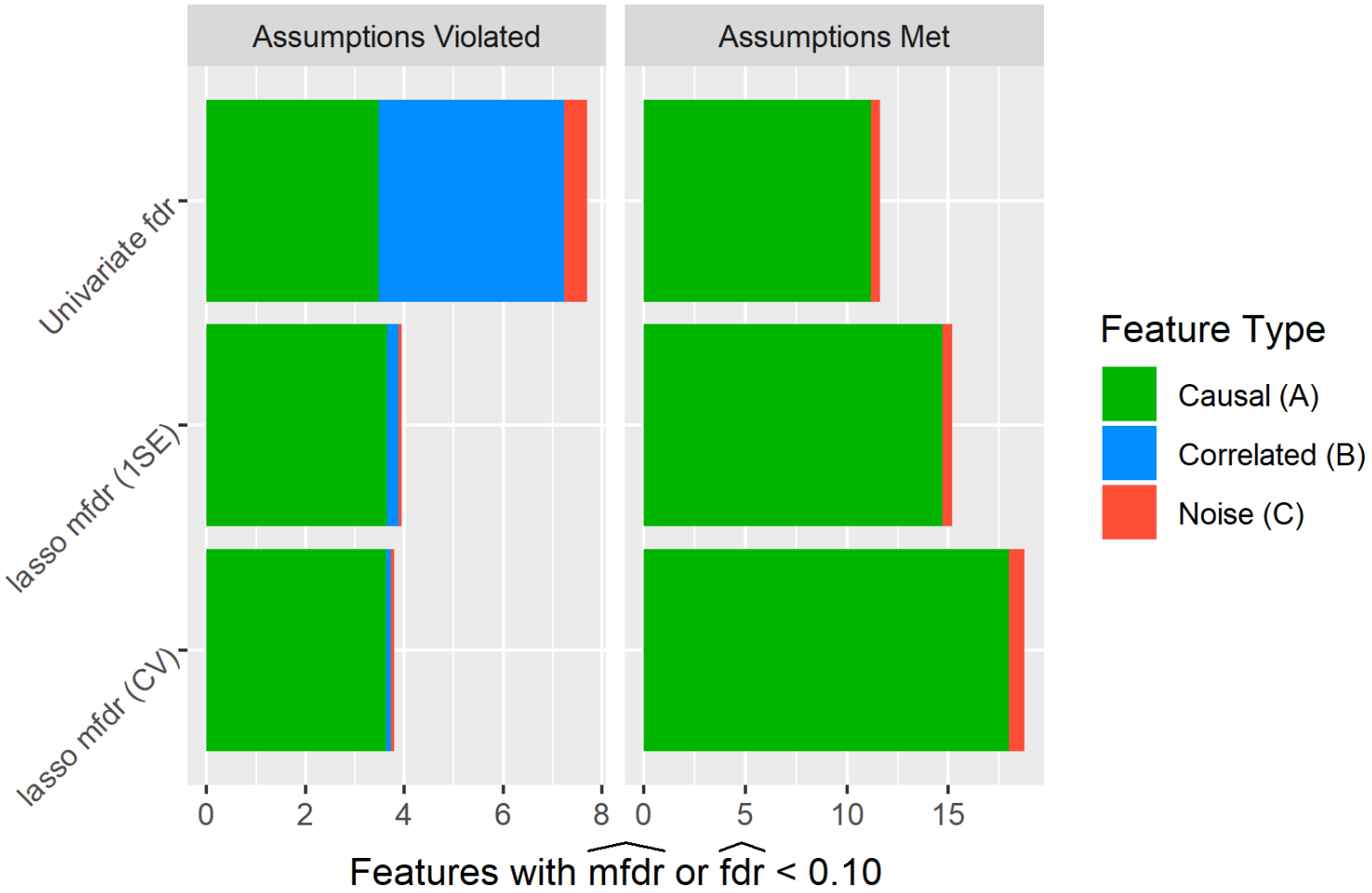
Simulation Studies - Comparisons

- For comparison, we also consider the following approaches:
  - Local false discovery rates in conjunction with large-scale univariate testing
  - Several related "selective inference" methods
  - Repeated sample splitting, or "multi-split"
  - The "knockoff filter"

Simulation Studies - Validity



Simulation Studies - Power



Simulation Studies - Power

Table 1: A comparison of selections by LASSO based approaches (targeting 10% Fdr)

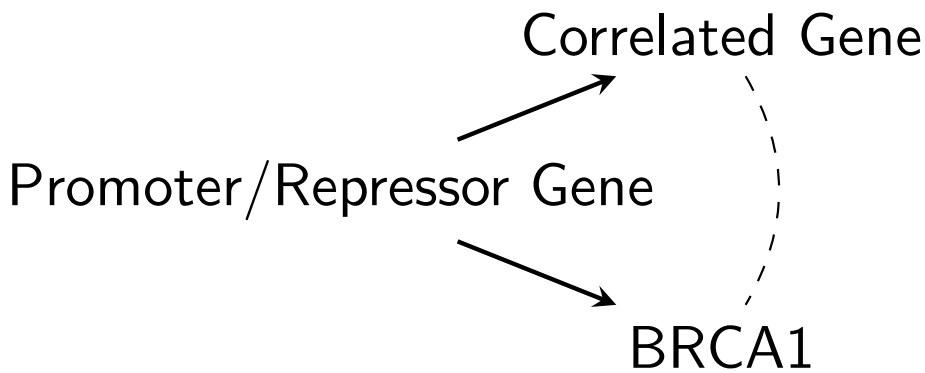
	Assumptions Met			Assumptions Violated	
Method	'A' (6)	'B' (54)	Rate of 'C'	'A' (60)	Rate of 'C'
mfdr (CV)	3.88	0.70	2.0%	25.87	2.3%
exact	0.92	0.06	0%	0.71	0%
spacing	1.60	0.07	0.4%	0.92	0%
mod-spacing	1.60	0.07	0.4%	0.92	0%
covtest	1.55	0.06	0.4%	0.89	0%
multi-split	1.93	0.03	0%	10.83	0.2%
knock-off	0.42	0.16	5%	3.5	0%

Application - BCRA1 Gene Expression

- BRCA1 is a well-known tumor suppressor gene with a strong relationship to breast cancer risk
  - Decreased BRCA1 expression, due to mutation or down-regulation, contributes to tumor formation
- Data from The Cancer Genome Atlas project (*http://cancergenome.nih.gov*)
- 17,814 gene expression measures for $n = 536$ breast cancer patients

Application - BCRA1 Gene Expression

Figure 2: Causal diagram depicting possible gene relationships with BRCA1



- Goal is to identify potential promoters/repressors of BRCA1
- Selecting some correlated genes is okay, but selecting a large number is undesirable

Application - BCRA1 Gene Expression

Gene	Location	Univariate $\widehat{\text{fdr}}$	$\widehat{\text{mfdr}}$ at λ_{CV}
C17orf53	17q21.31	<0.00001	1
TUBG1	17q21.2	<0.00001	1
DTL	1q31.3	<0.00001	0.00008
VPS25	17q21.2	<0.00001	0.02597
TOP2A	17q21.2	<0.00001	0.02138
PSME3	17q21.31	<0.00001	0.00288
TUBG2	17q21.2	<0.00001	1
TIMELESS	12q13.3	<0.00001	1
NBR2	17q21.31	<0.00001	<0.00001
CCDC43	17q21.31	<0.00001	1

Table 2: The top 10 selected genes from the univariate testing approach. Over 500 genes have $\widehat{\text{fdr}} < .1$ using the univariate approach, compared to only 16 having $\widehat{\text{mfdr}} < .1$ at λ_{CV}

Application - BCRA1 Gene Expression

- In this application, our mfdr method:
  - Appears to avoid selecting large numbers of correlated, non-causal genes
  - Retains promising genes with plausible biological relationships with BCRA1
- The method isn't perfect. Determining true causality is an inherently difficult for many statistical approaches (see NBR2)

Summary

- The KKT conditions can be used to inspire a “test statistic" for variable selection
- Local false discovery rates derived from these statistics are:
  - Reasonably well-calibrated when the data contains a challenging correlation structure
  - More powerful than large-scale testing methods and model-based methods
  - Useful when applied to real data
- The development of the $\widehat{\mathrm{mfdr}}$ method uses a marginal definition of false discoveries; stronger conditional definitions might be more relevant in certain situations

Closing Remarks and Future Work

- The methods described in this talk are implemented in the `R` package `ncvreg`, and can be accessed via the `summary` or `local_mfdr` functions

Closing Remarks and Future Work

- The methods described in this talk are implemented in the `R` package `ncvreg`, and can be accessed via the `summary` or `local_mfdr` functions
- Future work involves:
  - Extensions to the group LASSO, with a focus on reliably selecting non-linear effects
  - Extensions to fusion penalties, with a focus on hotspot or changepoint detection

Thank you!

Thank you!

References

- Breheny, P. and Huang, J. (2011) Coordinate descent algorithms for nonconvexpenalized regression with applications to biological feature selection *Annals of Applied Statistics*
- Breheny, P. (2019) Marginal false discovery rates for penalized regression models *Biostatistics*
- Di, L. (2010) Transcriptional regulation of BRCA1 expression by a metabolic switch *Nat Struct Mol Biology*
- Efron, B. (2012) *Large Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* Cambridge University Press.
- Efron, B. et al. (2015) locfdr: An R package for computing local false discovery rates
- Meinshausen, N. et al (2009) p-values for high-dimensional regression *Journal of the American Statistical Association*
- Miller, R. and Breheny, P. (2019) Marginal false discovery rate control for likelihood-based penalized regression models *Biometrical Journal*
- Miller, R. and Breheny, P. (2019) Feature-specific inference for penalized regression using local false discovery rates. *In Submission*
- Stephens, M. (2016) False discovery rates: a new deal *Biostatistics*
- Simon, N. (2011) Regularization paths for Cox's proportional hazards model via coordinate descent *Journal of Statistical Software*
- Story, J. et al (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach *Journal of the Royal Statistical Society: Series B*
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso *Journal of the Royal Statistical Society*
- Tibshirani, R. (2013) The lasso problem and uniqueness *Electronic Journal of Statistics*
- Tibshirani, R. et al (2016) Exact post selection inference for sequential regression procedures *Journal of the American Statistical Association*