Marginal False Discovery Rates for Group Lasso Regression

Ryan Miller

Goals

1. Review linear models in matrix notation
2. Introduce penalized regression, including the group lasso
3. Introduce marginal false discovery rates for the group lasso

Introduction

- A statistical model is a simplified representation of a complex phenomenon
  - I'll focus on models involving n independent observations of an outcome variable, Y

Introduction

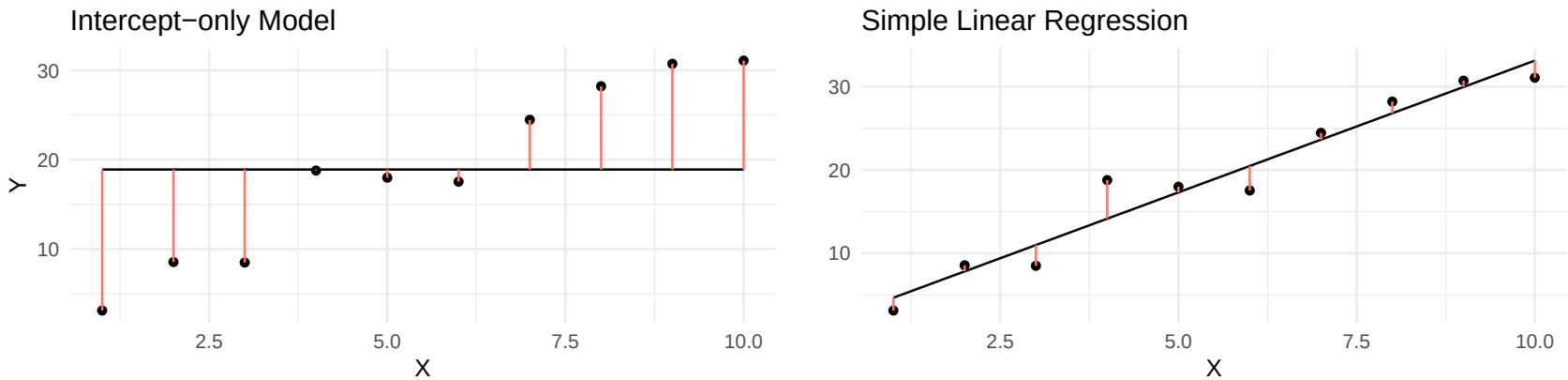
- A statistical model is a simplified representation of a complex phenomenon
  - I'll focus on models involving n independent observations of an outcome variable, Y
- Linear models are popular due to their ease of interpretation:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Every 1 unit change in X predicts a β_1 change in Y

Introduction (model fitting)

Statistical models involve unknown parameters (such as β_0 and β_1) that must be *estimated* using data



A popular estimation method is least squares, which optimizes:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, the model's predicted outcome

Introduction (multiple predictors)

Linear models can accommodate multiple explanatory variables:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \epsilon_i$$

Introduction (multiple predictors)

Linear models can accommodate multiple explanatory variables:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \epsilon_i$$

Matrix notation allows for a more compact expression:

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

- $\mathbf{y}$ is an n-dimensional vector of outcomes
- β is a p-dimensional vector of model coefficients
- $\mathbf{X}$ is an n by p *design matrix* of explanatory variables
- ϵ is an n-dimensional vector of random errors

This framework allows for a closed form expression of the least squares coefficient estimates: $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

High-dimensional data

It's now possible to collect large amounts of data (in terms of both n and p), but this creates several modeling challenges:

- When $p > n$, least squares estimation is inconsistent, $\mathbf{X}^T\mathbf{X}$ doesn't have a unique inverse
- When p is large, interpretation model coefficients becomes difficult
- When p is large, the potential for false discoveries is increased

Penalized regression

Least squares minimizes (with respect to β):

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2 \propto \frac{1}{2n}||\mathbf{y} - \mathbf{X}\beta||^2$$

Penalized regression introduces a penalty term capable of solving many of the issues associated with high-dimensional data:

$$\text{minimize } \frac{1}{2n}||\mathbf{y} - \mathbf{X}\beta||^2 + P_\lambda(\beta)$$

A popular choice is the *lasso penalty*:

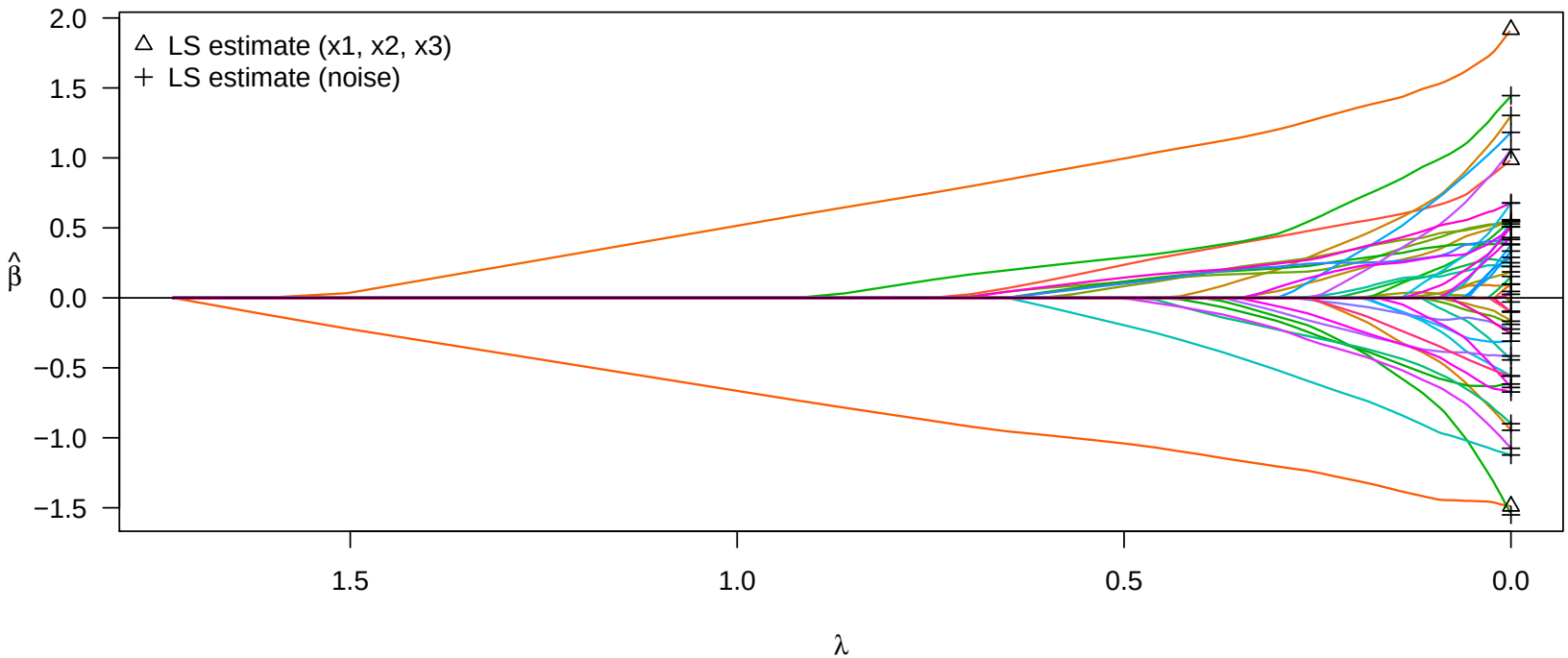
$$P_\lambda(\beta) = \lambda\sum_{k=1}^{p}|\beta_j| = \lambda||\beta||_1$$

Lasso regression (a simple example)

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

$$y_i = 1x_1 - 2x_2 + 2x_3 + \epsilon_i$$

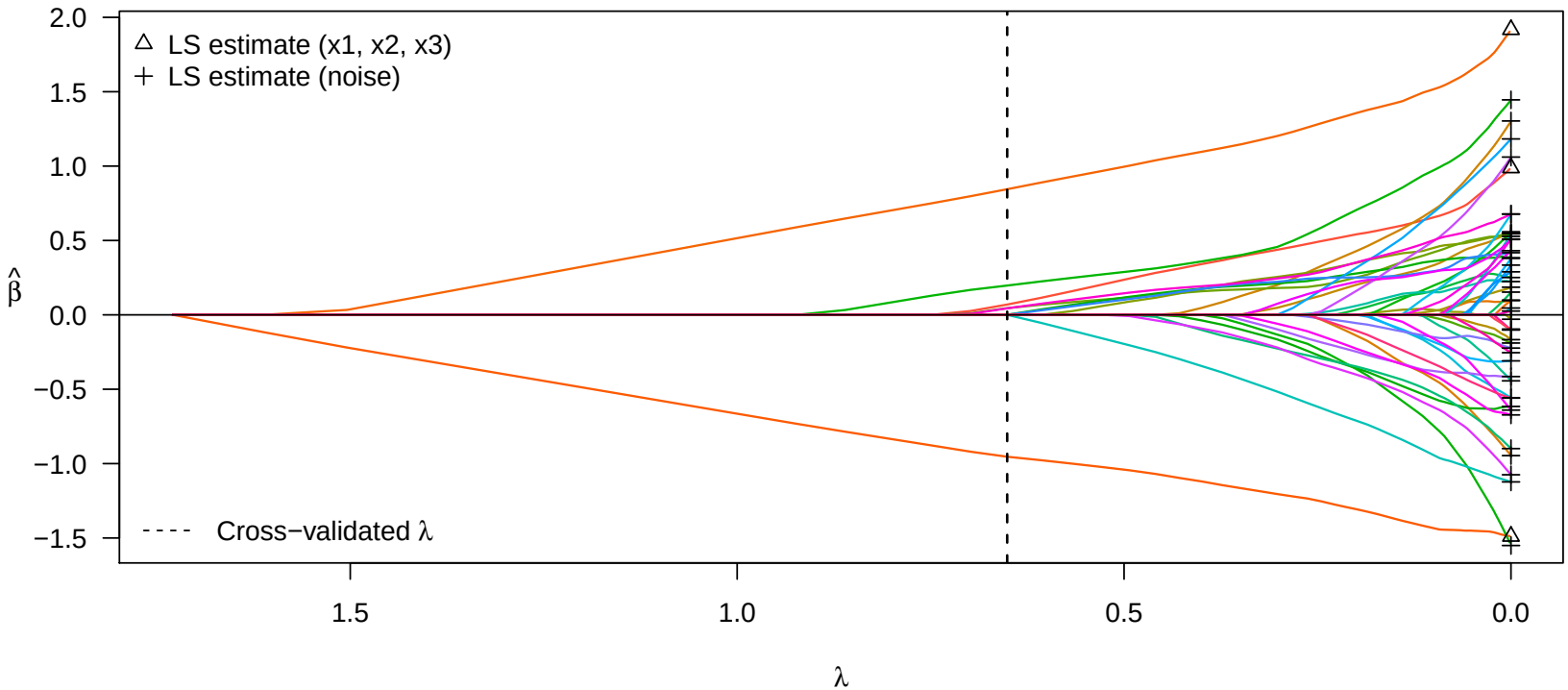
- $\epsilon_i \sim N(0, 4)$ and there are 47 predictors unrelated to Y as "noise"



Notice how least squares (right edge) yields a very bad model!

Lasso regression (choosing λ)

- Lasso regression produces a spectrum of models that vary according to the penalty parameter, λ
  - Cross-validation is a popular way to choose a single one of these models



However, the model favored by cross-validation contains numerous *false discoveries*

Grouped predictors

- Categorical predictors present another challenge for lasso regression
- The conventional approach is to express a nominal categorical predictor using a set of binary indicators

Y	group
8.5	B
11.6	C
9.0	C
9.1	A
8.0	C
9.7	A

Y	X1	X2	X3
8.5	0	1	0
11.6	0	0	1
9.0	0	0	1
9.1	1	0	0
8.0	0	0	1
9.7	1	0	0

- This means the lasso might select some categories, but not others, making it ambiguous as to whether variable as a whole was selected (is it a false discovery or not?)

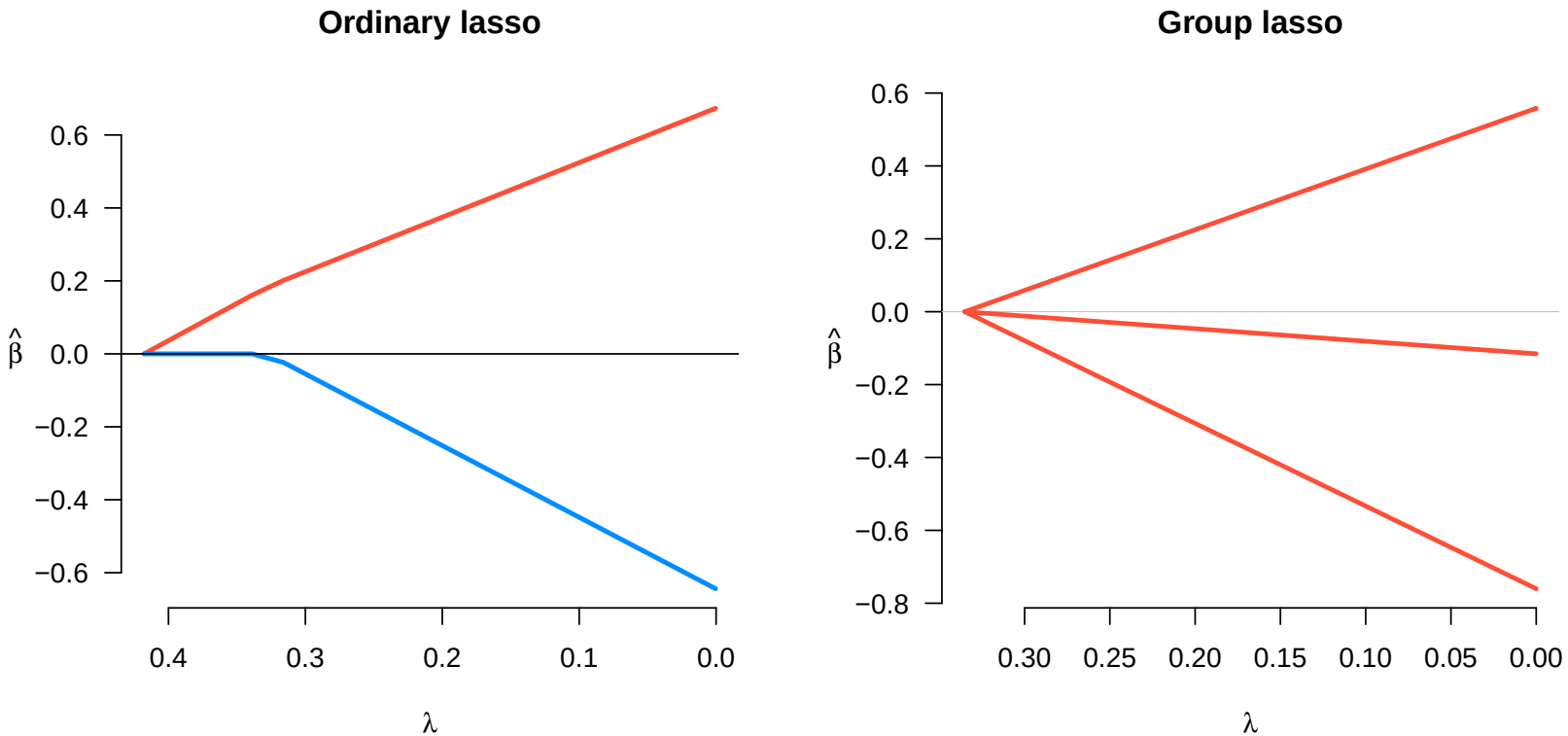
Group lasso regression

The issues surrounding categorical predictors can be addressed by the *group lasso penalty*:

$$P_\lambda(\beta) = \sum_{J=1}^{J} \lambda_j ||\beta_j||$$

- The group lasso penalizes entire groups of model coefficients, resulting in group-level selections
- Typically, $\lambda_j = \sqrt{K_j}\lambda$, where K_j is the dimension of group j

Group lasso (simple example)



Notice how the entire group is selected at once by the group lasso, but not by the ordinary lasso

Marginal false discovery rates (outline)

Quantifying the reliability of selections made by the group lasso:

1. Determine the mathematical conditions for a variable/group to be selected at a given value of λ
2. Use these conditions to estimate the probability of a variable/group that is marginally independent of the outcome being selected by random chance
3. Repeat and sum across all variables/groups under consideration

Marginal false discovery rates (Part 1)

The group lasso uses the following objective function:

$$Q(\beta) = \frac{1}{2n}||\mathbf{y} - \mathbf{X}\beta||^2 + \sum_{j=1}^{J} \lambda_j ||\beta_j||$$

Solving for $\hat{\beta}$ involves the *subdifferential* of Q with respect to β_j:

$$\begin{aligned} -\frac{1}{n}\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}_{-j}\beta_{-j}) + \beta_j + \lambda_j \frac{\beta_j}{||\beta_j||} \quad & \text{if } \beta_j \neq 0 \\ -\frac{1}{n}\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}_{-j}\beta_{-j}) + \lambda_j \mathbf{v} \quad & \text{if } \beta_j = 0 \end{aligned}$$

Where $\mathbf{v}$ is any vector satisfying $||\mathbf{v}|| < 1$

Marginal false discovery rates (Part 1)

For the j^{th} group to have a non-zero role in the model, it must be that the case that:

$$\frac{1}{n}\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}) - \hat{\boldsymbol{\beta}}_j = \lambda_j \frac{\hat{\boldsymbol{\beta}}_j}{||\hat{\boldsymbol{\beta}}_j||}$$

Algebraic manipulation leads to the following selection condition:

$$\frac{1}{n}||\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}_{-j})||^2 > \lambda_j^2$$

The goal now is to estimate the probability that the left-hand side exceeds λ_j^2 by random chance

Marginal false discovery rates (Part 2)

Expanding the left-hand side leads to the following:

$$\begin{aligned}\frac{1}{n}||\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\hat{\beta}_{-j})||^2 &= \frac{1}{n}||\mathbf{X}_j^T(\mathbf{X}\beta+\epsilon-\mathbf{X}_{-j}\hat{\beta}_{-j})||^2 \\ &= \frac{1}{n}||\mathbf{X}_j^T\epsilon-\mathbf{X}_j^T\mathbf{X}_{-j}(\beta_{-j}-\hat{\beta}_{-j})||^2\end{aligned}$$

Marginal false discovery rates (Part 2)

Expanding the left-hand side leads to the following:

$$\begin{aligned}\frac{1}{n}||\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\hat{\beta}_{-j})||^2 &= \frac{1}{n}||\mathbf{X}_j^T(\mathbf{X}\beta+\epsilon-\mathbf{X}_{-j}\hat{\beta}_{-j})||^2 \\ &= \frac{1}{n}||\mathbf{X}_j^T\epsilon-\mathbf{X}_j^T\mathbf{X}_{-j}(\beta_{-j}-\hat{\beta}_{-j})||^2\end{aligned}$$

Noting $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, it can be shown that:

$$\frac{1}{n\sigma^2}||\mathbf{X}_j^T(\mathbf{y}-\mathbf{X}_{-j}\hat{\beta}_{-j})||^2 \xrightarrow{d} \chi^2_{K_j}$$

when $\frac{1}{\sqrt{n}}\mathbf{X}_j^T\mathbf{X}_{-j}(\beta_{-j}-\hat{\beta}_{-j}) \xrightarrow{p} \mathbf{0}$, which occurs when the variables in group j are uncorrelated with those in all other groups

Marginal false discovery rates (Part 3)

- The previous result suggests the probability of any group that is marginally independent of the model's outcome being selected by random chance can be estimated using a Chi-squared distribution

Marginal false discovery rates (Part 3)

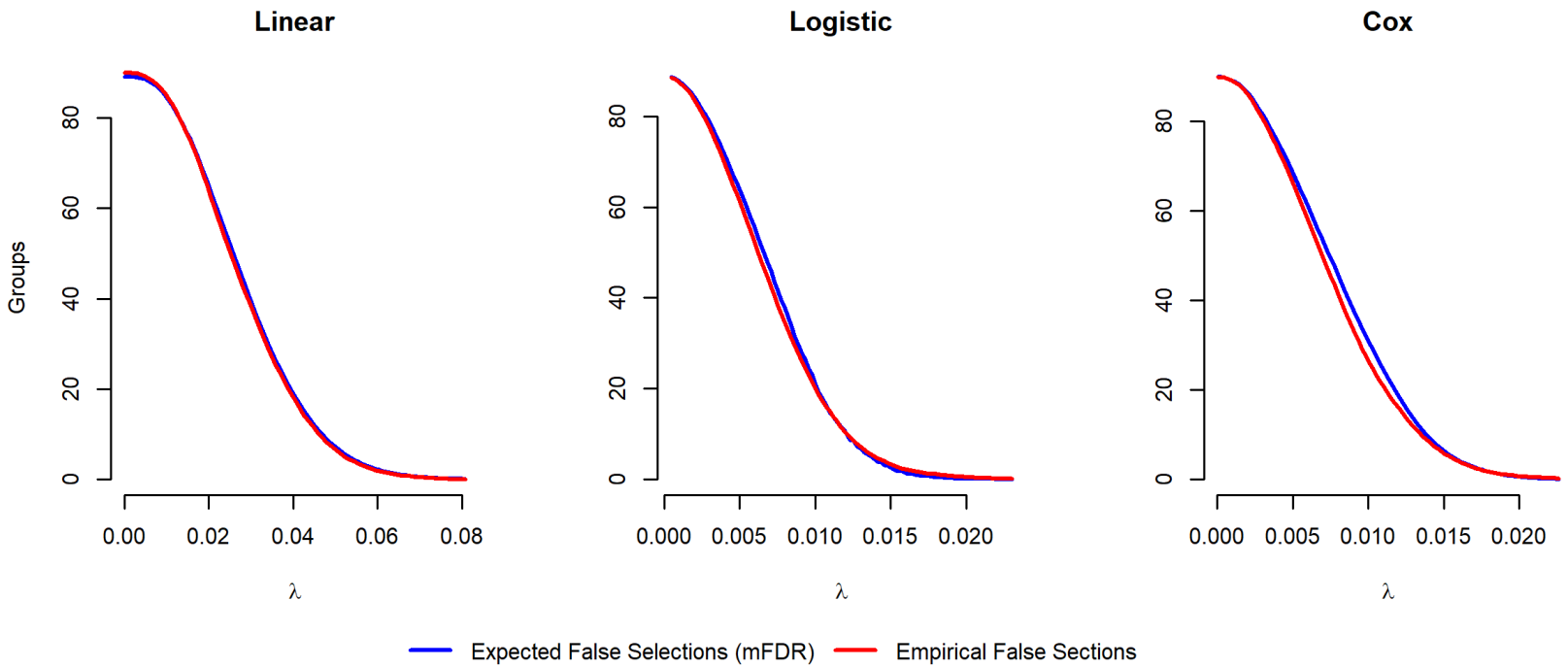
- The previous result suggests the probability of any group that is marginally independent of the model's outcome being selected by random chance can be estimated using a Chi-squared distribution
  - The total number of expected false discoveries can then be expressed as:

$$\widehat{FD} = \sum_{j=1}^{J} \Pr\left(\chi^2_{K_j} > \frac{n\lambda_j^2}{\sigma^2}\right)$$

- Dividing $\widehat{FD}$ by the number of groups that were actually selected provides an estimate of the *marginal false discovery rate*

Marginal false discovery rates (empirical validation)

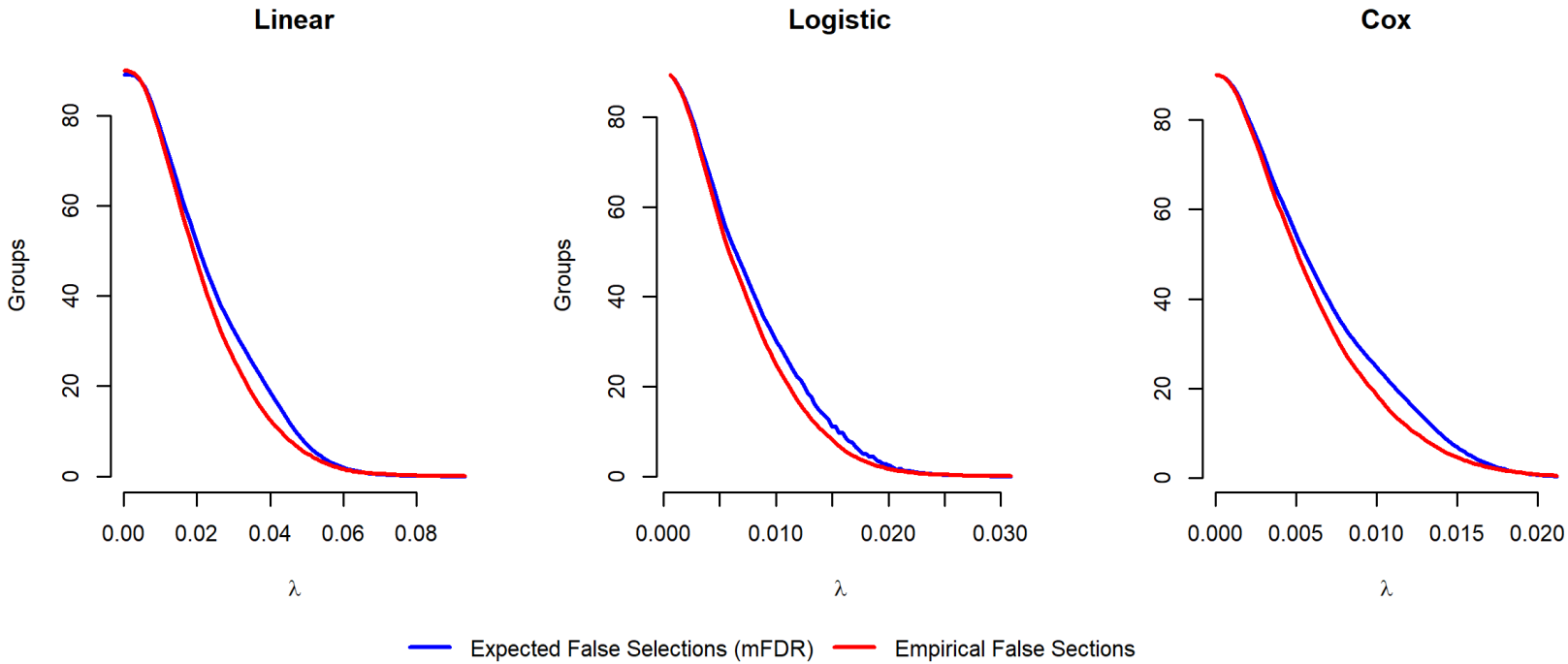
Displayed below are the estimated and actual false discoveries of different group lasso fits averaged across 100 simulated datasets (100 groups, 90 of which are "noise"):



In these simulations all variables are generated to be uncorrelated (satisfying the method's primary assumption)

Marginal false discovery rates (robustness)

Next variables are generated under an autoregressive correlation structure where $cor(\mathbf{x}_a, \mathbf{x}_b) = \rho^{|a-b|}$, for all $a, b \in \{1, \ldots, p\}$:



Correlation makes the method *conservative* in the sense that it will overestimate the false discovery rate (but it will still control it!)

Marginal false discovery rates (application)

- Lung cancer is among the leading causes of death worldwide, partially due to a lack of diagnostic tools for when the disease is in its early stages
- Spira et al (2007) collected RNA expression data for $p = 22,215$ genetic features from the bronchial epitheliums of 102 cases with lung cancer, and 90 controls without lung cancer
- The goal was to identify RNA features that could be used in the early diagnosis of lung cancer

Marginal false discovery rates (application)

Another strength of the group lasso is that it can be used to detect non-linear relationships (expressed using basis expanasion splines):

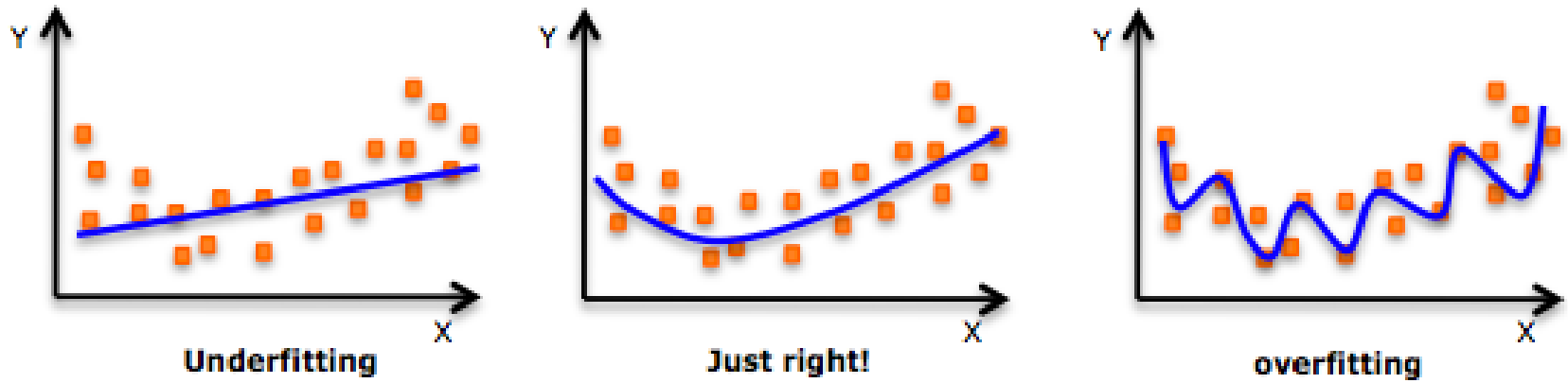


Image credit: https://www.analyticsvidhya.com/blog/2018/03/introduction-regression-splines-python-codes/

Marginal false discovery rates (application)

The table below showcases the downsides of existing approaches relative to using Mfdr in conjunction with the group lasso and basis expansions:

Method	Design Matrix	λ	S	mFDR	MCE
lasso	$\mathbf{X}$	CV	55	100%	24.5%
lasso	$\mathbf{X}$	mFDR	10	7.8%	31.8%
group lasso	$\tilde{\mathbf{X}}$	CV	45	53.4%	25.5%
group lasso	$\tilde{\mathbf{X}}$	mFDR	21	5.6%	27.1%
large-scale testing	$\tilde{\mathbf{X}}$	-	12,902	10.0%	-
large-scale testing	$\mathbf{X}$	-	2,426	10.0%	-

Future work with potential for student involvement

1) Extension to Poisson likelihood - great project for anyone with some familiarly with the Poisson distribution and working with likelihoods

Future work with potential for student involvement

1) Extension to Poisson likelihood - great project for anyone with some familiarly with the Poisson distribution and working with likelihoods
2) Comparisons versus other methods - great project for a student interested in working on `R` implementations of theoretical work that's already been done

Future work with potential for student involvement

1) Extension to Poisson likelihood - great project for anyone with some familiarly with the Poisson distribution and working with likelihoods
2) Comparisons versus other methods - great project for a student interested in working on `R` implementations of theoretical work that's already been done
3) Local false discovery rates - nice mix of theoretical work and programming implementation in `R`, great possibility to experience the design and execution of simulation studies

Future work with potential for student involvement

1) Extension to Poisson likelihood - great project for anyone with some familiarly with the Poisson distribution and working with likelihoods
2) Comparisons versus other methods - great project for a student interested in working on `R` implementations of theoretical work that's already been done
3) Local false discovery rates - nice mix of theoretical work and programming implementation in `R`, great possibility to experience the design and execution of simulation studies

References

1. Breheny, P. and Huang, J. (2012). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Statistics and Computing
2. Breheny, P. J. (2019). Marginal false discovery rates for penalized regression models. Biostatistics
3. Liu, H. and Zhang, J. (2009). Estimation consistency of the group lasso and its applications. In Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics
4. Miller, R. E. and Breheny, P. (2019). Marginal false discovery rate control for likelihood based penalized regression models. Biometrical Journal
5. Simon, N. and Tibshirani, R. (2004). Standardization and the group lasso penalty. Stat. Sinica
6. Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y.-M., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. E. and Brody, J. S. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat. Med.
7. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B
8. Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc. B
9. Zhou, Q. and Min, S. (2017). Estimator augmentation with applications in high-dimensional group inference. Electronic Journal of Statistics

Thank you

Thank you!