Marginal false discovery rates for penalized regression models (Biostatistics, 2019)

Patrick J. Breheny

Chao Cheng

School of Statistics and Management
Shanghai University of Finance and Economics, China

May 17, 2021

Outline

1. Introduction
2. Marginal False Discovery Rates
   - Model Setting
   - Orthonormal Case
   - General Case
3. Numerical Studies
   - Simulation
   - Real Data Analysis
4. Acknowledgement

Outline

1. Introduction
2. Marginal False Discovery Rates
   - Model Setting
   - Orthonormal Case
   - General Case
3. Numerical Studies
   - Simulation
   - Real Data Analysis
4. Acknowledgement

Multiple Inferences

	Accept	Reject	Total
Null is True	$h_0 - V$	V	h_0
Null is False	$h_1 - U$	U	h_1
Total	$h - R$	R	h

Table 1: A summary of multiple comparison procedure results

To control the multiplicity effect:

- FamilyWise Error Rate(FWER): Aim to control the probability of committing any Type-I error in families of comparisons under simultaneous consideration.
- False Discovery Rate(FDR): Consider the number of erroneous rejections, not just whether any error was made.
- Formal definition of FDR varies in literature.

FWER

- Aim to control $P(V \geq 1)$.
- Bonferroni method: For an overall level-α conclusion, each individual test should be considered at α/h.
- Classical procedures that control FWER tend to be so conservative that the power is much lower.
- FWER is important when the conclusion would be erroneous if any individual conclusion is wrong. But this is too restrictive in many cases.

FDR (Benjamini and Hochberg, 1995)

- The overall conclusion need not to be erroneous even some of the null hypotheses are falsely rejected.
- Define FDR to be the expectation

$$\mathrm{E}(V/R) = \mathrm{E}(\mathrm{E}(V/R|R)) = P(R > 0)\,\mathrm{E}(V/R|R > 0)$$

- Control of FWER is control of FDR

$$FDR = \mathrm{E}(V/R) \leq \mathrm{E}(1_{V>0}) = P(V > 0) = FWER$$

- The larger the number of non-true null hypotheses, the larger difference between FDR and FWER, hence potential higher power over FWER.

Introduction

The idea of false discovery is complicated in regression models due to various kinds of conditional independence.

Fully conditional perspective

- $X_j \perp Y | \{X_k\}_{k \neq j}$
- Split the data into two or more parts. (Wasserman and Roeder, 2009), (Dezeure et al., 2015).

Pathwise conditional perspective

- $X_j \perp Y | \{X_k : k \in M_j\}$
- Test the significance of adding a variable along the solution path. (Lockhart et al., 2014), (Tibshirani et al., 2016) and (G'Sell et al., 2015).
- The definition of null hypothesis is constantly changing with λ, which makes its results hard to interpretate.

Introduction

Marginal perspective

- A selected feature j is false if it is marginally independent of the outcome.
- $X_j \perp Y$.
- A simpler definition makes it easier to compute the mFDR.
- A perspective other than conditional FDR, no inclusion nor exclusion implied.

Outline

1. Introduction
2. Marginal False Discovery Rates
   - Model Setting
   - Orthonormal Case
   - General Case
3. Numerical Studies
   - Simulation
   - Real Data Analysis
4. Acknowledgement

Model Setting

- Linear model with normally distributed errors:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon; \quad \varepsilon_i \overset{i.i.d}{\sim} N\left(0, \sigma^2\right),$$

  where $\boldsymbol{X}$ is the $n \times p$ design matrix and $\boldsymbol{y}$ denotes the response.
- Response and covariates are centered so that the intercept term can be ignored.
- The features are standardized so that $\frac{1}{n}\sum_i x_{ij}^2 = 1$ for all j.
- The lasso estimator $\hat{\beta}$ is defined as

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_1 .$$

Orthonormal Case

For a given λ, let $\hat{\boldsymbol{r}} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}$ denote the residuals.

KKT conditions

$$
\begin{aligned}
\frac{1}{n}\boldsymbol{x}_j^T\hat{\boldsymbol{r}} &= \lambda \operatorname{sign}\left(\hat{\beta}_j\right) && \text{for all } \hat{\beta}_j \neq 0 \\
\frac{1}{n}\left|\boldsymbol{x}_j^T\hat{\boldsymbol{r}}\right| &\leq \lambda && \text{for all } \hat{\beta}_j = 0
\end{aligned}
$$

Orthonormal Case

Remove the jth feature, we have $\boldsymbol{X}_{-j}$, β_{-j} and $\hat{\boldsymbol{r}}_j = \boldsymbol{y} - \boldsymbol{X}_{-j}\hat{\beta}_{-j}$.

KKT conditions imply that

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

Therefore the probability that jth variable is selected is

$$P\left(\hat{\beta}_j \neq 0\right) = P\left(\frac{1}{n}\left|\boldsymbol{x}_j^T\hat{\boldsymbol{r}}_j\right| > \lambda\right)$$

Orthonormal Case

In the case of orthonormal design $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$:

$$\begin{aligned}
\frac{1}{n}\boldsymbol{x}_j^T\hat{\boldsymbol{r}}_j &= \frac{1}{n}\boldsymbol{x}_j^T\left(\boldsymbol{y} - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}\right) \\
&= \frac{1}{n}\boldsymbol{x}_j^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j} + \boldsymbol{x}_j\beta_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}\right) \\
&= \frac{1}{n}\boldsymbol{x}_j^T\varepsilon + \beta_j \\
&\sim\ N\left(\beta_j, \frac{\sigma^2}{n}\right).
\end{aligned}$$

Thus for $\beta_j = 0$, we have

$$P\left(\hat{\beta}_j \neq 0\right) = P\left(\frac{1}{n}\left|\boldsymbol{x}_j^T\hat{\boldsymbol{r}}_j\right| > \lambda\right) = 2\Phi\left(-\frac{\sqrt{n}\lambda}{\sigma}\right).$$

Orthonormal Case

Let $\mathcal{S} = \left\{ j : \hat{\beta}_j \neq 0 \right\}$ denote the set of selected features and $\mathcal{N} = \{ j : \beta_j = 0 \}$ the set of null features, for any value of λ. We have

Theorem (Expected number of false discoveries)

In orthonormal case, $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$, *we have*

$$\mathrm{E}\left|\mathcal{S} \cap \mathcal{N}\right| = 2\left|\mathcal{N}\right| \Phi\left(-\frac{\sqrt{n}\lambda}{\sigma}\right).$$

In practice, we have to estimate $|\mathcal{N}|$ and σ.

Orthonormal Case

One straightforward approach is to replace $|\mathcal{N}|$ with p and estimate σ^2 by

$$\hat{\sigma^2} = \frac{\boldsymbol{r}^T \boldsymbol{r}}{n - |\mathcal{S}|}.$$

Then the expected number of false discoveries is

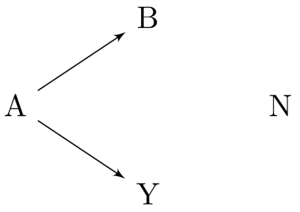
$$\hat{\mathrm{FD}} = 2p\Phi\left(-\sqrt{n}\lambda/\hat{\sigma}\right)$$

and

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

This estimate is somewhat conservative.

General Case



Definition of a False Discovery

- A should never be considered as a false discovery.
- N should always be considered as a false discovery.
- Variable B occupies a gray area as far as false discoveries are concerned.
- Different perspectives lead to different conclusions.

Independent Noise Features

Let $\mathcal{A}, \mathcal{N}$ partition $\{1, \cdots, p\}$ such that $\beta_j = 0$ for all $j \in \mathcal{N}$. And the following condition holds:

$$\lim_{n \to \infty} \frac{1}{n} \boldsymbol{X}^T \boldsymbol{X} = \begin{pmatrix} \Sigma_{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & \Sigma_{\mathcal{N}} \end{pmatrix}.$$

Theorem

Suppose $\Sigma_{\mathcal{N}} = \boldsymbol{I}$. Then for any $j \in \mathcal{N}$ and for λ_n such that $\sqrt{n}\lambda_n$ is bounded,

$$\frac{1}{\sqrt{n}} \boldsymbol{x}_j^T \hat{\boldsymbol{r}}_j \xrightarrow{D} N\left(0, \sigma^2\right).$$

Genralization of the theorem (Miller and Breheny, 2019)

Consider a general penalized likelihood optimization where $\hat{\beta}$ is found by minimizing the objective function

$$Q(\beta|\boldsymbol{X}, \boldsymbol{y}) = -\frac{1}{n} L(\beta|\boldsymbol{X}, \boldsymbol{y}) + P_\lambda(\beta),$$

where $L(\beta|\boldsymbol{X}, \boldsymbol{y})$ is the log-likelihood function and the score function is defined as

$$\boldsymbol{\mu}(\beta) = \nabla L(\beta).$$

The likelihood depneds on $\boldsymbol{X}$ and β often through a linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\beta$. Then we can define $f(\boldsymbol{\eta}|\boldsymbol{y}) = -L(\beta|\boldsymbol{X}, \boldsymbol{y})$ and its second derivative as

$$\boldsymbol{W} = \nabla^2 f.$$

Thus $\nabla^2 L(\beta) = -\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X}$.

Genralization of the theorem (Miller and Breheny, 2019)

Regularity conditions

(A1) Asymptotic normality of the score function: $\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1/2}\boldsymbol{\mu}\left(\beta_0\right) \xrightarrow{d} N(0, \boldsymbol{I})$.

(A2) Vanishing correlation: $\frac{1}{n}\boldsymbol{x}_j^T\boldsymbol{W}\boldsymbol{X}_{-j} \xrightarrow{p} \mathbf{0}$.

(A3) Estimation consistency: $\sqrt{n}\left(\hat{\beta} - \beta_0\right)$ is bounded in probability.

For the original lasso problem, we have $L(\beta) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)$. Therefore $\boldsymbol{\mu}(\beta) = \boldsymbol{X}^T(\boldsymbol{X}\beta - \boldsymbol{y})$ and $\boldsymbol{W} = -\boldsymbol{I}_p$.

- (A2) is not trival and unlikely to be truly satisfied by most features in practice. But in practice, (A2) serves as a worst-case scenario for the correlation structure.

Genralization of the theorem (Miller and Breheny, 2019)

Let $v_j = \boldsymbol{x}_j^T \boldsymbol{W} \boldsymbol{x}_j$.

Theorem

For any solution $\hat{\beta}$ of the lasso-penalized objective, we have $\hat{\beta}_j \neq 0$ if and only if

$$\frac{1}{n}\left|\mu_i\left(\hat{\beta}\right) + v_j\hat{\beta}_j\right| > \lambda.$$

Furthermore, provided that feature j satisfies (A1)-(A3) and $\beta_j = 0$, we have

$$\frac{\mu_j\left(\hat{\beta}\right) + v_j\hat{\beta}_j}{\sqrt{v_j}} \xrightarrow{d} N(0,1).$$

Correlated Noise Features

Difficulty

It's significantly less mathematically tractable, making it harder to construct a rigorous proof.

Observation

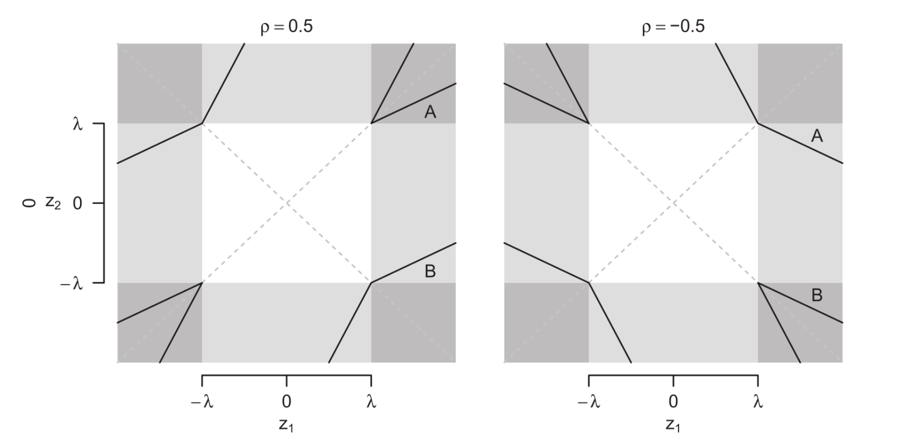
When $\Sigma_{\mathcal{N}} \neq \boldsymbol{I}$, the quantity $\frac{1}{\sqrt{n}}\boldsymbol{x}_j^T\hat{\boldsymbol{r}}_j$ converges to a distribution with thinner tails than $N\left(0, \sigma^2\right)$.

Intuition

If two noise features are correlated, a regression-based method will tend to select a single feature instead of both.

Correlated Noise Features

Illustration of how correlation affects selection in the bivariate case



Correlated Noise Features

Permuting the outcome

1. Randomly permute the outcome $\boldsymbol{y}$, creating new outcomes $\tilde{\boldsymbol{y}}^{(b)}$ for $b = 1, 2, \cdots, B$.
2. For each permutation b, solve for the lasso path $\tilde{\beta}^{(b)}(\lambda; \tilde{\boldsymbol{y}}^{(b)}, \boldsymbol{X})$.
3. Estimate the average number of noise features included in the model for a given λ

$$\hat{\mathrm{FD}}(\lambda) = \frac{\sum_b \# \left\{ \tilde{\beta}_j^{(b)}(\lambda) \neq 0 \right\}}{B},$$

- By permuting $\boldsymbol{y}$, all features belong to $\mathcal{N}$.
- Considerably less conservative than the analytic approach. (doubtable statement)
- Overestimate the noise present in the model.

Correlated Noise Features

Permuting the residual

- Permute the residuals, $\boldsymbol{r}(\lambda) = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}(\lambda)$, of the original lasso fit.
- The method is otherwise identical to that permuting the outcome.

- Only permutate the residuals, not the signals.
- Residuals depend on λ, thus B seperate lasso solutions must be calculated at each value of λ. This substaintially increase the computational burden.
- Essentially unbiased estimate except at very small λ.

Outline

1. Introduction
2. Marginal False Discovery Rates
   - Model Setting
   - Orthonormal Case
   - General Case
3. Numerical Studies
   - Simulation
   - Real Data Analysis
4. Acknowledgement

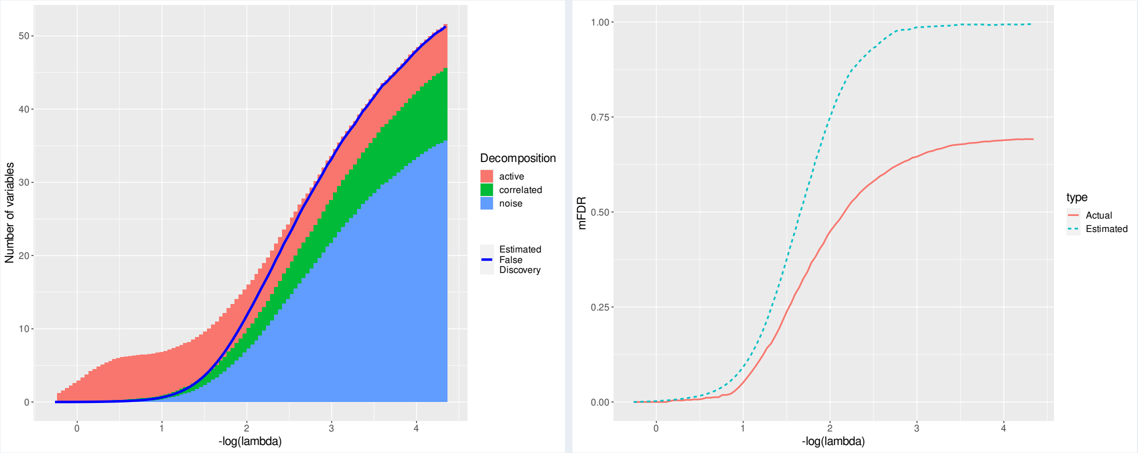
Orthonormal Case

Simulation settings

- In each case, sample size $n = 100$.
- Causative: Six variables with $\beta_j = 1$.
- Correlated: Each causative feature is correlated($\rho = 0.5$) with m other features; $m = 2$ for low-dimensional case and $m = 9$ for high-dimensional case.
- Noise: Independent noise features are added to bring the total number of variables up to 60 in the low-dimensional case and 600 in the high dimensional case.

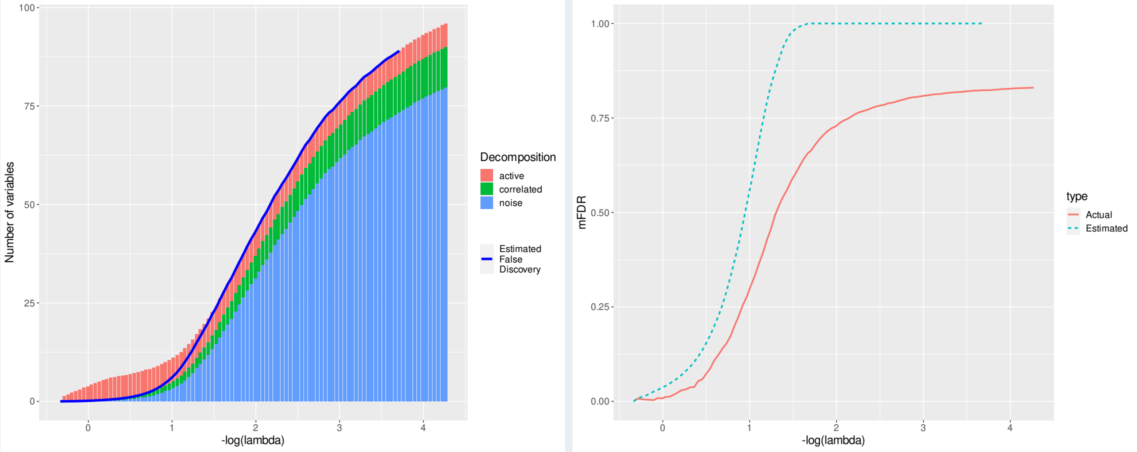
Orthonormal Case

Low-dimensional case



Orthonormal Case

High-dimensional case



Orthonormal Case

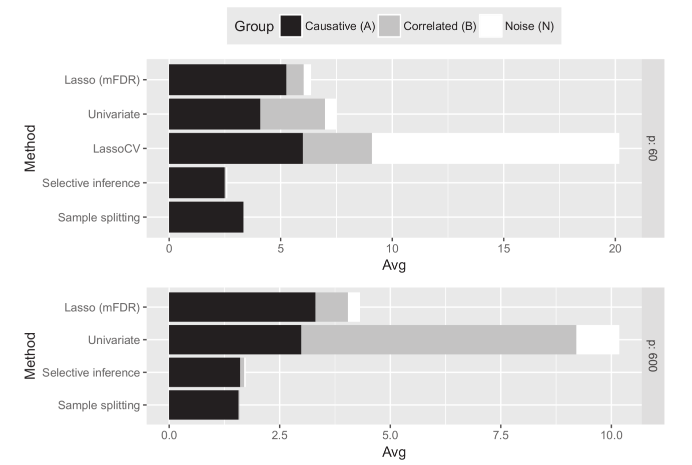
Candidate methods

- Lasso(mFDR): the proposed method.
- Univariate: marginal regression.
- Lasso(CV): model selection based on cross-validation.
- Sample splitting: fully conditional perspective, using (Dezeure et al., 2015).
- Selective inference: pathwise conditional perspective, using (Tibshirani et al., 2016).

When comparing these methods, the nominal false discovery rates were set to 10%.

Orthonormal Case

Comparison against other method

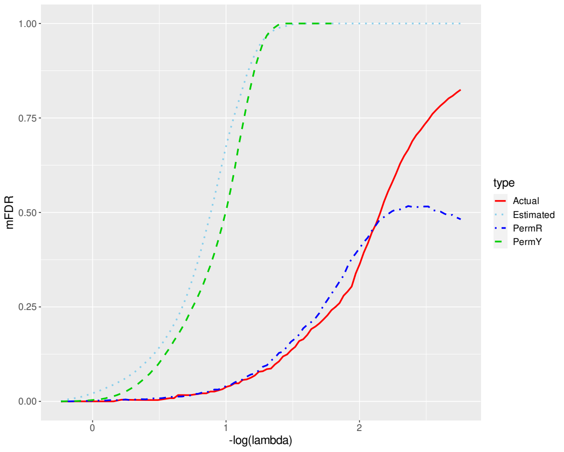


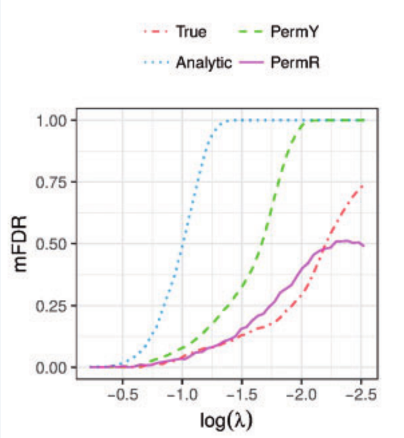
General Case

Simulation settings

- $n = 100$, $p = 500$, $|\mathcal{A}| = 6$, $|\mathcal{B}| = 0$, $|\mathcal{N}| = 494$.
- Autoregressive correlation structure: $\mathrm{cov}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = 0.8^{|i-j|}$, for all $i, j \in \mathcal{N}$.
- Exchangeable correlation structure: $\mathrm{cov}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = 0.8$, for all $i, j \in \mathcal{N}$.

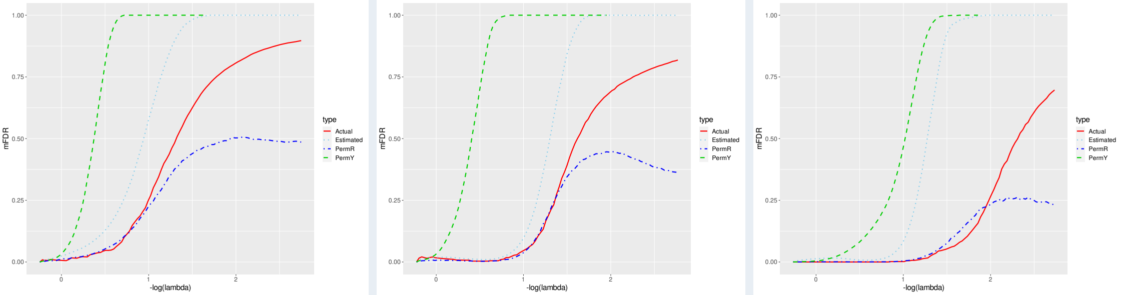
Exchangeable correlation structure





General Case

Autoregressive/Exchangeable correlation structure

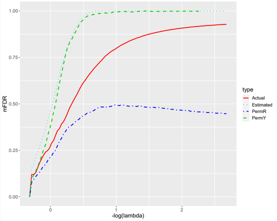


(a) autoregressive(lasso) (b) autoregressive(MCP) (c) exchangeable(MCP)

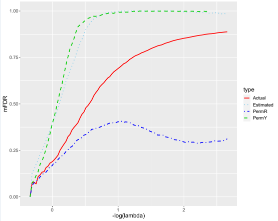
General Case

simulation with a 1 : 1 signal-to-noise ratio($\sigma = \sqrt{6}$)

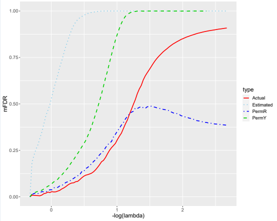
Autoregressive/Exchangeable correlation structure



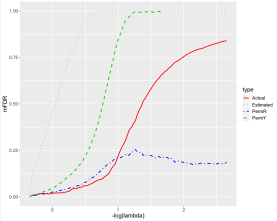
(a) autoregressive(lasso)



(b) autoregressive(MCP)



(c) exchangeable(lasso)



(d) exchangeable(MCP)

Breast Cancer Gene Expression Study

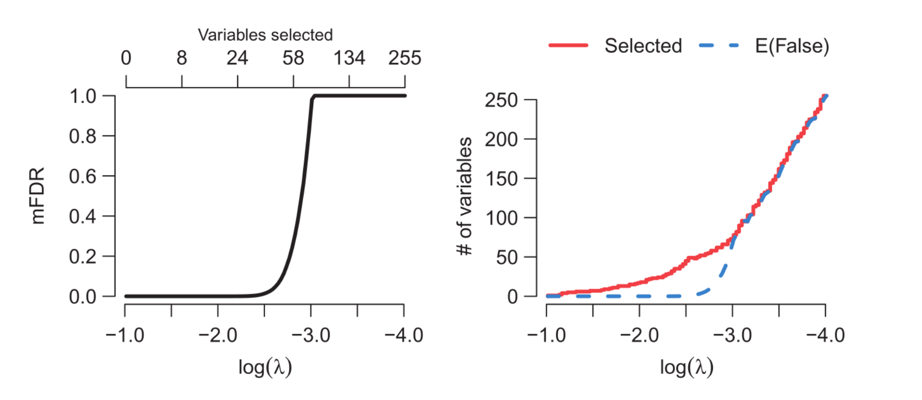
- The BRCA dataset from The Cancer Genome Atlas(TCGA) project.
- There are 536 patients, 17322 genes as predictors and BRCA1 is the gene of interest.

Number of identified genes at 10% FDR

Lasso(mFDR)	Univariate	Sample splitting	Selective inference
55	7903	1	1

Breast Cancer Gene Expression Study

FDR estimates applied to BRCA data



Outline

1. Introduction
2. Marginal False Discovery Rates
   - Model Setting
   - Orthonormal Case
   - General Case
3. Numerical Studies
   - Simulation
   - Real Data Analysis
4. Acknowledgement

Acknowledgement

Thank you all for your attention!

Reference I

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57, 289–300. doi:doi: 10.1111/j.2517-6161.1995.tb02031.x.

Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N., 2015. High-dimensional inference: Confidence intervals, p-values and r-software hdi. Statistical Science 30. doi:doi: 10.1214/15-sts527.

G'Sell, M.G., Wager, S., Chouldechova, A., Tibshirani, R., 2015. Sequential selection procedures and false discovery rate control. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78, 423–444. doi:doi: 10.1111/rssb.12122.

Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R., 2014. A significance test for the lasso. The Annals of Statistics 42. doi:doi: 10.1214/13-aos1175.

Reference II

Miller, R.E., Breheny, P., 2019. Marginal false discovery rate control for likelihood-based penalized regression models. Biometrical Journal 61, 889–901. doi:doi: 10.1002/bimj.201800138.

Tibshirani, R.J., Taylor, J., Lockhart, R., Tibshirani, R., 2016. Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association 111, 600–620. doi:doi: 10.1080/01621459.2015.1108848.

Wasserman, L., Roeder, K., 2009. High-dimensional variable selection. The Annals of Statistics 37. doi:doi: 10.1214/08-aos646.