Regularization methods in multiple regression

Malgorzata Bogdan

University of Wroclaw

March 10, 2023

High dimensional regression

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, \sigma^2 I)$$

High dimensional regression

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, \sigma^2 I)$$

$Y = (Y_1, \ldots, Y_n)^T$ - wektor of trait values for n individuals

High dimensional regression

$$Y_{nx1} = X_{nxp}\beta_{px1} + z_{nx1}, \quad z \sim N(0, \sigma^2 I)$$

$Y = (Y_1, \ldots, Y_n)^T$ - wektor of trait values for n individuals

$X_{n\times p}$ - matrix of regressors

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β does not exist

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β does not exist

Ridge regression:

$$\hat{\beta} = argmin_{\beta \in R^p} L(b) \quad \text{, where } \quad L(b) = ||Y - Xb||^2 + \gamma ||b||^2$$

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β does not exist

Ridge regression:

$$\hat{\beta} = argmin_{\beta \in R^p} L(b) \quad \text{, where} \quad L(b) = ||Y - Xb||^2 + \gamma ||b||^2$$

$$\frac{\partial L(b)}{\partial b} = -2X'(Y - Xb) + 2\gamma b = 0$$

Ridge regression (1)

When $n > p$ but p is large (say $n/2$) the variance of LS estimates may be very large

When $p > n$ the matrix $X'X$ is singular and the LS estimate of β does not exist

Ridge regression:

$$\hat{\beta} = argmin_{\beta \in R^p} L(b) \quad \text{, where} \quad L(b) = ||Y - Xb||^2 + \gamma ||b||^2$$

$$\frac{\partial L(b)}{\partial b} = -2X'(Y - Xb) + 2\gamma b = 0$$

$$-X'Y + (X'X + \gamma I)b = 0 \quad \Leftrightarrow \quad b = (X'X + \gamma I)^{-1} X'Y$$

Ridge regression (1)

$$\hat{\beta} = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

Ridge regression (1)

$$\hat{\beta} = (X'X + \gamma I)^{-1} X'Y, \text{ where } \gamma > 0$$

$$\hat{Y} = X\hat{\beta} = MY, \text{ with } M = X(X'X + \gamma I)^{-1} X'$$

Ridge regression (1)

$$\hat{\beta} = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$\hat{Y} = X\hat{\beta} = MY, \text{ with } \quad M = X(X'X + \gamma I)^{-1} X'$$

$$Tr[M] = Tr\left[(X'X + \gamma I)^{-1} X'X\right]$$

Ridge regression (1)

$$\hat{\beta} = (X'X + \gamma I)^{-1} X'Y, \quad \text{where} \quad \gamma > 0$$

$$\hat{Y} = X\hat{\beta} = MY, \text{ with } \ M = X(X'X + \gamma I)^{-1} X'$$

$$Tr[M] = Tr\left[(X'X + \gamma I)^{-1} X'X\right]$$

$$Tr[M] = \sum_{i=1}^{p} \lambda_i(M), \quad \text{where} \quad \lambda_1(M), \ldots, \lambda_n(M) \quad \text{are eigenvalues of} \quad M$$

Eigenvalues of M

$$X'Xu = \lambda u$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \;\; Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \;\; Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \;\; Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Eigenvalues of M

$$X'Xu = \lambda u$$

$$(X'X + \gamma I)u = (\lambda + \gamma)u, \;\; (X'X + \gamma I)^{-1}u = \frac{1}{\lambda + \gamma}u$$

$$(X'X + \gamma I)^{-1}X'Xu = \frac{\lambda}{\lambda + \gamma}u, \;\; Tr(M) = \sum_{i=1}^{n} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

$$\hat{P}E = RSS + 2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i(X'X)}{\lambda_i(X'X) + \gamma}$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma} X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma} X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

Ridge regression - orthogonal design

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma} X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

$$E(\hat{\beta}_i - \beta_i)^2 = E\left(\frac{1}{1+\gamma}\beta_i - \beta_i + \frac{1}{1+\gamma} Z_i\right)^2$$

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma} X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

$$E(\hat{\beta}_i - \beta_i)^2 = E\left(\frac{1}{1+\gamma}\beta_i - \beta_i + \frac{1}{1+\gamma} Z_i\right)^2$$

$$= \frac{\gamma^2}{(1+\gamma)^2}\beta_i^2 + \frac{\sigma^2}{(1+\gamma)^2}$$

$$X'X = I, \quad \hat{\beta} = \frac{1}{1+\gamma}X'Y = \frac{1}{1+\gamma}(\beta + X'\epsilon)$$

$$Z = X'\epsilon \sim N(0, \sigma^2 I)$$

$$E(\hat{\beta}_i - \beta_i)^2 = E\left(\frac{1}{1+\gamma}\beta_i - \beta_i + \frac{1}{1+\gamma}Z_i\right)^2$$

$$= \frac{\gamma^2}{(1+\gamma)^2}\beta_i^2 + \frac{\sigma^2}{(1+\gamma)^2}$$

$$E||\hat{\beta} - \beta||^2 = \frac{\gamma^2}{(1+\gamma)^2}||\beta||^2 + \frac{p\sigma^2}{(1+\gamma)^2}$$

When rigde is better than LS ?

Ridge regression - orthogonal design (2)

When rigde is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

Ridge regression - orthogonal design (2)

When rigde is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

Ridge is always better than LS when $||\beta||^2 < p\sigma^2$

Ridge regression - orthogonal design (2)

When rigde is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

Ridge is always better than LS when $||\beta||^2 < p\sigma^2$

Otherwise, when

$$|||\beta||^2 < \frac{\gamma + 2}{\gamma} p\sigma^2$$

When rigde is better than LS ?

$$\frac{\gamma^2||\beta||^2 + p\sigma^2}{(1+\gamma)^2} < p\sigma^2$$

Ridge is always better than LS when $||\beta||^2 < p\sigma^2$

Otherwise, when

$$|||\beta||^2 < \frac{\gamma+2}{\gamma} p\sigma^2$$

$$\gamma < \frac{2p\sigma^2}{||\beta||^2 - p\sigma^2}$$

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover β by minimizing $||b||_1 = \sum_{i=1}^{n} |b_i|$ subject to $Y = Xb$.

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover β by minimizing $||b||_1 = \sum_{i=1}^{n} |b_i|$ subject to $Y = Xb$.

BP can recover β if it is *identifiable* with respect to L_1 norm, i.e.

$$\text{If} \quad X\gamma = X\beta \text{ and } \gamma \neq \beta \ \text{ then } \ \|\gamma\|_1 > \|\beta\|_1.$$

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover β by minimizing $||b||_1 = \sum_{i=1}^{n} |b_i|$ subject to $Y = Xb$.

BP can recover β if it is *identifiable* with respect to L_1 norm, i.e.

$$\text{If} \quad X\gamma = X\beta \text{ and } \gamma \neq \beta \ \text{ then } \ \|\gamma\|_1 > \|\beta\|_1.$$

$$k = ||\beta||_0 = \#\{i : \ \beta_i \neq 0\}$$

Basis Pursuit

$$Y = X\beta$$

Basis Pursuit (Chen and Donoho, 1994): when $p > n$ recover β by minimizing $||b||_1 = \sum_{i=1}^n |b_i|$ subject to $Y = Xb$.

BP can recover β if it is *identifiable* with respect to L_1 norm, i.e.

$$\text{If} \quad X\gamma = X\beta \text{ and } \gamma \neq \beta \ \text{ then } \ \|\gamma\|_1 > \|\beta\|_1.$$

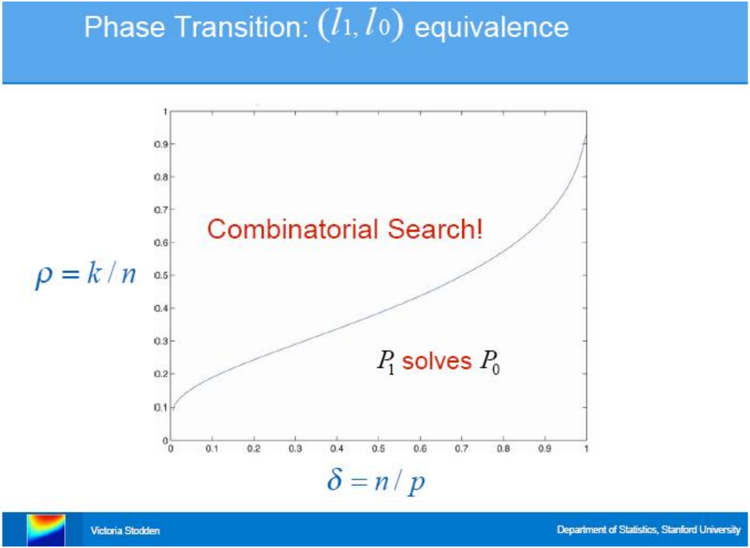
$$k = ||\beta||_0 = \#\{i: \ \beta_i \neq 0\}$$

Basis Pursuit can recover β if k is small enough.

Transition curve (Donoho and Tanner, 2005)

Let's assume than $p \to \infty$, $n/p \to \delta$ and $k/n \to \epsilon$.

If X_{ij} are iid $N(0, \tau^2)$ then the probability that BP recovers β converges to 1 if $\epsilon < \rho(\delta)$ and to 0 if $\epsilon > \rho(\delta)$, where $\rho(\delta)$ is the *transition curve*.


Phase Transition: (l_1, l_0) equivalence
Combinatorial Search!
P_1 solves P_0
$\rho = k/n$
$\delta = n/p$
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Victoria Stodden
Department of Statistics, Stanford University

Noisy case - multiple regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Noisy case - multiple regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Convex program: Minimize $||b||_1$ subject to $||Y - Xb||_2^2 \leq \epsilon$

Or alternatively: $\min_{b\in R^p} ||y - Xb||_2^2 + \lambda ||b||_1$

Noisy case - multiple regression

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + z_{n\times 1}, \quad z \sim N(0, \sigma I)$$

Convex program: Minimize $||b||_1$ subject to $||Y - Xb||_2^2 \leq \epsilon$

Or alternatively: $\min_{b\in R^p} ||y - Xb||_2^2 + \lambda ||b||_1$

BPDN (Chen and Donoho, 1994) or LASSO (Tibshirani, 1996)

Selection of the tuning parameter for LASSO

- General rule: the reduction of λ_L results in identification of more elements from the true support (true discoveries) but at the same time it produces more falsely identified variables (false discoveries)
- The choice of λ_L is challenging- e.g. crossvalidation typically leads to many false discoveries
- When $X^T X = I$ Lasso selects X_j iff $|\hat{\beta}_j^{LS}| > \lambda$
- Selection $\lambda = \sigma\Phi^{-1}(1 - \alpha/(2p)) \approx \sigma\sqrt{2\log p}$ corresponds to Bonferroni correction and controls FWER.

Irrepresentability condition

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$

Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\overline{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$.

Irrepresentability condition

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$

Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\overline{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$.

Irrepresentable condition:

$$\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

Irrepresentability condition

The sign vector of β is defined as
$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$,
where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$

Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\overline{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$.

Irrepresentable condition:

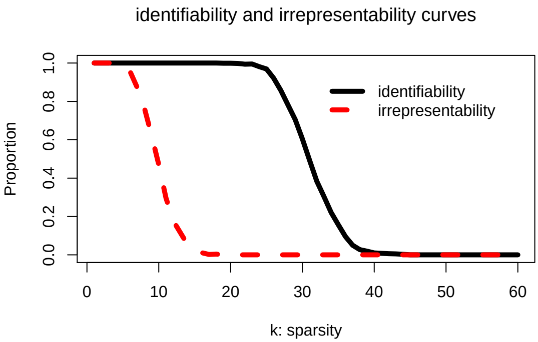
$$\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

When

$$\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty > 1$$

then probability of the support recovery by LASSO is smaller than 0.5 (Wainwright, 2009).

n=100, p=300, elements of X were generated as iid N(0,1)



Identifiability condition

Definition (Identifiability)

Let X be a $n \times p$ matrix. The vector $\beta \in R^p$ is said to be identifiable with respect to the l^1 norm if the following implication holds

$$X\gamma = X\beta \text{ and } \gamma \neq \beta \Rightarrow \|\gamma\|_1 > \|\beta\|_1. \tag{1}$$

Theorem (Tardivel, Bogdan, 2019)

For any $\lambda > 0$ LASSO can separate well the causal and null features if and only if vector β is identifiable with respect to l_1 norm and $min_{i \in I}|\beta_i|$ is sufficiently large.

Modifications of LASSO

Corollary

Appropriately thresholded LASSO can properly identify the sign of sufficiently large β if and only if β is identifiable with respect to l_1 norm.

Conjecture

Adaptive (reweighted) LASSO can properly identify the sign of sufficiently large β if and only if β is identifiable with respect to l_1 norm.

Adaptive LASSO

Adaptive LASSO [Zou, *JASA* 2006], [Candès, Wakin and Boyd, *J. Fourier Anal. Appl.* 2008]

$$\beta_{aL} = argmin_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda \sum_{i=1}^{p} w_i |b|_i \right\}, \qquad (2)$$

where $w_i = \frac{1}{\hat{\beta}_i}$, and $\hat{\beta}_i$ is some consistent estimator of β_i.

Adaptive LASSO

Adaptive LASSO [Zou, *JASA* 2006], [Candès, Wakin and Boyd, *J. Fourier Anal. Appl.* 2008]

$$\beta_{aL} = argmin_b \left\{ \frac{1}{2} \|y - Xb\|_2^2 + \lambda \sum_{i=1}^{p} w_i |b|_i \right\}, \tag{2}$$

where $w_i = \frac{1}{\hat{\beta}_i}$, and $\hat{\beta}_i$ is some consistent estimator of β_i.
Reduces bias and improves model selection properties

AMP theory for LASSO, (Bayati and Montanari, 2012)

$$X_{ij} \sim \mathcal{N}(0, 1/n), \quad z_i \sim \mathcal{N}(0, \sigma^2)$$

AMP theory for LASSO, (Bayati and Montanari, 2012)

$X_{ij} \sim \mathcal{N}(0, 1/n), \quad z_i \sim \mathcal{N}(0, \sigma^2)$

$\beta_1, \ldots, \beta_p$: iid, distributed as the random variable Π, such that $\mathbb{E}\,\Pi < \infty, \quad \mathbb{P}(\Pi \neq 0) = \epsilon \in (0, 1)$.

AMP theory for LASSO, (Bayati and Montanari, 2012)

$X_{ij} \sim \mathcal{N}(0, 1/n), \quad z_i \sim \mathcal{N}(0, \sigma^2)$

$\beta_1, \ldots, \beta_p$: iid, distributed as the random variable Π, such that $\mathbb{E}\,\Pi < \infty, \quad \mathbb{P}(\Pi \neq 0) = \epsilon \in (0, 1)$.

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\Big(\eta_{\alpha\tau}(\Pi + \tau Z) - \Pi\Big)^2,$$

$$\lambda = \Big(1 - \frac{1}{\delta}\mathbb{P}(|\Pi + \tau Z| > \alpha\tau)\Big)\alpha\tau.$$

AMP theory for LASSO, (Bayati and Montanari, 2012)

Theorem

For any pseudo-Lipschitz function φ*, the lasso solution* $\hat{\beta}$ *with fixed* λ *obeys*

$$\frac{1}{p}\sum_{i=1}^{p}\varphi(\hat{\beta}_i,\beta_i) \longrightarrow \mathbb{E}\varphi(\eta_{\alpha\tau}(\Pi+\tau Z),\Pi)$$

AMP formulas for FDR and Power

$\widehat{\mathcal{S}}$ - set of variables selected by LASSO

$$\mathsf{FDP} \equiv \frac{|\widehat{\mathcal{S}} \cap \mathcal{H}_0|}{|\widehat{\mathcal{S}}|}$$

$$FDR = E(FDP)$$

AMP formulas for FDR and Power

$\widehat{\mathcal{S}}$ - set of variables selected by LASSO

$$\text{FDP} \equiv \frac{|\widehat{\mathcal{S}} \cap \mathcal{H}_0|}{|\widehat{\mathcal{S}}|}$$

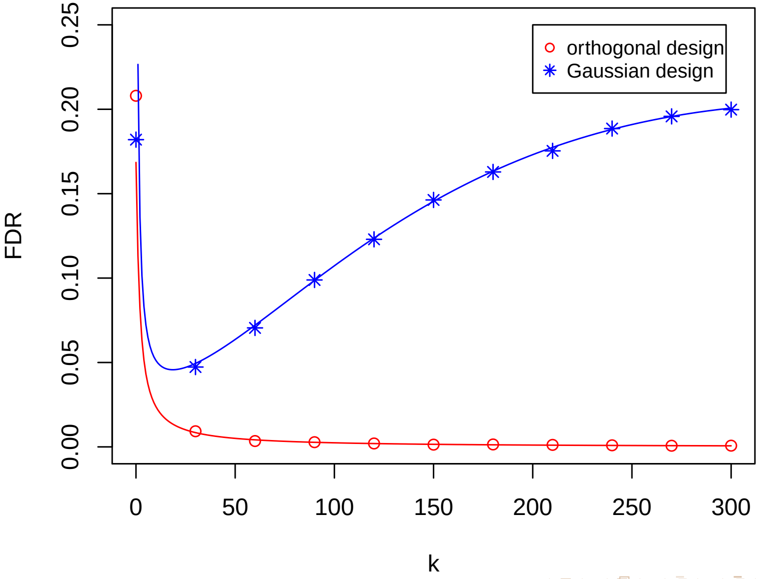
$$FDR = E(FDP)$$

Bogdan, van den Berg, Su and Candés, 2013

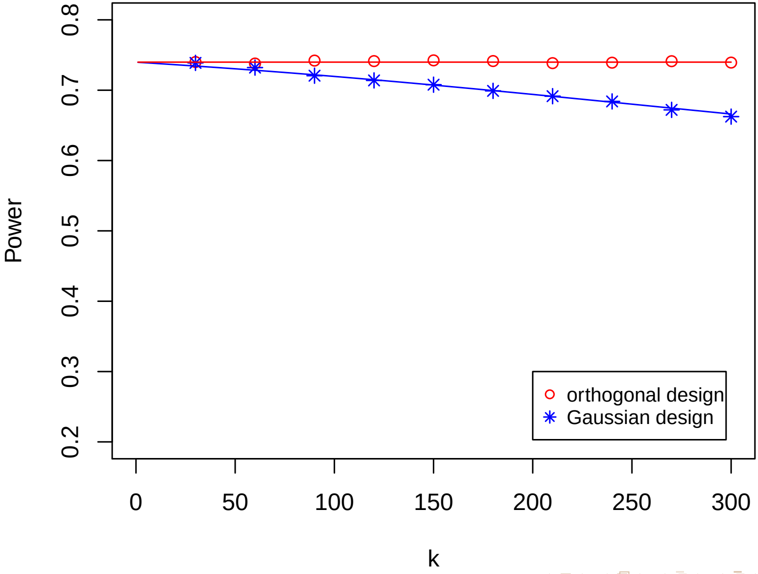
$$\text{FDR} \to \frac{2\mathbb{P}(\Pi = 0)\Phi(-\alpha)}{\mathbb{P}(|\Pi + \tau Z| > \alpha\tau)} \quad ,$$

$$\text{Power} \to \mathbb{P}(|\Pi + \tau Z| > \alpha\tau | \Pi \neq 0).$$

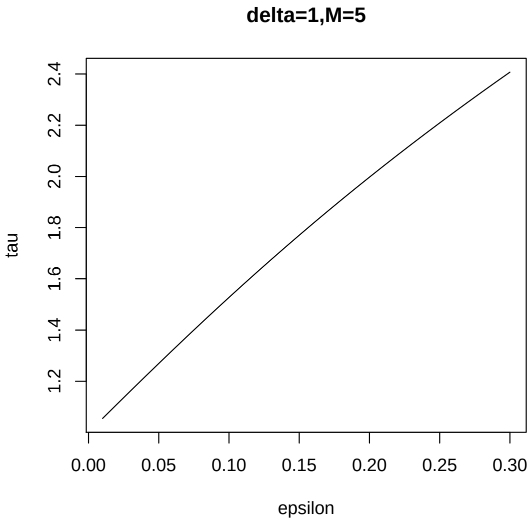
FDR - illustration



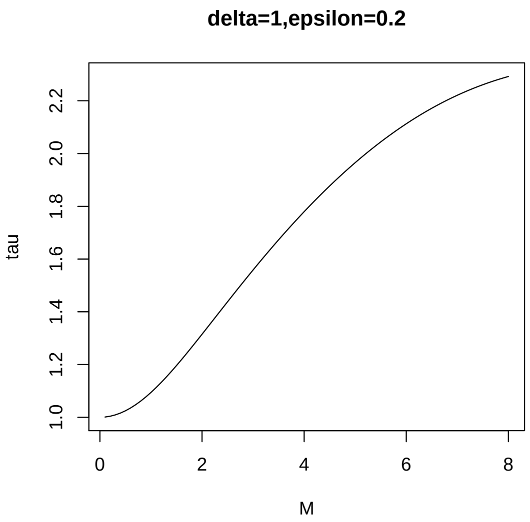
Power - illustration



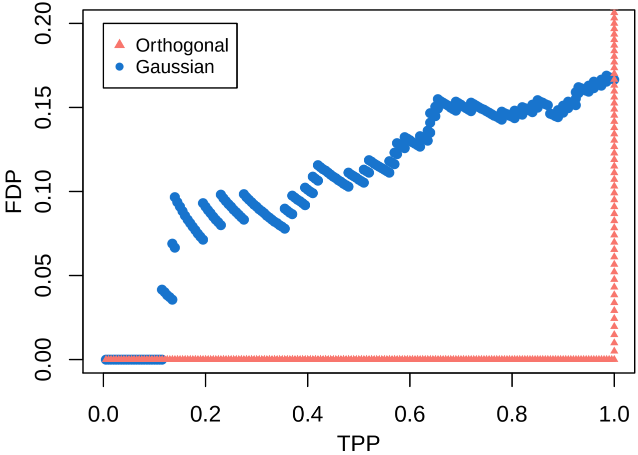
Magnitude of additional noise (1)



Magnitude of additional noise (2)



Su, Bogdan and Candes, (2017), $\delta = 1$, $\epsilon = 0.2$



FDP-Power tradeoff

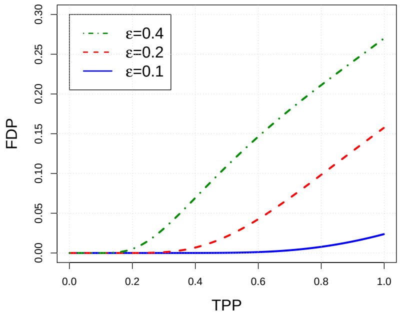
Theorem (Su, Bogdan, Candes, 2017)

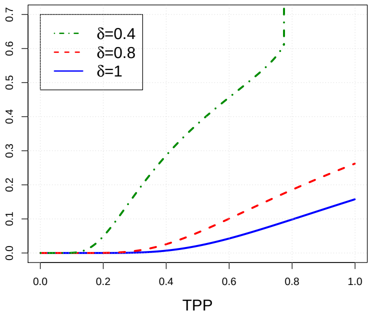
Fix $\delta \in (0, \infty)$ *and* $\epsilon \in (0, 1)$. *Then the event*

$$\bigcap_{\lambda \geq 0.01} \left\{ FDP(\lambda) \geq q^{\star}\left(TPP(\lambda)\right) - 0.001 \right\} \tag{3}$$

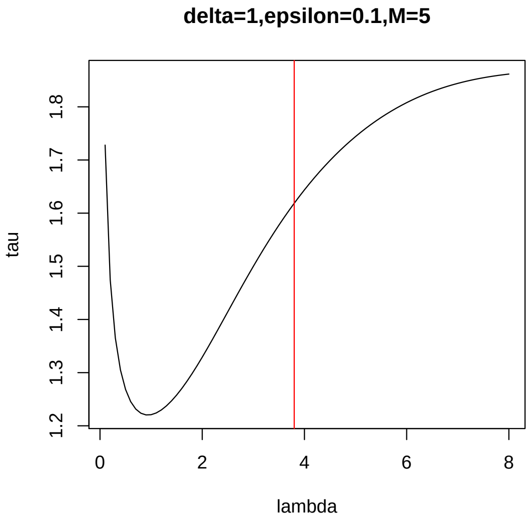
holds with probability tending to one.

FDR-Power trade-off (2)

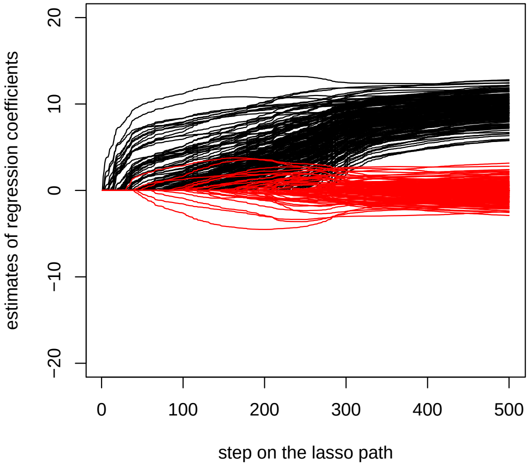




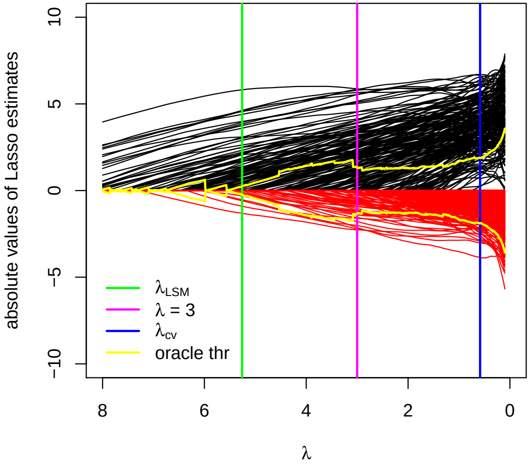
Magnitude of noise



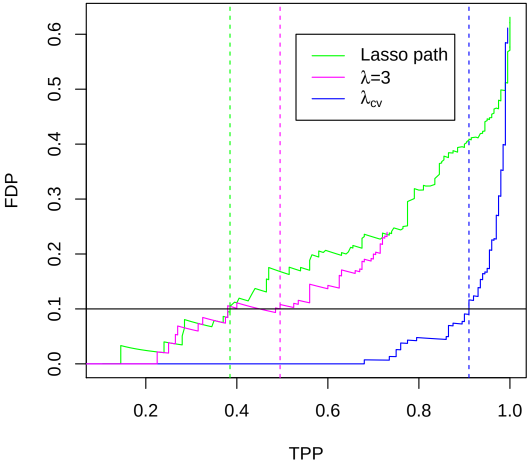
Thresholded LASSO (1)



Thresholded LASSO (2)



Thresholded LASSO (3)



False Discoveries in Random Designs

Candès, Fan, Janson and Lv (2017) - model free knockoffs

False Discoveries in Random Designs

Candès, Fan, Janson and Lv (2017) - model free knockoffs

Consider n i.i.d random vectors $(Y_i, X_{1i}, \ldots, X_{pi})$

False Discoveries in Random Designs

Candès, Fan, Janson and Lv (2017) - model free knockoffs

Consider n i.i.d random vectors $(Y_i, X_{1i}, \ldots, X_{pi})$

Variable X_j is a null variable, if $Y \perp X_j \mid X_{-j}$, where X_{-j} denotes the remaining $p-1$ variables excluding X_j

Construct a set of "fake" covariates $\tilde{X} = (\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p)$ which satisfy:

1. *Exchangeability:* for any subset $S \subset \{1, \ldots, p\}$,
$$(X, \tilde{X})_{\text{swap}(S)} \stackrel{d}{=} (X, \tilde{X}), \tag{4}$$
where $swap(S)$ is obtained by swapping the entries X_j and $\tilde{X}_j$ for each $j \in S$.
2. *Unimportant variables:* $\tilde{X} \perp Y | X$, which can be guaranteed if $\tilde{X}$ is constructed without looking at Y.

Model Free Knockoffs for Gaussian Designs (1)

If $X \sim \mathcal{N}(0, \Sigma)$, then a joint distribution of $(X, \tilde{X})$ can be:

$$(X, \tilde{X}) \sim \mathcal{N}(0, G), \quad \text{where } G = \begin{bmatrix} \Sigma & \Sigma - \mathsf{diag}(s) \\ \Sigma - \mathsf{diag}(s) & \Sigma \end{bmatrix}, \tag{5}$$

with any choice of the diagonal matrix $\mathsf{diag}(s)$ *s.t.* G is positive semidefinite. Possible choice of s:

$$s_j = 2\lambda_{\min}(\Sigma) \wedge 1, \ \forall j \,, \tag{6}$$

Model Free Knockoffs for Gaussian Designs (2)

Sample $\tilde{X}$ from its conditional distribution:

$$\tilde{X} \mid X \overset{d}{=} \mathcal{N}(\mu_c, \Sigma_c),$$

where

$$\begin{aligned} \mu_c &= X - X\Sigma^{-1}\mathrm{diag}(s) \\ \Sigma_c &= 2\mathrm{diag}(s) - \mathrm{diag}(s)\Sigma^{-1}\mathrm{diag}(s). \end{aligned} \tag{7}$$

Knockoff statistics (1)

$$W = (W_1, \ldots, W_p)\ , \quad W_j = w_j([X, \tilde{X}], Y)$$

Knockoff statistics (1)

$$W = (W_1, \ldots, W_p)\ ,\ \ W_j = w_j([X, \tilde{X}], Y)$$

flip-sign property:

$$w_j([X, \tilde{X}]_{swap(S)}, Y) = \begin{cases} w_j([X, \tilde{X}], Y), & j \notin S \\ -w_j([X, \tilde{X}], Y), & j \in S \end{cases}$$

Knockoff statistics (2)

Example:

$$T = (Z, \tilde{Z}) = t([X, \tilde{X}], Y)$$

Knockoff statistics (2)

Example:

$$T = (Z, \tilde{Z}) = t([X, \tilde{X}], Y)$$

$$(Z, \tilde{Z})_{swap(S)} = t([X, \tilde{X}]_{swap(S)}, Y)$$

Knockoff statistics (2)

Example:

$$T = (Z, \tilde{Z}) = t([X, \tilde{X}], Y)$$

$$(Z, \tilde{Z})_{swap(S)} = t([X, \tilde{X}]_{swap(S)}, Y)$$

$$w_j = f_j(Z_j, \tilde{Z}_j), \quad f(v, u) = -f(u, v)$$

Knockoff statistics (2)

Example:

$$T = (Z, \tilde{Z}) = t([X, \tilde{X}], Y)$$

$$(Z, \tilde{Z})_{swap(S)} = t([X, \tilde{X}]_{swap(S)}, Y)$$

$$w_j = f_j(Z_j, \tilde{Z}_j), \quad f(v, u) = -f(u, v)$$

Lasso Coefficient Difference (LCD) statistic:

$$W_j = |\hat{\beta}_j(\lambda)| - |\hat{\beta}_{j+p}(\lambda)|$$

Knockoff filter

Define a random threshold as

$$\hat{t}(\lambda) = \min\left\{ t > 0 : \frac{1 + \#\{j : W_j(\lambda) \le -t\}}{\#\{j : W_j(\lambda) \ge t\}} \le q \right\}$$

and select

$$\widehat{\mathcal{S}(\lambda)} = \{j : W_j(\lambda) \ge \hat{t}(\lambda)\},$$

Knockoff filter

Define a random threshold as

$$\hat{t}(\lambda) = \min\left\{t > 0 : \frac{1 + \#\{j : W_j(\lambda) \leq -t\}}{\#\{j : W_j(\lambda) \geq t\}} \leq q\right\}$$

and select

$$\widehat{\mathcal{S}(\lambda)} = \{j : W_j(\lambda) \geq \hat{t}(\lambda)\},$$

Candès, Fan, Janson and Lv (2017) - The above knockoff procedure $KN(\lambda, q)$ controls FDR at the level q.

Breaking through FDR-Power diagram

Weinstein, Su, Bogdan, Barber, Candès (2023, to appear in AOS)

Definition

A random variable Π is said to be ϵ-sparse if $\mathbb{E}\,\Pi^2 < \infty$ and $\mathbb{P}(\Pi \neq 0) = \epsilon$.

Theorem

Assume that $\epsilon/2 < \epsilon_{\mathsf{DT}}(\delta/2)$, where $\epsilon_{\mathsf{DT}}(\delta)$ is a point on the Donoho-Tanner transition curve. Then for any fixed $0 < \lambda_1 < \lambda_2, 0 < q < 1$, and any $\nu > 0$, there exists an ϵ-sparse prior Π and n' such that

$$\mathbb{P}\left(\inf_{\lambda_1 \le \lambda \le \lambda_2} (\lambda, \Pi, q, n, p) > 1 - \nu\right) \geq 1 - \nu$$

if $n \geq n'$.

How to pick λ

Given that

$$\hat{\beta}_i \sim \tau\eta_\alpha\left(\frac{\Pi}{\tau} + Z\right)$$

the "best" ordering of $\hat{\beta}_i$ occurs when τ is minimal.

How to pick λ

Given that

$$\hat{\beta}_i \sim \tau\eta_\alpha\left(\frac{\Pi}{\tau} + Z\right)$$

the "best" ordering of $\hat{\beta}_i$ occurs when τ is minimal.

Bayati and Montanari (2012):

$$\frac{1}{p}||\hat{\beta} - \beta||^2 \rightarrow \delta(\tau^2 - \sigma^2)$$

How to pick λ

Given that

$$\hat{\beta}_i \sim \tau\eta_\alpha\left(\frac{\Pi}{\tau} + Z\right)$$

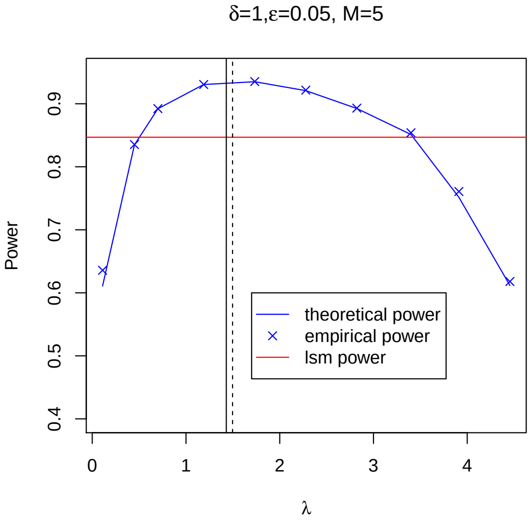
the "best" ordering of $\hat{\beta}_i$ occurs when τ is minimal.

Bayati and Montanari (2012):

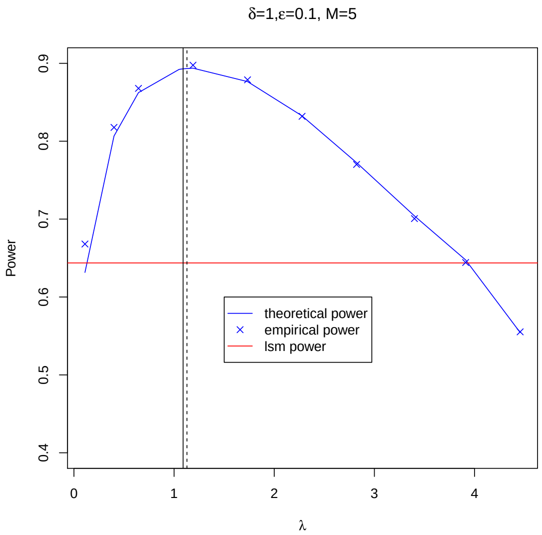
$$\frac{1}{p}||\hat{\beta} - \beta||^2 \to \delta(\tau^2 - \sigma^2)$$

Thus minimizing τ corresponds to minimizing the prediction error.
Optimal τ can be identified through crossvalidation

Gain in power over LSM



Gain in power over LSM



Other examples of applications of the AMP theory or the mean field asymptotics

G. Reeves, 2017, neural networks

Other examples of applications of the AMP theory or the mean field asymptotics

G. Reeves, 2017, neural networks

P.Sur and E.J.Candès, 2018, maximum likelihood estimators in logistic regression

The method relies on two de-randomizaton steps. First, the knockoff threshold value τ is calculated based on many knockoff test statistics:

$$W_{mj} = Z_{mj} - \tilde{Z}_{mj}, \qquad m = 1, 2, \cdots, M \text{ and } j = 1, 2, \cdots, p,$$

$$\tau = \min\left\{ t : \frac{1}{M} \sum_{m=1}^{M} \frac{\#\{j : W_{mj} \leq -t\} + c}{\#\{j : W_{mj} \geq t\} \vee 1} \leq q \right\} \tag{8}$$

with $c = \frac{1}{m}$.
In the second step for each $j = 1, 2, \cdots, p$ we calculate the median $med(W_j) = median(W_{mj})$ over $m = 1, \ldots, M$ and reject the j-th variable if $med(W_j) \geq \tau$.

Theorem

Consider the single knockoffs procedure, which rejects $H_{0j} : \beta_j = 0$ if the feature statistics W_j satisifes $W_j > t$ and let

$$FDR(t) = \mathbb{E}\left[\frac{\#\{j \in H_0 : W_j \geq t\}}{1 \vee \#\{j : W_j \geq t\}}\right] \quad . \tag{9}$$

If for each $i \in 1, \ldots, m$ it holds that the signs of the feature statistics W_{mj} , $j \in \{1, \ldots, p\}$ are i.i.d coin flips then we have:

$$\mathbb{E}\left(\frac{1}{M}\sum_{m=1}^{M}\frac{\#\{j : W_{mj} \leq -t\}}{\#\{j : W_{mj} \geq t\} \vee 1}\right) \geq FDR(t)$$