Regularization methods in multiple regression

Malgorzata Bogdan

University of Wroclaw

March 13, 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:

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

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)

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

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

Ridge regression (1)

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

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

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

Ridge regression (1)

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

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

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

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

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

$$X'Xu = \lambda u$$

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

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

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

$$X'Xu = \lambda u$$

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

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

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

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

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

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

Ridge regression - orthogonal design

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

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

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

Ridge regression - orthogonal design

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

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

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

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

Ridge regression - orthogonal design

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

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

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

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

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

Ridge regression - orthogonal design (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

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

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

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

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

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 $Y = Xb$.

$$Y = X\beta$$

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.

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

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

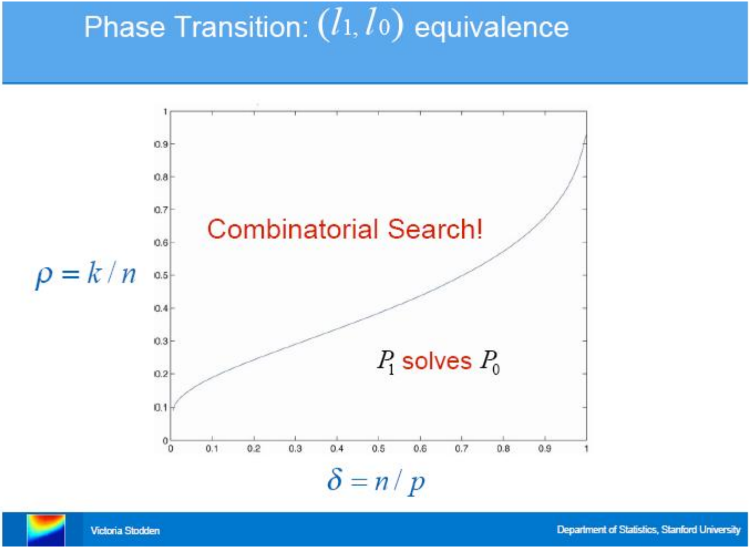
Basis Pursuit can recover β if k is small enough.

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*.

Transition curve (2)



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)

Least Absolute Shrinkage and Selection Operator (LASSO)

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

Least Absolute Shrinkage and Selection Operator (LASSO)

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

$$\hat{\beta}_L = argmin_{b \in R^p} ||y - Xb||_2^2 + \lambda ||b||_1$$

Least Absolute Shrinkage and Selection Operator (LASSO)

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

$$\hat{\beta}_L = argmin_{b \in R^p} ||y - Xb||_2^2 + \lambda ||b||_1$$

For a convex function $f : R^p \to R$ we define the subdifferential as

$$\partial_f(b) = \{v \in \mathbb{R}^p : f(z) - f(b) \geq v'(z - b) \ \forall z \in \mathbb{R}^p\}.$$

.

Least Absolute Shrinkage and Selection Operator (LASSO)

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

$$\hat{\beta}_L = argmin_{b \in R^p} ||y - Xb||_2^2 + \lambda ||b||_1$$

For a convex function $f : R^p \to R$ we define the subdifferential as

$$\partial_f(b) = \{v \in \mathbb{R}^p : f(z) - f(b) \geq v'(z - b) \; \forall z \in \mathbb{R}^p\}.$$

The convex function $f(x)$ attains a minimum at x_0 if and only if $0 \in \partial_f(x_0)$.

LASSO for the orthogonal design $X'X = I$

$$\beta^{LS} = Y'X, \;\; ||Y - Xb||^2 + \lambda||b||_1 = Y'Y + \sum_{i=1}^{p} f_i(b_i)$$

$$f_i(x) = x^2 - 2\beta_i^{LS} x + \lambda|x|$$

LASSO for the orthogonal design $X'X = I$

$$\beta^{LS} = Y'X, \ \ ||Y - Xb||^2 + \lambda||b||_1 = Y'Y + \sum_{i=1}^{p} f_i(b_i)$$

$$f_i(x) = x^2 - 2\beta_i^{LS} x + \lambda|x|$$

$$\partial_{f_i}(x_0) = 2x_0 - 2\beta_i^{LS} + \lambda\partial_{|x|}(x_0)$$

LASSO for the orthogonal design $X'X = I$

$$\beta^{LS} = Y'X, \ \ ||Y - Xb||^2 + \lambda||b||_1 = Y'Y + \sum_{i=1}^{p} f_i(b_i)$$

$$f_i(x) = x^2 - 2\beta_i^{LS} x + \lambda|x|$$

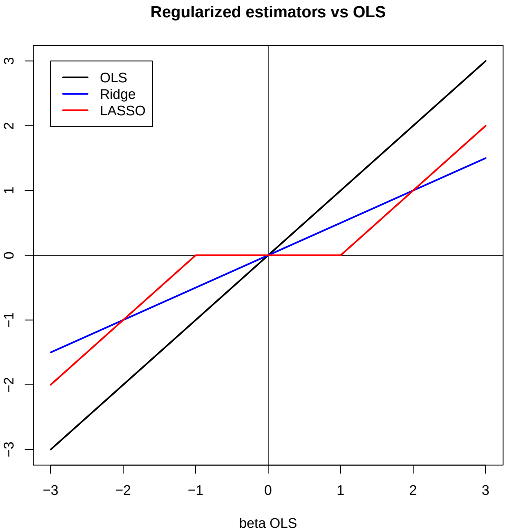
$$\partial_{f_i}(x_0) = 2x_0 - 2\beta_i^{LS} + \lambda\partial_{|x|}(x_0)$$

$$\lambda\partial_{|x|}(x_0) = \begin{cases} \lambda & \text{for} \quad x_0 > 0 \\ -\lambda & \text{for} \quad x_0 < 0 \\ < -\lambda, \lambda > & \text{for} \quad x_0 = 0 \end{cases}$$

LASSO for the orthogonal design

$$\hat{\beta}_i^L = \begin{cases} \beta^{LS} - \lambda/2 & \text{when} \quad \beta_i^{LS} > \lambda/2 \\ -\beta_i^{LS} + \lambda/2 & \text{when} \quad \beta_i^{LS} < -\lambda/2 \\ 0 & \text{when} \quad |\beta_i^{LS}| < \lambda/2 \end{cases}$$

Regularized estimators vs OLS



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 $(X_i)_{i \notin I}$.

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:

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

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:

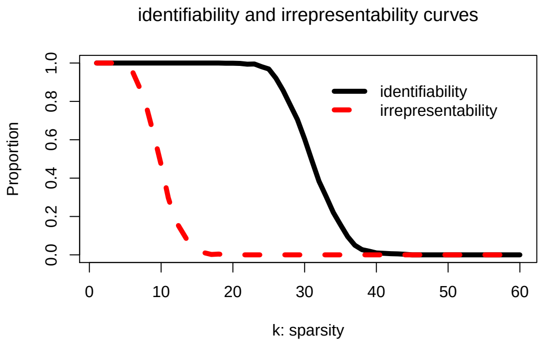
$$\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

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]

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

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)

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

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

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

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

$\beta_1, \ldots, \beta_p$: 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

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

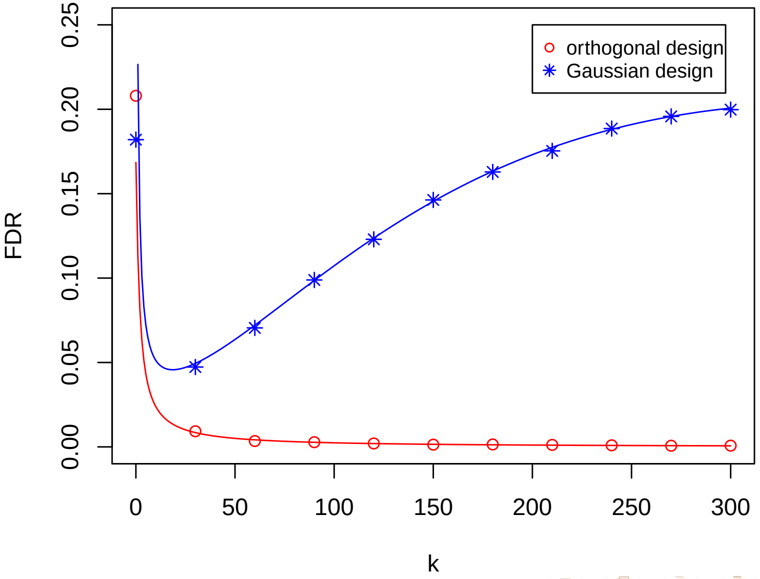
$$FDR = E(FDP)$$

Bogdan, van den Berg, Su and Candés, 2013

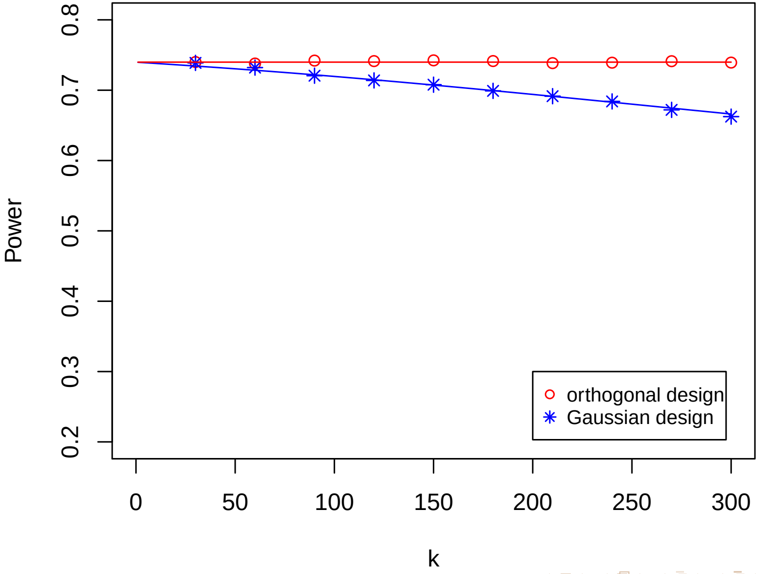
$$\text{FDR} \to \frac{2\mathbb{P}(\Pi = 0)\Phi(-\alpha)}{\mathbb{P}(|\Pi + \tau Z| > \alpha\tau)} \quad ,$$

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

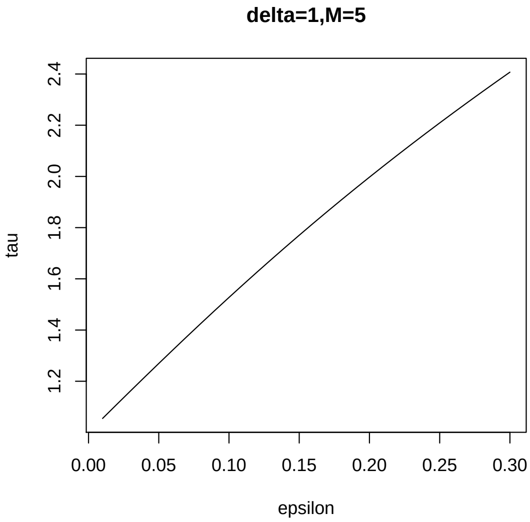
FDR - illustration



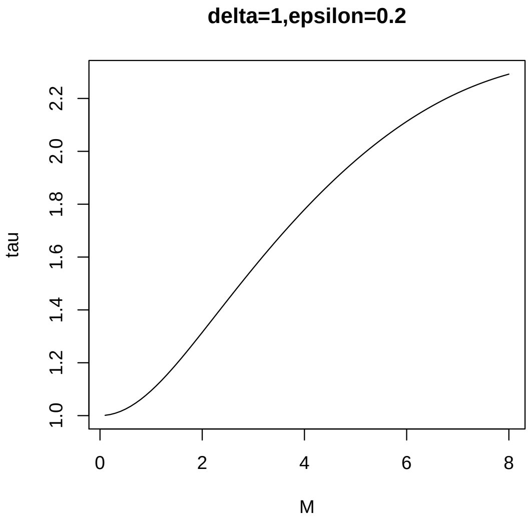
Power - illustration



Magnitude of additional noise (1)

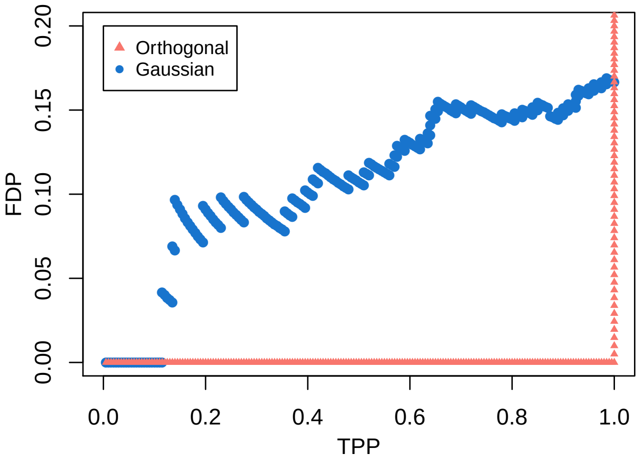


Magnitude of additional noise (2)



False Discoveries along the lasso path

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



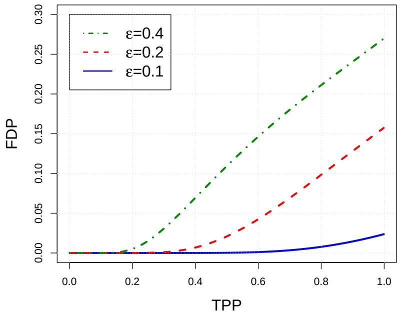
Theorem (Su, Bogdan, Candes, 2017)

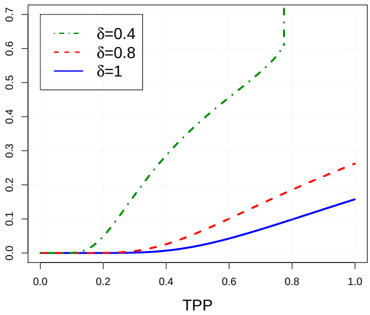
Fix $\delta \in (0, \infty)$ *and* $\epsilon \in (0, 1)$. *Then the event*

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

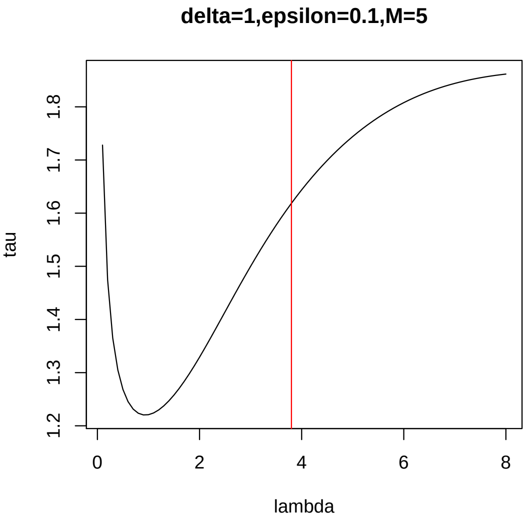
holds with probability tending to one.

FDR-Power trade-off (2)

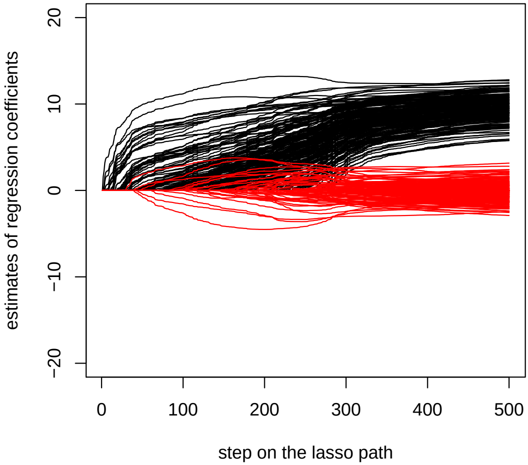




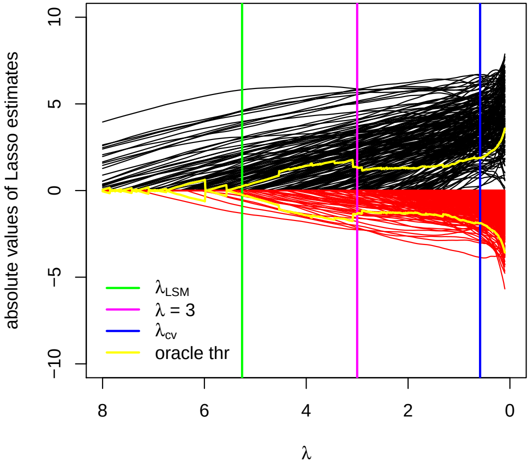
Magnitude of noise



Thresholded LASSO (1)



Thresholded LASSO (2)



Thresholded LASSO (3)