3.27pt

Regularization methods in multiple regression

Malgorzata Bogdan

University of Wroclaw

April 2, 2020

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

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)

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

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

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

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

Eigenvalues of M

$$X'Xu = \lambda u$$

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

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

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

Ridge regression - orthogonal design

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

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

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

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

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

Ridge regression - orthogonal design (2)

When rigde is better than LS ?

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

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

[Tardivel, Bogdan, 2019] BP can recover β if it is *identifiable* with respect to L_1 norm, i.e.

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

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

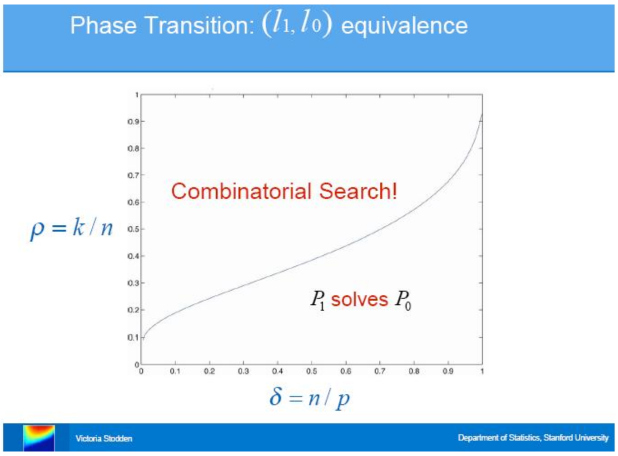
Basis Pursuit can recover β if k is small enough.

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

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

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 $X^TX = I$ 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 $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$
Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\bar{I}}$ be matrices whose columns are respectively $(X_i)_{i\in I}$ and $(X_i)_{i\notin I}$.
Irrepresentable condition:

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

When
$$\|X_{\bar{I}}'X_I(X_I'X_I)^{-1}S(\beta_I)\|_\infty > 1$$
then probability of the support recovery by LASSO is smaller than 0.5 (Wainwright, 2009).

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

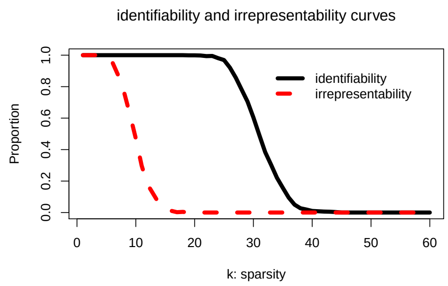
Appropriately thresholded LASSO can properly identify the sign of sufficiently large β if and only if β is identifiable with respect to l_1 norm.

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.

Irrepresentability and identifiability curves

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



Problem with shrinkage

Intuitive explanation:

$$\hat{\beta} = \eta_\lambda(\beta_i + X_i' z + v_i)$$

$$v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j) \rangle$$

$$\eta_\lambda(t) = sign(t)(|t| - \lambda)_+, \quad \text{applied componentwise}$$

If $X^T X = I$ then $X_i' z = Z_i \sim N(0,1)$, $v_i = 0$ and H_{0i} is rejected if $\beta_i + Z_i > \lambda$

When the design is not orthogonal: $v_i \neq 0$ - additional noise, dependent on λ (level of shrinkage), the level of sparsity and magnitude of true signals

Adaptive LASSO

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

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

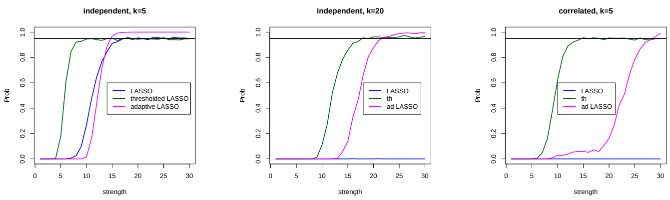
where $w_i = \frac{1}{\hat{\beta}_i}$, and $\hat{\beta}_i$ is some consistent estimator of β_i.

Reduces bias and improves model selection properties

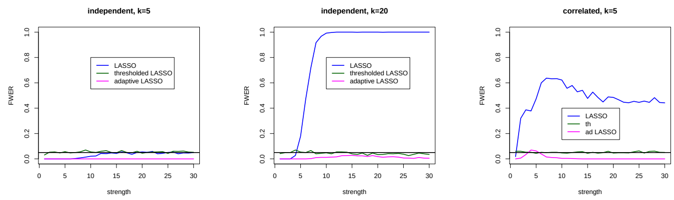
Numerical experiments

1. λ for LASSO selected as to control FWER at the level 0.05 for $k = 5$ (theoretical result in (Tardivel and Bogdan, 2019))
2. λ_{AMP} for thresholded LASSO and independent gaussian design selected according to AMP theory for LASSO (see e.g. (Wang, Weng, Maleki, 2018))
3. For correlated design (off diagonal covariance 0.9) we used 0.5 λ_{AMP}
4. For adaptive LASSO - weights based on LASSO estimator with λ as in 2 and 3, selection based on LASSO with λ as in 1
5. Threshold selected by using knockoff control variables (Foygel-Barber and Candès, 2015; Candès, Fan, Janson, Lv, 2016)

Probability of the sign recovery

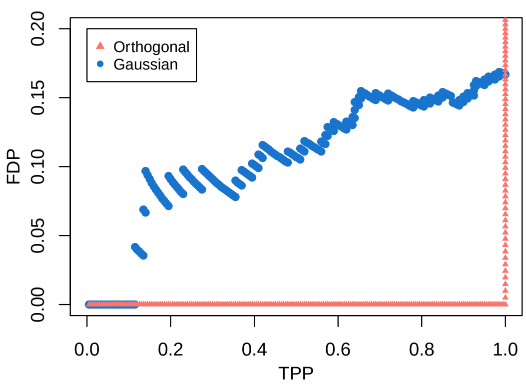


Family Wise Error Rate



False Discoveries along the lasso path

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



Reason - shrinkage

LASSO solution

$$\hat{\beta} = \eta_\lambda(\hat{\beta} - X'(X\hat{\beta} - y)) = \eta_\lambda(\hat{\beta} - X'X(\hat{\beta} - \beta) + X'z) \ ,$$

where $\eta_\lambda(t) = \mathrm{sgn}(t)(|t| - \lambda)_+$, applied componentwise

$$\hat{\beta}_i = \eta_\lambda(\beta_i + Z_i + v_i),$$

$$\text{where} \ \ v_i = \langle X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j)\rangle \ \ \text{and} \ \ Z_i \sim N(0, \sigma_i^2)$$

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

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

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

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

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

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

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

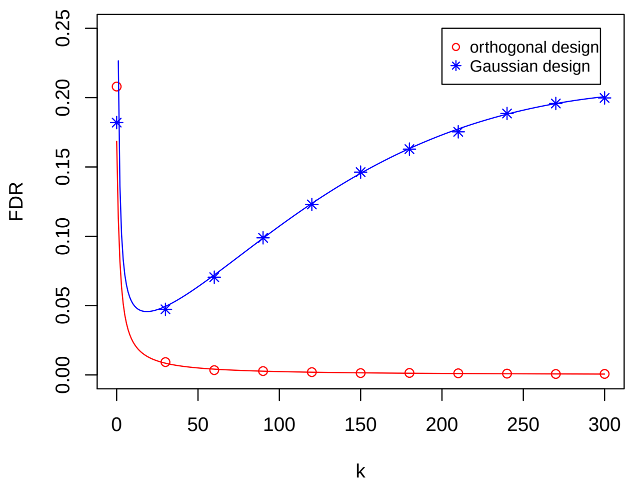
$$FDR = E(FDP)$$

Bogdan, van den Berg, Su and Candés, 2013

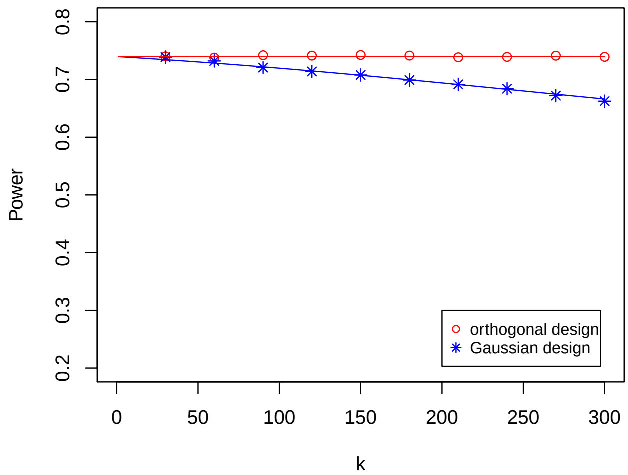
$$\mathrm{FDR} \to \frac{2\mathbb{P}(\Pi = 0)\Phi(-\alpha)}{\mathbb{P}(|\Pi + \tau Z| > \alpha\tau)},$$

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

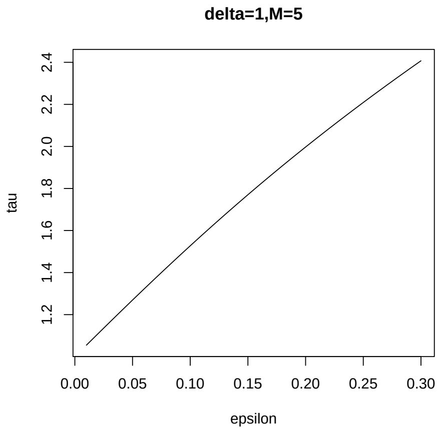
FDR - illustration



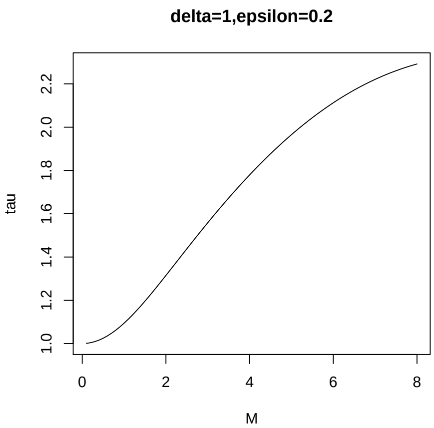
Power - illustration



Magnitude of additional noise (1)



Magnitude of additional noise (2)



FDP-Power tradeoff

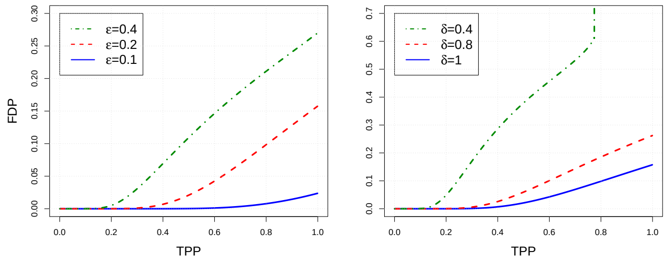
Theorem (Su, Bogdan, Candes, 2017)

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

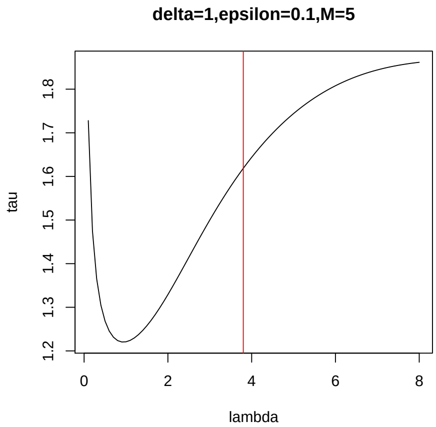
$$\bigcap_{\lambda \geq 0.01} \left\{ FDP(\lambda) \geq q^{\star}\left(TPP(\lambda)\right) - 0.001 \right\} \tag{3}$$

holds with probability tending to one.

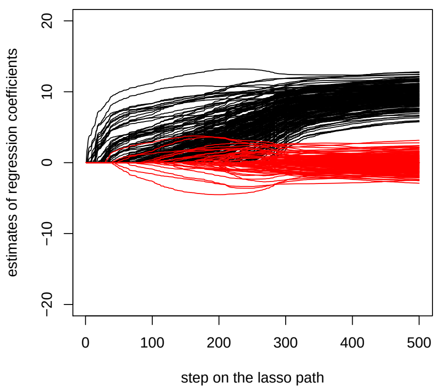
FDR-Power trade-off (2)



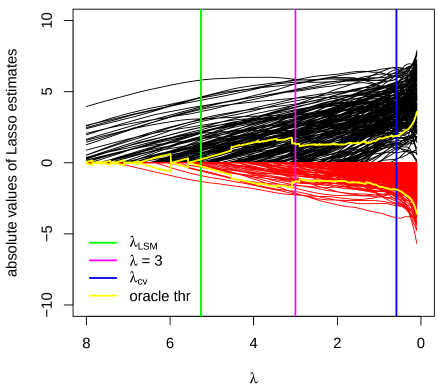
Magnitude of noise



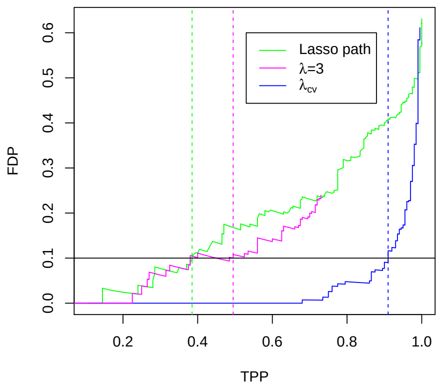
Thresholded LASSO (1)



Thresholded LASSO (2)



Thresholded LASSO (3)



Model X knockoffs and LCD statistics

Candès, Fan, Janson and Lv (2017) - augment X with the matrix $\tilde{X}$ of specifically constructed fake null variables

Necessary requirement:
$\Sigma_X = \Sigma_{\tilde{X}}$ and for $i \neq j$ $Cov(X_i, \tilde{X}_j) = Cov(X_i, X_j)$.
When X_{ij} are iid $N(0, 1/n)$ then $\tilde{X}_{ij}$ are also iid $N(0, 1/n)$.

$\hat{\beta}(\lambda)$ - vector of $2p$ estimates of regression coefficients by LASSO applied on the augmented design matrix $X_{aug} = [X, \tilde{X}]$

Function $w : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is *faithful* if it obeys
(I) w is antisymmetric, $w(v, u) = -w(u, v)$
(II) for any fixed c, $w(x, c)$ tends to infinity as $|x| \to \infty$.

$$W_j = w(\widehat{\beta}_j, \widehat{\beta}_{p+j})$$

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.

Example: Lasso coefficient difference statistics $LCD(\lambda, q)$

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

Breaking through FDR-Power diagram

Su, Weinstein, Bogdan, Candès (2018)

Theorem

Consider a fixed sparsity parameter ϵ such that $\epsilon/2 < \epsilon_{\mathrm{DT}}(\delta/2)$ and a sequence of signal distributions Π_m such that for any given constant $M > 0$ $P(|\Pi_m| > M | \Pi \neq 0) \to 1$ as $m \to \infty$. Then for any given $\lambda > 0$ and $q > 0$ it holds

$$\lim_{m\to\infty} \lim_{p\to\infty} Power(KN(\lambda, q)) \to 1.$$

Extension of AMP theory

We show that the triples $(\beta_j, \widehat{\beta}_j, \widehat{\beta}_{p+j})$ are independent, and each is distributed as $(\Pi, \eta_{\alpha\tau}(\Pi + \tau W), \eta_{\alpha\tau}(\tau\widetilde{W}))$, where W and $\widetilde{W}$ are independent $\mathcal{N}(0,1)$ random variables that are furthermore independent of Π; and (α, τ) are determined by λ as the solution to

$$\begin{aligned} \tau^2 &= \sigma^2 + \frac{1}{\delta}\mathbb{E}\left[\eta_{\alpha\tau}(\Pi + \tau W) - \Pi\right]^2 + \frac{1}{\delta}\mathbb{E}\,\eta_{\alpha\tau}(\tau W)^2 \\ \lambda &= \left[1 - \frac{1}{\delta}\mathbb{P}(|\Pi + \tau W| > \alpha\tau) - \frac{1}{\delta}\mathbb{P}(|\tau W| > \alpha\tau)\right]\alpha\tau. \end{aligned} \tag{4}$$

Main component of the proof

For fixed $\lambda > 0$, let $t^\infty = t^\infty(q) > 0$ be such that

$$\frac{\mathbb{P}(\omega(\eta_{\alpha\tau}(\Pi + \tau W), \tau\eta_\alpha(\widetilde{W})) \le -t^\infty)}{\mathbb{P}(\omega(\eta_{\alpha\tau}(\Pi + \tau W), \tau\eta_\alpha(\widetilde{W})) \ge t^\infty)} = q, \tag{5}$$

where (α, τ) is the solution to (4). Then

1. The quantity $t^\infty(q)$ exists and is unique for any $q \in (0,1)$. Furthermore, it has a limit as $q \to 0$ (and, for fixed λ, this limit depends on Π only).
2. Knockoff random threshold $\hat{t}$ satisfies $\hat{t} \to t^\infty(q)$ in probability.

The asymptotic power is given by

$$\text{TPP} \to \mathbb{P}(\omega(\eta_{\alpha\tau}(\Pi + \tau W), \tau\eta_\alpha(\widetilde{W}) \ge t^\infty | \Pi \ne 0)$$

How to pick λ

Given that

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

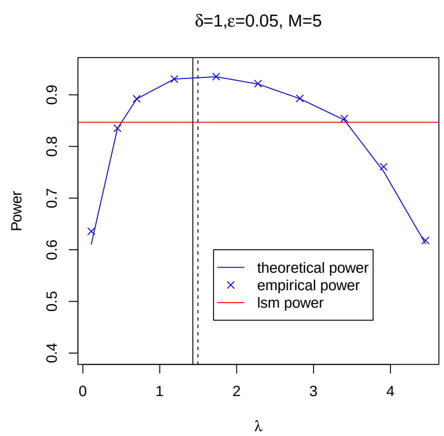
the "best" ordering of $\hat{\beta}_i$ occurs when τ is minimal.

Bayati and Montanari (2012):

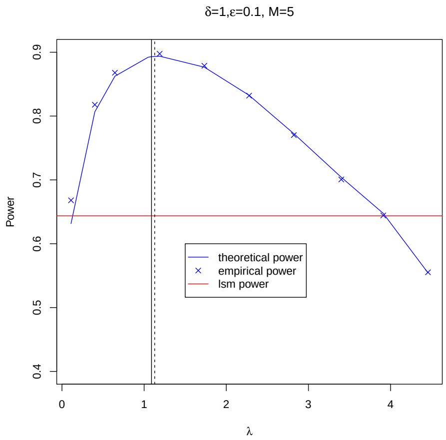
$$\frac{1}{p}||\hat{\beta} - \beta||^2 \to \delta(\tau^2 - \sigma^2)$$

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 AMP theory

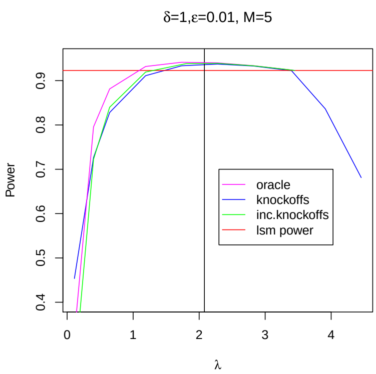
G. Reeves, 2017, neural networks

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

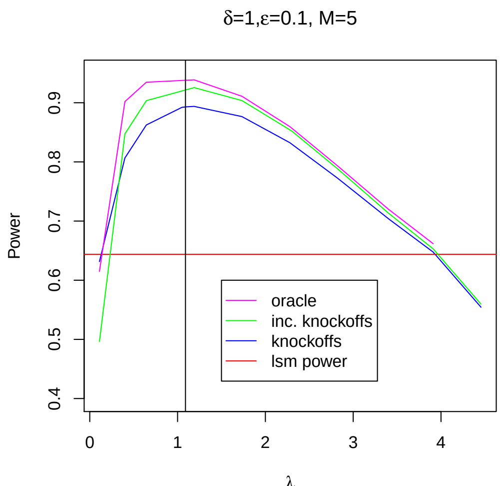
Incremental knockoffs

For $i = 1, \dots, p$ run LASSO on $[X, \tilde{X}_i]$ (only $p + 1$ columns) and calculate $W_i = |\hat{\beta}_i| - |\hat{\beta}_{p+1}|$

Conjecture: the procedure controls FDR when used with a fixed λ



Incremental knockoffs (2)



Incremental knockoffs (3)