Preconditioning for sparse inference

Karl Rohe (stat.wisc)

$$Y = X\beta^* + \epsilon$$

observe: $Y \in R^n,\ X \in R^{n \times p}$.
wish to estimate: $\beta^* \in R^p$.
contaminated by random variables (noise): $\epsilon \in R^n$.

Joint with Jinzhu Jia (Peking U)

When physical constraints restrict the design matrix or when the design matrix is *observed* (e.g. fMRI) we often fail to satisfy RIP or any other "condition for consistency."

- Preconditioning can sometimes correct the problems.

First, a teaser that relates to preconditioning.

Generalized Least Squares

In classical (low dimensional) set up, if the errors are not iid:

$$E(\epsilon\epsilon') = \Sigma$$

Use weighted least squares. Equivalently, "whiten the noise" by left multiplying:

$$\Sigma^{-1/2}Y = \Sigma^{-1/2}X\beta + \Sigma^{-1/2}\epsilon$$

Run OLS on $(\Sigma^{-1/2}Y, \Sigma^{-1/2}X)$ to get a BLUE estimator.

Could adapt this to use with the Lasso. Simply call Lasso with

$$(\Sigma^{-1/2}Y, \Sigma^{-1/2}X)$$

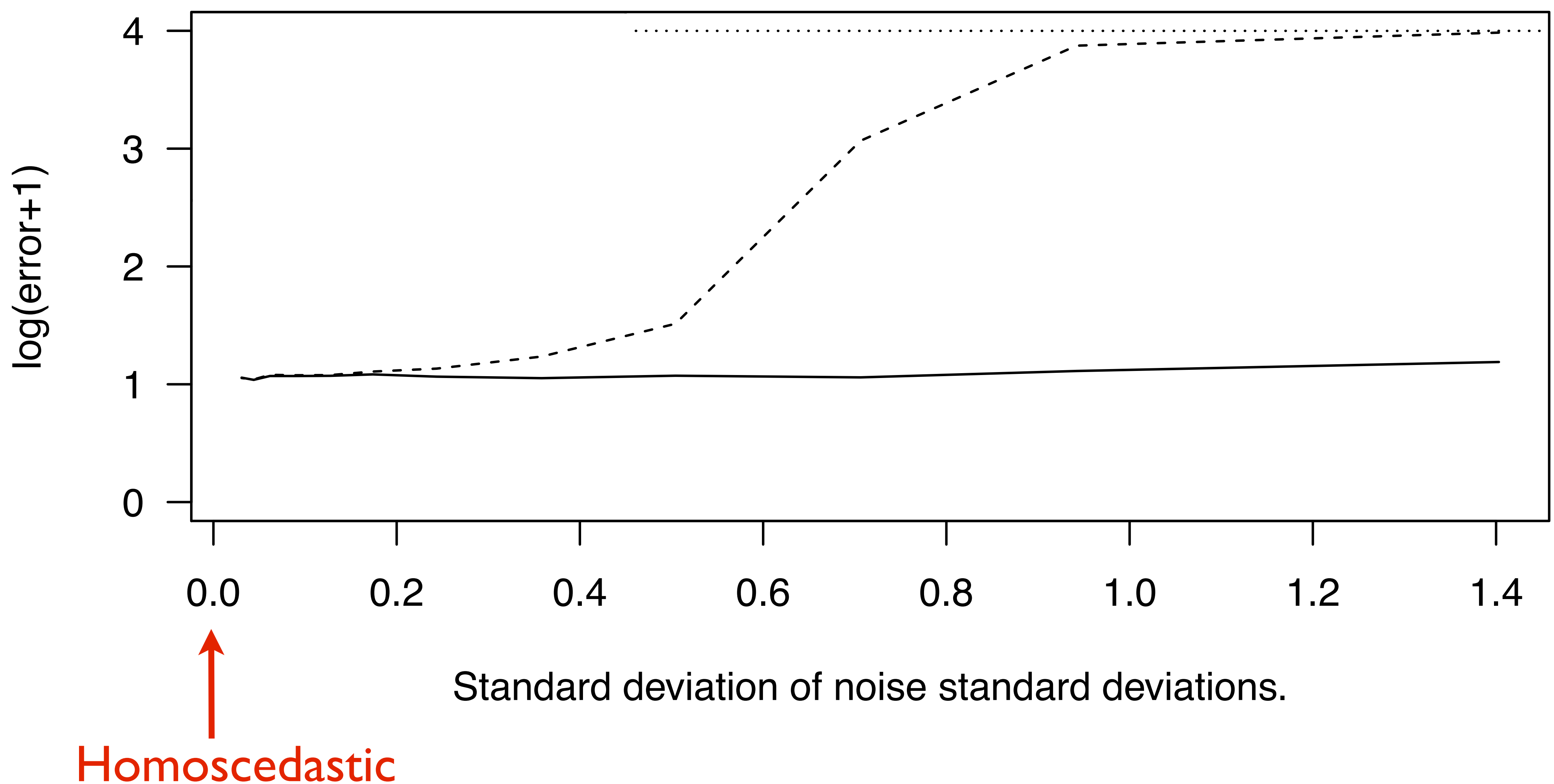
This gives the penalized MLE for non-iid (but still gaussian) errors.

GLS + Lasso vs. vanilla Lasso

- Simulation with p = 1000, n = 200, |S| = 20.
- X contains iid gaussians.
- tune with oracle.
- To get a heteroskedastic model, the error standard deviations are generated from Gamma distribution.
- x-axis controls the standard deviation of this Gamma: higher values are more heteroskedastic.
- y-axis is

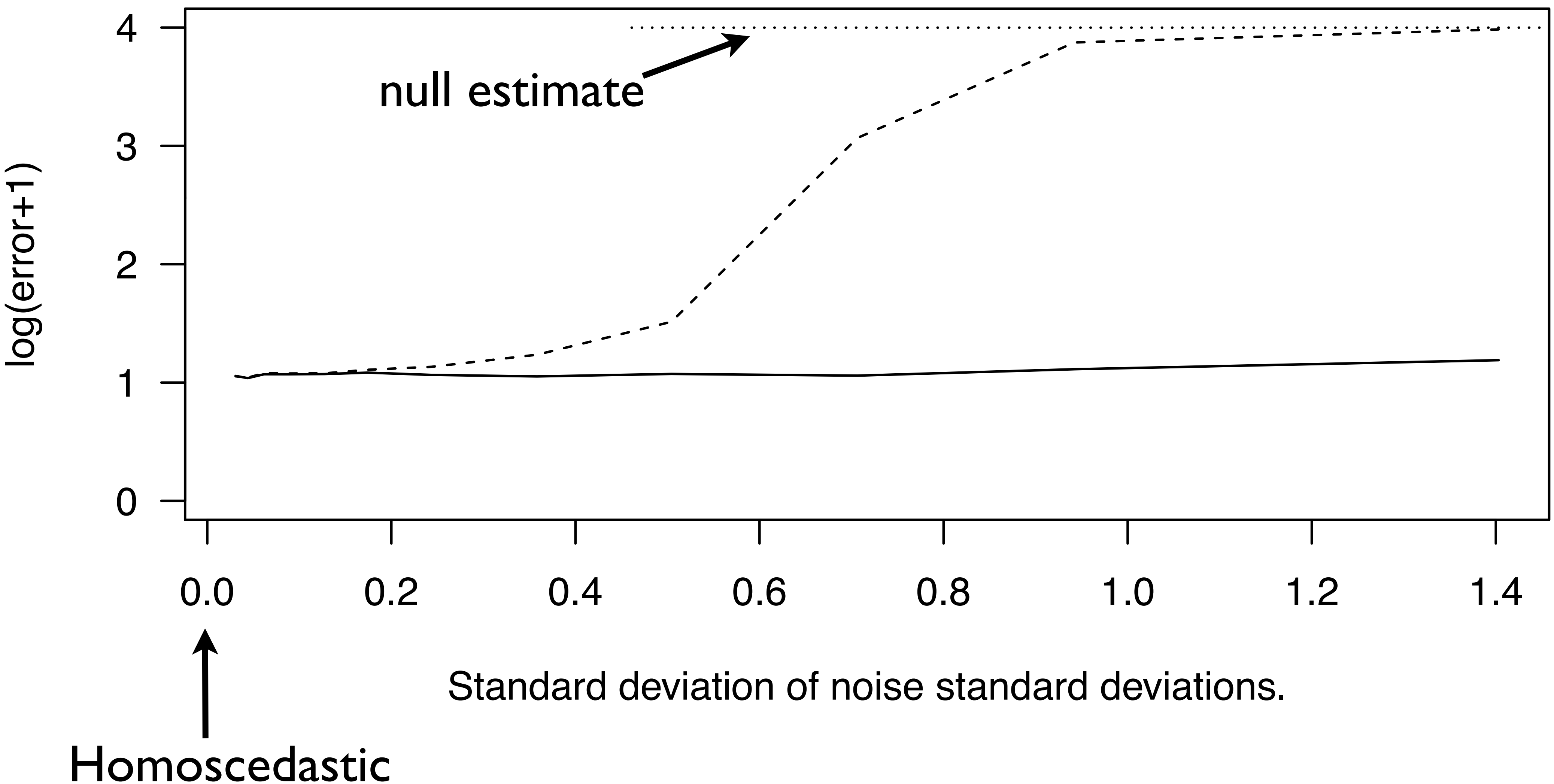
$$\log(1 + \|\beta - \hat{\beta}(\lambda)\|_2)$$

GLS + Lasso vs. vanilla Lasso



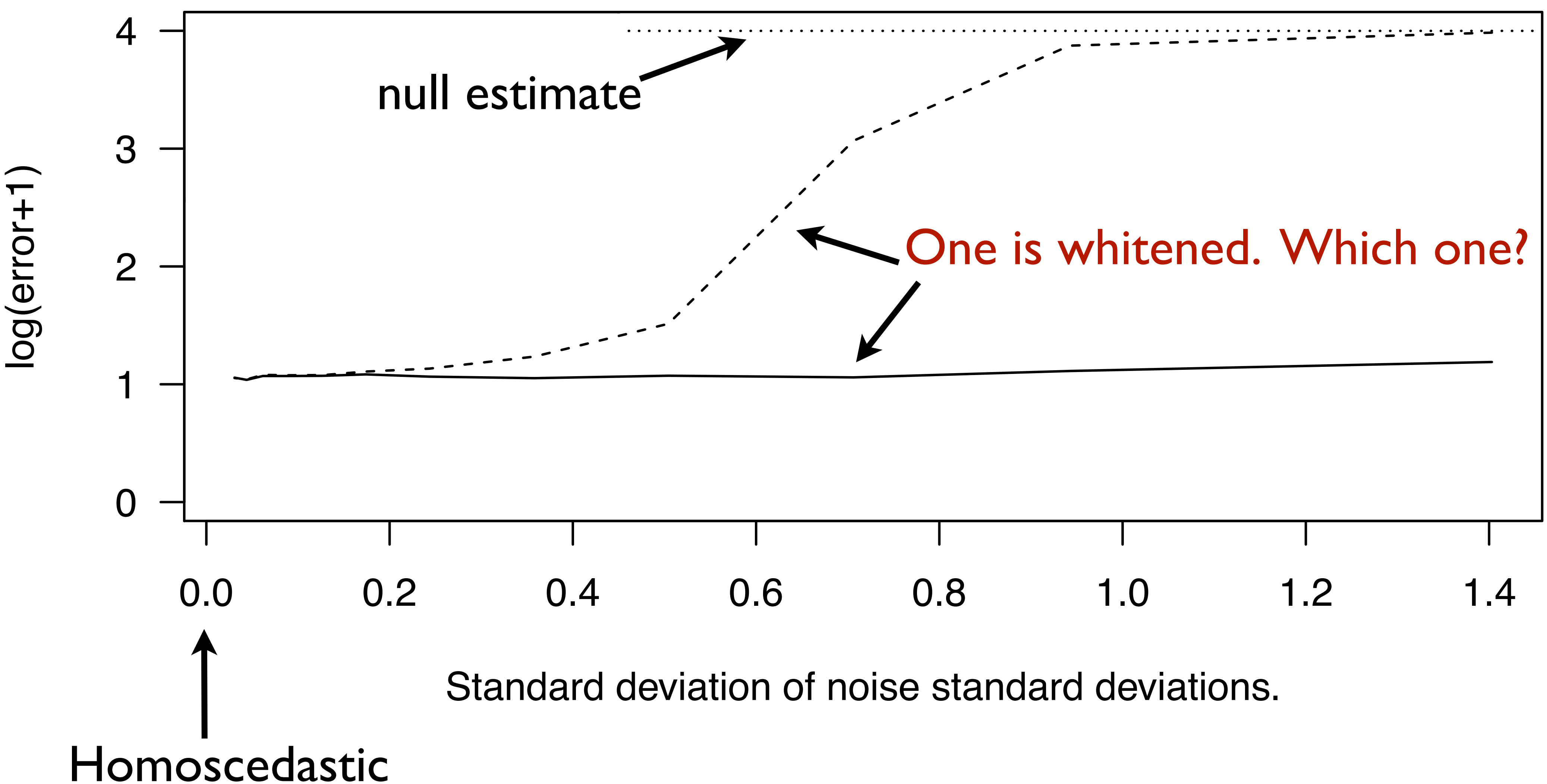
GLS + Lasso vs. vanilla Lasso

As the generative model becomes heteroskedastic, one of the estimators fails completely. Which one?

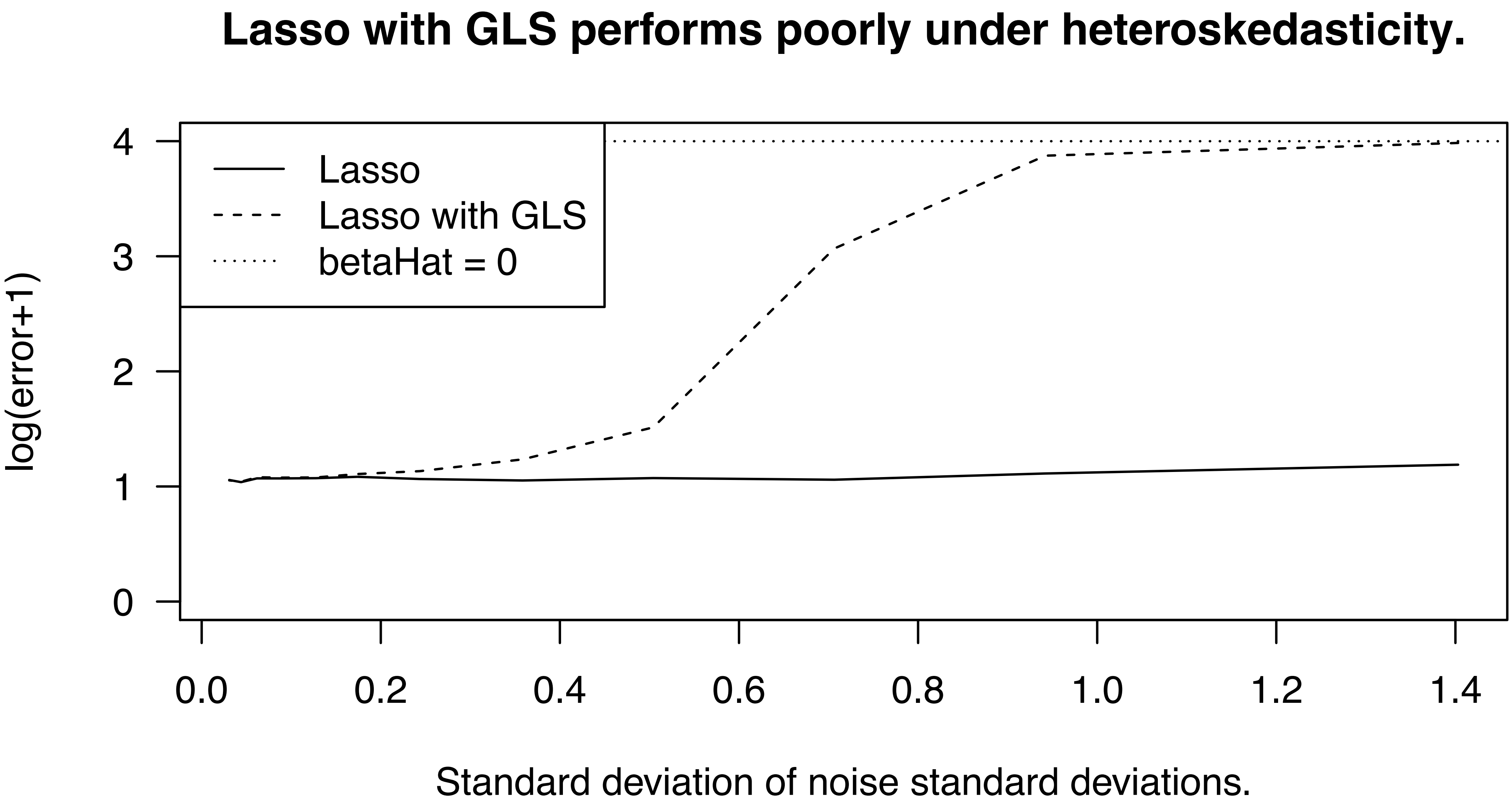


GLS + Lasso vs. vanilla Lasso

As the generative model becomes heteroskedastic, one of the estimators fails completely. Which one?



Whitening worsens estimation!



1) Left multiplying by Sigma^{-1/2} can make estimation worse. Why?

2) Are there other matrices that will *improve* estimation?

Outline

1) The Lasso and the conditions for consistency.
2) Preconditioning (and ill-conditioning).
   a) low dimensions.
   b) high dimensions.
3) Recap + open problems.

The current literature has explored several different assumptions on the design matrix.

- Restricted Isometry Principal (RIP)
- Nullspace property (NSP)
- Restricted Eigenvalue (RE) assumption
- Irrepresentable condition (IC)
- Mutual coherence

These conditions for consistency control the correlation between the columns of X is various ways.

Conditions are very sensitive to even a bit of correlation.

- Irrepresentable condition is a necessary and (almost) sufficient condition to consistently estimate the sign (-,0,+) of each element in beta.
- Want IC(X) less than one:

$$\max \left| X(S^c)^T X(S) \left(X(S)^T X(S) \right)^{-1} \mathrm{sign}(\beta^*(S)) \right| < 1 - \eta.$$

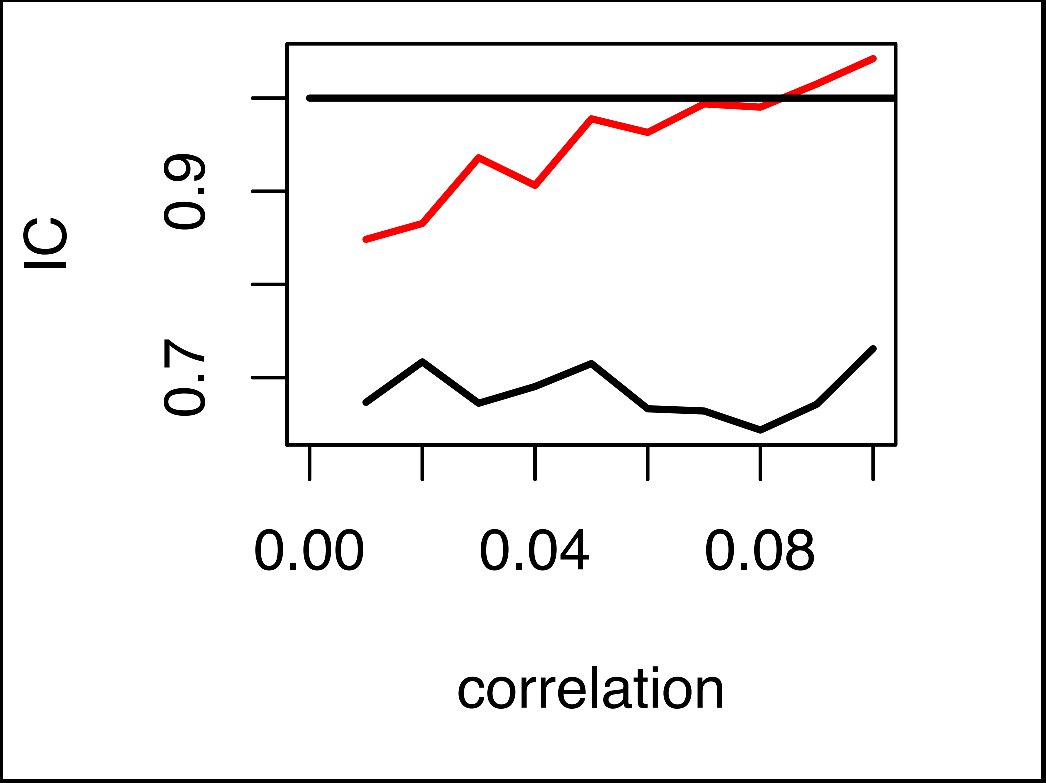
Just a bit of correlation can break the IC

n = 200, p = 1000, |S| = 10

$$X_i \sim N(0, \Sigma), iid$$
$$\Sigma_{ii} = 1,\ \Sigma_{i \neq j} = \rho$$

This increases on the horizontal axis.

$IC(X) > 1$ for $\rho > .08$



We will come back to the black line.

Current approaches adjust the penalty to side step the assumptions.

- SCAD (Fan and Li, 2001)
  - non-convex penalty, difficult to optimize.
- MC+ (Zhang, 2010)
  - non-convex penalty, fast algorithm, high probability results.
- Adaptive lasso (Zou 2006)
  - non-equal penalty for low dimensions
- OEM (Xiong, Dai, Qian 2011)
  - appends the data to get orthogonal design and then uses EM. computationally difficult in high dimensions.

Rest of the talk:

1) The Lasso, sign consistency, and the irrepresentable condition.

2) Preconditioning (and ill-conditioning).

   a) low dimensions.

   b) high dimensions.

3) Recap + open problems.

Preconditioning

Comes from numerical linear algebra.

System of equations: Given A and b, solve for x.

$$Ax = b \qquad (X\beta^* = Y)$$

$$A \in R^{n \times n}, x \in R^n, b \in R^n$$

The speed of most solvers depends on the condition number:

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

If this ratio is large, then the system is ill-conditioned. Slow algorithms.

Preconditioning

Left multiply the equation $Ax = b$ by a preconditioner.

$$TAx = Tb$$

T is designed so that $\kappa(TA) < \kappa(A)$

Connection to statistical ideas:
PCA on A discovers correlation when A
has large eigenvalues... ill conditioned.

The preconditioner **decorrelates** the columns.

The preconditioner **decorrelates** the columns.

- The conditions for consistency depend on the correlation between columns.
- If we decorrelate the columns, then we can potentially circumvent these assumptions!
- By preconditioning, we can potentially improve the performance of the Lasso for design matrices with correlated columns.

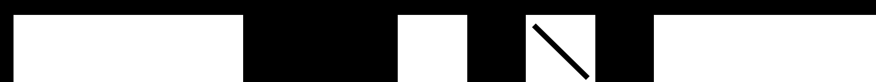
How to precondition for the Lasso.

The Puffer Transformation

$$Y = X\beta^* + \epsilon$$

Take the SVD of X.

$$X = UDV'$$



Define the Puffer Transformation. $F = UD^{-1}U'$

$$FX = (UD^{-1}U')(UDV') = UIV'$$

The nonzero singular values are all one.

The Puffer Transformation

$$Y = X\beta^* + \epsilon$$

$$FY = FX\beta^* + F\epsilon$$

Instead of running the Lasso on (X,Y), run it on (FX, FY).
Still estimates the same beta!

Rest of the talk:

1) The Lasso, sign consistency, and the irrepresentable condition.

2) Preconditioning (and ill-conditioning).

   a) low dimensions.

   b) high dimensions.

3) Recap + open problems.

Low dimensions (n>p)

Assume X is full rank.

$$FX = UV' \text{ and } (FX)'FX = VU'UV' = I$$

- Orthogonal columns!
- Trivially satisfies the conditions for consistency.

Low dimensional theorem (Jia, R 2012)

Under the regression equation,

$$Y_n = X_n\beta^* + \epsilon_n$$

with $Y_n \in R^n$, $X_n \in R^{n\times p}$, and $\epsilon_n \sim N(0, \sigma^2 I_n) \in R^n$ with p fixed and $n \to \infty$, take the SVD of $n^{-1/2}X_n = UDV'$ and define the preconditioner $F_n = UD^{-1}U'$. Define $d_{\min}^{(n)}$ is the smallest eigenvalue of $X_n' X_n/n$. If $d_{\min}^{(n)}$ is bounded from below and $\lambda_n = n^{-1/4}$, then the Lasso estimator computed on the preconditioned variables,

$$\hat\beta(\lambda_n) = \arg\min_{\beta\in R^p} \frac{1}{2n}\|F_nY_n - F_nX_n\beta\|_2^2 + \lambda_n\|\beta\|_1,$$

is sign consistent

$$P(\text{sign}(\hat\beta(\lambda_n)) = \text{sign}(\beta^*)) \to 1.$$

This theorem does not have an irrepresentable condition.
Instead: a lower bound on the spectrum of X.

$$FY = FX\beta^* + \boxed{F\epsilon}$$

- The preconditioned errors are no longer iid!
- F can amplify the noise.
- Our proof technique needs to bound the spectral norm of the error covariance:

$$\|E(F\epsilon\epsilon' F')\| = \|FE(\epsilon\epsilon')F'\| = c\|FF'\| = \frac{c}{\sigma^2_{\min}(X)}$$

smallest singular value (squared) of X

Punch line (bad news?): when columns of X are correlated, then X has a small singular value and F amplifies the errors.

Simple fix: replace largest singular values of F with something else: zero, M, anything to bound it.

$$F = UD^{-1}U'$$

Setting the small values to zero is equivalent to discarding the weak directions, reducing the "sample size".

Rest of the talk:

1) The Lasso, sign consistency, and the irrepresentable condition.

2) Preconditioning (and ill-conditioning).

   a) low dimensions.

   b) high dimensions.

3) Recap + open problems.

In high dimensions, the columns cannot be orthogonal.

- In low dimensions, preconditioning ensures that the conditions for consistency are satisfied by making columns orthogonal.
- In high dimensions, preconditioning does not always satisfy the conditions for consistency.
  - If p >> n, then colrank(X) << p.
  - Columns cannot be orthogonal.

But it does have orthogonal rows!

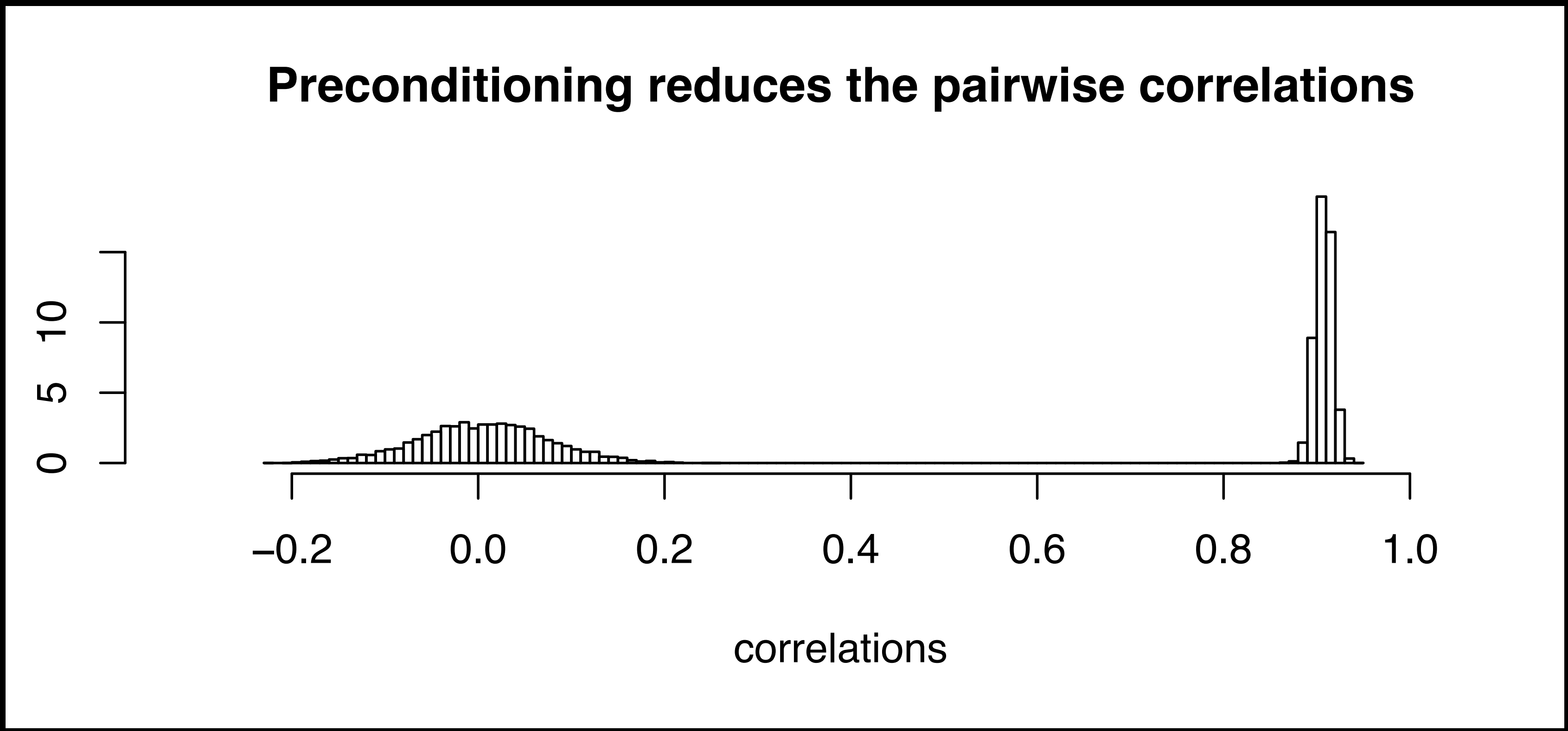
- Fan and Hoffman showed in 1955 that for any unitarily invariant norm, FX is the projection of X onto the Stiefel manifold (the set of n x p orthonormal matrices)

Existing literature suggests that if (orthogonal) rows are drawn from a nice distribution, then the design matrix is well conditioned.

- Rudelson, Vershynin 2006. “Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements.”
- Tropp 2007. “On the conditioning of random subdictionaries”
- Candes, Romberg 2007. “Sparsity and incoherence in compressive sampling”
- Candes, Plan 2010. “A probabilistic and RIPless Theory of Compressed Sensing”
- Others...

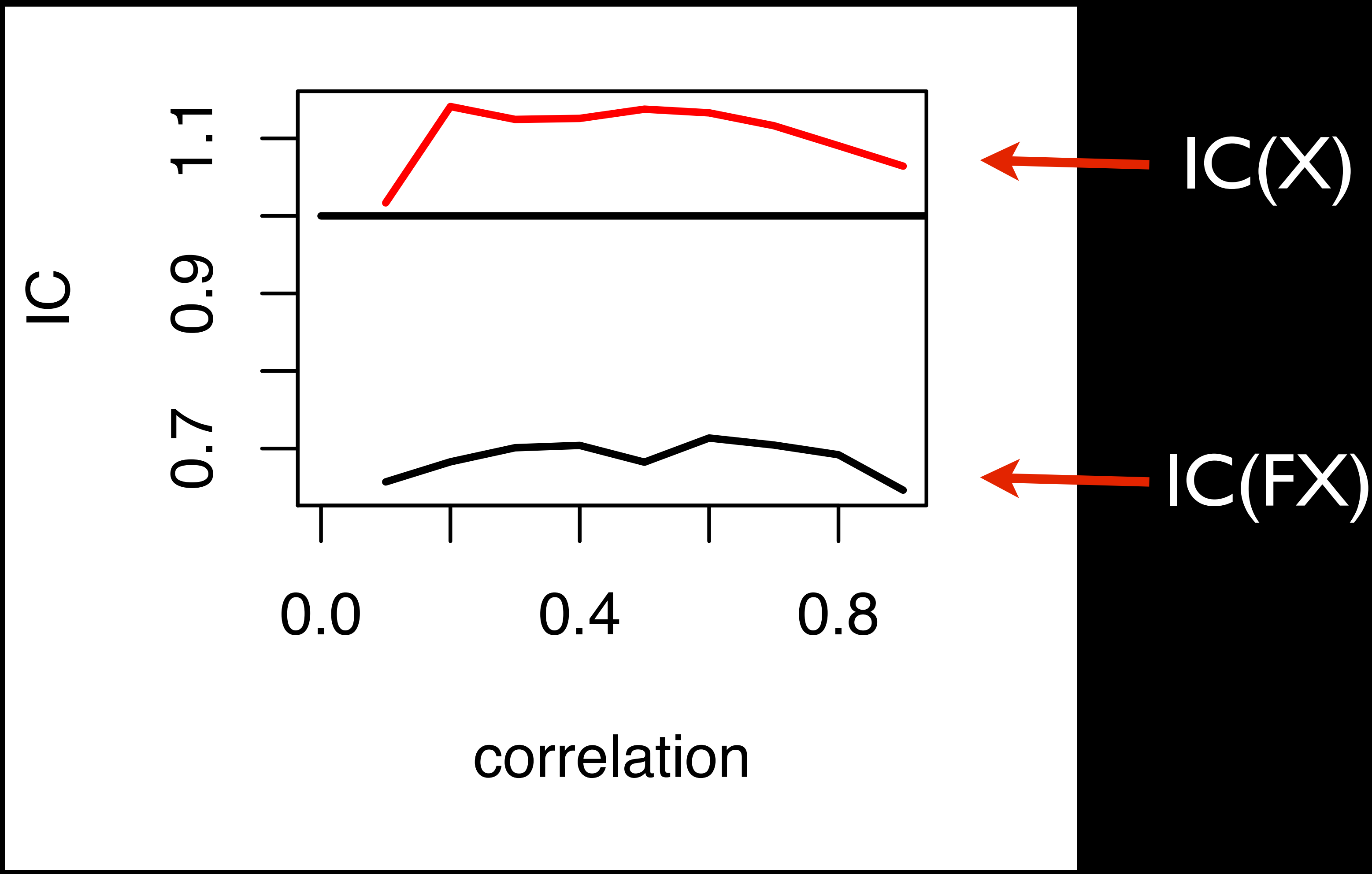
Preconditioning drastically reduces pairwise correlation between columns.

n = 200, p = 10,000



Rows of X are iid Gaussian with correlation .9.
Rows of FX have average correlation .005
with a standard deviation of .07.

Preconditioning circumvents the Irrepresentable Condition in high dimensions.



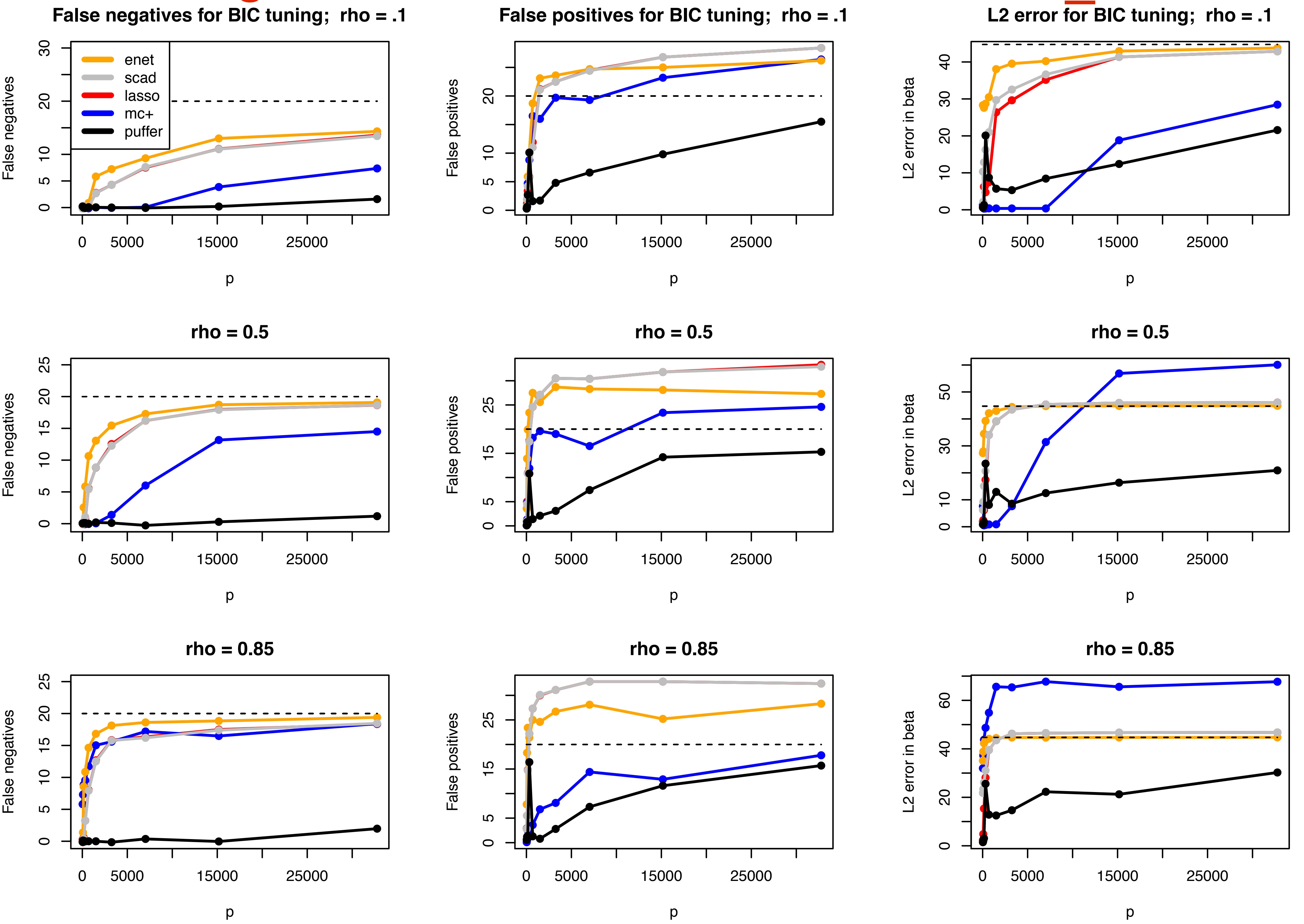
n = 200,
p = 1000,
|S| = 10

Preconditioning drastically reduces the IC value.
Below the cutoff of one.

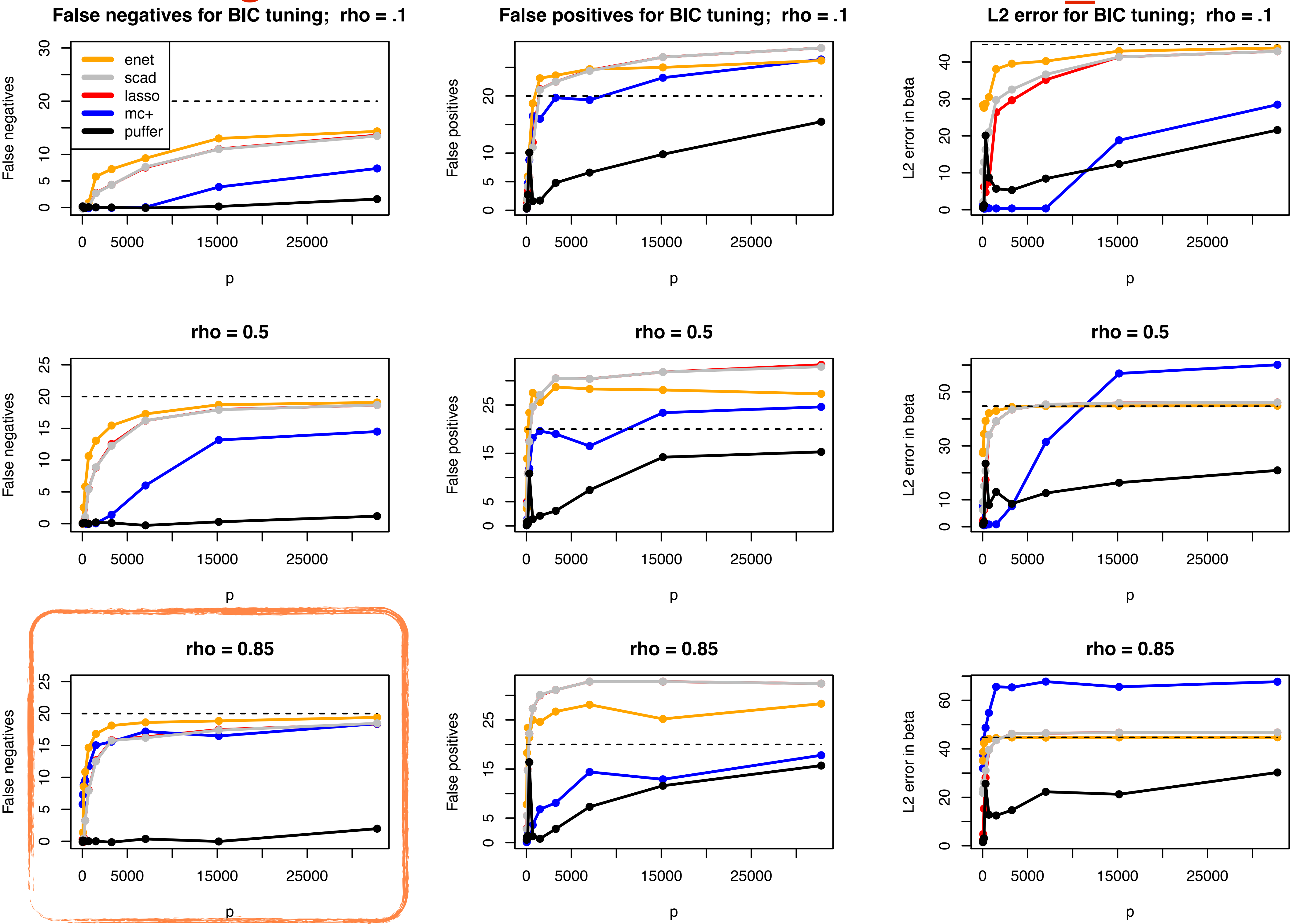
Simulations

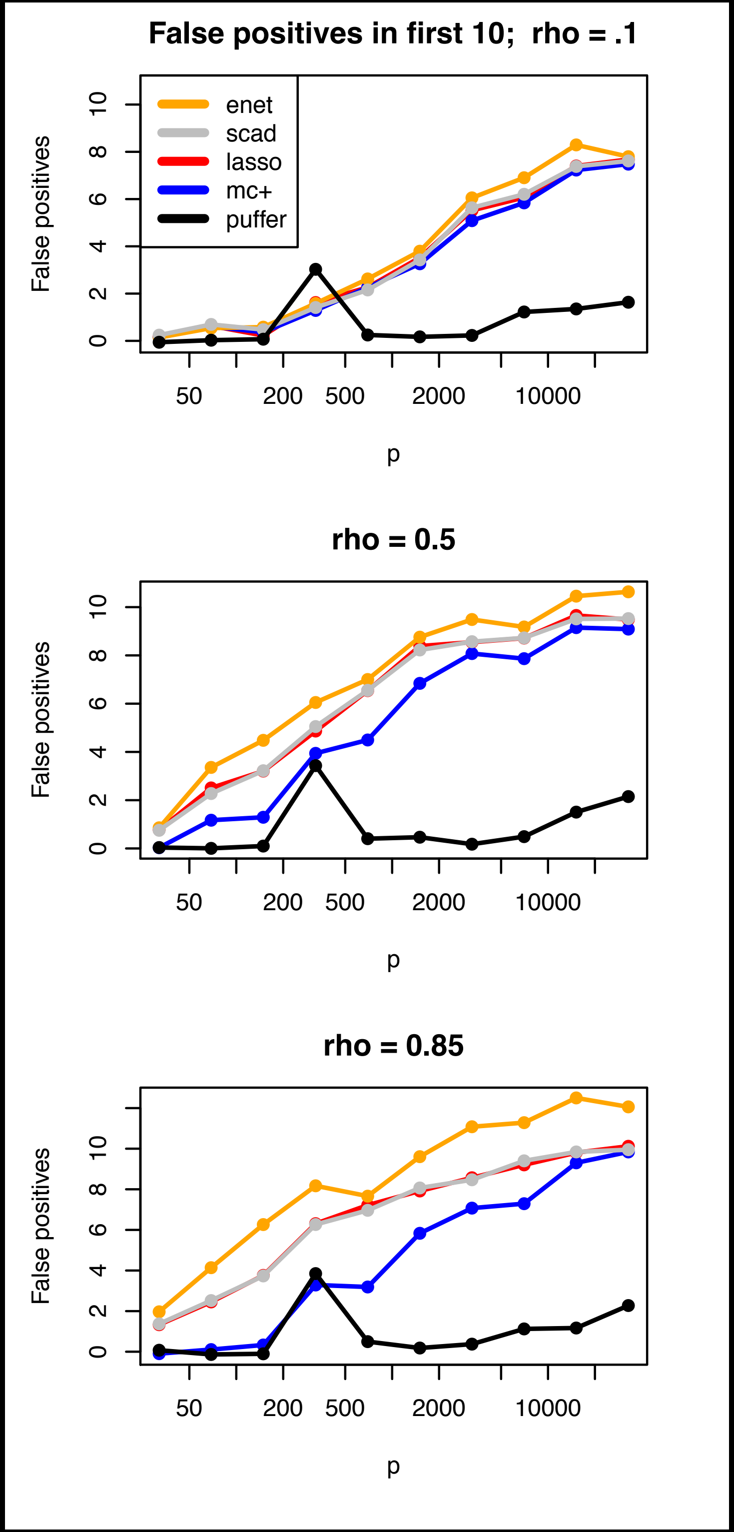
- n=250, |S| = 20, p grows from 32 to 32k.
- Rows of X are iid Gaussians with correlations increasing down the row of plots. .1, .5, .85
- First column of plots shows false negatives, (type II error for Ho: beta_j = 0)
- Second column shows false positives, (type I error)
- Third column shows $\|\hat{\beta}(\lambda) - \beta^*\|_2$
- Tuning parameter selected by choosing best of the first 40 models (in ols) by BIC.

False negatives False Positives ell_2 error



False negatives | False Positives | ell_2 error





- This simulation investigates if our method of choosing a tuning parameter influenced the result.
- Take the first 10 predicted coefficients.
- How many false positives out of 10?
- Lasso hits 8/10 false positives at p = 1500. Preconditioned Lasso barely has 1/10.
- (log scale)

Outline:

1) The Lasso, sign consistency, and the irrepresentable condition.
2) Preconditioning (and ill-conditioning).
   a) low dimensions.
   b) high dimensions.
3) Recap + open problems.

Remember GLS... it is preconditioning!

Preconditioning and GLS are two sides of the same coin.

$$FY = FX\beta^* + \boxed{F\epsilon}$$

$$\Sigma^{-1/2}Y = \Sigma^{-1/2}X\beta + \Sigma^{-1/2}\epsilon$$

GLS "fixes" the error distribution.
Preconditioning makes it worse!

Open problems

1. Real example(s).
2. Preconditioning in non regression settings (e.g. sparse inverse covariance estimation)
3. Fitting logistic regression. Left multiplies with IRLS. Can this hurt the conditioning? Better methods?
   - bootstrapping, generalized lasso also change the conditioning of the design matrix.
4. Can it make fitting faster? Anything else from the numerical linear algebra literature?

Key points

1. Lasso fails when columns of X are weakly correlated.
2. Preconditioning can (often) fix this problem.
3. Can give error terms a strange distribution.
   - Easy fix: threshold the spectral norm of F.
4. Simulations: beats the competition when n>200 and p >1000. Far superior when p>20,000.
5. Several open problems.