Electronic Journal of Statistics
Vol. 9 (2015) 1150–1172
ISSN: 1935-7524
DOI: 10.1214/15-EJS1029

Preconditioning the Lasso for sign consistency[*]

Jinzhu Jia

School of Mathematical Sciences and Center for Statistical Science
Peking University, Beijing, 100871, China
e-mail: jzjia@math.pku.edu.cn
url: www.math.pku.edu.cn/teachers/jjia

and

Karl Rohe

Department of Statistics
University of Wisconsin-Madison, WI 53706, USA
e-mail: karlrohe@stat.wisc.edu
url: http://www.stat.wisc.edu/~karlrohe/

Abstract: Sign consistency of the Lasso requires the stringent irrepresentable condition. This paper examines whether preconditioning can circumvent this condition. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^n$ satisfy the standard linear regression equation. Instead of computing the Lasso with $(\mathbf{X}, Y)$, preconditioning first left multiplies by $F \in \mathbb{R}^{n \times n}$ and then computes the Lasso with $(F\mathbf{X}, FY)$.

While others have proposed preconditioning for other purposes, we provide the first results that show $F\mathbf{X}$ can satisfy the irrepresentable condition even when $\mathbf{X}$ fails to satisfy the condition. Preconditioning the Lasso creates a new estimator that is sign consistent in a wider variety of settings. Importantly, left multiplying the regression equation by F does not change β, the vector of unknown coefficients. However, left multiplying this equation by F often inflates the variance of the errors. We propose a class of preconditioners to balance these costs and benefits.

Keywords and phrases: Preconditioning, irrepresentable condition, sign consistency.

Received July 2014.

[*] The authors contributed equally and are listed in alphabetical order. This work is partially supported by the National Basic Research Program of China (973 Program 2011CB809105), the National Science Foundation of China (61121002, 11101005), DPHEC-20110001120113, the Wisconsin Alumni Research Foundation, and by the US National Science Foundation (DMS-1309998). Thank you to Bin Yu and Peter Qian for fruitful conversations while this idea was still unformed. Important steps of this research were conducted at the Institute for Mathematics and its Applications' (IMA's) conference "Machine Learning: Theory and Computation." Thank you to IMA for both the conference and the financial support. Thank you to Nicolai Meinshausen for sharing code that contributed to Section 2.2. Thank you to Sara Fernandes-Taylor for helpful comments.

1. Introduction

Recent breakthroughs in information technology have provided new experimental capabilities in astronomy, biology, chemistry, neuroscience, and several other disciplines. Many of these new measurement devices create data sets with many more “measurements” than units of observation. For example, due to experimental constraints, both fMRI and microarray experiments often include tens or hundreds of people. However, the fMRI and microarray technologies can simultaneously measure thousands to millions of different pieces of information for each individual. Sparse high dimensional regression aims to select a small set of measurements that relate to an outcome of interest.

The Lasso (Tibshirani, 1996) is one of the most popular techniques for sparse high dimensional regression because it is the solution to a convex optimization problem, allowing for fast algorithms and assurances of global optimality. A rich theoretical literature describes the conditions for the Lasso to consistently estimate the regression coefficients (Bühlmann and van de Geer, 2011). Because of the Lasso’s ability to select a sparse solution, it is of particular interest to understand when the Lasso can select the true nonzero coefficients in the linear regression model. Stated loosely, the Lasso performs well in this respect when the columns of $\mathbf{X}$ are weakly correlated. This concept is formalized with sign consistency and the irrepresentable condition (see Section 2).

It is well known that the Ordinary Least Squares (OLS) estimator performs poorly when the columns of the design matrix are highly correlated. However, more samples overcome this problem; OLS is still consistent. With the Lasso, the detrimental effects of correlation are more severe. If the columns of the design matrix are correlated in a way that violates the irrepresentable condition, then the Lasso will fail to estimate the correct signs and the estimation performance will not improve by increasing the number of samples or increasing the signal to noise ratio. This paper demonstrates that, for the purposes of the Lasso, the correlation in the design matrix is malleable and can be diminished (at the expense of marginally more variance) by preconditioning, a classical technique to accelerate solvers of systems of equations. This paper demonstrates that in many sparse regression settings, preconditioning the Lasso produces a better estimator of the sparsity pattern. The next section gives a surprising simulation that shows how correcting for heteroskedasticity with generalized least squares (GLS) can act as a bad preconditioner and degrade the estimation performance of the Lasso. This contrasts with the classical results that demonstrate how GLS improves upon the estimation performance of ordinary least squares (OLS).

1.1. Some notation and a surprising simulation

Suppose the regression model

$$Y = \mathbf{X}\beta^* + \epsilon \tag{1}$$

where $Y \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\beta^* \in \mathbb{R}^p$, and $\epsilon \sim N(0, \Sigma)$. By observing Y and $\mathbf{X}$, we are interested in estimating the support of β^*,

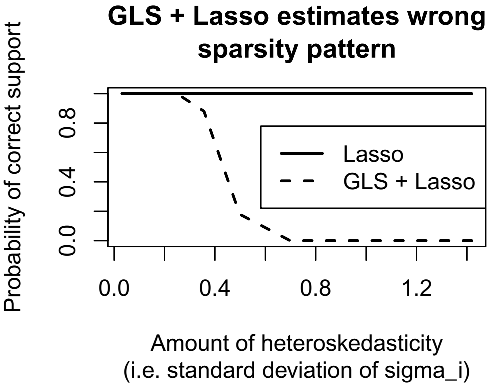

Fig 1. *GLS acts as a bad preconditioner, making the design matrix ill-conditioned. Thus, correcting for the heteroskedasticity degrades the estimation performance. In this simulation,* $n = 200$*,* $p = 1000$ *and there are 10 nonzero elements in* β^**. Appendix D contains further details on this simulation.*

$$S = \{j : \beta_j^* \neq 0\} \subset \{1, \ldots, p\}. \tag{2}$$

The rest of the paper assumes $\Sigma = \sigma^2 I_n$ (where $I_n \in R^{n\times n}$ is the identity matrix). However, this motivating simulation uses a heteroskedastic model where Σ is a diagonal matrix with diagonal entries σ_i^2. Each σ_i is an independent draw from the Gamma distribution. In all simulations $E(\sigma_i) = 1$. The horizontal axis in Figure 1 represent the standard deviation of σ_i (i.e. the amount of heteroskedasticity).

Figure 1 compares two techniques under this heteroskedastic model. The first estimator does not correct for heteroskedasticity. It is the standard Lasso estimator that we study in the rest of the paper

$$\hat{\beta}(\lambda) = \arg\min_b \frac{1}{2}\|Y - \mathbf{X}b\|_2^2 + \lambda\ \|b\|_1 \tag{3}$$

where $\|x\|_r = (\sum_{i=1}^k |x_i|^r)^{1/r}$ for $x \in \mathbb{R}^p$. To correct for the heteroskedastic or correlated errors, GLS left multiplies the regression equation (1) by $\Sigma^{-1/2}$,

$$\Sigma^{-1/2}Y = \Sigma^{-1/2}X\beta^* + \Sigma^{-1/2}\epsilon. \tag{4}$$

Then, the error term becomes a vector of iid normal variables, $\Sigma^{-1/2}\epsilon \sim N(0, I)$. Instead of computing the Lasso with $(\mathbf{X}, Y)$ as in Equation (3), use

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

and define the resulting estimator as $\hat{\beta}_{GLS+Lasso}(\lambda)$.

The vertical axis of Figure 1 reports the proportion of fifty simulations in which there exists a tuning parameter λ such that the support of the estimator aligns perfectly with the true support S. At the very left, the model is homoskedastic and the estimators are equivalent. As the heteroskedasticity increases, one expects the $\hat{\beta}_{GLS+Lasso}(\lambda)$ to outperform $\hat{\beta}(\lambda)$. However, the performance of $\hat{\beta}_{GLS+Lasso}(\lambda)$ quickly degrades, failing to estimate the correct sparsity pattern. In this simulation, correcting for heteroskedasticity *degrades* estimation. This surprising result happens because $\Sigma^{-1/2}$ in Equation (4) acts as a preconditioner, *a bad preconditioner*. It makes the design matrix ill-conditioned. In least squares regression, an ill-conditioned design matrix does not create bias. As such, GLS can improve the estimation performance in the classical setting. However, in penalized least squares regression, an ill-conditioned design matrix prevents support recovery; any decrease in the variance is offset by an increase in the support recovery bias.

Just as there are several settings where the original data has heteroskedastic or correlated errors, there are several settings where the original data contains an ill-conditioned design matrix. The techniques described in this paper correct for ill-conditioned design matrices. For example, perhaps some rows of $\mathbf{X}$ have a much larger ℓ_2 length than some other rows. Where GLS "whitens" the errors (with the side effect of making the design ill-conditioned), preconditioning "whitens" the design matrix (with the side effect of making the errors heteroskedastic and correlated). In this sense, GLS and the preconditioning are opposite transformations. Figure 1 gives a simulation setting where penalized regression (1) can accommodate heteroskedastic errors and (2) cannot accommodate an ill-conditioned design matrix. This suggests that the classical intuition, which ignores the conditioning of the design and focus exclusively on the distribution of the errors, does not extend to the modern settings. The rest of this paper focuses on preconditioning matrices that make the design matrix well conditioned, thus improving the estimation performance of the Lasso.

In this simulation, Σ does not describe heteroskedastic behavior in $\mathbf{X}$. This is why $\Sigma^{-1/2}$ acts as a poor preconditioner. However, in many applications, the underlying structure which leads to covariance or heteroskedasticity in ϵ (e.g. spatial dependence) may create covariance or heteroskedasticity in the rows of $\mathbf{X}$. In these situations, $\Sigma^{-1/2}$ will act as a *good* preconditioner because it will decorrelate and normalize the rows of $\mathbf{X}$.

2. Preconditioning to circumvent the irrepresentable condition

This section defines sign consistency and the irrepresentable condition, a necessary and almost sufficient condition for the Lasso to be sign consistent.

Let $Y = (Y_1, \dots, Y_n)^T$ and $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$. For $T \subset \{1, \dots, p\}$ with $|T| = t$, define $\mathbf{X}_T \in \mathbb{R}^{n \times t}$ to contain the columns of $\mathbf{X}$ indexed by T. For any vector $x \in \mathbb{R}^p$, define $x_T = (x_j)_{j \in T}$. Define $s = |S|$, the cardinality of S in Equation (2). To define an appropriate measure of selection consistency, define

$$\operatorname{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

For a vector $b \in \mathbb{R}^p$, $\operatorname{sign}(b) \in \mathbb{R}^p$ is defined elementwise: $[\operatorname{sign}(b)]_i = \operatorname{sign}(b_i)$.

Definition 1. The Lasso is **sign consistent** if there exists a sequence λ_n such that,

$$P\left(\text{sign}(\hat{\beta}(\lambda_n)) = \text{sign}(\beta^*)\right) \to 1, \text{ as } n \to \infty.$$

This implies that $\hat{\beta}(\lambda)$ can asymptotically identify the relevant and irrelevant variables when it is sign consistent. Several authors, including Meinshausen and Bühlmann (2006); Zou (2006); Zhao and Yu (2006), have studied the sign consistency property and found a sufficient condition for sign consistency. Zhao and Yu (2006) call this assumption the irrepresentable condition. For a vector x, denote $\|x\|_\infty = \max_i |x_i|$.

Definition 2. The design matrix $\mathbf{X}$ satisfies the **irrepresentable condition** for β^* if, for some constant $\eta \in (0, 1]$,

$$\left\| \mathbf{X}_{S^c}^T \mathbf{X}_S \left(\mathbf{X}_S^T \mathbf{X}_S\right)^{-1} \text{sign}(\beta_S^*) \right\|_\infty \leq 1 - \eta. \tag{5}$$

In the above sufficient condition, $\eta > 0$. If this is replaced with $\eta \geq 0$, then it is a necessary condition for sign consistency (Zhao and Yu, 2006; Zou, 2006). This condition is difficult to check because it relies on the unknown set S. Section 2 of Zhao and Yu (2006) gives several sufficient conditions. For example, if $|\text{cor}(X_i, X_j)| \leq c/(2s-1)$ for a constant $0 \leq c < 1$, then the irrepresentable condition holds for any S.

The extant literature has proposed several ways of circumventing the irrepresentable condition. The two methods that have received the most attention both focus on refining the ℓ_1 penalty term in the Lasso objective function (3). Fan and Li (2001) and Zhang (2010) proposed making the penalty function concave. The adaptive Lasso is another popular approach (Zou, 2006); this applies a different ℓ_1 penalty to each element of the coefficient vector; these penalty weights come from an initial run of OLS. van de Geer et al. (2011) illustrated how this can be extended to the high dimensional setting by using an initial run of the Lasso instead of OLS. While these previous approaches alter the penalty function, preconditioning instead changes the shape of the least squares contours in the data fidelity term $\|Y - Xb\|_2^2$. Similar to work presented here, Xiong et al. (2011) also propose adjusting the data fidelity term to avoid the irrepresentable condition. However, instead of preconditioning, they proposed a procedures which (1) makes the design matrix orthogonal by adding rows, and (2) applies an EM algorithm, with concave penalty SCAD, to estimate the outcomes corresponding to the additional rows in the design matrix.

2.1. The Puffer transformation

In this paper, we always assume that the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has rank $d = \min\{n, p\}$. From the singular value decomposition, there exist matrices $U \in \mathbb{R}^{n\times d}$ and $V \in \mathbb{R}^{p\times d}$ with $U^T U = V^T V = I_d$ and diagonal matrix $D \in \mathbb{R}^{d\times d}$ such that $\mathbf{X} = UDV^T$. Define the **Puffer transformation** $F = UD^{-1}U^T$. The preconditioned design matrix $F\mathbf{X}$ has the same singular vectors as $\mathbf{X}$. However, all of the nonzero singular values of $F\mathbf{X}$ are set to unity: $F\mathbf{X} = UV'$. When $n \geq p$, the columns of $F\mathbf{X}$ are orthonormal. When $n \leq p$, the rows of $F\mathbf{X}$ are orthonormal.

After left multiplying the regression equation

$$Y = \mathbf{X}\beta^* + \epsilon$$

by the matrix F, the transformed regression equation becomes

$$FY = (F\mathbf{X})\beta^* + F\epsilon.$$

If $\epsilon \sim N(0, \sigma^2 I_n)$, then $F\epsilon \sim N(0, \tilde{\Sigma})$ where $\tilde{\Sigma} = \sigma^2 UD^{-2}U^T$. The parenthesis around $(F\mathbf{X})$ emphasize that preconditioning is transforming $\mathbf{X}$, not β^*. Just as in GLS (Equation 4), β^* remains unchanged after left multiplying the regression equation.

The scale of $\tilde{\Sigma}$ depends on the diagonal matrix D, which contains the d singular values of $\mathbf{X}$. If any singular values of $\mathbf{X}$ approach zero, the corresponding elements of D^{-2} grow, amplifying the noise $F\epsilon$. This increased noise can quickly overwhelm the benefits of a well conditioned design matrix. For this reason, Section 3.3 proposes a slightly modified preconditioner that bounds the spectral norm of $\tilde{\Sigma}$.

In numerical linear algebra, the objective is speed, and there is a trade off between the time spent computing the preconditioner vs. solving the system of equations. Better preconditioners make the resulting system of equations easier to solve. However, these preconditioners themselves can be time consuming to compute. In our setting, the objective is inference, not speed per se, and the tradeoff is between a well behaved design matrix and a well behaved error term. Preconditioning can aid statistical inference if it can balance these two constraints.

2.1.1. Previous literature on preconditioning for sparse inference

This paper contributes to the existing literature by studying when preconditioning can circumvent the irrepresentable condition. For other reasons, preconditioning the Lasso has been proposed elsewhere.

Paul et al. (2008) estimate a type of latent factor model; theoretical and simulation results suggest that their preconditioning technique improved estimation in settings with a low signal to noise ratio. However, Paul et al. (2008) do not study the relationship between preconditioning and the irrepresentable condition.

More recently, Huang and Jojic (2011) use preconditioning to remove the effects of confounding in high-throughput biological experiments and are motivated by empirical observations in genome wide association studies. In such biological studies, Alter et al. (2000) show how the leading singular vectors from $\mathbf{X}$ often “represent additive or multiplicative noise, experimental artifacts, or even irrelevant biological processes.” As such, several papers have studied techniques that screen out the (typically large) singular vectors of $\mathbf{X}$; see Yang et al.

(2014) for a further references and discussion. These empirical observations, not the irrepresentable condition, motivated Huang and Jojic (2011) to emphasize the bottom singular values in $\mathbf{X}$. Although they accomplish this through preconditioning, their motivation is focused on biological experiments. With data analysis and biologically motivated simulations, they show that that preconditioning improves the model selection performance of the Lasso.

Most recently, Rauhut and Ward (2011) study interpolation with orthogonal polynomials. They precondition the polynomials with a *diagonal* preconditioner to satisfy the restricted isometry principal with high probability. In the current paper, we employ non-diagonal preconditioning which drastically increases the class of design matrices that benefit from preconditioning. Moreover, we demonstrate how preconditioning alters the error term, creating a statistical tradeoffs between a well conditioned design matrix and a well behaved error term.

The technical report for this paper introduced the Puffer Transformation. This led to two pieces of follow up research. First, Qian and Jia (2012) demonstrate the benefits of the Puffer transformation for the fused Lasso, a sparse high dimensional regression problem that is particularly plagued by correlation in the design matrix. Second, Wauthier et al. (2013) compares the Puffer transformation to the two two previous techniques in Paul et al. (2008) and Huang and Jojic (2011). Their analysis assumes that there exists a range of $\lambda \in [\lambda_l, \lambda_u]$ for which the standard Lasso estimates the correct sign. They study when preconditioning increases the ratio λ_u/λ_l, thus making sign estimation more robust to the choice of tuning parameter λ. Their results highlight the fact that the preconditioners in Paul et al. (2008) and Huang and Jojic (2011) project onto rank deficient subspaces. Wauthier et al. (2013) goes on to present a specific model for the design matrix $\mathbf{X}$ under which the Puffer transformation deterministically scales λ_u/λ_l. Under their model, if the largest singular vectors have small values in the positions of S, then the Puffer transformation will increase λ_u/λ_l. Otherwise, the Puffer transformation will decrease λ_u/λ_l. In this paper, we are particularly interested in situations where the design matrix fails to satisfy the irrepresentable condition before preconditioning (i.e. $\lambda_u < \lambda_l$).

Section 4.3 gives two brief simulations that compare the Puffer transformation to the preconditioners defined in Paul et al. (2008) and Huang and Jojic (2011).

In the sparse regression literature, other papers have considered left multiplying the regression equation for alternative reasons. Bootstrapping techniques such as Chatterjee and Lahiri (2011) left multiply the regression equation by a random diagonal matrix. Penalized generalized linear models are fit by iteratively reweighted least squares, which is equivalent to left multiplying by a diagonal matrix at each iteration; van de Geer (2008) highlights how it is the conditioning of the final iteration that matters for sign consistency. The subbagging technique in Bradic (2013) concludes by solving the Lasso with a random diagonal weighting matrix on each sub-Lasso problem. These weights are chosen to obtain a random approximation for the solution of the original unweighted problem, not to adjust for the irrepresentable condition.

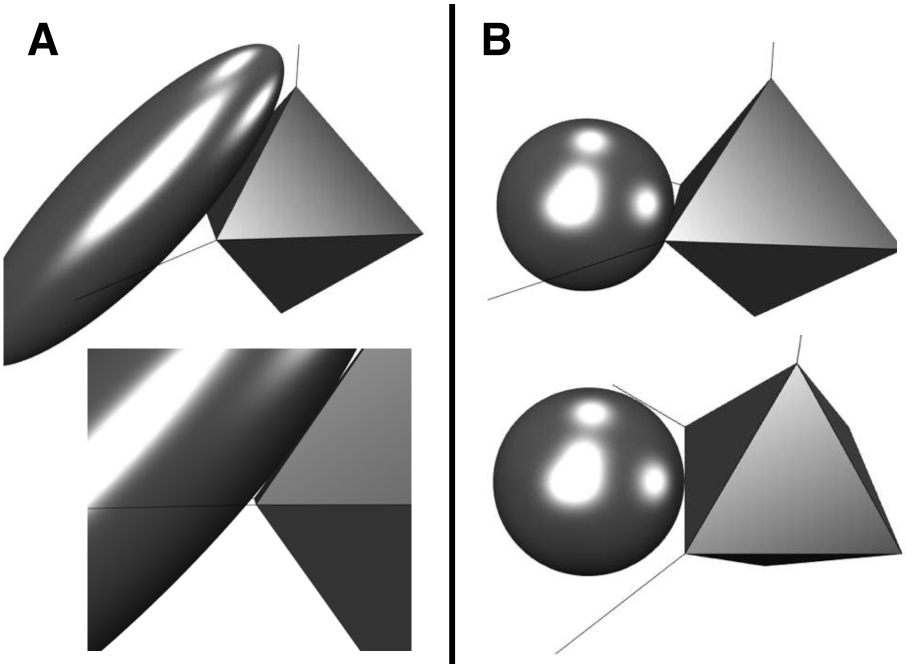

Fig 2. *Panel A illustrates the Lasso without preconditioning. In panel B, preconditioning turns the ellipse from the ℓ_2 loss into a sphere. Here, the Lasso correctly selects the true model. These figures were drawn with the R library RGL (Adler et al., 2003).*

2.2. *Geometrical representation of preconditioning and the irrepresentable condition*

Figure 2 displays the geometry of the Lasso before and after the Puffer transformation. This figure (i) demonstrate what happens when the irrepresentable condition is not satisfied, (ii) reveal how the Puffer transformation circumvents the irrepresentable condition, and (iii) illustrate why we call F the Puffer transformation.

The figures in this section are derived from the following optimization problem which is equivalent to the Lasso, $\hat{\beta}(c) = \arg\min_{b:\|b\|_1 \le c} \|Y - \mathbf{X}b\|_2^2$. The definition of $\hat{\beta}(c)$ abuses notation. In fact, there is a one-to-one function $\phi(c) = \lambda$ to make the Lagrangian form of the Lasso (Equation 3) equivalent to the constrained form of the Lasso denoted by $\hat{\beta}(c)$. Given the constraint set $\|b\|_1 \le c$ and a continuum of sets $\|Y - \mathbf{X}b\|_2^2 \le x$ for $x \ge 0$, define

$$\mathcal{I}(c, x) = \{b : \|b\|_1 \le c\} \cap \{b : \|Y - \mathbf{X}b\|_2^2 \le x\}.$$

Let x^* be the smallest x such that $\mathcal{I}(c, x)$ is nonempty. Then, $\hat{\beta}(c) \in \mathcal{I}(c, x^*)$. Under certain conditions on $\mathbf{X}$ (e.g. full column rank), the solution is unique and $\hat{\beta}(c) = \mathcal{I}(c, x^*)$. In Figure 2, the constraint set $\{b : \|b\|_1 \le c\}$ appears as a diamond shaped polyhedron and the level set of the loss function $\{b : \|Y - \mathbf{X}b\|_2^2 < x^*\}$ appears as an ellipse. The rows of X are sampled as three dimensional Gaussian vectors with mean zero. The first two elements are independent Gaussians and the third element has correlation .6 with both the first and second elements. To highlight the effects of preconditioning, the noise is very small and $n = 10{,}000$.

In panel A, the design matrix is not preconditioned. In panel B, the problem has been preconditioned, and the ellipse represents the set $\|FY - F\mathbf{X}b\|_2^2 \leq x^*$; preconditioning turns the oblong ellipse in panel A into the sphere in panel B.

In this simulation, $\beta^* = (1, 1, 0)$ and in all illustrations, the third dimension is represented by the axis that points up and down. Thus, the Lasso estimates the correct sign if the ellipse intersects the constraint set in the (horizontal) plane formed by the first two dimensions. The design matrix in panel A fails the irrepresentable condition because the elongated ellipse forces $\hat{\beta}(c)$ off of the true plane. This is shown in the bottom illustration in panel A.

In panel B, the design matrix $F\mathbf{X}$ satisfies the irrepresentable condition because the elongated direction of the ellipse shrinks down and the ellipse is puffed out into a sphere. Because of this, $\hat{\beta}(\lambda)$ lies in the true plane. When $n > p$ (as in these figures) preconditioning with F makes the ellipse a sphere. When $p > n$, preconditioning with F can make low dimensional projections of the ellipse more spherical. The name Puffer transformation comes from the pufferfish. As Figure 2 illustrates, the Puffer transformation inflates the smallest singular values of the design matrix, making the contours of $\|FY - F\mathbf{X}b\|_2^2 \leq x^*$ more spherical.

2.3. *Low dimensional results*

If $n \geq p$ and $\mathbf{X}$ is full rank, then the preconditioned design matrix is orthonormal.

$$(F\mathbf{X})^T F\mathbf{X} = VDU^TUD^{-1}U^TUD^{-1}U^TUDV^T = I$$

Orthogonal matrices trivially satisfy the irrepresentable condition and other conditions such as the restricted eigenvalue condition for ℓ_2 consistency (Bickel et al., 2009). Theorem 1 proves that the preconditioned Lasso is sign consistent, so long as the smallest eigenvalue of $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ is bounded away from zero.

Theorem 1. *Suppose that data* $(\mathbf{X}, Y)$ *follows the linear model described in Equation (1) with iid Gaussian noise* $\epsilon \sim N(0, \sigma^2 I_n)$. *Define the singular value decomposition of* $\mathbf{X}$ *as* $\mathbf{X} = UDV^T$. *Suppose that* $n \geq p$ *and* $\mathbf{X}$ *has rank* p. *Further assume that* $\Lambda_{\min}(\frac{1}{n}\mathbf{X}^T\mathbf{X}) \geq \tilde{C}_{\min} > 0$. *Define the* ***Puffer transformation****,* $F = UD^{-1}U^T$. *Let* $\tilde{\mathbf{X}} = F\mathbf{X}$ *and* $\tilde{Y} = FY$. *Define* $\tilde{\beta}(\lambda) = \arg\min_b \frac{1}{2}\|\tilde{Y} - \tilde{\mathbf{X}}b\|_2^2 + \lambda\|b\|_1$.

If $\min_{j \in S} |\beta_j^*| \geq 2\lambda$, *then* $\tilde{\beta}(\lambda) =_s \beta^*$ *with probability greater than*

$$1 - 2p \exp\left\{-\frac{n\lambda^2\tilde{C}_{\min}}{2\sigma^2}\right\}.$$

A proof can be found in Appendix A in the supplementary material (Jia and Rohe, 2015) on page 5. The proof is very similar to the standard result for the homogenous linear model. The difference here is that after the Puffer Transformation, the vector ϵ contains correlated entries. To overcome this difficulty, the proof relies on a Gaussian comparison result that does not need any assumptions on the correlations of Gaussian random variables. Instead, it depends on the maximum among a set of Gaussian (or sub-Gaussian) random variables.

Remark 1. The loss function defined in the Lasso estimator,

$$\tilde{\beta}(\lambda) = \arg\min_b \frac{1}{2}\|\tilde{Y} - \tilde{\mathbf{X}}b\|_2^2 + \lambda\|b\|_1,$$

is slightly different from the classical definition, which uses $\frac{1}{2n}\|\tilde{Y} - \tilde{\mathbf{X}}b\|_2^2$. In fact, the the Puffer Transformation accounts for this change of scale because it depends on the SVD of X, instead of $\frac{1}{n}X^TX$. As such, it changes the scale of the loss function.

Remark 2. Suppose that $\tilde{C}_{\min} > 0$ is a constant. If $p, \min_{j\in S}|\beta_j^*|$ and σ^2 do not change with n, then choosing λ such that $\lambda \to 0$ and $\lambda^2 n \to \infty$, ensures that $\tilde{\beta}(\lambda)$ is sign consistent. One possible choice is $\lambda = \sqrt{\frac{\log n}{n}}$.

In classical linear regression, increasing the correlation between columns of $\mathbf{X}$ amplifies the variance of the standard OLS estimator; correlated predictors make estimation more difficult. Without preconditioning, this intuition does not hold for the standard Lasso; increased correlation in $\mathbf{X}$ creates an increasingly biased estimator. Theorem 1 shows that after preconditioning, the intuition from OLS again translates; increasing the correlation between the columns of $\mathbf{X}$ decreases the smallest singular value of $\mathbf{X}$, increasing the spectral norm of F and the variance of the noise terms. Importantly, a large sample size n can overcome the additional noise induced by preconditioning.

Theorem 1 applies to the more general class of penalized least squares methods

$$\arg\min_b \frac{1}{2}\|Y - \mathbf{X}b\|_2^2 + pen(b,\lambda) \tag{6}$$

for some type of penalty function $pen(b) : \mathbb{R}^p \to \mathbb{R}$, e.g. Lasso, SCAD, and MCP (Fan and Li, 2001; Zhang, 2010). After preconditioning, the design matrix $F\mathbf{X}$ is orthogonal and several convenient facts follow. First, if the penalty decomposes, $pen(b,\lambda) = \sum_{j=1}^p pen_j(b_j,\lambda)$ so that pen_j does not rely on b_k for $k \neq j$, then the penalized least squares methods admit closed form solutions. If it is also true that all the pen_j's are identical functions that have a cusp at zero (e.g. Lasso, SCAD, MCP), then the solution to the preconditioned penalized least squares problem selects the same sequence of models as preconditioned correlation screening (i.e. select X_j if $|cor(FY, FX_j)| \geq \lambda$) (Fan and Lv, 2008). Theorem 1 implies that all such methods are sign consistent. These observations rely on the fact that $F\mathbf{X}$ is an orthogonal matrix. In high dimensions, $F\mathbf{X}$ is no longer orthogonal. So, the various methods could potentially estimate different models.

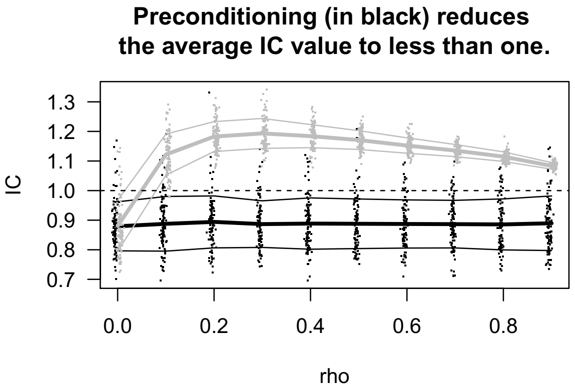

FIG 3. *As the correlation ρ increases, most values of $IC_{\beta^*}(F\mathbf{X})$ (in black) remain below the dashed line corresponding to the critical threshold at 1. Without preconditioning, $IC_{\beta^*}(\mathbf{X})$ (in grey) quickly surpasses the critical threshold. Each point corresponds to one design matrix. The thick black and grey lines pass through the average IC value for each setting of ρ. The thin solid lines correspond to $+/-$ one standard deviation.*

3. High dimensional results

Subsection 3.1 gives a motivating simulation that illustrates the benefits of preconditioning in the high dimensional setting. Theorem 2 in Subsection 3.2 shows that $F\mathbf{X}$ satisfies the irrepresentable condition for many design matrices $\mathbf{X}$. Subsection 3.3 proposes a class of generalized Puffer transformations and Theorem 3 proves that the Lasso with a specific preconditioner in this class can be sign consistent with arbitrarily small singular values in $\mathbf{X}$.

3.1. Motivating simulation

Figure 3 presents an illustrative numerical simulation to prime our intuition on preconditioning in high dimensions. In this simulation, $n = 200, p = 10{,}000$, and each row of $\mathbf{X}$ is an independent Gaussian vector with mean zero and covariance matrix Σ. The diagonal of Σ is all ones and the off diagonal elements are all ρ; ρ varies on the horizontal axis of Figure 3. The vertical axis plots the values

$$IC_{\beta^*}(\mathbf{X}) = \left\| \mathbf{X}_{S^c}^T \mathbf{X}_S \left(\mathbf{X}_S^T \mathbf{X}_S \right)^{-1} \mathrm{sign}(\beta_S^*) \right\|_\infty \tag{7}$$

where $S = \{1, \ldots, 10\}$ and the nonzero elements of β^* are all positive. Along with $IC_{\beta^*}(\mathbf{X})$, the figure also contains $IC_{\beta^*}(F\mathbf{X})$ and a horizontal line at 1. Recall that if $IC_{\beta^*}(\mathbf{X}) < 1$, then $\mathbf{X}$ satisfies the irrepresentable condition.

The figure shows that $IC_{\beta^*}(\mathbf{X})$ quickly exceeds 1, while $IC_{\beta^*}(F\mathbf{X}) < 1$ for all values of ρ. The reason that this happens is that preconditioning drastically reduces the correlation between the columns. For example, for $\rho = .9$, the pairwise correlations between the columns of $\mathbf{X}$ have an average of .90 with a standard deviation of .01. After the transformation, the average correlation is .005, and the standard deviation is .07. By reducing the pairwise correlations, preconditioning helps the design matrix satisfy the irrepresentable condition.

3.2. Uniform distribution on the Stiefel manifold

When $p \geq n$ and $\mathbf{X}$ is full rank the *rows* of $F\mathbf{X}$ are orthogonal. It lies in the Stiefel manifold,

$$F\mathbf{X} \in V(n,p) = \{V \in \mathbb{R}^{n\times p} : VV^T = I_n\}.$$

Moreover, when $p \geq n$, F can be computed as $(\mathbf{X}\mathbf{X}')^{-1/2}$ and $F\mathbf{X} = (\mathbf{X}\mathbf{X}')^{-1/2}\mathbf{X}$ is the projection of $\mathbf{X}$ onto $V(n,p)$ under any unitarily invariant norm (Fan and Hoffman, 1955). Denote the orthogonal group of matrices as $O(p,\mathbb{R}) = V(p,p)$.

Definition 3 (Chikuse (2003))**.** A random matrix V is uniformly distributed on $V(n,p)$, written $V \sim \mathit{uniform}(V(n,p))$, if the distribution of V is equal to the distribution of VO for any fixed O in the orthogonal group of matrices $O(p,\mathbb{R})$.

Theorem 2 shows that if $F\mathbf{X} \sim \mathit{uniform}(V(n,p))$, then the matrix satisfies the irrepresentable condition with high probability. Propositions 1 and 2 give two examples of random design matrices $\mathbf{X}$ where $F\mathbf{X}$ is uniformly distributed on $V(n,p)$.

Theorem 2. *Suppose that $V \sim \mathit{uniform}(V(n,p))$ and let $\mathbf{X} = UDV^T$ for any $U \in O(p,\mathbb{R})$ and diagonal matrix D. If $p - s \geq n, p > 9n$ and $n > 400(s+1)^2$, then*

$$P\left[IC_{\beta^*}(F\mathbf{X}) \geq 4/5\right] \leq 10p \exp\left\{-\frac{n}{800s}\right\}.$$

Section C in the supplementary material contains a proof for this theorem. The proof is on page 17 and restated as Theorm C.2 in the supplementary material. The first step of the proof is to relate the matrix $F\mathbf{X}$, drawn uniformly from Stifle manifold, to matrices that contain iid $N(0,1)$ elements. Then, results for random matrix theory control the spectral norm of the Gaussian random matrix and provide the result (Davidson and Szarek, 2001). A similar argument obtains a similar result for a non-preconditioned design matrix $\mathbf{X}$ with iid $N(0,1)$ entries; this is included in Theorem B.2 in the supplementary material. Propositions 1 and 2 give two models for $\mathbf{X}$ that make $F\mathbf{X} \sim \mathit{uniform}(V(n,p))$.

Proposition 1. *If the elements of $\mathbf{X}$ are independent $N(0,1)$ random variables, then $F\mathbf{X} \sim \mathit{uniform}(V(n,p))$.*

Proposition 2. *Suppose that $U_\Sigma \in \mathbb{R}^{p\times p}$ is drawn uniformly from $O(p,\mathbb{R})$ and $D_\Sigma \in \mathbb{R}^{p\times p}$ is a diagonal matrix with positive entries. Define $\Sigma = U_\Sigma D_\Sigma U_\Sigma^T$ and suppose the rows of $\mathbf{X}$ are drawn independently from $N(0,\Sigma)$, then $F\mathbf{X} \sim \mathit{uniform}(V(n,p))$.*

The proofs for these propositions are in the supplementary material, Section C.

3.3. Generalized Puffer transformation

In the preconditioned regression equation, the noise ϵ becomes $F\epsilon$. Since the spectral norm of F is unbounded as the smallest nonzero singular value of $\mathbf{X}$ approaches zero, the preconditioned noise $F\epsilon$ has unbounded variance. To diminish the increase in variance from preconditioning, this section studies a generalized form of the Puffer transformation.

Definition 4. Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be a design matrix with singular value decomposition $\mathbf{X} = UDV^T$. For the matrix $\mathbf{X}$, the generalized Puffer transformation with $g : \mathbb{R}^2 \to \mathbb{R}$ and $\tau \in \mathbb{R}$ is $F_{g,\tau} = U\hat{D}U^T,$ where $\hat{D}_{ii} = g(D_{ii}, \tau)/D_{ii}$.

This definition implies that $F_{g,\tau}\mathbf{X} = U\tilde{D}V^T$ where $\tilde{D}_{ii} = g(D_{ii}, \tau)$. Here, g is a function of the singular values of $\mathbf{X}$ and a tuning parameter τ. The Puffer transformation is $F = F_{1,\tau}$ where $1(D_{ii}, \tau) = 1$. To illustrate the potential benefits from this generalized preconditioner, define the hard thresholding function as

$$h(x, \tau) = 1 \text{ if } x \geq \tau \text{ and zero otherwise.} \tag{8}$$

The spectral norm of $F_{h,\tau}$ is bounded by $1/\tau$, limiting the amount that the preconditioner amplifies the noise. This next theorem studies this preconditioner under a model where the singular values of $\mathbf{X}$ are potentially very small and assumes that $V \sim$ *uniform*$(V(n, p))$, where V contains the right singular vectors of $\mathbf{X}$. This highlights the tradeoff between (a) satisfying the irrepresentable condition and (b) limiting the amount of additional noise created by preconditioning.

Theorem 3. *Suppose that $V \sim$ uniform$(V(n, p))$ and let $\mathbf{X} = UDV^T$ for any $U \in O(p, \mathbb{R})$ and diagonal matrix D. Suppose $Y = \mathbf{X}\beta^* + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I_n)$, independent of $\mathbf{X}$. For $\tau_n > 0$, let $\tilde{n}$ be the number of D_{ii}'s greater than or equal to τ_n.*

Define the hard thresholding function $h(x, \tau_n)$ as in Equation (8) and the generalized preconditioner F_{h,τ_n} as in Definition 4. Define $\tilde{Y} = F_{h,\tau_n}Y$, $\tilde{\mathbf{X}} = F_{h,\tau_n}\mathbf{X}$, and $\tilde{\beta}(\lambda) = \arg\min_b \frac{1}{2}\|\tilde{Y} - \tilde{\mathbf{X}}b\|_2^2 + \lambda\|b\|_1$. Suppose that $p - s \geq \tilde{n}$, $p > 9\tilde{n}$ and $\tilde{n} > 400(s + 1)^2$. If $\min_{j\in S} |\beta_j^*| \geq 2\lambda\sqrt{9sp/(5\tilde{n})}$, then

$$P\big(\tilde{\beta}(\lambda) =_s \beta^*\big) \geq \left[1 - 10p\exp\left\{-\frac{\tilde{n}}{800s}\right\} - 5\exp\left\{-\frac{\tilde{n}}{800}\right\}\right]\left[1 - 2p\exp\left\{-\frac{\lambda^2\tau_n^2}{50\sigma^2}\right\}\right].$$

A proof can be found in Section C in the supplementary material. The proof is on page 20 and restated as Theorm C.4 in the supplementary material. The proof for this result relies on the previous result in Theorem 2 saying that with high probably the irrepresentable condition holds.

The assumption on $\min_{j\in S} |\beta_j^*|$ appears restrictive. However, the scale of λ, τ_n, and D play an essential role. After accounting for these terms, this condition is comparable to previous results. For the probability bound to converge to one, $\lambda^2\tau_n^2$ must grow faster than $\log p$. So, it is necessary to consider how τ_n grows in a standard setting. In the situation where the elements of $\mathbf{X}$ contain iid random variables with constant variance, the average element of D is $O(\sqrt{p})$. If τ_n grows at this rate, then choosing $\lambda^2 = p^{-1}\log n \log p$ ensures the last term in the probability bound converges to one. This yields the condition

$$\min_{j \in S} |\beta_j^*| \geq c\frac{\log n}{\sqrt{\tilde{n}}},$$

which is comparable to previous results. If τ_n is smaller, then $\tilde{n}$ is larger. However, λ^2 must also be larger. As a result, but the lower bound on $\min_{j\in S}|\beta_j^*|$ becomes more strict. To ensure that this lower bound is not growing, τ_n should grow faster than $\sqrt{p/n}$.

This theorem does not assume that $\mathbf{X}$ satisfies the irrepresentable condition. Instead, it supposes that $V \sim \mathit{uniform}(V(n,p))$ and only presumes that D has sufficiently many values greater than τ. Several previous papers have also studied the Lasso (without preconditioning) under generative models for $\mathbf{X}$ (e.g. (Rudelson and Vershynin, 2006; Candes and Romberg, 2007)). The previous literature has constructed these random designs in a few different ways. For example, containing independent and identically distributed elements (e.g. binary or Gaussian) or by taking an orthonormal basis $O \in V(p,p)$ (e.g. Fourier transform) and sampling n elements of this basis uniformly at random; these n elements are then concatenated to form an $n \times p$ design matrix. In all previous cases, these matrices will be well conditioned (i.e. the smallest non-zero singular value of $\mathbf{X}$ has the same order of magnitude as the largest singular value). However, if the experimental design or physical constraints restrict the sampling mechanism in $\mathbf{X}$, then $\mathbf{X}$ will likely be ill conditioned and thus fail the irrepresentable condition. Theorem 3 allows for such design matrices by not making any assumptions on the smallest elements in D, showing that the Lasso can still be sign consistent with a generalized preconditioner.

4. Simulations

This section contains two simulations that study the performance of the Puffer transformation and the generalized Puffer transformation. The first simulation compares the model selection and ℓ_2 estimation performance of the Puffer transformed Lasso with the standard Lasso, Elastic Net, SCAD, and MC+ (Zou and Hastie, 2005; Fan and Li, 2001; Zhang, 2010). The second simulation illustrates a situation where the generalized preconditioner improves upon the Puffer transformation.

4.1. Preconditioning with F

After preconditioning, the noise vector $F\epsilon$ contains statistically dependent terms that are no longer exchangeable. This complicates many of the standard methods of tuning parameter selection (e.g. CV, AIC, BIC). We use the following OLS-BIC procedure. Appendix D gives an additional simulation that ensures this procedure does not differentially favor the preconditioned Lasso.

OLS-BIC; To choose a model in a path of models Starting from the null model, select the first model along the solution path with nz nonzero elements, for $nz = 1, \ldots, 40$. For each value of nz, use the selected nz features to fit an OLS model with the un-preconditioned data. Compute the BIC for the resulting OLS model. Finally, select the tuning parameter that corresponds to the model with the lowest OLS-BIC score. The OLS models were fit with the R function `lm` and the BIC was computed with the R function `BIC`.

In this simulation, $n = 250$, $s = 20$, and p grows along the horizontal axis of the figures (from $2^5 = 32$ to $2^{15} = 32{,}768$). All nonzero elements in β^* equal three and $\sigma^2 = 1$. The rows of $\mathbf{X}$ are mean zero Gaussian vectors with constant correlation ρ. In the top row of plots in Figure 4 and 5, $\rho = .1$. In the middle and bottom rows, $\rho = .5$ and $.85$ respectively.

The first column of plots in Figure 4 corresponds to the number of false negatives. The second column corresponds to the number of false positives. Figure 5 plots the ℓ_2 error $\|\hat{\beta}(\lambda) - \beta^*\|_2$ on the right. Each data point in every plot comes from an average of ten simulation runs.

In many settings, across both p and ρ, the preconditioned Lasso simultaneously admits fewer false positives and fewer false negatives than the competing methods. The number of false negatives when $\rho = .85$ (displayed in the bottom left plot of Figure 4) gives the starkest example. In particular, as the correlation increases or the number of predictors grows, the preconditioned Lasso has the best relative performance. When $p \approx n = 250$, the preconditioned Lasso performs poorly; this is because the singular values of $\mathbf{X}$ follow the Marchenko-Pastur law and when $p \approx n$, this distribution has mass around zero. As a result, F has large spectral norm leading to excessive noise. This would be an appropriate regime to explore the use of a generalized preconditioner.

All simulations in this section were deployed in R with the packages LARS (for the Lasso), PLUS (for SCAD and MC+), and `glmnet` (for the elastic net) (Efron et al., 2004; Zhang, 2010; Friedman et al., 2010).

4.2. Bounded preconditioning

This simulation compares the Puffer transformation to the generalized preconditioner $F_{g,.1}$ with $g(x, \tau) = \min(1, \tau^{-1}x)$. Note that $\|F_{g,\tau}\| \le \tau^{-1}$, where $\|\cdot\|$ is the spectral norm.

The design matrix is simulated as $\mathbf{X}_{ij} = (G_i/\alpha)Z_{ij}$, where Z_{ij} are iid $N(0,1)$ random variables and the G_i are independent Gamma random variables with shape α and rate one. The Gamma random variables make the rows of X have heterogeneous lengths. The horizontal axis of Figure 6 represents the standard deviation of (G_i/α). As $\alpha \to \infty$, G_i/α concentrates around one. So, large values of α are on the left. As $\alpha \to 0$, the standard deviation of G_i/α grows; these values are plotted on the right.

The top plot in Figure 6 shows that $IC(\mathbf{X})$ quickly surpasses the critical threshold of one. As such, $\mathbf{X}$ is much less likely to satisfy the irrepresentable condition when the standard deviation of G_i/α is large. The middle plot in Figure 6 shows that as the standard deviation of row length increases, the Puffer

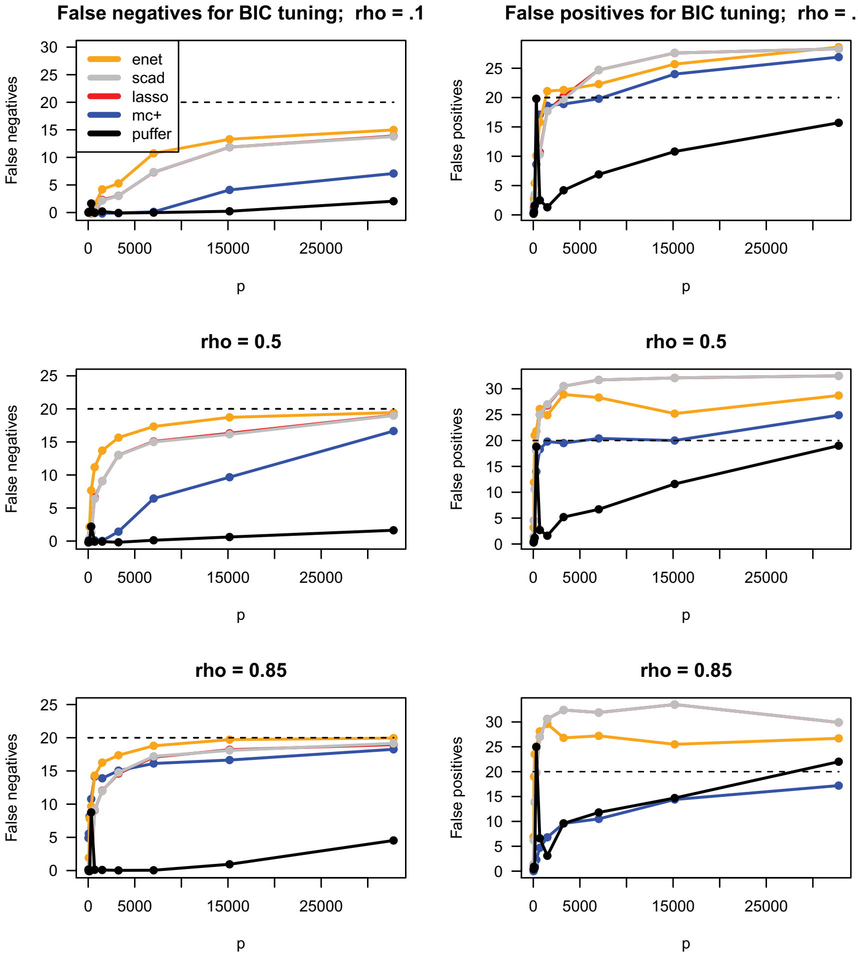

FIG 4. *From top to bottom, the correlation in* $\mathbf{X}$ *increases. The results from the preconditioned Lasso appear as a solid black line. Note that the number of false negatives cannot exceed* $s = 20$. *In the plots on the left side, a dashed horizontal line at 20 represents this limit. For scale, this dashed line is also included in the false positive plots. For both* $\rho = .5$ *and .85, the competing methods miss a significant fraction of the true nonzero coefficients.*

transformation drastically reduces the signal to noise ratio, where SNR_{dB} is defined as

$$SNR_{dB}(\mathbf{X}\beta^*, e) = \log_{10} \frac{\|\mathbf{X}\beta^*\|_2}{\|e\|_2}. \tag{9}$$

After preconditioning with F, the SNR_{dB} becomes $SNR_{dB}(F\mathbf{X}\beta^*, Fe)$.

The top plot in Figure 6 shows that $F_{g,\tau}$ retains many of the advantages of preconditioning by drastically expanding the region of design matrices that can satisfy the irrepresentable condition. At the same time, it drastically increases

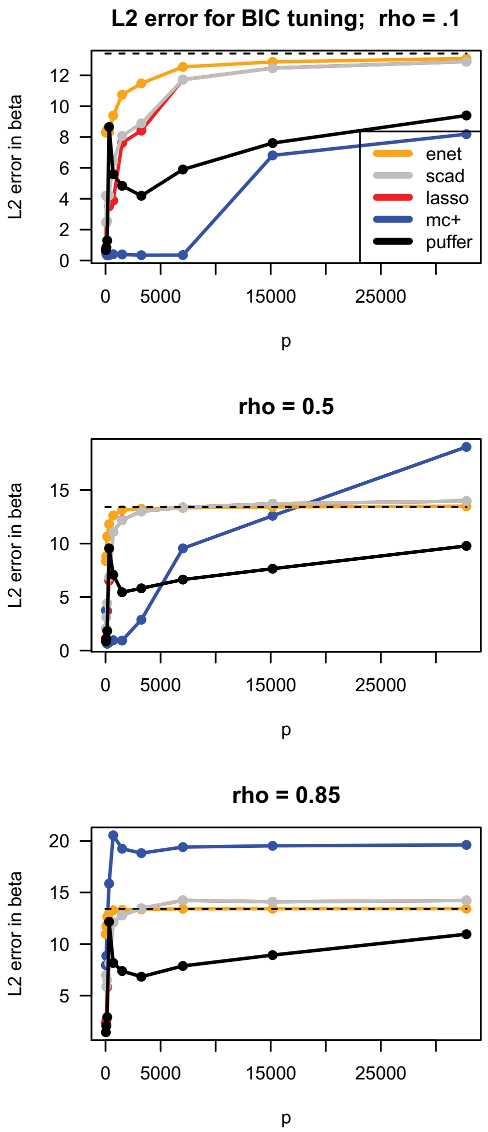

Fig 5. *From top to bottom, the correlation in* $\mathbf{X}$ *increases. The results from the preconditioned Lasso appear as a solid black line. The dashed line corresponds to the* ℓ_2 *error for the estimate* $\hat{\beta} = 0$.

the signal to noise ratio (compared to the Puffer transformation). As a result, $F_{g,\tau}$ a yields better sign estimator than both the standard Lasso and the Puffer preconditioned Lasso (bottom plot). In all simulations in Figure 6, there are $s = 10$ nonzero elements in β^* and each nonzero element is 30. The error terms are iid $N(0,1)$, $n = 200$, and $p = 1000$. The tuning parameter is $\tau = .05\sqrt{p}$.

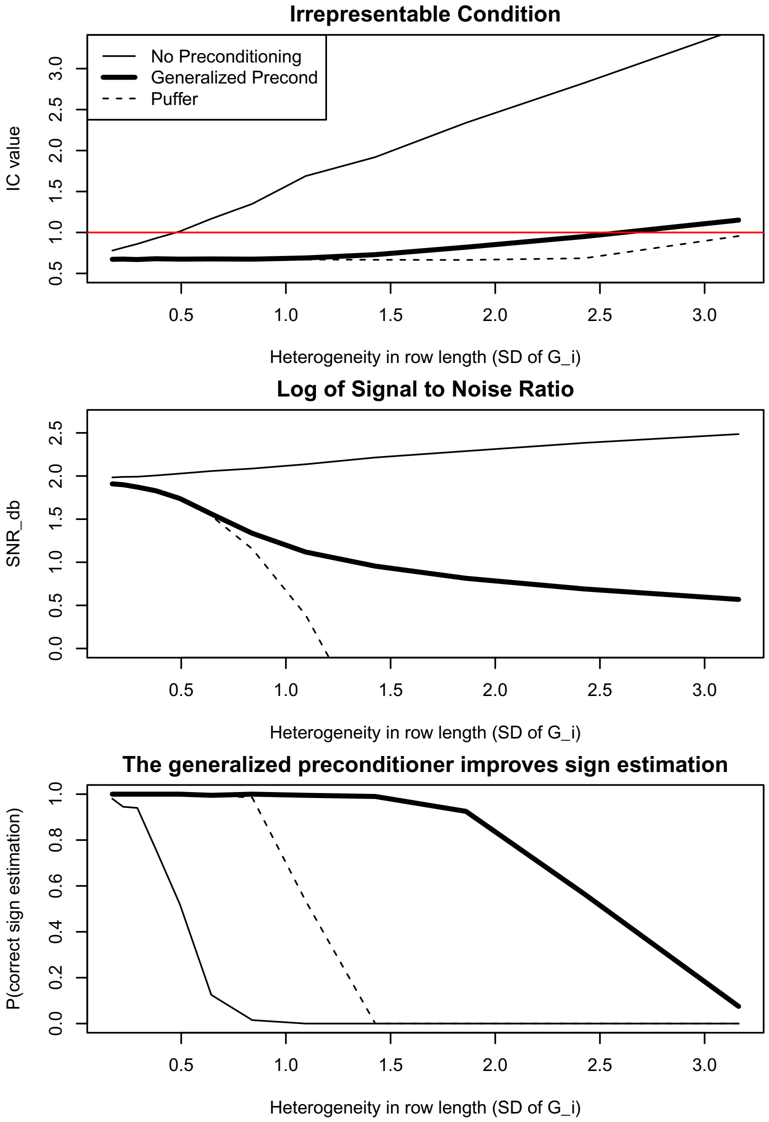

Fig 6. *The horizontal axis controls the amount of heterogeneity in the row lengths of* $\mathbf{X}$*. As this increases, the irrepresentable condition evaluated with* $\mathbf{X}$ *quickly fails by surpassing the red line. Simultaneously, the signal to noise ratio (defined in equation 9) for* $F\mathbf{X}$ *converges to* $-\infty$ *because the spectrum of* $\mathbf{X}$ *decays faster as the row heterogeneity increases. The generalized preconditioner* $F_{g,\tau}$ *balances these trade-offs and improves sign estimation.*

4.3. Comparing Lasso preconditioners

This simulation compares four different preconditioning methods, investigating their ability to (1) satisfy the irrepresentable condition and (2) select the correct model. In addition to the Puffer Transformation, this simulation investigates the following three techniques:

- **Row Normalization.** This preconditioner is a diagonal matrix D, with D_{ii} equal to the ℓ_2 length of the ith row of of $\mathbf{X}$. Preconditioning with D creates a design matrix with equal row lengths.
- **Correlation Sifting.** Huang and Jojic (2011) suggests a preconditioner that projects $\mathbf{X}$ and Y onto the $n - K$ *smallest* principal components of $\mathbf{X}$. To define the preconditioner, take the SVD $\mathbf{X} = UDV^T$ and define $U_A \in R^{n \times n-k}$ to contain the $n - K$ smallest left singular vectors of $\mathbf{X}$. The preconditioner is $U_A U_A^T$.
- **Latent Model.** To estimate a Gaussian latent variable model, Paul et al. (2008) propose the following preconditioning technique: First, identify the q columns of $\mathbf{X}$ that are maximally correlated with Y and place these columns into a matrix $\mathbf{X}_S$. Then, project Y onto the K *largest* principal components of $\mathbf{X}_S$. In this routine, $\mathbf{X}$ is not preconditioned.

Latent Model preconditioning differs from the others in two important respects. First, it only preconditions Y. So, it does not alter the IC value of the design matrix (see Equation 7). Second, both **Correlation Sifting** and the Puffer transformation remove the effect of the largest singular vectors in $\mathbf{X}$. Meanwhile, **Latent Model** emphasizes these directions.

The simulation in Figure 7 samples each row of $\mathbf{X} \in R^{300 \times p}$ independently from a multivariate normal distribution, with mean zero and covariance matrix

$$\Sigma = (1 - \rho)I + \rho \mathbf{1}\mathbf{1}^T.$$

The value of ρ is represented in the horizontal axis in each of the four plots. The first simulation is "very-sparse," with $p = 10{,}000$ columns in $\mathbf{X}$ and $q = 10$ nonzero elements in β^*. The second simulation is "semi-sparse," with $p = 500$ and $q = 50$. Here, we use q to denote both the true number of nonzeros in β^* and also the number of variables screened for the **Latent Model** preconditioner. The nonzero elements of β^* are all 10 and the noise variance is 1. Because Σ is a rank one perturbation of the identity, this simulation uses $K = 1$ for both **Correlation Sifting** and **Latent Model** preconditioning.

The top left panel in Figure 7 displays the IC values (7) under the very-sparse setting; recall that $F\mathbf{X}$ satisfies the irrepresentable condition when $IC(F\mathbf{X}) < 1$. In this plot, as the correlation increases from 0 to .09, the IC values of both **Correlation Sifting** and Puffer preconditioning are unchanged. In fact, both techniques are insensitive to ρ all the way through $\rho = .95$ (not shown). However, as ρ exceeds .08, $\mathbf{X}$ begins to fail the irrepresentable condition. In the top plots, **Latent Model** is over plotted by "no preconditioning" because it does not precondition $\mathbf{X}$. The lower left plot shows that both **Correlation Sifting** and Puffer preconditioning estimate the correct model for $\rho \in [0, .09]$. They select the correct model all the way through $\rho = .95$ (not shown). The lower left plot shows that for all levels of $\rho \in [0, .09]$, **Latent Model** preconditioning has worse model selection performance than the Lasso without any preconditioning.

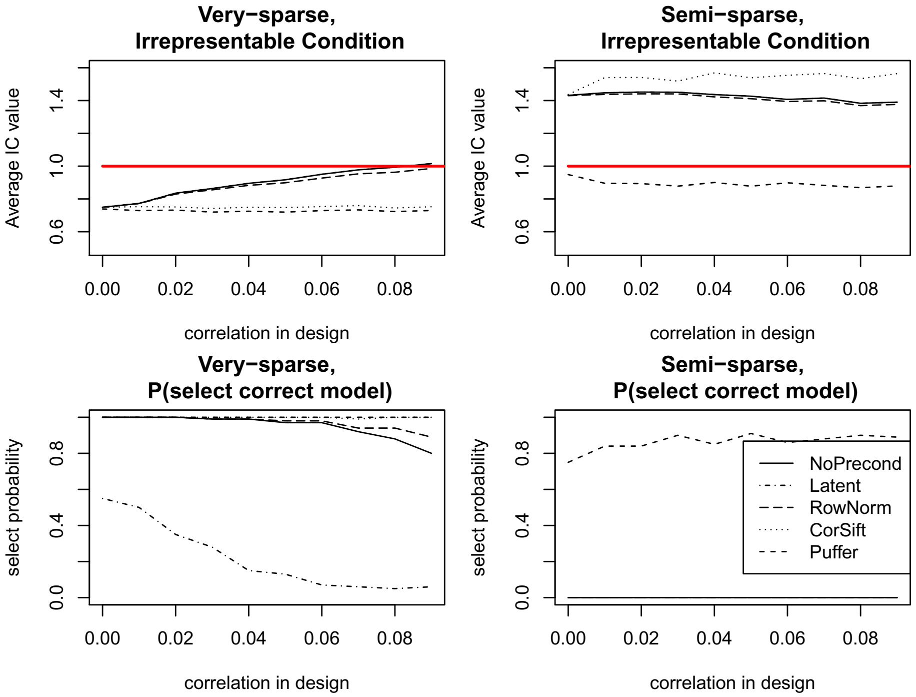

FIG 7. *In both the very- and semi-sparse settings, Puffer preconditioning creates a well conditioned design matrix that allows for good model selection performance.*

The top right panel in Figure 7 shows the conditioning performance under the semi-sparse setting. In this setting, Puffer preconditioning is again insensitive to ρ and satisfies the irrepresentable condition for all values of $\rho \in [0, .09]$. This performance translates into far superior model selection performance (shown in the bottom right panel). These simulations were created with ρ taking the values $0, .01, .02, \ldots, .09$. For each of these values, both $\mathbf{X}$ and Y were sampled 100 different times. The lines connect the average of 100 points. A technique is deemed to "select the correct model" if there exists a value of λ such that $\hat{\beta}(\lambda)$ has the same support as β^*.

5. Discussion

This paper shows that preconditioning has the potential to circumvent the irrepresentable condition in several sparse regression settings. This means that a preprocessing step can make the Lasso, and several other methods, sign consistent with fewer restrictions on the design matrix. Furthermore, this preprocessing step is easy to implement and it is motivated by a wide body of research in numerical linear algebra. The preconditioning described in this paper left multiplies the design matrix $\mathbf{X}$ and the response Y by a matrix $F = UD^{-1}U^T$, where

U and D are derived from the SVD of $\mathbf{X} = UDV^T$. This preprocessing step makes the columns of the design matrix less correlated; while the original design matrix $\mathbf{X}$ might fail the irrepresentable condition, the new design matrix $F\mathbf{X}$ can satisfy it. In low dimensions, the Puffer transformation, ensures that the design matrix always satisfies the irrepresentable condition. In high dimensions, the Puffer transformation projects the design matrix onto the Stiefel manifold, and Theorem 2 shows that in the high dimensional asymptote, most matrices on the Stiefel manifold satisfy the irrepresentable condition. Section 3.3 introduces the generalized Puffer transformation. Theorem 3 proves that one type of generalized Puffer transformation makes the Lasso sign consistent under drastically reduced assumptions on the singular values of $\mathbf{X}$.

In our simulation settings, the Puffer transformation drastically improves the Lasso's estimation performance, particularly in high dimensions. This opens the door to several other important questions (theoretical, methodological, and applied) on how preconditioning can aid sparse high dimensional inference. For example, can preconditioning be formulated in a way that it both whitens the design matrix similarly to the Puffer transformation and also allows for fast computation?

This is the first paper to demonstrate how preconditioning the standard linear regression equation can circumvent the irrepresentable condition. This represents a computationally straightforward fix for the Lasso inspired by an extensive numerical linear algebra literature. The algorithm easily extends to high dimensions and, in our simulations, demonstrates a selection advantage and improved ℓ_2 performance over previous techniques in very high dimensions.

Supplementary Material

Supplement to "Preconditioning the Lasso for sign consistency"
(doi: 10.1214/15-EJS1029SUPP; .pdf).

References

Adler, D., Nenadic, O., and Zucchini, W., Rgl: A r-library for 3d visualization with opengl. In *Proceedings of the 35th Symposium of the Interface: Computing Science and Statistics, Salt Lake City*, volume 35, 2003.

Alter, O., Brown, P. O., and Botstein, D., Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106, 2000.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B., Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. MR2533469

Bradic, J., Efficient support recovery via weighted maximum-contrast subagging. arXiv:1306.3494, 2013.

Bühlmann, P. and van de Geer, S., *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer-Verlag, New York Incorporated, 2011. MR2807761

Candes, E. and Romberg, J., Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007. MR2329927

Chatterjee, A. and Lahiri, S. N., Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011. MR2847974

Chikuse, Y., *Statistics on Special Manifolds*, volume 174. Springer Verlag, 2003. MR1960435

Davidson, K. R. and Szarek, S. J., Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–366. Elsevier, Amsterdan, NL, 2001. MR1863696

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. MR2060166

Fan, J. and Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. MR1946581

Fan, J. and Lv, J., Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008. MR2530322

Fan, K. and Hoffman, A. J., Some metric inequalities in the space of matrices. In *Proc. Amer. Math. Soc*, volume 6, pages 1–116, 1955. MR0067841

Friedman, J., Hastie, T., and Tibshirani, R., Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Huang, J. C. and Jojic, N., Variable selection through correlation sifting. In *Research in Computational Molecular Biology*, pages 106–123. Springer, 2011.

Jia, J. and Rohe, K., Supplement to "Preconditioning the Lasso for sign consistency". DOI: 10.1214/15-EJS1029SUPP, 2015.

Meinshausen, N. and Bühlmann, P., High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. MR2278363

Paul, D., Bair, E., Hastie, T., and Tibshirani, R., "Preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, 36(4):1595–1618, 2008. MR2435449

Qian, J. and Jia, J., On pattern recovery of the fused lasso. arXiv:1211.5194, 2012.

Rauhut, H. and Ward, R., Sparse recovery for spherical harmonic expansions. arXiv:1102.4097, 2011.

Rudelson, M. and Vershynin, R., Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In *40th Annual Conference on Information Sciences and Systems, 2006*, pages 207–212. IEEE, 2006.

Tibshirani, R., Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. MR1379242 van de Geer, S., High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008. MR2396809

van de Geer, S., Bühlmann, P., and Zhou, S., The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics*, 5:688–749, 2011. MR2820636

Wauthier, F. L., Jojic, N., and Jordan, M., A comparative framework for preconditioned lasso algorithms. In *Advances in Neural Information Processing Systems*, pages 1061–1069, 2013.

Xiong, S., Dai, B., and Qian, P. Z. G., Orthogonalizing penalized regression. arXiv:1108.0185, 2011.

Yang, F., Doksum, K., and Tsui, K.-W., Principal component analysis for high dimensional data. PCA is dead. Long live PCA. In *Proc. for Workshop on Persp. on High Dim. Data Analysis II*, Montreal, 2014. MR3289747

Zhang, C. H., Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. MR2604701

Zhao, P. and Yu, B., On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(2):2541, 2006. MR2274449

Zou, H., The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. MR2279469

Zou, H. and Hastie, T., Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. MR2137327