PRECONDITIONING TO COMPLY WITH THE IRREPRESENTABLE CONDITION

By Jinzhu Jia*†

Peking University

AND

By Karl Rohe*†

University of Wisconsin–Madison

Preconditioning is a technique from numerical linear algebra that can accelerate algorithms to solve systems of equations. In this paper, we demonstrate how preconditioning can circumvent a stringent assumption for sign consistency in sparse linear regression. Given $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^n$ that satisfy the standard regression equation, this paper demonstrates that even if the design matrix $\mathbf{X}$ does not satisfy the irrepresentable condition for the Lasso, the design matrix $F\mathbf{X}$ often does, where $F \in \mathbb{R}^{n\times n}$ is a preconditioning matrix defined in this paper. By computing the Lasso on $(F\mathbf{X}, FY)$, instead of on $(\mathbf{X}, Y)$, the necessary assumptions on $\mathbf{X}$ become much less stringent.

Our preconditioner F ensures that the singular values of the design matrix are either zero or one. When $n \geq p$, the columns of $F\mathbf{X}$ are orthogonal and the preconditioner always circumvents the stringent assumptions. When $p \geq n$, F projects the design matrix onto the Stiefel manifold; the rows of $F\mathbf{X}$ are orthogonal. We give both theoretical results and simulation results to show that, in the high dimensional case, the preconditioner helps to circumvent the stringent assumptions, improving the statistical performance of a broad class of model selection techniques in linear regression. Simulation results are particularly promising.

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 10's or 100's of thousands

*The authors contributed equally and are listed in alphabetical order.

†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 2.1. Thank you to Sara Fernandes-Taylor for helpful comments.

Keywords and phrases: Preconditioning, Lasso, Sign consistency

of different pieces of information for each individual. Classical statistical inference in such "high-dimensional" or $p \gg n$ regimes is often impossible. Successful experiments must rely on some type of sparsity or low-dimensional structure. Several statistical techniques have been developed to exploit various types of structure, including sparse high dimensional regression.

Sparse high dimensional regression aims to select the few measurements (among the 10's of thousands) that relate to an outcome of interest; these techniques can screen out the irrelevant variables. A rich theoretical literature describes the consistency of various sparse high dimensional regression techniques, highlighting several potential pitfalls (e.g. Knight and Fu [2000], Fan and Li [2001], Greenshtein and Ritov [2004], Donoho et al. [2006], Meinshausen and Bühlmann [2006], Tropp [2006], Zhao and Yu [2006], Zou [2006], Zhang and Huang [2008], Fan and Lv [2008], Wainwright [2009], Meinshausen and Yu [2009], Bickel et al. [2009], Zhang [2010], Shao and Deng [2012]). In this literature, one of the most popular measures of asymptotic performance is sign consistency, which implies that the estimator selects the correct set of predictors asymptotically. One of the most popular methods in sparse regression, the Lasso (defined in Section 1.1), requires a stringent "irrepresentable condition" to achieve sign consistency [Tibshirani, 1996, Zhao and Yu, 2006]. The irrepresentable condition restricts the correlation between the columns of the design matrix in a way made explicit in Section 1.1.

It is well known that the Ordinary Least Squares (OLS) estimator performs poorly when the columns of the design matrix are highly correlated. However, this problem can be overcome by more samples; 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 not be sign consistent (i.e. statistical estimation will not improve with more samples).

To avoid the irrepresentable condition, several researchers have proposed alternative penalized least square methods that use a different penalty from the Lasso penalty. For example, Fan and Li [2001] propose SCAD, a concave penalty function; Zhang [2010] proposes the minimax concave penalty (MCP), another concave penalty function, and gives high probability results for PLUS, an optimization algorithm. Unfortunately, these concave penalties lead to nonconvex optimization problems. Although there are algorithmic approximations for these problems and some high probability results [Zhang, 2010], the estimator that these algorithms compute is not necessarily the estimator that optimizes the penalized least squares objective. The Adaptive Lasso provides another alternative penalty which is a data adaptive and heterogeneous [Zou, 2006]. Unfortunately, its statistical performance degrades in high dimensions.

In penalized least squares, there is both a penalty and a data fidelity term (which makes the estimator conform to the data). The papers cited in the previous paragraph adjust the type of sparse penalty. In this paper, we precondition the data, which is equivalent to adjusting the data fidelity term. Other researchers have previously proposed alternative ways of measuring data fidelity [Van De Geer, 2008], but their alternatives are meant

to accommodate different error distributions, not to avoid the irrepresentable condition. Similar to work presented here, Xiong et al. [2011] also propose adjusting the data fidelity term to avoid the irrepresentable condition. They proposed a procedures which (1) makes the design matrix orthogonal by adding rows, and (2) applies an EM algorithm, with SCAD, to estimate the outcomes corresponding to the additional rows in the design matrix. Although this algorithm performs well in the low dimensional case, it is computationally expensive in high dimensional problems. The procedure proposed in this paper adjusts the data fidelity term by preconditioning, a preprocessing step. Relative to these alternative methods, preconditioning is easier to implement, requiring only a couple lines of code before calling any standard Lasso package. Furthermore, this type of preprocessing is widely studied in a related field, numerical linear algebra.

Preconditioning describes a popular suite of techniques in numerical linear algebra that stabilize and accelerate algorithms to solve systems of equations (e.g. Axelsson [1985], Golub and Van Loan [1996]). In a system of equations, one seeks the vector x that satisfies $Ax = b$, where $A \in \mathbb{R}^{n\times n}$ is a given matrix and $b \in \mathbb{R}^n$ is a given vector. The speed of most solvers is inversely proportional to the condition number of the matrix A, the ratio of its largest eigenvalue over its smallest eigenvalue. When the matrix A has both large eigenvalues and small eigenvalues, the system $Ax = b$ is "ill-conditioned." For example, if the matrix A has highly correlated columns, it will be ill-conditioned. One can "precondition" the problem by left multiplying the system by a matrix T, $TAx = Tb$; the preconditioner T is designed to shrink the condition number of A thereby accelerating the system solver. If the columns of A are highly correlated, then preconditioning decorrelates the columns.

The system of equations $Ax = b$ has many similarities with the linear regression equation

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

where we observe $Y \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$, and $\epsilon \in \mathbb{R}^n$ contains unobserved iid noise terms with $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2 I_n$. From the system of equations $Ax = b$, the regression equation adds an error term and allows for the design matrix to be rectangular. Where numerical linear algebraists use preconditioners for algorithmic speed, this paper shows that preconditioning can circumvent the irrepresentable condition, improving the statistical performance of the Lasso.

Just as preconditioning sidesteps the difficulties presented by correlation in systems of equations, preconditioning can sidestep the difficulties in sparse linear regression. Numerical algebraists precondition systems of linear equations to make algorithms faster and more stable. In this paper, we show that preconditioning the regression equation (Equation 1) can circumvent the irrepresentable condition. For a design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, we study a specific preconditioner $F \in \mathbb{R}^{n\times n}$ that is defined from the singular value decomposition of $\mathbf{X} = UDV'$. We call $F = UD^{-1}U'$ the Puffer Transformation because it inflates the smallest nonsingular values of the design matrix (Section 2.1 discusses this in more detail). This paper demonstrates why the matrix $F\mathbf{X}$ can satisfy the irrepresentable condition, while the matrix $\mathbf{X}$ may not; in essence, the preconditioner makes the columns of $\mathbf{X}$ less

correlated. When $n \geq p$, the columns of $F\mathbf{X}$ are exactly orthogonal, trivially satisfying the irrepresentable condition. When $n < p$ and the columns of $\mathbf{X}$ are moderately or highly correlated, F can greatly reduce the the pairwise correlations between columns, making the design matrix $\mathbf{X}$ more amenable to the irrepresentable condition.

In a paper titled *"Preconditioning" for feature selection and regression in high-dimensional problems*, the authors propose projecting the outcome Y onto the top singular vectors of $\mathbf{X}$ before running the Lasso [Paul et al., 2008]. They leave $\mathbf{X}$ unchanged. The current paper preconditions the entire regression equation, which reduces the impact of the top singular vectors of $\mathbf{X}$, and thus reduces the correlation between the columns of $\mathbf{X}$. Whereas the preprocessing step in Paul et al. [2008] performs noise reduction, the Puffer Transform makes the design matrix conform to the irrepresentable condition.

The outline of the paper is as follows: Section 1.1 gives the necessary mathematical notation and definitions. Section 2 introduces the Puffer Transformation and gives a geometrical interpretation of the transformation. Section 3 discusses the low dimensional setting ($p \leq n$), where the Puffer Transformation makes the columns of $F\mathbf{X}$ orthogonal. Theorem 1 gives sufficient conditions for the sign consistency of the preconditioned Lasso when $p \leq n$; these sufficient conditions do not include an irrepresentable condition. Section 4 discusses the high dimensional setting ($p > n$), where the Puffer Transformation projects the design matrix onto the Stiefel manifold. Theorem 2 shows that most matrices on the Stiefel manifold satisfy the irrepresentable condition. Theorem 3 gives sufficient conditions for the sign consistency of the preconditioned Lasso in high dimensions. This theorem includes an irrepresentable condition. Section 5 shows promising simulations that compare the preconditioned Lasso to several other (un-preconditioned) methods. Section 6 describes four data analysis techniques that incidentally precondition the design matrix with a potentially harmful preconditioner; just as a good preconditioner can improve estimation performance, a bad preconditioner can severely detract from performance. Users should be cautious when using the four techniques described in Section 6. Section 7 concludes the paper.

1.1. *Preliminaries.* To define the Lasso estimator, suppose the observed data are independent pairs $\{(x_i, Y_i)\} \in \mathbb{R}^p \times \mathbb{R}$ for $i = 1, 2, \ldots, n$ following the linear regression model

$$Y_i = x_i^T \beta^* + \epsilon_i, \tag{2}$$

where x_i^T is a row vector representing the predictors for the ith observation, Y_i is the corresponding ith response variable, and ϵ_i's are independent, mean zero noise terms with variance σ^2. The unobserved coefficients are $\beta^* \in \mathbb{R}^p$. Use $\mathbf{X} \in \mathbb{R}^{n \times p}$ to denote the $n \times p$ design matrix with $x_k^T = (\mathbf{X}_{k1}, \ldots, \mathbf{X}_{kp})$ as its kth row and with $X_j = (\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn})^T$ as

its jth column, then

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = (X_1, X_2, \ldots, X_p).$$

Let $Y = (Y_1, \ldots, Y_n)^T$ and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T \in \mathbb{R}^n$.

For penalty function $pen(b) : \mathbb{R}^p \to \mathbb{R}$, define the penalized least squares objective function,

$$\ell(b, pen, \lambda) = \frac{1}{2}\|Y - \mathbf{X}b\|_2^2 + pen(b, \lambda). \tag{3}$$

The Lasso estimator uses the ℓ_1 penalty,

$$\hat{\beta}(\lambda) = \arg\min_b \frac{1}{2}\|Y - \mathbf{X}b\|_2^2 + \lambda \, \|b\|_1. \tag{4}$$

where for some vector $b \in \mathbb{R}^p$, $\|b\|_r = (\sum_{i=1}^k |x_i|^r)^{1/r}$.

The popularity of the Lasso (and other sparse penalized least squares methods) stems from the fact that, for large enough values of λ, the estimated coefficient vectors contain several zeros. If one is willing to assume the linear regression model, then the Lasso estimates which columns in $\mathbf{X}$ are conditionally independent of Y given the other columns in $\mathbf{X}$.

1.1.1. *Sign consistency and the irrepresentable condition.* For $T \subset \{1, \ldots, p\}$ 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 $x(T) = (x_j)_{j \in T}$. $S \subset \{1, \ldots, p\}$, the support of β^*, is defined

$$S = \{j : \beta_j^* \neq 0\}.$$

Define $s = |S|$. In order to define sign consistency, define

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

and for a vector b, sign(b) is defined as a vector with the ith element sign(b_i).

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

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

In other words, $\hat{\beta}(\lambda)$ 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], Yuan and Lin [2007], have studied the sign consistency property and found a sufficient condition for sign consistency. Zhao and Yu [2006] called this assumption the "irrepresentable condition" and showed that it is almost necessary for sign consistency.

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

$$\left\| X(S^c)^T X(S) \left(X(S)^T X(S) \right)^{-1} sign(\beta^*(S)) \right\|_\infty \leq 1 - \eta, \tag{5}$$

where for a vector x, $\|x\|_\infty = \max_i |x_i|$.

In practice, 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, their Corollary 2 shows that if $|\text{cor}(X_i, X_j)| \leq c/(2s-1)$ for a constant $0 \leq c < 1$, then the irrepresentable condition holds. Theorem 2 in Section 4 of this paper relies on their corollary.

2. Preconditioning to circumvent the stringent assumption. We will always assume that the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ has rank $d = \min\{n, p\}$. From 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 $\mathbf{X} = UDV'$. Define the **Puffer Transformation**,

$$F = UD^{-1}U^T. \tag{6}$$

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.

Define $\tilde{Y} = FY$, $\tilde{\mathbf{X}} = F\mathbf{X}$, and $\tilde{\epsilon} = F\epsilon$. After left multiplying the regression equation $Y = \mathbf{X}\beta^* + \epsilon$ by the matrix F, the transformed regression equation becomes

$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\epsilon} \tag{7}$$

If $\epsilon \sim N(0, \sigma^2 I_n)$, then $\tilde{\epsilon} \sim N(0, \tilde{\Sigma})$ where $\tilde{\Sigma} = \sigma^2 UD^{-2}U^T$.

The scale of $\tilde{\Sigma}$ depends on the diagonal matrix D, which contains the d singular values of $\mathbf{X}$. As the singular values of $\mathbf{X}$ approach zero, the corresponding elements of D^{-2} grow very quickly. This increased noise can quickly overwhelm the benefits of a well conditioned design matrix. For this reason, it might be necessary add a Tikhonov regularization term to the diagonal of D. The simulations in Section 5 show that when $p \approx n$, the transformation

can harm estimation. Future research will examine if a Tikhonov regularizer resolves this issue.

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 original problem easier to solve. However, these preconditioners themselves can be time consuming to compute. Successful preconditioners balance these two costs to provide a computational advantage. 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. *Geometrical representation.* The figures in this section display the geometry of the Lasso before and after the Puffer Transformation. These figures (a) demonstrate what happens when the irrepresentable condition is not satisfied, (b) reveal how the Puffer Transformation circumvents the irrepresentable condition, and (c) 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.[1]

$$\hat{\beta}(c) = \arg \min_{b:\|b\|_1 \leq c} \|Y - \mathbf{X}b\|_2^2 \tag{8}$$

Given the constraint set $\|b\|_1 \leq c$ and a continuum of sets $\|Y - Xb\|_2^2 \leq x$ for $x \geq 0$, define

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

When c is small enough, $\mathcal{I}(c, 0)$ is an empty set, implying that there is no b with $\|b\|_1 \leq c$ such that $Y = Xb$. To find $\hat{\beta}(c)$, increase the value of x until $\mathcal{I}(c, x)$ is no longer an empty set. 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 X (e.g. full column rank), the solution is unique and $\hat{\beta}(c) = \mathcal{I}(c, x^*)$.

Figures 1 and 2 below give a graphical representation of this description of the Lasso before and after preconditioning. The constraint set $\{b : \|b\|_1 \leq c\}$ appears as a diamond shaped polyhedron and the level set of the loss function $\{b : \|Y - Xb\|_2^2 < x\}$ appears as an ellipse. Starting from $x = 0$, x increases, dilating the ellipse, until the ellipse intersects the constraint set. The first point of intersection represents the solution to the Lasso. This point $\hat{\beta}(c) \in \mathbb{R}^p$ is the element of the constraint set which minimizes $\|Y - Xb\|_2^2$.

In both Figure 1 and Figure 2, the rows of X are independent Gaussian vectors with mean zero and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0 & .6 \\ 0 & 1 & .6 \\ .6 & .6 & 1 \end{pmatrix}$$

[1]In an abuse of notation, the left hand side of Equation 8 is $\hat{\beta}(c)$. In fact, there is a one-to-one function $\phi(c) = \lambda$ to make the Lagrangian form of the Lasso (Equation 4) equivalent to the constrained form of the Lasso (Equation 8).

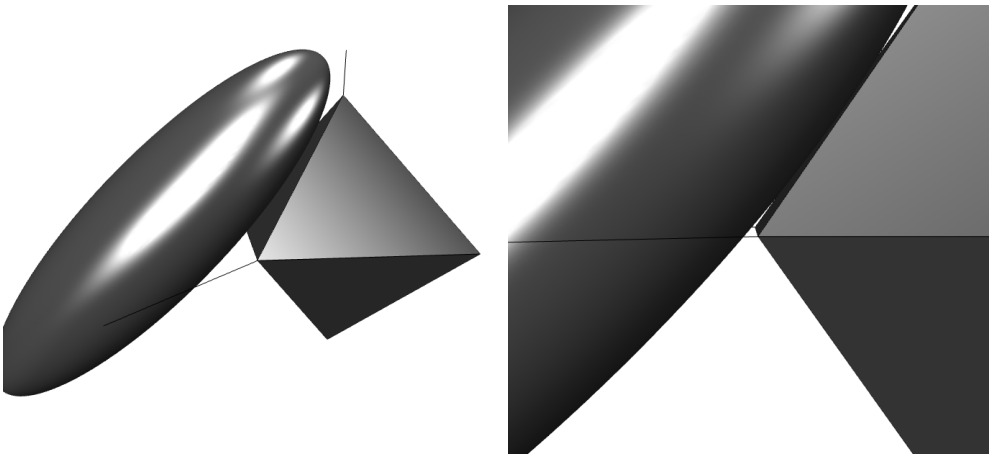

FIG 1. *When the problem is not preconditioned, the elongated ellipse (representing the ℓ_2 loss) intersects the ℓ_1 ball off of the plane created from the axes that point to the left. The Lasso fails to select the true model.*

To highlight the effects of preconditioning, the noise is very small and $n = 10,000$. In Figure 1, the design matrix is not preconditioned. In Figure 2, the problem has been preconditioned, and the ellipse represents the set $\|FY - FXb\|_2^2 \leq x$ for some value of x. The preconditioning turns the ellipse into a sphere.

In both Figure 1 and Figure 2, $\beta^* = (1, 1, 0)$. In both figures, the third dimension is represented by the axis that points up and down. Thus, if the ellipse intersects the constraint set in the (horizontal) plane formed by the first two dimensions, then the Lasso estimates the correct sign. In Figure 1, the design matrix fails the irrepresentable condition and the elongated shape of the ellipse forces $\hat{\beta}(c)$ off of the true plane. This is illustrated in the right panel of Figure 1. High dimensional regression techniques that utilize concave penalties avoid this geometrical misfortune by changing the shape of the constraint set. Where the ℓ_1 ball has a flat surface, the non-convex balls flex inward, dodging any unfortunate protrusions of the ellipse. As preconditioning acts on the opposite set, it restricts the protrusions in the ellipse.

In Figure 2, the elongated direction of the ellipse shrinks down, and the ellipse is puffed out into a sphere; it then satisfies the irrepresentable condition, and $\hat{\beta}(\lambda)$ lies in the true plane. Therefore, in this example, preconditioning circumvents the stringent condition for sign consistency. When $n > p$ preconditioning makes the ellipse a sphere. When $p > n$, preconditioning can drastically reduce the pairwise correlations between columns, thus making low dimensional projections of the ellipse more spherical. Both Figure 1 and Figure 2 were drawn with the R library rgl.

Figure 2 illustrates why the Puffer Transformation is so named. We call F the Puffer Transformation in reference to the pufferfish. In its relaxed state, a pufferfish looks similar

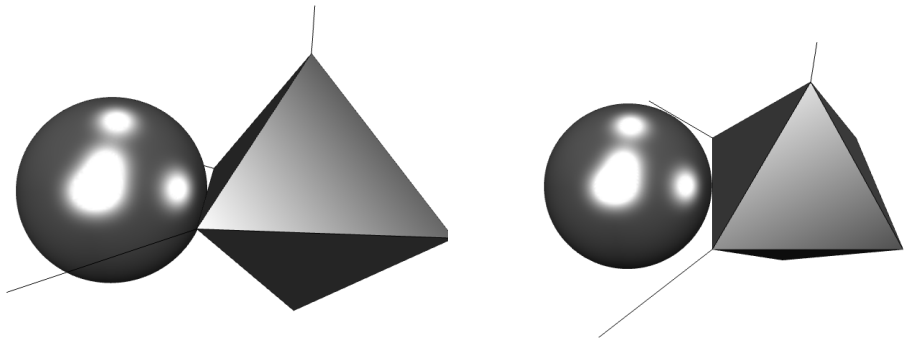

FIG 2. *After preconditioning, the ℓ_2 loss ellipse turns to a sphere. This ensures that it intersects the polyhedron in the plane. After preconditioning, the Lasso correctly selects the true model.*

to most other fish, oval and oblong. However, when a pufferfish senses danger, it defends itself by inflating its body into a sphere. In regression, if the columns of the design matrix are correlated, then the contours of the loss function $\ell_2(b) = \|Y - \mathbf{X}b\|^2$ are oval and oblong as illustrated in Figure 1. If a data analyst has reason to believe that the design matrix might not satisfy the irrepresentable condition, then she could employ the Puffer Transformation to "inflate" the smallest singular values, making the contours of $\ell_{2,F}(b) = \|FY - F\mathbf{X}b\|^2$ spherical. Although these contours are not faithful to the covariance structure of the standard OLS estimator, the spherical contours are more suited to the geometry of the Lasso in many settings.

3. Low dimensional Results. This section demonstrates that for $n \geq p$, after the Puffer Transformation, the Lasso is sign consistent with a minimal assumption on the design matrix $\mathbf{X}$. When $n \geq p$, the preconditioned design matrix is orthonormal.

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

The irrepresentable condition makes a requirement on

$$X(S^c)'X(S)\left(X(S)'X(S)\right)^{-1}$$

Since the Puffer Transformation makes the columns of the design matrix orthonormal, $(FX(S^c))FX(S) = 0$, satisfying irrepresentable condition.

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 X as*

$\mathbf{X} = UDV'$. *Suppose that* $n \geq p$ *and* $\mathbf{X}$ *has rank* p. *We further assume that* $\Lambda_{\min}(\frac{1}{n}X'X) \geq \tilde{C}_{\min} > 0$. *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 with probability greater than*

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

$\tilde{\beta}(\lambda) =_s \beta^*$.

Remarks. 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 (1) $\lambda \to 0$ and (2) $\lambda^2 n \to \infty$ ensures that $\tilde{\beta}(\lambda)$ is sign consistent. One possible choice is $\lambda = \sqrt{\frac{\log n}{n}}$.

From classical linear regression, we know that increasing the correlation between columns of $\mathbf{X}$ increases the variance of the standard OLS estimator; correlated predictors make estimation more difficult. This intuition translates to preconditioning with the Lasso; increasing the correlation between the columns of $\mathbf{X}$ decreases the smallest singular value of $\mathbf{X}$, increasing the variance of the noise terms $cov(F\epsilon) = UD^{-2}U'$, and weakening the bound in Theorem 1. When the columns of $\mathbf{X}$ are correlated, then $\tilde{C}_{\min}$ is small and Theorem 1 gives a smaller probability of sign consistency.

Theorem 1 applies to many more sparse regression methods that use penalized least squares. After preconditioning, the design matrix 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 method admits a closed form solution. If it is also true that all the pen_j's are equivalent and have a cusp at zero (like the Lasso penalty), then the method selects the same sequence of models as the Lasso and correlation screening (i.e. select X_j if $|cor(Y, X_j)| \geq \lambda$) [Fan and Lv, 2008]. For example, both SCAD and MCP satisfy these conditions. If a method selects the same models as the Lasso, and the Lasso is sign consistent, then this method is also sign consistent. Thus, Theorem 1 implies that (1) preconditioned correlation screening, (2) preconditioned SCAD, and (3) preconditioned MCP are sign consistent under some mild conditions (similar to the conditions in Theorem 1). However, in high dimensions, FX is no longer orthogonal. So, the various methods could potentially estimate different models.

4. High dimensional Results. The results for $p > n$ are not as straightforward because the columns of the design matrix cannot be orthogonal. However, the results in this section suggest that for many high dimensional design matrices $\mathbf{X}$, the matrix $F\mathbf{X}$ satisfies the stringent assumptions of the Lasso. In fact, the simulation results in the following section suggest that preconditioning offers dramatic improvements in high dimensions.

Before introducing Theorem 2, 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 of Σ are all .9. The histogram in Figure 3 includes both the distribution of pairwise correlations between the columns of $\mathbf{X}$ and the distribution of pairwise correlations between the columns of $F\mathbf{X}$. Before the transformation, 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. Figure 3 shows this massive reduction in correlation. The histogram has two modes; the left mode corresponds to the distribution of pairwise correlations in $F\mathbf{X}$, and the right mode corresponds to the distribution of correlations in $\mathbf{X}$.

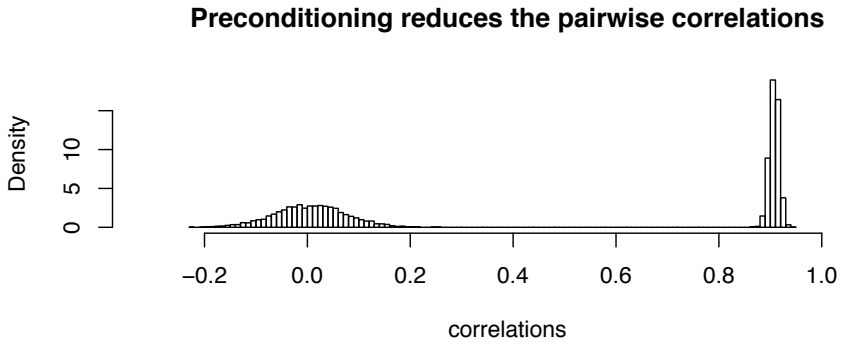

FIG 3. *This figure displays the pairwise correlations between the columns of* $F\mathbf{X}$ *(on the left) and the pairwise correlations between the columns of* $\mathbf{X}$ *(on the far right). In this simulation, the rows of* $\mathbf{X}$ *are iid multivariate Gaussians with mean zero and covariance matrix* Σ*,* $\Sigma_{ii} = 1$ *and* $\Sigma_{i \neq j} = .9$*. The figure was created by first sampling 10,000* (i, j) *pairs without replacement such that* $i \neq j$ *and then computing* $cor(FX_i, FX_j)$ *and* $cor(X_i, X_j)$ *for each of these 10,000 pairs. Before preconditioning, the pairwise correlations are much larger. In this setting, preconditioning reduces the pairwise correlations.*

By reducing the pairwise correlations, preconditioning helps the design matrix satisfy the irrepresentable condition. For example, if the first twenty elements of β^* are positive and the other 9,980 elements equal zero, then the Puffer Transformation makes the design matrix satisfy the irrepresentable condition in this simulation. Recall that the irrepresentable

condition bounds the quantity

$$\left\| X(S^c)^T X(S) \left(X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty < 1 - \eta \tag{9}$$

for $\eta > 0$. From the simulated data displayed in Figure 3, the left hand side of Equation (9) evaluated with the matrix $\mathbf{X}$ equals 1.09; evaluated with $F\mathbf{X}$, it equals .84. In this example, the Puffer Transformation circumvents the irrepresentable condition.

4.1. *Uniform distribution on the Stiefel manifold.* When $n \geq p$, the columns of $F\mathbf{X}$ are orthogonal. When $p \geq n$, the *rows* are orthogonal. $F\mathbf{X}$ lies in the Stiefel manifold,

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

Further, under any unitarily invariant norm, $F\mathbf{X}$ is the projection of $\mathbf{X}$ onto $V(n,p)$ [Fan and Hoffman, 1955]. Since $V(n,p)$ is a bounded set, it has a uniform distribution.

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

$$O(p,\mathbb{R}) = \{O \in \mathbb{R}^{p \times p} : OO' = I_p\}.$$

In this section, Theorem 2 shows that if V is uniformly distributed on $V(n,p)$, then after normalizing the columns of V to have equal length, the matrix satisfies the irrepresentable condition with high probability in certain regimes of n, p, and s. 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 $X \in \mathbb{R}^{n \times p}$ is distributed $uniform(V(n,p))$ and $n > cs^4$, where s is the number of relevant predictors and c is some constant, then asymptotically, the irrepresentable condition holds for normalized version of X with probability no less than*

$$1 - 4p^2 \exp(-n^{1/2}),$$

where the normalized version of X, denoted $\tilde{X}$ is defined as $\tilde{X}_{ij} = X_{ij} / \sqrt{\sum_{i=1}^n X_{ij}^2}$.

The proof of Theorem 2 relies on the fact that if $|\text{cor}(X_i, X_j)| \leq c/(2s-1)$ for all pairs (i,j), then $\mathbf{X}$ satisfies the irrepresentable condition. This line of argument requires that $n > cs^4$. A similar argument applied to a design matrix $\mathbf{X}$ with iid $N(0,1)$ entries only requires that $n > cs^2$ (this is included in Theorem 8 in the appendix Section B.2). If both of these are tight bounds, then it suggests that the irrepresentable condition for $V \sim uniform(V(n,p))$ is potentially more sensitive to the size of s than a design matrix with iid $N(0,1)$ entries. The final simulation in Section 5 suggests that this difference is

potentially an artifact of our proof. Both distributions are almost equally sensitive to s. If anything, $uniform(V(n,p))$ has a slight advantage in our simulation settings.

Propositions 1 and 2 give two models for $\mathbf{X}$ that makes $F\mathbf{X}$ uniformly distributed on the Stiefel manifold.

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

PROOF. Define U, V, D by the SVD of $\mathbf{X}$, $\mathbf{X} = UDV'$. For a fixed $O \in O(p,\mathbb{R})$, $\tilde{V}' = V'O$ is an element of $V(n,p)$. Therefore, the SVD of $\mathbf{X}O = UD\tilde{V}'$. This implies that both $\mathbf{X}$ and $\mathbf{X}O$ have the same Puffer Transformation $F = UD^{-1}U'$. This yields the result when combined with the fact that $\mathbf{X} \stackrel{d}{=} \mathbf{X}O$,

$$F\mathbf{X} \stackrel{d}{=} F\mathbf{X}O.$$

□

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

PROOF. Notice that if $Z \in \mathbb{R}^{n\times p}$ has iid $N(0,1)$ elements and Z is independent of U_Σ, then $\mathbf{X} \stackrel{d}{=} Z\Sigma^{1/2}$. The following equalities in distribution follow from the fact that for any $O \in O(p,\mathbb{R})$, $U_\Sigma \stackrel{d}{=} OU_\Sigma$ and $ZO \stackrel{d}{=} Z$.

$$\begin{aligned} F\mathbf{X} &= (\mathbf{X}\mathbf{X}^T)^{-1/2}\mathbf{X} \\ &\stackrel{d}{=} (Z\Sigma Z^T)^{-1/2}Z\Sigma^{1/2} \\ &\stackrel{d}{=} (ZOU_\Sigma D_\Sigma U_\Sigma^T O^T Z^T)^{-1/2}ZOU_\Sigma D_\Sigma^{1/2}U_\Sigma^T O^T \\ &\stackrel{d}{=} (ZU_\Sigma D_\Sigma U_\Sigma^T Z^T)^{-1/2}ZU_\Sigma D_\Sigma^{1/2}U_\Sigma^T O^T \\ &= F\mathbf{X}O \end{aligned}$$

Thus, $F\mathbf{X} \sim uniform(V(n,p))$. □

Proposition 2 presents a scenario in which the distribution of $F\mathbf{X}$ is independent of the eigenvalues of Σ. Of course, if Σ has a large condition number, then the transformation $F = (\mathbf{X}\mathbf{X}^T)^{-1/2}$ could potentially induce excessive variance in the noise.

In practice, one usually gets a single design matrix $\mathbf{X}$ and a single preconditioned matrix $F\mathbf{X} \in V(n,p)$. It might be difficult to argue that $F\mathbf{X} \sim uniform(V(n,p))$. Instead, one should interpret Theorem 2 as saying that nearly all matrices in $V(n,p)$ satisfy the irrepresentable condition.

So far, this section has illustrated how preconditioning circumvents the irrepresentable condition in high dimensions. This next theorem assumes $F\mathbf{X}$ satisfies the irrepresentable condition and studies when the preconditioned Lasso estimator is sign consistent, highlighting when F might induce excessive noise.

THEOREM 3. *Suppose that data $(\mathbf{X}, Y)$ follows the linear model described in Equation (1) with iid normal noises $\epsilon \sim N(0, \sigma^2 I_n)$. Define the singular value decomposition of X as $\mathbf{X} = UDV^T$. Suppose that $p \geq n$. We further assume that $\Lambda_{\min}(\frac{1}{n}X(S)^T X(S)) \geq \tilde{C}_{\min}$ and $\min_i(D_{ii}^2) \geq pd_{\min}$ with constants $\tilde{C}_{\min} > 0$ and $d_{\min} > 0$. For $F = UD^{-1}U^T$, define $\tilde{Y} = FY$ and $\tilde{\mathbf{X}} = F\mathbf{X}$. Define*

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

Under the following three conditions,

1. $\left\|\tilde{\mathbf{X}}(S^c)^T\tilde{\mathbf{X}}(S)\left(\tilde{\mathbf{X}}(S)^T\tilde{\mathbf{X}}(S)\right)^{-1}\overrightarrow{b}\right\|_\infty \leq 1 - \eta,$
2. $\Lambda_{\min}\left(\tilde{\mathbf{X}}(S)^T\tilde{\mathbf{X}}(S)\right) \geq \frac{n}{cp}$
3. $\min_{j\in S}|\beta_j^*| \geq 2\lambda\sqrt{scp/n}$

where c is some constant; we have

$$P\left(\tilde{\beta}(\lambda) =_s \beta^*\right) > 1 - 2p\exp\left\{-\frac{p\lambda^2\eta^2 d_{\min}}{2\sigma^2}\right\}. \tag{10}$$

This theorem explicitly states when preconditioning will lead to sign consistency. Note that $d_{\min}$ in the exponent of Equation (10) corresponds to the amount of additional noise induced by the preconditioning. For $p \gg n$, the assumption $\min_i(D_{ii}^2) \geq pd_{min}$ is often satisfied. For example, it holds if $\mathbf{X} \sim N(0, \Sigma)$ and the eigenvalues of Σ are lower bounded. To see this, define $\mathbf{X} = Z\Sigma^{1/2}$ for $Z \in \mathbb{R}^{n\times p}$ with $Z_{ij} \sim_{i.i.d.} N(0,1)$. Then,

$$D_{ii}^2 \geq \min_{w\in\mathbb{R}^n, \|w\|_2=1}\|w'\mathbf{X}\|_2^2 = \min_{w\in\mathbb{R}^n, \|w\|_2=1}\|w'Z\Sigma^{1/2}\|_2^2 \geq \min_{w\in\mathbb{R}^n, \|w\|_2=1}\|w'Z\|_2^2\lambda_{\min}(\Sigma),$$

where $\lambda_{\min}(\Sigma)$ is the smallest eigenvalue of Σ. With high probability, $\frac{1}{p}\|w'Z\|_2^2$ is bounded below [Davidson and Szarek, 2001]. Thus, $\min_i(D_{ii}^2) \geq pd_{min}$ holds for some constant $d_{\min}$ with high probability.

The enumerated conditions in Theorem 3 correspond to standard assumptions for the Lasso to be sign consistent. Condition (1) is the irrepresentable condition applied to $\tilde{\mathbf{X}}$. Condition (2) ensures first, that the columns in $\tilde{\mathbf{X}}(S)$ are not too correlated and second, that the columns in $\tilde{\mathbf{X}}(S)$ do not become too short since F rescales the lengths of the columns. From Section B.2 and the discussion of the uniform distribution on the Stiefel manifold, most matrices in $V(n,p)$ satisfy condition (1) as long as $s = o(n^{1/4})$ and $p^2 = o(\exp(\sqrt{n}))$.

Similarly, most matrices satisfy condition (2); Theorem 7 in Appendix Section B.2 states that the diagonals of $\tilde{\mathbf{X}}(S)^T\tilde{\mathbf{X}}(S)$ concentrate around n/p and the off-diagonals concentrate around 0. Condition (3) in Theorem 3 ensures that the signal strength does not decay too fast relative to λ. The next corollary chooses a sequence of λ's.

COROLLARY 1. *Under the conditions of Theorem 3, if* $\min_{j\in S}|{\beta^*}_j|$ *is a constant and* $n/(s\log p)\to\infty$, *then setting* $\lambda^2=\sqrt{n\log(p)/(sp^2)}$ *ensures sign consistency.*

Theorem 3 highlights the tradeoff between (a) satisfying the irrepresentable condition and (b) limiting the amount of additional noise created by preconditioning. Interestingly, even if $\mathbf{X}$ satisfies the irrepresentable condition, the Puffer Transformation might still improve sign estimation performance as long as the preconditioner can balance the tradeoff. To illustrate this, presume that both $\tilde{\mathbf{X}}$ and $\mathbf{X}$ satisfy the irrepresentable condition with constants $\tilde{\eta}$ and η respectively. Preconditioning improves the bound in Equation 10 if $\tilde{\eta}^2 d_{\min}>\eta^2$. Alternatively, if $d_{\min}$ is very small, then the Puffer Transformation will induce excess noise and $\tilde{\eta}^2 d_{\min}<\eta^2$, weakening the upper bound.

5. Simulations. The first two simulations in this section compare the model selection performance of the preconditioned Lasso to the standard Lasso, Elastic Net, SCAD, and MC+. The third simulation compares the uniform distribution on $V(n,p)$ to the distribution that places iid Gaussian ensemble distribution, estimating how often these distributions satisfy the irrepresentable condition.

5.1. *Choosing* λ *with BIC.* After preconditioning, the noise vector $F\epsilon$ contains statistically dependent terms that are no longer exchangeable. This creates problems for selecting the tuning parameter in the preconditioned Lasso. For example, if one wants to use cross-validation, then the test set should not be preconditioned. After several attempts, we were unable to find a cross-validation procedure that led to adequate model selection performance. In this section, there are two sets of simulations that correspond to two ways of choosing λ. The first set of simulations choose λ using a BIC procedure. To ensure that the results are not due to the BIC procedure, the second set of simulations choose the λ such that it selects the first model with ten degrees of freedom.

The first set of simulations choose λ with the following procedure.

OLS-BIC; to choose a model in a path of models.
1) For df = 1:40

a) starting from the null model, select the first model along the solution path with df degrees of freedom.
b) use the df model to fit an OLS model. The OLS model is fit with the original (un-preconditioned) data.
c) compute the BIC for the OLS model.

end

2) 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 and that of Section 5.2, $n = 250$, $s = 20$, and p grows along the horizontal axis of the figures (from 32 to 32 768). All nonzero elements in β^* equal ten and $\sigma^2 = 1$. The rows of $\mathbf{X}$ are mean zero Gaussian vectors with constant correlation ρ.

In Figure 4, there are three columns of plots and three rows of plots. The rows correspond to different levels of correlation ρ; .1 in the top row, .5 in the middle row, and .85 on the bottom row. The columns correspond to different measurements of the estimation error; the number of false negatives on the left, the number of false positives in the middle, and the ℓ_2 error $\|\hat{\beta}(\lambda) - \beta^*\|_2$ on the right. Each data point in every figure comes from an average of ten simulation runs. In each run, both $\mathbf{X}$ and Y are resampled.

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. However, the results are not uniformly better. For example, when $p \approx n = 250$, the preconditioned Lasso performs poorly; this behavior around $p \approx n$ appeared in several other simulations that are not reported in this paper. Further, when the design matrix has fewer predictors p and lower correlations ρ, MC+ has smaller ℓ_2 error than the preconditioned Lasso. However, as the correlation increases or the number of predictors grows, the preconditioned Lasso appears to outperform MC+ in ℓ_2 error.

5.2. *Ensuring that OLS-BIC gives a fair comparison.* To ensure that the results in Figure 4 are not an artifact of the OLS-BIC tuning, Figure 5 tunes procedure by taking the first model to contain at least ten predictors. The horizontal axes displays p on the log scale. The vertical axes report the number of false positives divided by ten. Each dot in the figure represents an average of ten runs of the simulation, where each run resamples $\mathbf{X}$ and Y.

Figure 5 shows that as p grows large, the standard Lasso, Elastic Net (enet), SCAD, and MC+ perform in much the same way, whereas the preconditioned Lasso drastically out-

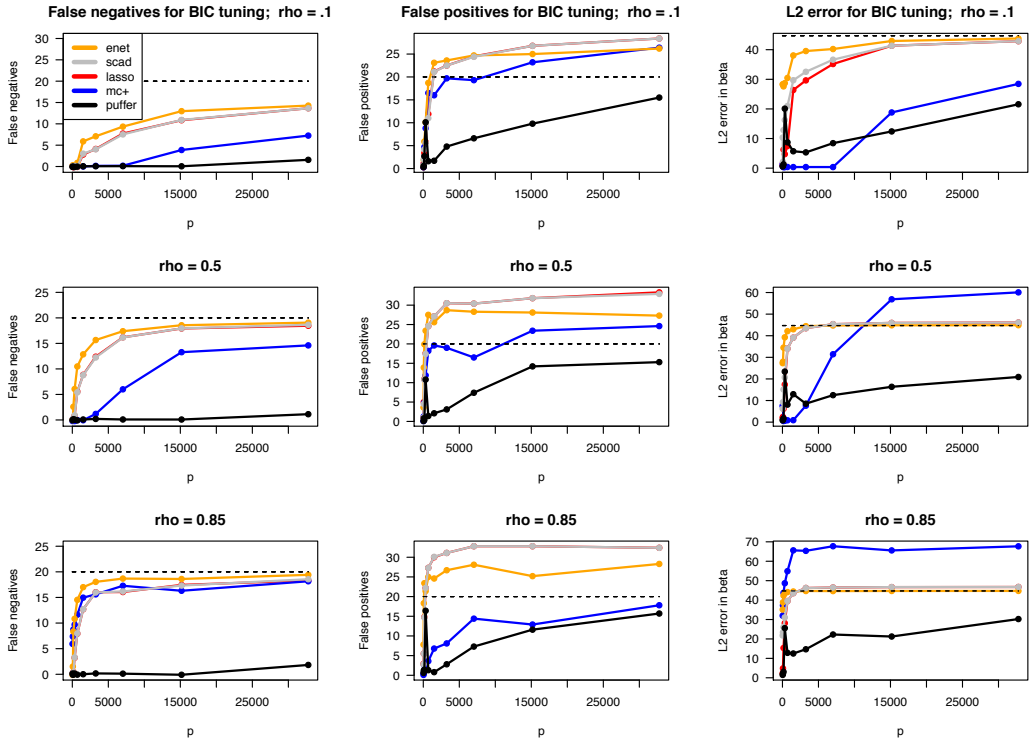

FIG 4. *In all figures, $n = 250$ and p grows along the horizontal axis. Each row of $\mathbf{X}$ is an independent multivariate Gaussian with constant correlation ρ in the off diagonal of the correlation matrix. The figure above has three rows and three columns of plots. The three rows correspond to different values of ρ. From top to bottom, ρ is .1, .5, and .85. The three columns correspond to the number of false negatives (on left), the number of false positives (in middle), and $\|\hat{\beta}(\lambda) - \beta^*\|_2$ (on right). The tuning parameter for each method is selected with the OLS-BIC procedure described in Section 5.1. 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. In the ℓ_2 error plots, the dashed line corresponds to the ℓ_2 error for the estimate $\hat{\beta} = 0$. For both $\rho = .5$ and .85, the competing methods miss a significant fraction of the true nonzero coefficients and only at the very end, for $p > 32,000$, does the preconditioned lasso start to miss any of the true coefficients. At the same time, the preconditioned Lasso accepts fewer false positives than the alternative methods.*

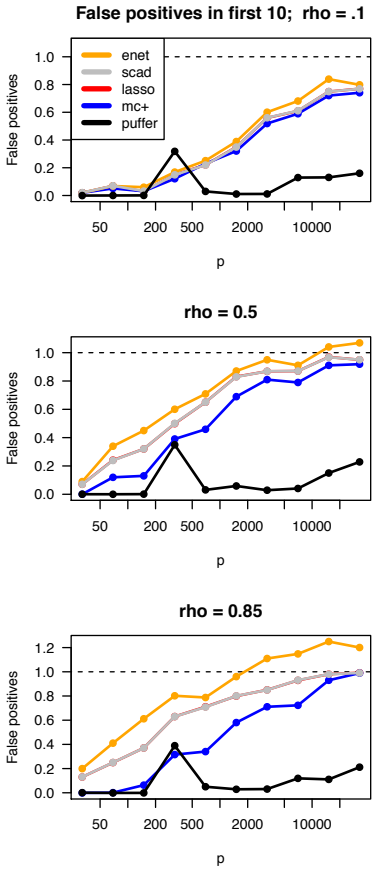

FIG 5. *To ensure that the results in Figure 4 are not due to the choice of tuning parameter, this figure chooses the tuning parameter in a much simpler fashion. Each tuning parameter corresponds to the first estimated model that contains at least ten predictors. Along the horizontal axis, p increases on the log scale. The vertical axis displays the number of false positives divided by ten. The line for the elastic net can exceed one because the default settings of glmnet sometimes fail to select a model with exactly 10 predictors. So, the tuning method selects the next biggest model. Both β^*, and σ^2 are unchanged from the simulations displayed in Figure 4 and* $\mathbf{X}$ *comes from the same distribution. The results in this figure are comparable to the results in Figure 4. This suggests that the results from OLS-BIC provide a fair comparison of the methods.*

performs all of them. For the largest p in the middle and lower graphs, the preconditioned Lasso yields two or three false positives out of ten predictors; all of the other methods have at least eight false positives out of ten predictors. These results suggest that the results in Figure 4 were not an artifact of the OLS-BIC tuning.

The preconditioned Lasso does not perform well when p is close to $n = 250$. This is potentially due to the instability of the spectrum of F when p and n are comparable. Results in random matrix theory suggest that in this regime, $C_{\min}$ (the smallest nonzero singular value of $\mathbf{X}$) can be very close to zero [Silverstein, 1985]. When the spectral norm of the Puffer Transform $(1/C_{\min})$ is large, for example when $p \approx n$, preconditioning can greatly amplify the noise, leading to poor estimation performance. As previously mentioned, it is an area for future research to investigate if a Tikhonov regularization of F might improve the performance in this regime.

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). All packages were run with

their default settings.

5.3. *Satisfying the irrepresentable condition on $V(n,p)$.* Theorem 2 shows that, under certain conditions, if $V \sim uniform(V(n,p))$ then V satisfies the irrepresentable condition with high probability. One of the more undesirable conditions is that it requires $n > cs^4$, where c is a constant and s is the number of relevant predictors. As is discussed after Theorem 2, if the design matrix contains iid $N(0,1)$ entries, then it is only required that $n > cs^2$. The simulation displayed in Figure 6 compares the irrepresentable condition scores:

$$\|IC(X,S) = \mathbf{X}(S^c)'\mathbf{X}(S)\left(\mathbf{X}(S)'\mathbf{X}(S)\right)^{-1}\mathbf{1}_s\|_\infty, \tag{11}$$

where $\mathbf{1}_s \in \mathbb{R}^s$ is a vector of ones, for $\mathbf{X}$ from the iid $N(0,1)$ distribution and for $\mathbf{X} \sim uniform(V(n,p))$. In these simulations, $n = 200$ and $p = 10,000$.

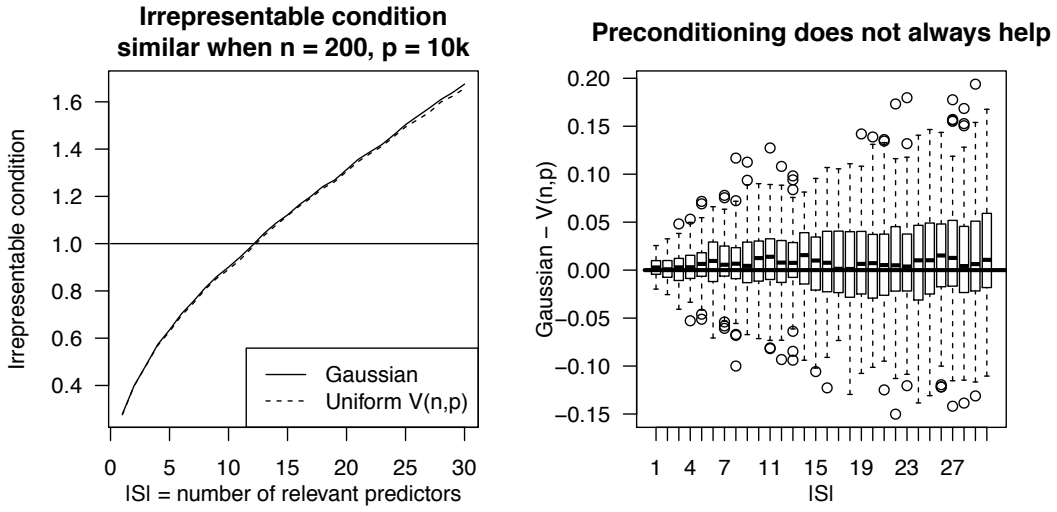

FIG 6. *The left panel displays a Monte Carlo estimate of $E(IC(\mathbf{X},S))$ where IC is defined in Equation 11. The solid line corresponds to random matrices $\mathbf{X}$ with iid $N(0,1)$ entries. The dashed line corresponds to $uniform(V(n,p))$. The lines are nearly identical, suggesting that $uniform(V(n,p))$ design matrices behave similarly to iid Gaussian designs with respect to the irrepresentable condition. The right panel displays the quantities $IC(\mathbf{X}^{(i)},S) - IC(V^{(i)},S)$, where $V^{(i)}$ is the projection of $\mathbf{X}^{(i)}$ onto $V(n,p)$. The variability displayed in the right panel implies that preconditioning only helps for a certain type of design matrix; it does not always help. In this example, the columns of $\mathbf{X}^{(i)}$ are not highly correlated and sometimes preconditioning decreases $IC(\cdot,S)$ and sometimes it increases $IC(\cdot,S)$.*

This simulation sampled 100 matrices $\mathbf{X}^{(i)} \in \mathbb{R}^{200\times 10,000}$with iid $N(0,1)$ entries, with $i = 1,\ldots,100$. Each of these matrices $\mathbf{X}^{(i)}$ were then projected onto $V(n,p)$ with the Puffer Transformation; $V^{(i)} = F^{(i)}\mathbf{X}^{(i)}$, where $F^{(i)}$ is the Puffer Transformation for $\mathbf{X}^{(i)}$. By Proposition 1, $V^{(i)} \sim uniform(V(n,p))$. Before doing any of the remaining calculations, each matrix was centered to have mean zero and scaled to have standard deviation one.

For each of these (centered and scaled) matrices $\mathbf{X}^{(i)}$ and $V^{(i)}$, the values $IC(\cdot, S)$ are computed for $S = \{1, \ldots, q\}$ for $q = \{1, \ldots, 30\}$. (So, the matrices $\mathbf{X}^{(i)}$ and $V^{(i)}$ are recycled 30 times, once for each value of s.) The left panel of Figure 6 plots the average $IC(\cdot, S)$ for each distribution, for each value of s. The solid line corresponds to the $\mathbf{X}^{(i)}$'s, and the dashed line corresponds to the $V^{(i)}$'s. These lines are nearly identical. This suggests that $uniform(V(n,p))$ and the iid Gaussian design perform similarly with respect to the irrepresentable condition.

The right plot in Figure 6 displays the quantities $IC(\mathbf{X}^{(i)}, S) - IC(V^{(i)}, S)$ as a function of s. The boxplots illustrate that, even between the paired samples $\mathbf{X}^{(i)}$ and $V^{(i)}$, there is significant variation in $IC(\cdot, S)$. Further, there are several cases where it is negative, implying

$$IC(\mathbf{X}^{(i)}, S) < IC(V^{(i)}, S).$$

In these cases, preconditioning makes $IC(\cdot, S)$ larger (in addition to make the noise terms dependent). This suggests that only a certain type of data can benefit from preconditioning.

6. Relationship to other data analysis techniques. This final section identifies four statistical techniques that incidentally precondition the design matrix. Just as a good preconditioning matrix can improve the conditioning of $\mathbf{X}$, a bad preconditioning matrix can worsen the conditioning of $\mathbf{X}$. Any processing or preprocessing step which is equivalent to multiplying $\mathbf{X}$ by another matrix could potentially lead to an ill conditioned design matrix. It is an area for future research to assess if/how these methods affect the conditioning and if/how these issues cascade into estimation performance.

Bootstrapping. In the most common form of the bootstrap, the "non-parametric" bootstrap, one samples pairs of observations (Y_i, x_i) with replacement from all n observations [Efron and Tibshirani, 1993]. Each bootstrap sampled data set is equivalent to left multiplying the regression equation by a diagonal matrix with a random multinomial vector down the diagonal. This notion is generalized in the weighted bootstrap [Mason and Newton, 1992], which has been suggested for high dimensional regression [Arlot et al., 2010]. The weighted bootstrap replaces the multinomial vector of random variables with any sequence of exchangeable random variables. In either case, such a diagonal matrix will make the singular values of the design matrix more dispersed, which could lead to potential problems in satisfying the irrepresentable condition. If $\mathbf{X}$ satisfies the irrepresentable condition with a constant η, then a large proportion of the the bootstrap samples of $\mathbf{X}$ might not satisfy the irrepresentable condition. Or, if they satisfy the irrepresentable condition, it might be with a reduced value of η, severely weakening their sign estimation performance. El Karoui [2010] encountered a similar problem when using the bootstrap to study the smallest eigenvalue of the sample covariance matrix.

Generalized least squares. In classical linear regression,

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

with $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $p << n$, if the error terms have expectation zero and covariance $\Sigma_\epsilon = \sigma^2 I$, then the ordinary least squares (OLS) estimator is the best linear unbiased estimator. If the covariance of the noise vector is a known matrix Σ_ϵ which is not proportional to the identity matrix, Aitken [1935] proposed what is commonly called generalized least squares (GLS). In GLS, the regression equation is preconditioned:

$$\Sigma_\epsilon^{-1/2} Y = \Sigma_\epsilon^{-1/2} \mathbf{X} \beta^* + \Sigma_\epsilon^{-1/2} \epsilon.$$

This ensures that the covariance of the noise term is proportional to the identity matrix, $\operatorname{cov}(\Sigma_\epsilon^{-1/2} \epsilon) = I$. Applying OLS to the transformed equation gives a best linear unbiased estimator.

Huang et al. [2010] proposed ℓ_1 penalized GLS for spatial regression. In effect, they preconditioned the regression equation with $\Sigma_\epsilon^{-1/2}$. This is potentially inadvisable. The harm from an ill conditioned design matrix potentially overwhelms the gains provided by uncorrelated error term. Similarly, if a data set has heteroskedastic noise, it is inadvisable to reweight the matrix to make the noise terms homoscedastic. This was observed in Simulation 3 of Jia et al.[2]

Generalized linear models. In a generalized linear model, the outcome Y_i is generated from a distribution in the exponential family and $E(Y_i) = g^{-1}(x_i \beta^*)$ for some link function g [Nelder and Wedderburn, 1972]. In the original Lasso paper, Tibshirani [1996] proposed applying a Lasso penalty to generalized linear models. To estimate these models, Park and Hastie [2007] and Friedman et al. [2010] both proposed iterative algorithms with fundamental similarities to Iteratively Reweighted Least Squares. At every iteration of these algorithms, $\mathbf{X}$ and Y are left multiplied by a different re-weighting matrix. It is an area for future research to examine if the sequence of re-weighting matrices might lead to an ill conditioned design and whether there are additional algorithmic steps that might ameliorate this issue.

Generalized Lasso. Where the standard Lasso penalizes with $\|b\|_1$, the generalized Lasso penalizes $\|Db\|_1$ for some predefined matrix D Tibshirani [2011]. For example, D may exploit some underlying geometry in the data generating mechanism. If D is invertible, then the generalized Lasso problem

$$\min_b \|Y - \mathbf{X} b\|_2^2 + \lambda \|Db\|_1$$

can be transformed into a standard Lasso problem for $\theta = Db$,

$$\min_\theta \|Y - \mathbf{X} D^{-1} \theta\|_2^2 + \lambda \|\theta\|_1.$$

[2]It was the research in Jia et al. that prompted the current paper.

In this problem, the design matrix is right multiplied by D^{-1}. In this paper, we only consider left preconditioning, where the matrix $\mathbf{X}$ is left multiplied by the preconditioner. However, in numerical linear algebra, there is also right preconditioning. Just as a bad left preconditioner can make the design matrix ill conditioned, so can a bad right preconditioner. In practice, the statistician should understand if their matrix D makes their design ill conditioned.

7. Discussion. The information technology revolution has brought on a new type of data, where the number of predictors, or covariates, is far greater than the number of observations. To make statistical inferences in this regime, one needs to assume that the "truth" lies in some low dimensional space. This paper addresses the Lasso for sparse regression. To achieve sign consistency, the Lasso requires a stringent irrepresentable condition. To avoid the irrepresentable condition, Fan and Li [2001], Zou [2006], Zhang [2010] suggest alternative penalty functions, Xiong et al. [2011] proposed an EM algorithm, and others, including Greenshtein and Ritov [2004] and Shao and Deng [2012], have described that alternative forms of consistency that do not require stringent assumptions on the design matrix. In this paper, we show that it is not necessary to rely on nonconvex penalties, nor is it necessary to move to alternative forms of consistency. We show that preconditioning has the potential to circumvent the irrepresentable condition in several settings. This means that a preprocessing step can make the Lasso sign consistent. 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 $F = UD^{-1}U'$, where U and D come from the SVD of $\mathbf{X} = UDV'$. 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, our preconditioner, 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 most matrices on the Stiefel manifold satisfy the irrepresentable condition. In our simulation settings, the Puffer Transformation drastically improves the Lasso's estimation performance, particularly in high dimensions. We believe that this opens the door to several other important questions (theoretical, methodological, and applied) on how preconditioning can aid sparse high dimensional inference. We have focused on preconditioning and the sign consistency of the Lasso. However, preconditioning also has the potential to benefit several other areas as well, including Basis Pursuit and the Restricted Isometry Principle [Chen and Donoho, 1994, Candès, 2008] and forms of ℓ_2 consistency for the Lasso and the restricted eigenvalue condition [Bickel et al., 2009].

This is the first paper to demonstrate how preconditioning the standard linear regression equation can circumvent the stringent 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.

Appendix

We prove our theorems in the appendix.

APPENDIX A: LOW DIMENSIONAL SETTINGS

We first give a well-known result that makes sure the Lasso exactly recovers the sparse pattern of β^*, that is $\hat{\beta}(\lambda) =_s \beta^*$. The following Lemma gives necessary and sufficient conditions for $\text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*)$. Wainwright [2009] gives these conditions that follow from KKT conditions.

LEMMA 1. *For linear model $Y = \mathbf{X}\beta^* + \epsilon$, assume that the matrix $X(S)^T X(S)$ is invertible. Then for any given $\lambda > 0$ and any noise term $\epsilon \in \mathbb{R}^n$, there exists a Lasso estimate $\hat{\beta}(\lambda)$ described in Equation (4) which satisfies $\hat{\beta}(\lambda) =_s \beta^*$, if and only if the following two conditions hold*

$$\left| X(S^c)^T X(S)(X(S)^T X(S))^{-1} \left[X(S)^T \epsilon - \lambda sign(\beta^*(S)) \right] - X(S^c)^T \epsilon \right| \leq \lambda, \tag{12}$$

$$sign\left(\beta^*(S) + (X(S)^T X(S))^{-1} \left[X(S)^T \epsilon - \lambda sign(\beta^*(S)) \right] \right) = sign(\beta^*(S)), \tag{13}$$

where the vector inequality and equality are taken elementwise. Moreover, if the inequality (12) holds strictly, then

$$\hat{\beta} = (\hat{\beta}^{(1)}, 0)$$

is the unique optimal solution to the Lasso problem in Equation (4), where

$$\hat{\beta}^{(1)} = \beta^*(S) + (X(S)^T X(S))^{-1} \left[X(S)^T \epsilon - \lambda sign(\beta^*) \right]. \tag{14}$$

Remarks. As in Wainwright [2009], we state sufficient conditions for (12) and (13). Define

$$\overrightarrow{b} = \text{sign}(\beta^*(S)),$$

and denote by e_i the vector with 1 in the ith position and zeroes elsewhere. Define

$$U_i = e_i^T (X(S)^T X(S))^{-1} \left[X(S)^T \epsilon - \lambda \overrightarrow{b} \right],$$

$$V_j = X_j^T \left\{ X(S)(X(S)^T X(S))^{-1} \lambda \overrightarrow{b} - \left[X(S)(X(S)^T X(S))^{-1} X(S)^T - I \right] \epsilon \right\}.$$

By rearranging terms, it is easy to see that (12) holds strictly if and only if

$$\mathcal{M}(V) = \left\{ \max_{j \in S^c} |V_j| < \lambda \right\} \tag{15}$$

holds. If we define $M(\beta^*) = \min_{j\in S} |\beta_j^*|$ (recall that $S = \{j : \beta_j^* \neq 0\}$ is the sparsity index), then the event

$$\mathcal{M}(U) = \left\{ \max_{i\in S} |U_i| < M(\beta^*) \right\}, \tag{16}$$

is sufficient to guarantee that condition (13) holds.

With the previous lemma, we now have our main result for the Lasso estimator on data from the linear model with *correlated noise terms*. It requires some regularity conditions. Suppose the irrepresentable condition holds. That is, for some constant $\eta \in (0, 1]$,

$$\left\| X(S^c)^T X(S) \left(X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty \leq 1 - \eta. \tag{17}$$

In addition, assume

$$C_{\min} = \Lambda_{\min} \left(X(S)^T X(S) \right) > 0. \tag{18}$$

where $\Lambda_{\min}$ denotes the minimal eigenvalue, and

$$\max_j \|X_j\|_2 \leq 1 \tag{19}$$

We also need the following Gaussian Comparison result for any mean zero Gaussian random vector.

LEMMA 2. *For any mean zero Gaussian random vector $(X_1, \ldots, X_n)$, and $t > 0$, we have*

$$P(\max_{1\leq i\leq n} |X_i| \geq t) \leq 2n \exp\left\{ -\frac{t^2}{2 \max_i E(X_i^2)} \right\} \tag{20}$$

Define

$$\Psi(\mathbf{X}, \beta^*, \lambda) = \lambda \left[\frac{\eta}{\sqrt{C_{\min}}} + \left\| \left(X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty \right].$$

LEMMA 3. *Suppose that data $(\mathbf{X}, Y)$ follow the linear model $Y = \mathbf{X}\beta^* + \epsilon$, with Gaussian noise $\epsilon \sim N(0, \Sigma_\epsilon)$. Assume that regularity conditions (17), (18), and (19) hold. If λ satisfies*

$$M(\beta^*) > \Psi(\mathbf{X}, \beta^*, \lambda),$$

then with probability greater than

$$1 - 2p \exp\left\{ -\frac{\lambda^2 \eta^2}{2\Lambda_{\max}(\Sigma_\epsilon)} \right\},$$

the Lasso has a unique solution $\hat{\beta}(\lambda)$ with $\hat{\beta}(\lambda) =_s \beta^*$.

PROOF. This proof is divided into two parts. First we analyze the probability of event $\mathcal{M}(V)$, and then we analyze the event of $\mathcal{M}(U)$.

Analysis of $\mathcal{M}(V)$: Note from (15) that $\mathcal{M}(V)$ holds if and only if $\frac{\max_{j\in S^c}|V_j|}{\lambda} < 1$. Each random variable V_j is Gaussian with mean

$$\mu_j = \lambda X_j^T X(S)(X(S)^T X(S))^{-1}\overrightarrow{b}.$$

Define $\tilde{V}_j = X_j^T\left[I - X(S)(X(S)^T X(S))^{-1}X(S)^T\right]\epsilon$, then $V_j = \mu_j + \tilde{V}_j$. Using the irrepresentable condition (17), we have $|\mu_j| \leq (1-\eta)\lambda$ for all $j \in S^c$, from which we obtain that

$$\frac{1}{\lambda}\max_{j\in S^c}|\tilde{V}_j| < \eta \Rightarrow \frac{\max_{j\in S^c}|V_j|}{\lambda} < 1.$$

Recall $s = |S|$. By the Gaussian comparison result (20) stated in Lemma 2, we have

$$P\left[\frac{1}{\lambda}\max_{j\in S^c}|\tilde{V}_j| \geq \eta\right] \leq 2(p-s)\exp\{-\frac{\lambda^2\eta^2}{2\max_{j\in S^c}E(\tilde{V}_j^2)}\}.$$

Since

$$E(\tilde{V}_j^2) = X_j^T H\Sigma_e H X_j,$$

where $H = I - X(S)(X(S)^T X(S))^{-1}X(S)^T$ which has maximum eigenvalue equal to 1. An operator bound yields

$$E(\tilde{V}_j^2) \leq \Lambda_{\max}(\Sigma_\epsilon)\|X_j\|_2^2 \leq \Lambda_{\max}(\Sigma_\epsilon).$$

Therefore,

$$P\left[\frac{1}{\lambda}\max_{j}|\tilde{V}_j| \geq \eta\right] \leq 2(p-s)\exp\left\{-\frac{\lambda^2\eta^2}{2\Lambda_{\max}(\Sigma_\epsilon)}\right\}.$$

So,

$$\begin{aligned} P\left[\frac{1}{\lambda}\max_{j}|V_j| < 1\right] &\geq 1 - P\left[\frac{1}{\lambda}\max_{j}|\tilde{V}_j| \geq \eta\right] \\ &\geq 1 - 2(p-s)\exp\left\{-\frac{\lambda^2\eta^2}{2\Lambda_{\max}(\Sigma_\epsilon)}\right\}. \end{aligned}$$

Analysis of $\mathcal{M}(U)$:

$$\max_i |U_i| \leq \|(X(S)^T X(S))^{-1}X(S)^T\epsilon\|_\infty + \lambda\|(X(S)^T X(S))^{-1}\overrightarrow{b}\|_\infty$$

Define $Z_i := e_i^T (X(S)^T X(S))^{-1} X(S)^T \epsilon$. Each Z_i is a normal Gaussian with mean 0 and variance

$$\begin{aligned} var(Z_i) &= e_i^T (X(S)^T X(S))^{-1} X(S)^T \Sigma_\epsilon X(S) (X(S)^T X(S))^{-1} e_i \\ &\leq \Lambda_{\max}(\Sigma_\epsilon) e_i^T (X(S)^T X(S))^{-1} X(S)^T X(S) (X(S)^T X(S))^{-1} e_i \\ &\leq \frac{\Lambda_{\max}(\Sigma_\epsilon)}{C_{\min}}. \end{aligned}$$

So for any $t > 0$, by (20)

$$P(\max_{i \in S} |Z_i| \geq t) \leq 2s \exp\left\{ -\frac{t^2 C_{\min}}{2\Lambda_{\max}(\Sigma_\epsilon)} \right\},$$

by taking $t = \frac{\lambda\eta}{\sqrt{C_{\min}}}$, we have

$$P(\max_{i \in S} |Z_i| \geq \frac{\lambda\eta}{\sqrt{C_{\min}}}) \leq 2s \exp\left\{ -\frac{\lambda^2\eta^2}{2\Lambda_{\max}(\Sigma_\epsilon)} \right\}.$$

Recall the definition of $\Psi(\mathbf{X}, \beta^*, \lambda) = \lambda \left[\frac{\eta}{\sqrt{C_{\min}}} + \left\| \left(X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty \right]$. We now have

$$P(\max_i |U_i| \geq \Psi(\mathbf{X}, \beta^*, \lambda)) \leq 2s \exp\left\{ -\frac{\lambda^2\eta^2}{2\Lambda_{\max}(\Sigma_\epsilon)} \right\}.$$

By condition $M(\beta^*) > \Psi(\mathbf{X}, \beta^*, \lambda)$, we have

$$P(\max_i |U_i| < M(\beta^*)) \geq 1 - 2s \exp\left\{ -\frac{\lambda^2\eta^2}{2\Lambda_{\max}(\Sigma_\epsilon)} \right\}.$$

At last, we have

$$P\left[\mathcal{M}(V) \& \ \mathcal{M}(U) \right] \geq 1 - 2p \exp\left\{ -\frac{\lambda^2\eta^2}{2\Lambda_{\max}(\Sigma_\epsilon)} \right\}.$$

□

We are now ready to prove Theorem 1 in Section 3. For convenience, the theorem is repeated here.

THEOREM 4. *Suppose that data $(\mathbf{X}, Y)$ follows the linear model described in Equation (1) with iid normal noises $\epsilon \sim N(0, \sigma^2 I_n)$. Define the singular value decomposition of X as $\mathbf{X} = UDV^T$. Suppose that $n \geq p$ and $\mathbf{X}$ has rank p. We further assume that $\Lambda_{\min}(\frac{1}{n} X^T X) \geq \tilde{C}_{\min} > 0$. Define the* ***Puffer Transformation****, $F = UD^{-1}U^T$. Let $\tilde{\mathbf{X}} = F\mathbf{X}, \tilde{Y} = FY$ and $\tilde{\epsilon} = F\epsilon$. 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 with probability greater than*

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

we have $\tilde{\beta}(\lambda) =_s \beta^*$.

PROOF. Data after transformation $(\tilde{\mathbf{X}}, \tilde{Y})$ follows the following linear model:

$$\tilde{Y} = \tilde{\mathbf{X}}\beta^* + \tilde{e},$$

with $\tilde{e}$ having co-variance matrix $\Sigma_{\tilde{e}} = \sigma^2 F^T F = \sigma^2 U D^{-2} U^T$.

Since

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \mathbf{X}'F'F\mathbf{X} = [VDU^T][UD^{-2}U^T][UDV^T] = I_{p\times p}.$$

So the irrepresentable condition (17) holds with $\eta = 1$. To apply Lemma 3, we first calculate $\Psi(\tilde{\mathbf{X}}, \beta^*, \lambda)$.

$$\Psi(\tilde{\mathbf{X}}, \beta^*, \lambda) = \lambda\left[\frac{\eta}{\sqrt{C_{\min}}\max_{j\in S^c}\|\tilde{X}_j\|_2} + \left\|\left(\tilde{X(S)}^T\tilde{X(S)}\right)^{-1}\overrightarrow{b}\right\|_\infty\right].$$

Notice that

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = I_{p\times p}.$$

So, $C_{\min} = 1$, $\|\tilde{X}_j\|_2 = 1$ and $\left\|\left(\tilde{X(S)}^T\tilde{X(S)}\right)^{-1}\overrightarrow{b}\right\|_\infty = 1$ and, consequently,

$$\begin{aligned}\Psi(\tilde{\mathbf{X}}, \beta^*, \lambda) &= \lambda\left[\frac{\eta}{\sqrt{C_{\min}}} + \left\|\left(\tilde{X(S)}^T\tilde{X(S)}\right)^{-1}\overrightarrow{b}\right\|_\infty\right] \\ &= 2\lambda.\end{aligned}$$

Now we calculate the lower bound probability:

$$1 - 2p\exp\left\{-\frac{\lambda^2\eta^2}{2\Lambda_{\max}(\Sigma_{\tilde{e}})}\right\}$$

Notice that $\Sigma_{\tilde{e}} = \sigma^2 U D^{-2} U'$. So $\Lambda_{\max}(\Sigma_{\tilde{e}}) = \frac{\sigma^2}{(\min_i D_{ii})^2}$. From $\Lambda_{\min}(\frac{1}{n}X^T X) \geq \tilde{C}_{\min}$, we see that $\tilde{C}_{\min} \leq \Lambda_{\min}(\frac{1}{n}X^T X) = \frac{1}{n}\Lambda_{\min}(VD^2V^T) = \frac{1}{n}\min_i(D_{ii}^2)$. This is to say $\min_i D_{ii}^2 \geq n\tilde{C}_{\min}$.

$$
\begin{aligned}
&1-2p\exp\left\{-\frac{\lambda^2\eta^2}{2[\Lambda_{\max}(\Sigma_{\tilde{\epsilon}})]}\right\}\\
=\ &1-2p\exp\left\{-\frac{\lambda^2\min_i(D_{ii}^2)}{2\sigma^2}\right\}\\
\geq\ &1-2p\exp\left\{-\frac{n\lambda^2\tilde{C}_{\min}}{2\sigma^2}\right\}
\end{aligned}
$$

□

Next, we prove Theorem 3 in Section 4. The theorem is repeated here for convenience.

THEOREM 5. *Suppose that data $(\mathbf{X}, Y)$ follows the linear model described at (1) with iid normal noises $\epsilon \sim N(0, \sigma^2 I_n)$. Define the singular value decomposition of X as $\mathbf{X} = UDV^T$. Suppose that $p \geq n$. We further assume that $\Lambda_{\min}(\frac{1}{n}X(S)^T X(S)) \geq \tilde{C}_{\min}$ and $\min_i(D_{ii}^2) \geq pd_{\min}$ with constants $\tilde{C}_{\min} > 0$ and $d_{\min} > 0$. For $F = UD^{-1}U^T$, define $\tilde{Y} = FY$ and $\tilde{\mathbf{X}} = F\mathbf{X}$. Define*

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

Under the following three conditions,

1. $\left\|\tilde{\mathbf{X}}(S^c)^T\tilde{\mathbf{X}}(S)\left(\tilde{\mathbf{X}}(S)^T\tilde{\mathbf{X}}(S)\right)^{-1}\overrightarrow{b}\right\|_\infty \leq 1-\eta,$
2. $\Lambda_{\min}\left(\tilde{\mathbf{X}}(S)^T\tilde{\mathbf{X}}(S)\right) \geq \frac{n}{cp}$
3. $\min_{j\in S}|\beta_j^*| \geq 2\lambda\sqrt{scp/n}$

where c is some constant; we have

$$P\left(\tilde{\beta}(\lambda) =_s \beta^*\right) > 1-2p\exp\left\{-\frac{p\lambda^2\eta^2 d_{\min}}{2\sigma^2}\right\}. \tag{21}$$

PROOF. After transformation, $(\tilde{\mathbf{X}}, \tilde{Y})$ follows the following linear model:

$$\tilde{Y} = \tilde{\mathbf{X}}\beta^* + \tilde{e},$$

with $\tilde{e}$ having co-variance matrix $\Sigma_{\tilde{e}} = \sigma^2 F^T F = \sigma^2 UD^{-2}U^T$.

To apply Lemma 3, first calculate $\Psi(\tilde{\mathbf{X}}, \beta^*, \lambda)$.

$$\Psi(\tilde{\mathbf{X}}, \beta^*, \lambda) = \lambda\left[\frac{\eta}{\sqrt{C_{\min}}} + \left\|\left(X(\tilde{S})^T X(\tilde{S})\right)^{-1}\overrightarrow{b}\right\|_\infty\right],$$

where $C_{\min} = \Lambda_{\min}\left(X\tilde{(S)}^T X\tilde{(S)}\right)$. By condition (3), we have $C_{\min} \geq \frac{n}{cp}$.

$$\left\|\left(X\tilde{(S)}^T X\tilde{(S)}\right)^{-1}\overrightarrow{b}\right\|_\infty \leq \sqrt{\frac{s}{C_{\min}}} \leq \sqrt{scp/n}$$

and, consequently,

$$\begin{aligned}\Psi(\tilde{\mathbf{X}}, \beta^*, \lambda) &= \lambda\left[\frac{\eta}{\sqrt{C_{\min}}} + \left\|\left(X\tilde{(S)}^T X\tilde{(S)}\right)^{-1}\overrightarrow{b}\right\|_\infty\right] \\ &\leq \lambda\left[\frac{\eta\sqrt{cp}}{\sqrt{n}} + \sqrt{scp/n}\right] \\ &\leq 2\lambda\sqrt{scp/n}.\end{aligned}$$

Now we calculate the lower bound probability:

$$1 - 2p\exp\left\{-\frac{\lambda^2\eta^2}{2\Lambda_{\max}(\Sigma_{\tilde{\epsilon}})}\right\}$$

Notice that $\Sigma_{\tilde{\epsilon}} = \sigma^2 U D^{-2} U'$. So $\Lambda_{\max}(\Sigma_{\tilde{\epsilon}}) = \frac{\sigma^2}{(\min_i D_{ii})^2}$.

$$\begin{aligned}&1 - 2p\exp\left\{-\frac{\lambda^2\eta^2}{2\Lambda_{\max}(\Sigma_{\tilde{\epsilon}})}\right\} \\ =\ &1 - 2p\exp\left\{-\frac{\lambda^2\eta^2 \min_i D_{ii}^2}{2\sigma^2}\right\} \\ \geq\ &1 - 2p\exp\left\{-\frac{p\lambda^2\eta^2 d_{\min}}{2\sigma^2}\right\}\end{aligned}$$

□

APPENDIX B: UNIFORM DISTRIBUTION ON THE STIEFEL MANIFOLD

To prove Theorem 2, we first present some results related to Beta distributions.

B.1. Beta distribution. The density function for the Beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$ is

$$f(x) = cx^{\alpha-1}(1-x)^{\beta-1},$$

for $x \in (0, 1)$. c is a normalization constant. If X follows a Beta distribution with parameters (α, β), it is denoted $X \sim Beta(\alpha, \beta)$.

PROPOSITION 3. *If $X \sim Beta(\alpha, \beta)$, then*

$$E(X) = \frac{\alpha}{\alpha+\beta} \quad \textit{and} \quad var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The next two inequalities for the χ^2 distributions can be found from Laurent and Massart [2000] (pp. 1352).

LEMMA 4. *Let X be a χ^2 distribution with n degrees of freedom. Then, for any positive x, we have*

$$P(X - n \geq 2\sqrt{nx} + 2x) \leq \exp(-x),$$
$$P(X - n \leq -2\sqrt{nx}) \leq \exp(-x).$$

By taking $x = \sqrt{n}$, from the above inequality we immediately have

COROLLARY 2. *Let X be a χ^2 distribution with n degrees of freedom. There exists a constant c such that for a large enough n,*

$$P(|X - n| \geq cn^{3/4}) \leq \exp\left(-n^{1/2}\right).$$

Corollary 2 says that $\chi^2(n) = n + o(n)$ with high probability. These large deviation results can give concentration inequalities for Beta distributions. This is because a Beta distribution can be expressed via χ^2 distributions.

LEMMA 5. *Suppose that $X \sim \chi^2(\alpha)$, $Y \sim \chi^2(\beta)$ and X is independent of Y, then $\frac{X}{X+Y} \sim Beta(\frac{\alpha}{2}, \frac{\beta}{2})$.*

With the relationship constructed between a Beta distribution and χ^2 distributions, we can have the following inequalities.

THEOREM 6. *Suppose $Z \sim Beta(\frac{n}{2}, \frac{m}{2})$. When both m and n are big enough, there exists some constant c,*

$$P\left(Z \geq \frac{n + cn^{3/4}}{m + n - c(m+n)^{3/4}}\right) \leq \exp\{-n^{1/2}\} + \exp\{-(m+n)^{1/2}\}$$

$$P\left(Z \leq \frac{n - cn^{3/4}}{m + n + c(m+n)^{3/4}}\right) \leq \exp\{-n^{1/2}\} + \exp\{-(m+n)^{1/2}\}$$

PROOF. Let $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$ are independent. Then $\frac{X}{X+Y}$ has the same distribution as Z.

By Corollary 2, we have

$$P(|X-n| > cn^{3/4}) \leq \exp\{-n^{1/2}\},$$

$$P\left(|X+Y-(m+n)| > c(m+n)^{3/4}\right) \leq \exp\{-(m+n)^{1/2}\}.$$

If $\frac{X}{X+Y} \geq \frac{n+cn^{3/4}}{m+n-(m+n)^{3/4}}$, then $X > n+cn^{3/4}$ or $X+Y < m+n-(m+n)^{3/4}$. So

$$\begin{aligned} P\left(\frac{X}{X+Y} \geq \frac{n+cn^{3/4}}{m+n-c(m+n)^{3/4}}\right) &\leq P\left(X > n+cn^{3/4}\right) \\ &\quad + P\left(X+Y < m+n-c(m+n)^{3/4}\right) \\ &\leq \exp\{-n^{1/2}\} + \exp\{-(m+n)^{1/2}\} \end{aligned}$$

The same way, we have

$$P\left(\frac{X}{X+Y} \leq \frac{n-cn^{3/4}}{m+n+c(m+n)^{3/4}}\right) \leq \exp\{-n^{1/2}\} + \exp\{-(m+n)^{1/2}\}$$

□

COROLLARY 3. *Suppose $Z \sim Beta(\frac{n}{2}, \frac{m}{2})$ with $m > n$ and $n/m \to c_3$ $(0 \leq c_3 < 1)$. There exists a constant c_1 such that the following inequality holds when both m and n are big enough,*

$$P\left(\left|Z - \frac{n}{m+n}\right| \geq \frac{c_1 n^{3/4}}{m+n}\right) \leq 2\exp\{-n^{1/2}\}$$

Corollary 3 states that if $Z \sim Beta(\frac{n}{2}, \frac{m}{2})$(with $m > n$), then $Z = \frac{n}{m+n} + O(\frac{n^{3/4}}{m+n})$ with high probability.

PROOF. From Theorem 6, we only have to prove that, for any constant c_0

$$\left(\frac{n-c_0 n^{3/4}}{m+n+c_0(m+n)^{3/4}} - \frac{n}{m+n}\right) \Big/ \left(\frac{n^{3/4}}{m+n}\right) \to c_2,$$

where $c_2 \in \mathbb{R}$ is a constant with the same sign as c_0.

$$
\begin{aligned}
& \left(\frac{n - c_0 n^{3/4}}{m+n+c_0(m+n)^{3/4}} - \frac{n}{m+n}\right) \Bigg/ \left(\frac{n^{3/4}}{m+n}\right) \\
= & \frac{-c_0 n^{3/4}(m+n) - c_0 n(m+n)^{3/4}}{[m+n+c_0(m+n)^{3/4}](m+n)} \times \frac{m+n}{n^{3/4}} \\
= & \frac{-c_0 n^{3/4}(m+n) - c_0 n(m+n)^{3/4}}{[m+n+c_0(m+n)^{3/4}]} \times \frac{1}{n^{3/4}} \\
= & \frac{-c_0 - c_0 n^{1/4}(m+n)^{-1/4}}{[1+c_0(m+n)^{-1/4}]} \\
\to & -c_0 - \frac{c_0 c_3}{1+c_3} \text{ as } m, n \to \infty.
\end{aligned}
$$

□

With the previous results, we can prove results for a random vector uniformly distributed on the Stiefel manifold.

B.2. Stiefel manifold. Suppose that $V \in \mathbb{R}^{n \times p}$ with $p > n$, which satisfies $VV' = I_n$ — the rows are orthogonal. All of these matrices V form $V(n,p)$, called the Stiefel manifold [Downs, 1972]. We seek to examine the properties of a matrix V that is uniformly distributed on $V(n,p)$. Specifically, we show that any two columns of V are nearly orthogonal.

We suppose that V comes uniformly from Stiefel manifold. Let $X = [V_j, V_k] \in \mathbb{R}^{n \times 2}$ be two columns of V. Then, from Khatri [1970], if $p > n+1$, the marginal density of X is

$$c|I_n - XX'|^{p-2-n-1},$$

where c is a normalization constant. The density of $X'X = A \in \mathbb{R}^{2 \times 2}$ is given by

$$c|A|^{(n-3)/2}|I_2 - A|^{(p-n-3)/2}, 0 < A < I,$$

where c is a normalization constant. A is distributed as multivariate Beta of type I.

From Khatri and Pillai [1965],

$$E(A) = \begin{pmatrix} n/p & 0 \\ 0 & n/p \end{pmatrix}.$$

This result shows that any two columns of V are orthogonal in expectation.

In Mitra [1970], a symmetric matrix $U \sim B_k(\frac{n_1}{2}, \frac{n_2}{2})$ with $\min(n_1, n_2) \geq k$ has the density

$$f(U) = c|U|^{(n_1-k-1)/2}|I - U|^{(n_2-k-1)/2}.$$

So, A defined above follows $B_2(\frac{n}{2}, \frac{p-n}{2})$. From Mitra (1970), we have the following results.

LEMMA 6 (Mitra (1970)). *If $U \sim B_k(\frac{n_1}{2}, \frac{n_2}{2})$, then for each fixed non-null vector a,*

$$a'Ua/a'a \sim Beta(\frac{n_1}{2}, \frac{n_2}{2}).$$

By taking $a = (1,0)', (0,1)'$ and $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})'$ respectively, we have the following results:

COROLLARY 4. *If $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2} \sim B_2(\frac{n}{2}, \frac{p-n}{2})$, we have*

1. $A_{11} \sim Beta(\frac{n}{2}, \frac{p-n}{2})$
2. $A_{22} \sim Beta(\frac{n}{2}, \frac{p-n}{2})$
3. $\frac{1}{2}A_{11} + \frac{1}{2}A_{22} + A_{12} \sim Beta(\frac{n}{2}, \frac{p-n}{2})$

Now, concentration results in Corollary 3 can bound A_{11}, A_{12}, and A_{22}. This yields an inequality to bound $V_i'V_j/(\|V_i\|_2\|V_j\|_2)$ which describes the linear relationship between two columns V_i and V_j.

THEOREM 7. *Suppose that $V \in \mathbb{R}^{n\times p}$ with $p \gg n$ uniformly from the Stiefel manifold. For a large enough n and p, there exist some constants c_1 and c_2, such that for any two different columns of V — V_j and V_k, the following results hold:*

$$(1)\ P\left(|V_j'V_k| \geq \frac{2c_1 n^{3/4}}{p}\right) \leq 6\exp\{-n^{1/2}\},\ \textit{and}$$

$$(2)\ P\left(\frac{|V_j'V_k|}{\|V_j\|_2\|V_k\|_2} \geq c_2 n^{-1/4}\right) \leq 8\exp\{-n^{1/2}\}$$

PROOF. **We first prove (1).** From Corollary 3, for large enough n and p,

$$P\left(\left|A_{11} - \frac{n}{p}\right| \geq \frac{c_1 n^{3/4}}{p}\right) \leq 2\exp\{-n^{1/2}\},$$

$$P\left(\left|A_{22} - \frac{n}{p}\right| \geq \frac{c_1 n^{3/4}}{p}\right) \leq 2\exp\{-n^{1/2}\},\ \text{and}$$

$$P\left(\left|(A_{11}+A_{22})/2 + A_{12} - \frac{n}{p}\right| \geq \frac{c_1 n^{3/4}}{p}\right) \leq 2\exp\{-n^{1/2}\}.$$

Since

$$\begin{aligned} |A_{12}| &\leq \left|(A_{11}+A_{22})/2+A_{12}-\frac{n}{p}\right|+\left|(A_{11}+A_{22})/2-\frac{n}{p}\right| \\ &\leq \left|(A_{11}+A_{22})/2+A_{12}-\frac{n}{p}\right|+\left|\frac{A_{11}}{2}-\frac{n}{2p}\right|+\left|\frac{A_{22}}{2}-\frac{n}{2p}\right|, \end{aligned}$$

it implies that, if $|A_{12}| \geq \frac{2c_1 n^{3/4}}{p}$, then

$$\begin{aligned} \left|A_{11}-\frac{n}{p}\right| &\geq \frac{c_1 n^{3/4}}{p} \text{ or} \\ \left|A_{22}-\frac{n}{p}\right| &\geq \frac{c_1 n^{3/4}}{p} \text{ or} \\ \left|(A_{11}+A_{22})/2+A_{12}-\frac{n}{p}\right| &\geq \frac{c_1 n^{3/4}}{p}. \end{aligned}$$

So,

$$\begin{aligned} P\left(|A_{12}| \geq \frac{2c_1 n^{3/4}}{p}\right) &\leq P\left(\left|(A_{11}+A_{22})/2+A_{12}-\frac{n}{p}\right| \geq \frac{c_1 n^{3/4}}{p}\right) \\ &+ P\left(\left|A_{11}-\frac{n}{p}\right| \geq \frac{c_1 n^{3/4}}{p}\right)+P\left(\left|A_{22}-\frac{n}{p}\right| \geq \frac{c_1 n^{3/4}}{p}\right) \\ &\leq 6\exp\{-n^{1/2}\}. \end{aligned}$$

Now we prove (2). If

$$\frac{|V_j' V_k|}{\sqrt{\|V_j\|_2^2 \|V_k\|_2^2}} \geq \frac{\frac{2c_1 n^{3/4}}{p}}{\sqrt{[\frac{n}{p}-\frac{c_1 n^{3/4}}{p}]^2}},$$

we must have $|V_j' V_k| \geq \frac{2c_1 n^{3/4}}{p}$ or $\|V_j\|_2^2 \leq \frac{n}{p}-\frac{c_1 n^{3/4}}{p}$ or $\|V_k\|_2^2 \leq \frac{n}{p}-\frac{c_1 n^{3/4}}{p}$.

So,

$$\begin{aligned} P\left(\frac{|V_j' V_k|}{\sqrt{\|V_j\|_2^2 \|V_k\|_2^2}} \geq \frac{\frac{2c_1 n^{3/4}}{p}}{\frac{n}{p}-\frac{c_1 n^{3/4}}{p}}\right) &\leq P\left(|V_j' V_k| \geq \frac{2c_1 n^{3/4}}{p}\right) \\ &+ P\left(\|V_j\|_2^2 \leq \frac{n}{p}-\frac{c_1 n^{3/4}}{p}\right) \\ &+ P\left(\|V_k\|_2^2 \leq \frac{n}{p}-\frac{c_1 n^{3/4}}{p}\right) \\ &\leq 8\exp\{-n^{1/2}\}. \end{aligned}$$

Note that

$$\frac{\frac{2c_1 n^{3/4}}{p}}{\frac{n}{p} - \frac{c_1 n^{3/4}}{p}} = O(n^{-1/4}).$$

So, for large enough n, there exists a constant c_2 such that

$$P\left(\frac{|V_j' V_k|}{\sqrt{\|V_j\|_2^2 \|V_k\|_2^2}} \geq c_2 n^{-1/4}\right) \leq 8\exp\{-n^{1/2}\}.$$

□

COROLLARY 5. *Suppose that $V \in \mathbb{R}^{n\times p}$ is uniformly distributed on the Stiefel manifold. After normalizing the columns of V to have equal length in ℓ_2, the irrepresentable condition holds for s relevant variables with probability no less than*

$$1 - 4p(p-1)e^{-n^{1/2}},$$

as long as n is large enough and $n > (2c_2)^4(2s-1)^4$.

PROOF. Let C be the normalized Gram matrix of V defined as

$$C_{jk} = \frac{|V_j' V_k|}{\sqrt{\|V_j\|_2^2 \|V_k\|_2^2}}.$$

From Theorem 7,

$$P\left(|C_{jk}| \geq c_2 n^{-1/4}\right) \leq 8\exp\{-n^{1/2}\}.$$

Using the union bound,

$$P\left(\max_{jk} |C_{jk}| \geq c_2 n^{-1/4}\right) \leq 4p(p-1)\exp\{-n^{1/2}\}.$$

By Corollary 2 of Zhao and Yu (2006), when $\max|C_{ij}| \leq \frac{c}{2s-1}$ for some $0 < c < 1$, the irrepresentable condition holds. So, if $c_2 n^{-1/4} \leq \frac{1}{2(2s-1)}$, that is $n > (2c_2)^4(2s-1)^4$, we have

$$P\left(\max_{jk} |C_{jk}| \geq \frac{1}{2(2s-1)}\right) \geq 1 - 4p(p-1)\exp\{-n^{1/2}\}.$$

□

THEOREM 8. *Suppose that $X \in \mathbb{R}^{n\times p}$ is a random matrix with each element X_{ij} drawn iid from $N(0,1)$. Then, after normalizing the columns of X to have equal length in ℓ_2, the irrepresentable condition holds for s relevant variables with probability no less than*

$$1 - \frac{1}{2}p(p-1)e^{-\frac{nc^2}{16(2s-1)^2}} - 3p(p-1)e^{-\frac{n}{16}},$$

for any $0 < c < 1$.

This implies that n must grow faster than $s^2 \log(p)$ for X to satisfy the irrepresentable condition.

PROOF. Let C be the empirical correlation matrix of X

$$C_{jk} = \frac{1/n \sum_{\ell=1}^n X_{j\ell} X_{k\ell}}{\sqrt{1/n \sum_{\ell=1}^n X_{j\ell}^2 \times 1/n \sum_{\ell=1}^n {X_{k\ell}}^2}}.$$

By Corollary 2. of Zhao and Yu (2006), when $\max |C_{ij}| \leq \frac{c}{2s-1}$, IC holds. Now we try to bound $P\left(\max_{i \neq j} |C_{ij}| \leq \frac{c}{2s-1}\right)$.

Note that

$$1/n \sum_{\ell=1}^n X_{j\ell} X_{k\ell} \mid X_k \sim N\left(0, 1/n^2 \sum_\ell X_{k\ell}^2\right).$$

By a concentration inequality on χ^2 distribution,

$$P(1/2 < \frac{\chi^2(n)}{n} < 2) \geq 1 - 2e^{-\frac{n}{16}}.$$

For $Z \sim N(0, 2/n)$,

$$P\left(1/n \sum_{\ell=1}^n X_{j\ell} X_{k\ell} > t \mid \frac{1}{n} \sum_\ell X_{k\ell}^2 < 2\right) \leq P(Z > t) \leq e^{-\frac{nt^2}{4}}.$$

This holds because the variance increases. So,

$$\begin{aligned} P\left(1/n \sum_{\ell=1}^n X_{j\ell} X_{k\ell} > t\right) &= P\left(1/n \sum_{\ell=1}^n X_{j\ell} X_{k\ell} > t \mid \frac{1}{n} \sum_\ell X_{k\ell}^2 < 2\right) P\left(\frac{1}{n} \sum_\ell X_{k\ell}^2 < 2\right) \\ &\quad + P\left(1/n \sum_{\ell=1}^n X_{j\ell} X_{k\ell} > t \mid \frac{1}{n} \sum_\ell X_{k\ell}^2 > 2\right) P\left(\frac{1}{n} \sum_\ell X_{k\ell}^2 > 2\right) \\ &\leq P\left(1/n \sum_{\ell=1}^n X_{j\ell} X_{k\ell} > t \mid \frac{1}{n} \sum_\ell X_{k\ell}^2 < 2\right) + P\left(\frac{1}{n} \sum_\ell X_{k\ell}^2 > 2\right) \\ &\leq e^{-\frac{nt^2}{4}} + 2e^{-n/16}. \end{aligned}$$

Finally,

$$\begin{aligned} P(|C_{jk}| < a) &\geq P\left(|1/n \sum_{\ell=1}^n X_{j\ell} X_{k\ell}| < a/2, 1/n \sum_{\ell=1}^n X_{j\ell}^2 > 1/2, 1/n \sum_{\ell=1}^n X_{k\ell}^2 > 1/2\right) \\ &\geq P(1/n \sum_{\ell=1}^n X_{j\ell} X_{k\ell} < a/2) + P(1/n \sum_{\ell=1}^n X_{j\ell}^2 > 1/2) \end{aligned}$$

$$
\begin{aligned}
&+ P(1/n \sum_{\ell=1}^{n} X_{k\ell}^2 > 1/2) - 2 \\
\geq \;& 1 - e^{-\frac{na^2}{16}} - 2e^{-n/16} + [1 - 2e^{-\frac{n}{16}}] \times 2 - 2 \\
= \;& 1 - e^{-\frac{na^2}{16}} - 6e^{-\frac{n}{16}}.
\end{aligned}
$$

Taking $a = \frac{c}{2s-1}$,

$$P(|C_{jk}| < \frac{c}{2s-1}) \geq 1 - e^{-\frac{nc^2}{16(2s-1)^2}} - 3e^{-\frac{n}{16}}$$

and

$$P(\max_{j \neq k} |C_{jk}| < \frac{c}{2s-1}) \geq 1 - \frac{1}{2}p(p-1)e^{-\frac{nc^2}{16(2s-1)^2}} - 3p(p-1)e^{-\frac{n}{16}}.$$

□

REFERENCES

AC Aitken. On least squares and linear combinations of observations. Proceedings of the Royal Society of Edinburgh, 55:42–48, 1935.

S. Arlot, G. Blanchard, and E. Roquain. Some nonasymptotic results on resampling in high dimension, i: Confidence regions. The Annals of Statistics, 38(1):51–82, 2010.

O. Axelsson. A survey of preconditioned iterative methods for linear systems of algebraic equations. BIT Numerical Mathematics, 25(1):165–187, 1985.

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

E.J. Candès. The restricted isometry property and its implications for compressed sensing. Comptes Rendus Mathematique, 346(9):589–592, 2008.

S. Chen and D. Donoho. Basis pursuit. In Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on, volume 1, pages 41–44. IEEE, 1994.

Y. Chikuse. Statistics on special manifolds, volume 174. Springer Verlag, 2003.

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

D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. Information Theory, IEEE Transactions on, 52(1):6–18, 2006.

T.D. Downs. Orientation statistics. Biometrika, 59(3):665–676, 1972.

B. Efron and R. Tibshirani. An introduction to the bootstrap, volume 57. Chapman & Hall/CRC, 1993.

N. El Karoui. High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: Risk underestimation. The Annals of Statistics, 38(6):3487–3566, 2010.

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

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

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

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

G.H. Golub and C.F. Van Loan. Matrix computations, volume 3. Johns Hopkins Univ Pr, 1996.

E. Greenshtein and Y.A. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli, 10(6):971–988, 2004.

H.C. Huang, N.J. Hsu, D.M. Theobald, and F.J. Breidt. Spatial lasso with applications to gis model selection. Journal of Computational and Graphical Statistics, 19(4):963–983, 2010.

J. Jia, K. Rohe, and B. Yu. The lasso under poisson-like heteroscedasticity. Statistica Sinica, in press.

C. G. Khatri and K. C. S Pillai. Some results on the non-central multivariate beta distribution and moments of traces of two matrices. Ann. Math. Statist., pages 1511– 1520, 1965.

CG Khatri. A note on mitra's paper" a density free approach to the matrix variate beta distribution". Sankhyā: The Indian Journal of Statistics, Series A, pages 311–318, 1970.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. Annals of Statistics, pages 1356–1378, 2000.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. The annals of Statistics, 28(5):1302–1338, 2000.

D.M. Mason and M.A. Newton. A rank statistics approach to the consistency of a general bootstrap. The Annals of Statistics, 20(3):1611–1624, 1992.

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

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. The Annals of Statistics, 37(1):246–270, 2009.

Sujit Kumar Mitra. A density free approach to the matrix variate beta distribution. Sankhyā: The Indian

Journal of Statistics, Series A, pages 81–88, 1970.
J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. Journal of the Royal Statistical Society. Series A (General), pages 370–384, 1972.
M.Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(4):659–677, 2007.
D. Paul, E. Bair, T. Hastie, and R. Tibshirani. preconditioning for feature selection and regression in high-dimensional problems. The Annals of Statistics, 36(4):1595–1618, 2008.
J. Shao and X. Deng. Estimation in high-dimensional linear models with deterministic design matrices. The Annals of Statistics, 40(2):812–831, 2012.
J.W. Silverstein. The smallest eigenvalue of a large dimensional wishart matrix. The Annals of Probability, 13(4):1364–1368, 1985.
R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.
R.J. Tibshirani. The solution path of the generalized lasso. Stanford University, 2011.
J.A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. Information Theory, IEEE Transactions on, 52(3):1030–1051, 2006.
S.A. Van De Geer. High-dimensional generalized linear models and the lasso. The Annals of Statistics, 36 (2):614–645, 2008.
M.J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. Information Theory, IEEE Transactions on, 55(12):5728–5741, 2009.
S. Xiong, B. Dai, and P.Z.G. Qian. Orthogonalizing penalized regression. Arxiv preprint arXiv:1108.0185, 2011.
M. Yuan and Y. Lin. On the non-negative garrotte estimator. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2):143–161, 2007.
C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 38(2):894–942, 2010.
C.H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics, 36(4):1567–1594, 2008.
P. Zhao and B. Yu. On model selection consistency of lasso. Journal of Machine Learning Research, 7(2): 2541, 2006.
H. Zou. The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101 (476):1418–1429, 2006.

LMAM School of Mathematical Sciences
and Center for Statistical Science
Peking University
Beijing 100871, China
E-mail: jzjia@math.pku.edu.cn

Department of Statistics
University of Wisconsin-Madison
WI 53706, USA
E-mail: karlrohe@stat.wisc.edu