Rank-Extreme Association of Gaussian Vectors and Low-Rank Detection

Kai Zhang

May 7, 2014

Abstract

It is important to detect a low-dimensional linear dependency in high-dimensional data. We provide a perspective on this problem, called the rank-extreme (ReX) association, through studies of the maximum norm of a vector of \( p \) standard Gaussian variables that has a covariance matrix of rank \( d \leq p \). We find a simple asymptotic upper bound of such extreme values as \( \sqrt{d(1 - p^{-2/(d-1)})} \). This upper bound is shown to be sharp when the entries of the correlation matrix are generated by inner products of i.i.d. uniformly distributed unit vectors. This upper bound also takes on an interesting trichotomy phenomenon depending on the limit of \( d/\log p \). Based on this ReX approach, we propose several methods for high-dimensional inference. These applications include a test of the overall significance in regressions, a refinement of valid post-selection inference when the size of selected models is restricted, a classification of deficient ranks based on the magnitude of the extreme, and an inference method for low-ranks. One advantage of this approach is that the asymptotics are in the dimensions \( d \) and \( p \) but not in the sample size \( n \). Thus, the inference can be made even when \( n < d \leq p \), which allows fast detection of low-dimensional structure. Furthermore, the higher the dimension is, the more accurate the inference is. Therefore, these results can be regarded as a “blessing of dimensionality.”

1 Introduction

How many stocks are principal drivers of the whole stock market? How many medical tests are sufficient to describe a person’s health status? These are all important questions being asked nowadays, and these types of questions can often be formulated as questions about whether there exists a low-dimensional linear dependency in a high-dimensional dataset. To be
more specific, suppose we observe a collection of $p$ variables $X_1, \ldots, X_p$, each $\mathcal{N}(0, 1)$, but possibly correlated. Can we tell whether there exists a subset of the Gaussian variables $X_{j_1}, \ldots, X_{j_d}$ such that all other variables are almost simply linear combinations of them?

This question has been studied very carefully under the context of low-rank matrix approximation and subspace tracking, in particular under the setting when the number of variables is larger than the sample size, i.e., $n < p$. For a review of low-rank approximation methods, see [1]. For recent algorithms on subspace tracking, see [2], [3], [4], and [5]. One of the most important components of all these methods involves the estimation of the rank of the covariance matrix, or the dimension of the signal subspace, $d$. This estimation problem has been first studied by [6], [7], [8], and [9]. More recent approaches from statistics include [10], [11], and [12].

One principal technique in the statistical approaches to the rank estimation problem is eigenvalue thresholding based on the principal component analysis (PCA), where we look for the “cut off” among singular values of the covariance matrix when they drop to nearly 0. However, when $n < d$, PCA of the sample correlation matrix will return with $n$ positive eigenvalues, which is less than the truth $d$. Moreover, although we may get low-rank solutions to many problems, an important question hasn’t been completely addressed: how low is the rank, really? In other words, how can we justify that the low-rank results we obtain is in the correct dimension? Indeed, the inference on $d$ as a parameter was not clear. In the context of the above paragraphs, the question is that, given the data, can we make any probabilistic statements about $d$ as a parameter? Can we test if the correlation matrix of a joint distribution is low-rank? Can we build confidence intervals that covers $d$ with high probability? Even these basic questions have not been completely answered yet.

Our approach to this problem is through the study the low rank from the extreme value of the joint distribution. Consider a Gaussian vector $\mathbf{X} = (X_1, \ldots, X_p)^T \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma$ has 1’s on the diagonal, so that each $X_j$ is standard normally distributed. In this case, classical extreme value theory suggests that for a large collection of $\Sigma$’s, the magnitude of the extreme value $\max_j |X_j|$ is about $\sqrt{2 \log p}$ ([13] and [14]). Although it seems like that there is not much difference in this asymptotic rate among $\Sigma$’s, this is not the complete story yet. In [15], it is shown that the asymptotics of the extremes from a low-rank Gaussian distribution is below the square root of the rank, which can be far below the dimension of the distribution. This result suggests the dependence of the Gaussian extreme on the rank of the covariance matrix. A generalized version of such asymptotics is studied
here, and it leads to an explicit description of such dependence, which we call the rank-extreme (ReX) association.

Here are the heuristics of this ReX approach: Note that for any $p \times p$ covariance matrix $\Sigma$ that is positive semi-definite, has ones on the diagonal and has rank $d$. Through its eigen-decomposition, we can write $\Sigma = L^T L$, where $L = [l_1, \ldots, l_p]$ is a $d \times p$ matrix with columns $l_j$'s such that $\|l_j\|_2 = 1$. Thus, we can write $X = L^T Z$ where $Z \sim N(0, I_d)$. Based on this fact, we reduce the problem to one of sphere packing problems in $\mathbb{R}^d$, as described in Section 2. In this case, with the help from the classical arguments from coding theory and sphere packing, we first derive a universal asymptotic bound for the maximal correlation between any collection of $p$ directions and a uniformly distributed direction in $\mathbb{R}^d$. The bound is found to be $\sqrt{1 - p^{-2/(d-1)}}$. We show that the rate $\sqrt{1 - p^{-2/(d-1)}}$ takes on a trichotomy phenomenon to tend toward 1, $\sqrt{1 - e^{-2/\beta}}$, or 0, as the ratio $d(p)/\log p$ converges to 0, some fixed positive $\beta$, or $+\infty$ respectively. Based on these results, we propose a test of overall significance in regressions for general $n$ and $p$ as an extension of the universal threshold proposed in [16]. We further show that this bound is sharp and is attained when the $l_j$'s are i.i.d. uniformly distributed unit vectors over the sphere $S^{d-1}$.

As a consequence of the above sphere packing results, we derive a universal asymptotic upper bound for $\|X\|_\infty = \max_j |X_j|$, the Gaussian extreme with an arbitrary $\Sigma$ with $\text{rank}(\Sigma) = d \leq p$ as $\sqrt{d(1 - p^{-2/(d-1)})}$. This upper bound explicitly describes the relationship between the magnitude of the extreme and the rank of the correlation matrix. As in the sphere packing problem described before, we show that this bound is sharp in the sense that if $\Sigma$ is generated by $l_j$'s that are i.i.d. uniformly distributed over the sphere $S^{d-1}$, the bound is attained. We were unable to find a similar sharp bound for extremes from low-rank Gaussian distributions in the literature. Due to the dependence of this bound on the rank, we call the bound $\sqrt{d(1 - p^{-2/(d-1)})}$ the rank-extreme (ReX) asymptotic bound. This association inspires us to use the extremes in drawing information about the underlying rank $d$ as a parameter.

One interesting fact about the ReX rate $\sqrt{d(1 - p^{-2/(d-1)})}$ is that it takes on a trichotomy phenomenon too like its sphere packing counterpart. Indeed, the rate $\sqrt{d(1 - p^{-2/(d-1)})}$ is shown to grow below, at, or above $\sqrt{\log p}$ asymptotically according to whether the ratio $d(p)/\log p$ tends below, at, or above a fixed positive constant $\beta^\dagger = 1.255005 \ldots$, which is the unique solution to the equation $\beta(1 - e^{-2/\beta}) = 1$. Based on this ReX trichotomy, we also propose inference methods under this situation and show that finite
sample inference can be made about \( d \).

It is noteworthy that the ReX inference is based on the asymptotics in \( p \) rather than in \( n \). Thus, the method works even when \( n < d \), when PCA of the sample correlation matrix would return \( n < d \) positive eigenvalues. In the extreme, the inference can be made even when there is only one sample but billions of variables. The freedom of the inference for any \( n \) is one important advantage of the ReX approach. This key feature and the easiness of finding the maximal entry in each sample allow fast detection of low-rank correlation structure in high-dimensional data.

The paper is organized as follows. In Section 2 we introduce our sphere packing method and develop the ReX asymptotic bound of the \( \ell_\infty \) norm of Gaussian vectors. We discuss two applications of the ReX bound in Section 2.2 and Section 2.4. We show that the ReX bound is sharp when the correlation matrices are uniformly randomly generated in Section 3. In this case, we study the ReX trichotomy phenomenon of the \( \ell_\infty \) norm of Gaussian vectors in Section 3.1 and we propose an inference method for low-dimensional linear dependency in a high-dimensional dataset in Section 3.2. In Section 4 we consider the case when we observe the data with noise. We illustrate through simulations both the ReX trichotomy phenomenon and the performance of the ReX inference in Section 5. In Section 6 we conclude our findings and discuss possible future work along this direction. We provide proofs of the key theorems in the Appendix.

### 2 Sphere Packing Bounds and the ReX Asymptotic Bound

#### 2.1 Universal Asymptotic Bound for the Maximal Correlation to a Uniform Direction

Our key observation is that for any \( \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d) \), if we consider the spherical coordinates, then we have \( \mathbf{Z} = \|\mathbf{Z}\|_2 \mathbf{U} \) where \( \|\mathbf{Z}\|_2 \sim \chi_d \) and \( \mathbf{U} \sim \text{Unif}(S^{d-1}) \). Moreover, it is well-known that \( \|\mathbf{Z}\|_2 \) and \( \mathbf{U} \) are independent. Therefore,

\[
\max_j |X_j| = \max_j |l_j^T \mathbf{Z}| = \|\mathbf{Z}\|_2 \max_j |l_j^T \mathbf{U}|.
\]  

(2.1)

Thus, since the distribution \( \chi_d \) is well known, as long as we understand the distribution of \( \max_j |l_j^T \mathbf{U}| \), we will understand that of the absolute extremes \( \max_j |X_j| \).

An asymptotic upper bound for \( \max_j |l_j^T \mathbf{U}| \) is summarized below.
**Theorem 2.1 (Universal Bound for the Maximal Correlation)** For arbitrary unit vectors \( l_1, \ldots, l_p \) and a uniformly distributed unit vector \( U \) over \( S^{d-1} \), the random variable \( \max_j |l_j^T U| \) satisfies that for any fixed \( \epsilon > 0 \), uniformly for any \( d \geq 2 \),

\[
\lim_{p \to \infty} P \left( \max_j |l_j^T U|^2 / (1 - p^{-2/(d-1)}) > 1 + \epsilon \right) = 0. \tag{2.2}
\]

In particular, if \( d \to \infty \), then

\[
\lim_{p,d \to \infty} P \left( \max_j |l_j^T U|^2 / (1 - p^{-2/(d-1)}) \leq 1 \right) = 1. \tag{2.3}
\]

We shall show the sharpness of this bound in Section 3.1. Note that the inner product \( |l_j^T U| \) is the cosine of the angle between \( l_j \) and \( U \). Therefore, a geometric interpretation of the theorem above is that: If we have \( p \) directions in a \( d \)-dimension space, regardless of where and how they are located, the minimum acute angle between these directions to a uniformly random direction is at least \( \cos^{-1} \sqrt{1 - p^{-2/(d-1)}} \). Interestingly, this fact indicates that a lower bound for the sine of the smallest angle between any collection of \( p \) directions and a uniformly random direction in \( \mathbb{R}^d \) is about \( O(1/p^{1/(d-1)}) \). One intuitive explanation of this interesting angle is from the fact that the surface area of a spherical cap with half-angle \( \theta \) on \( S^{d-1} \) is in the order of \( \theta^d \). Thus, with \( p \) points on the sphere, the minimal angle between a uniform direction and these points should be in the order of the \( (d-1) \)-th root of \( p \).

We note also that the bound holds uniformly for any \( d \geq 2 \), which includes the case when the dimension \( d \) is larger than the number of vectors \( p \). In fact, the bound \( \sqrt{1 - p^{-2/(d-1)}} \) behaves differently according to whether the ratio \( d / \log p \) converges to 0, a fixed positive \( \beta \), or \( \infty \):

(i) If \( \lim_{p \to \infty} d / \log p = 0 \), then as \( p \to \infty \),

\[
\sqrt{1 - p^{-2/(d-1)}} \to 1. \tag{2.4}
\]

(ii) If \( \lim_{p \to \infty} d / \log p = \beta \) for fixed \( 0 < \beta < \infty \), then as \( p \to \infty \),

\[
\sqrt{1 - p^{-2/(d-1)}} \to \sqrt{1 - e^{-2/\beta}}. \tag{2.5}
\]

(iii) If \( \lim_{p \to \infty} d / \log p = +\infty \), then as \( p \to \infty \),

\[
\sqrt{1 - p^{-2/(d-1)}} \to 0. \tag{2.6}
\]
We can see that the above trichotomy cover the complete range of [0, 1].

As an intuitive explanation from the view of random packing, the above trichotomy can be attributed to two facts: (1) The more “evenly” distributed the \( l_j \)’s are on the sphere, the larger the \( \max_j |l_j^T U| \) is; (2) the number of orthants in \( \mathbb{R}^d \) grows exponentially as \( d \) grows. From (1), we consider the situation when we have an “evenly distributed mesh” of \( p \) points \( l_j \)’s on the sphere \( S^{d-1} \). Suppose we now generate a uniformly random direction \( U \). By (2), if the size of the “mesh” grows slower than the exponential rate, then the points on the “mesh” are so “sparse” that a random direction \( U \) can be almost orthogonal to them; if the size of the “mesh” does grow exponentially as the rate of the number of orthants, then a random direction \( U \) will stay around some angle to the points on the “mesh”; on the other hand, if the size of the “mesh” grows faster than the exponential rate, then the points on the “mesh” are so “dense” that a random direction will “hit” the mesh with high probability. Thus, the case \( d/\log p \to \beta \) is the balanced case in which \( \max_j |l_j^T U| \) converges to a constant between 0 and 1.

We shall note that the bound holds uniformly for any \( d \geq 2 \), which includes the case when the dimension \( d \) is larger than the number of vectors \( p \). In particular, in the limit \( d/\log p \to \infty \), this random direction will become almost orthogonal to any collection of directions, which includes all axes! This limit is consistent with the well-known result that a random direction is orthogonal to almost everything in high-dimensional Euclidean spaces.

### 2.2 A Test of the Overall Significance in Regressions

Consider a regression model with \( n \) i.i.d observations of \( p \) explanatory variables (\( p \) may be bigger than \( n \)) and one response variable. Suppose the distribution of the error in the regression model is spherically symmetric. The problem of the overall significance of the model is to test whether the coefficients of the explanatory variables are all zeros. In this case, we can view the problem as observing \( p+1 \) vectors in \( \mathbb{R}^n \). To apply Theorem 2.1, we standardize each vector so that they all have length 1 as directions \( l_1, \ldots, l_p \) for the covariates and \( U \) for the response variable. If the explanatory variables are “irrelevant” to the response, so that the distribution of \( U \) does not depend on \( l_1, \ldots, l_p \) and is uniform over \( S^{n-1} \), then Theorem 2.1 says the followings: no matter where and how the \( l_j \)’s are located, the minimal angle between these \( l_j \)’s and \( U \) is at least \( \cos^{-1} \sqrt{1 - p^{-2/(n-1)}} \), and the maximal correlation between the response and the covariates is at most \( \sqrt{1 - p^{-2/(n-1)}} \). This fact thus leads to a test of the overall significance in
regressions: The explanatory variables in the model are considered to be “relevant” to the response variable if and only if their maximal correlation is above $\sqrt{1 - \frac{p^2}{n-1}}$.

The geometric interpretation of this threshold rule is easy to see: Since the upper bound is for a uniformly distributed $U$ and arbitrary $l_1, \ldots, l_p$, if the correlations between $l_j$'s and $U$ are all below the threshold $\sqrt{1 - \frac{p^2}{n-1}}$, there is hardly evidence to believe that $U$ is different from a uniformly random direction to the $l_j$'s. Thus there is hardly evidence to believe that the response is correlated to the covariates in the regression model.

One interesting fact to note here is that if $\frac{n}{\log p} \to \infty$ as $p \to \infty$, then this threshold is approximately $\sqrt{\frac{2 \log p}{n}}$, which is the universal threshold [16]. Thus, the proposed threshold can be regarded as a generalization of the universal threshold for any $n$ and $p$. Moreover, it can be seen that this method works even when $n$ is finite and $p$ is large. Therefore, the threshold at $\sqrt{1 - \frac{p^2}{n-1}}$ provides a simple test of the overall significance in a regression model for finite samples with large $p$. We will report more details about this test in a future paper.

2.3 ReX Universal Asymptotic Bound of the Maximum Norm of Gaussian Vectors

Since $\max_j |X_j| = \max_j |l_j^T Z| = \|Z\|_2 \max_j |l_j^T U|$ and $\|Z\|_2/\sqrt{d} \to 1$ as $d \to \infty$, we derive the following ReX universal asymptotic bound for the $\ell_\infty$ norm of Gaussian vectors which holds uniformly over all correlation matrices.

**Theorem 2.2 (ReX Universal Bound for Gaussian Extremes)** For any vector of $p$ standard Gaussian variables $X \sim \mathcal{N}(0, \Sigma)$ with $\text{rank}(\Sigma) = d$, the random variable $\|X\|_\infty = \max_j |X_j|$ satisfies that for any fixed $\delta > 0$,

$$
\lim_{p,d \to \infty} P \left( \frac{\|X\|_\infty}{\sqrt{d(1 - \frac{p^2}{(d-1)})}} > 1 + \delta \right) = 0. \quad (2.7)
$$

In particular, if $\lim_{p \to \infty} d(\log \log p)^2/(\log p)^2 \to \infty$, then

$$
\lim_{p \to \infty} P \left( \frac{\|X\|_\infty}{\sqrt{d(1 - \frac{p^2}{(d-1)})}} \leq 1 \right) = 1. \quad (2.8)
$$

We shall show the sharpness of this bound in Section [3.1]. This theorem implies that for large $p$ and $d$, an arbitrary $p$-dimensional Gaussian vector will essentially stay within a hypercube which centers at the origin and has
a radius of $\sqrt{d(1 - p^{-2/(d-1)})}$, where $d$ is the rank of the correlation matrix. Geometrically speaking, this fact together with the fact that the $\ell_2$ norm of a $p$-dimensional Gaussian vector is about $\sqrt{p}$, indicate that a low-rank Gaussian vector will be very likely to appear around the “corners” of the hypercube where it intersects with the sphere of radius $\sqrt{p}$. Algorithms in search for low-dimensional structure can be designed based on this geometric observation.

We should also note that the function $\sqrt{d(1 - p^{-2/(d-1)})}$ is monotone increasing in both $d$ and $p$. Hence, the ReX bound confirms our intuition that the higher the rank is, the higher the extreme value of Gaussians can reach. On the other hand, if the extreme value is high, then its rank cannot be low. Thus Theorem 2.2 allows fast detection of a low-rank: If the maxima in the samples are large, then the underlying rank must be high or full; on the other hand, if the maxima are small, then the full rank assumption is questionable and the small magnitude of the maxima may be due to an underlying low-rank structure.

Another observation of the bound is that it takes on a trichotomy phenomenon too for different ranges of $d$, as its sphere packing counterpart:

(i) If $\lim_{p\to\infty} d/\log p = 0$, then as $p \to \infty$,  
$$
\sqrt{d(1 - p^{-2/(d-1)})} \sim \sqrt{d}. 
$$

(ii) If $\lim_{p\to\infty} d/\log p = \beta$ for fixed $0 < \beta < \infty$, then as $p \to \infty$,  
$$
\sqrt{d(1 - p^{-2/(d-1)})} \sim \sqrt{\beta(1 - e^{-2/\beta}) \log p}. 
$$

(iii) If $\lim_{p\to\infty} d/\log p = +\infty$, then as $p \to \infty$,  
$$
\sqrt{d(1 - p^{-2/(d-1)})} \sim \sqrt{2 \log p}. 
$$

From the above trichotomy, we can see that the bound $\sqrt{d(1 - p^{-2/(d-1)})}$ indeed cover the entire scope from constants to $\sqrt{2 \log p}$. In particular, since the range of the function $\sqrt{\beta(1 - e^{-2/\beta})}$ is $(0, 2)$ for $\beta \in (0, \infty)$, this case can be regarded as the intermediate step between the very low rank limits and the full rank limit $\sqrt{2 \log p}$. The bound $\sqrt{d(1 - p^{-2/(d-1)})}$ explicitly describes the relationship between the possible extreme value from a Gaussian vector and the rank of its covariance matrix. Due to the dependence of the bound of extremes and the rank, we call this phenomenon the ReX trichotomy.
We should note that the ReX asymptotic bound is not always sharp for all covariance matrices. For example, consider the equicorrelation matrices with 1’s in the diagonal entries and \( \rho \)'s as off diagonals. This correlation matrix has full rank \( p \) for any \(-1/(p - 1) < \rho < 1\). On the other hand, as \( \rho \) approaches 1, the extreme from the multivariate Gaussian with this correlation matrix is essentially \( O_p(1) \). Nevertheless, the ReX asymptotic bound reveals the fact that the extremes from Gaussian vectors contains much information about the rank, and we can utilize this information in search of low-rank structure.

2.4 Application to Post-Selection Inference

Statistical inference after variable selection in linear models has recently gained great interest. This interest is sparked by both the prevalence of data-driven variable selections and the severe bias that the selection process may impact on the inference in the selected submodel. For an excellent review on variable selection procedures, see [17]. For recent discussions on the problems of the conventional inference after variable selection, see [18], [19] and [20].

The PoSI method in [15] provides one way to protect the statistical inference in the selected model by considering simultaneous inference on the partial slopes in selected models. The simultaneity here is over the set of submodels that may be possible selected. Thus, the resulting inference based on the simultaneity is valid and conservative. Moreover, the important advantage of such inference is that it is universally valid over any data-driven variable selection rules. This robustness over selection procedures makes the PoSI method a practical way to achieve valid inference.

The PoSI inference is based on the PoSI constant which is essentially the maximal \( t \)-statistics over all possible combinations of variables and submodels. For a regression model with \( p \) explanatory variables, this set of all possible combinations consists of \( L = p^{2p-1} \) \( t \)-statistics. When the variance of the noise is known or when \( n - p \) is large, the \( t \)-statistics become standard Gaussian variables. Also, it is easily seen that the rank of the correlation matrix of these \( L \) Gaussian variables is at most \( p \). Thus, by Theorem 2.2, the ReX bound of the PoSI constant for any design matrix grows at the rate of

\[
\sqrt{p(1 - L^{-2/(p-1)})} \sim \sqrt{\frac{3}{4}p } = 0.866\sqrt{p} \tag{2.12}
\]
as \( p \rightarrow \infty \), as developed in [15].
More importantly, in practice people may set some integer \(1 \leq m \leq p\) and select a model with a size less than or equal to \(m\). When the number of variables in a selected submodel is restricted by \(m\), then combinatorics easily show that we have a total of \(O((\frac{m}{p})^m)\) Gaussian variables. Therefore, by Theorem 2.2, the ReX asymptotic bound for PoSI constants in this case, \(K_m\), follows that

\[
K_m \sim O\left(\sqrt{\frac{m \log p}{m}}\right)
\]

(2.13)

for any design matrix. Hence, if we only consider submodels with sizes \(m\) such that \(m/p \to 0\), then the PoSI method can be much more efficient than its upper bound rate \(0.866\sqrt{p}\). Some special cases of small \(m\)'s include:

1. \(m = C\) is a constant, \(K_m = O(\sqrt{\log p})\).
2. \(m = \log p\), \(K_m = O(\log p)\).
3. \(m = p^\gamma\) for fixed \(\gamma < 1\), \(K_m = O(\sqrt{p^\gamma \log p})\).

3 Random Packing and the Attainment of the ReX Asymptotic Bound

Once the ReX asymptotic bound is developed, it is of interest whether it is sharp. To this end, we study the following model in this section: The \(j\)-th variable \((1 \leq j \leq p)\) in the \(i\)-th observation \((1 \leq i \leq n)\), \(X_{ij}\), is generated as

\[
X_{ij} = l_j^T Z_i
\]

where

1. \(l_j\)'s are i.i.d. uniformly distributed over the \((d-1)\)-sphere \(S_{d-1}\), \(\forall j\).
2. \(Z_i \sim \mathcal{N}(0_d, I_d)\) and is independent of \(l_j\)'s.

To fix ideas, for now we set \(n = 1\) and assume that we only have one observation of many variables. Thus, in this section we suppress the subscript \(i\) and write \(X_{ij}\) as \(X_j\).

3.1 The ReX Trichotomy

The next theorem shows that in this case, the asymptotics magnitude of \(\max_j |l_j^T U|\) attains the upper bound:
Theorem 3.1 (Asymptotic Rate for the Maximal Uniform Correlation)
If \( l_j' \)'s and \( U \) are i.i.d. uniformly distributed over the \( (d-1) \)-sphere \( S^{d-1} \), \( \forall j \), then uniformly for all \( d \geq 2 \), as \( p \to \infty \),

\[
\max_j |l_j^T U| \sqrt{1 - p^{-2/(d-1)}} \xrightarrow{\text{prob}} 1.
\] (3.1)

This result shows the limits of \( \max_j |l_j^T U| \) when \( l_j' \)'s are i.i.d. uniformly distributed over the sphere:

(i) If \( \lim_{p \to \infty} d / \log p = 0 \), then

\[
\max_j |l_j^T U| \xrightarrow{\text{prob}} 1.
\] (3.2)

(ii) If \( \lim_{p \to \infty} d / \log p = \beta \) for fixed \( 0 < \beta < \infty \), then

\[
\max_j |l_j^T U| \xrightarrow{\text{prob}} \sqrt{1 - e^{-2/\beta}}.
\] (3.3)

(iii) If \( \lim_{p \to \infty} d / \log p = +\infty \), then

\[
\max_j |l_j^T U| \xrightarrow{\text{prob}} 0.
\] (3.4)

The theorem is proved through a random packing argument that is inspired by [21]. We also note here that a parallel version of the above asymptotics were independently developed in [22] and [23], in a related but different context. In these papers, the limiting distribution of \( \max_j |l_j^T U| \) is carefully studied. Therefore, we shall not go into further details of the convergence, but concentrate on the application of the limit.

The immediate application of this result is the uniform ReX rate of \( \|X\|_\infty = \max_j |X_j| \):

Theorem 3.2 (ReX Rate for Gaussian Extremes with Uniform Correlations)
If \( X \sim \mathcal{N}(0, \Sigma) \) where \( \Sigma_{ij} = l_i^T l_j \) with \( l_j' \)'s are i.i.d. uniformly distributed over the \( (d-1) \)-sphere \( S^{d-1} \), \( \forall j \), then as \( d \to \infty \) and \( p \to \infty \),

\[
\|X\|_\infty / \sqrt{d(1 - p^{-2/(d-1)})} \xrightarrow{\text{prob}} 1.
\] (3.5)

The ReX trichotomy with i.i.d. \( l_j' \)'s shows the following phase transition of \( \max_j |X_j| \):
(i) If $d \to \infty$ and $\lim_{p \to \infty} \frac{d}{\log p} = 0$, then
\[
\|X\|_{\infty}/\sqrt{d} \xrightarrow{\text{prob.}} 1.
\] (3.6)

(ii) If $\lim_{p \to \infty} \frac{d}{\log p} = \beta$ for fixed $0 < \beta < \infty$, then
\[
\|X\|_{\infty}/\sqrt{\log p} \xrightarrow{\text{prob.}} \sqrt{\beta(1 - e^{-2/\beta})}.
\] (3.7)

(iii) If $\lim_{p \to \infty} \frac{d}{\log p} = +\infty$, then
\[
\|X\|_{\infty}/\sqrt{2 \log p} \xrightarrow{\text{prob.}} 1.
\] (3.8)

For the case $d \sim \beta \log p$, the equation
\[
\beta(1 - e^{-2/\beta}) = 1
\] (3.9)
is of our particular interest. We should notice that the function $\beta(1 - e^{-2/\beta})$ is a monotone increasing function in $\beta$. Therefore, the solution to (3.9) is unique and it is $\beta^\dagger = 1.255005\ldots$. From (3.7), we see that if $d^\dagger = d^\dagger(p) = \beta^\dagger \log p$, then $\|X\|_{\infty}/\sqrt{\log p} \to 1$ in probability. In fact, $d^\dagger$ is the only asymptotic rate of $d$ such that $\|X\|_{\infty}/\sqrt{\log p} \xrightarrow{\text{prob.}} 1$. If $d/\log p \to \beta$ with $\beta < \beta^\dagger$, then for large $p$, $\|X\|_{\infty} < \sqrt{\log p}$ with probability tending to 1. On the other hand, if $d/\log p \to \beta$ with $\beta > \beta^\dagger$, then for large $p$, $\|X\|_{\infty} > \sqrt{\log p}$ with probability tending to 1. Indeed, for $\beta \to +\infty$, $\beta(1 - e^{-2/\beta}) \to 2$, hence we are going back to the classical extreme value result of $\sqrt{2 \log p}$.

The above facts provide us with the basis of a nature classification of low-ranks. In fact, if we simply set a threshold at $\sqrt{\log p}$, then by (3.7), if $d$ grows faster than $\log p$ or $d$ grows as $\beta \log p$ with $\beta > \beta^\dagger$, then with high probability $\|X\|_{\infty}$ will be above $\sqrt{\log p}$; on the other hand, if $d$ grows slower than $\log p$ or $d$ grows as $\beta \log p$ with $\beta < \beta^\dagger$, then with high probability $\|X\|_{\infty}$ will go below $\sqrt{\log p}$. With Theorem 3.2, we are able to make these judgements accurately as long as $p$ is large, based on only a few observations.

Based on the magnitude of the corresponding extremes, we call the case $\limsup_{p \to \infty} d/\log p < \beta^\dagger$ the “super-low rank” case, we call the case $\lim_{p \to \infty} d/\log p = \beta^\dagger$ the “exact-low rank” case, and we call the case when $\liminf_{p \to \infty} d/\log p > \beta^\dagger$ the “moderately-low rank” case. The solution to (3.9), $\beta^\dagger = 1.255005\ldots$, will be called the ReX separation constant.
3.2 ReX Inference on Low Dimensional Random Linear Dependency

In Section 3.1, we study the case when both \( p \to \infty \) and \( d \to \infty \). We now turn to the another question: if \( p \) is large and \( d \) is just a fixed number, can we make inference about \( d \)? The theorem below answers this question with a confirmation. The following ReX inference is a simple application of the Slutsky Theorem: Since \( \max_j |X_j| = ||Z||_2 \max_j |l_j^T U| \) and \( \max_j |l_j^T U|/\sqrt{1 - p^{-2/(d-1)}} \to 1 \) in probability as \( p \to \infty \), the limiting distribution of \( ||X||_\infty \) is \( \chi_d \).

**Theorem 3.3 (ReX Inference)** If \( X \sim N(0, \Sigma) \) where \( \Sigma_{ij} = l_i^T l_j \) with \( l_j \)’s are i.i.d. uniformly distributed over the \((d-1)\)-sphere \( S^{d-1} \), \( \forall j \), and if \( d \) is fixed as \( p \to \infty \), then

\[
\frac{||X||_\infty}{\sqrt{1 - p^{-2/(d-1)}}} \xrightarrow{\text{dist.}} \chi_d. \tag{3.10}
\]

With this theorem, we convert the inference about \( d \) to a simple inference problem on the degrees of freedom of a \( \chi \) distribution. It is important to note here that the asymptotics in this theorem are in \( p \) rather than in the sample size \( n \). Thus, the classical finite sample theories can be directly applied here. In this case, many inference methods about \( d \) are readily available in literature, either in the context of the \( \chi \)-distribution or as a problem of inference about the shape parameter in the Gamma distribution. Furthermore, we can utilize the square-root transformation as the variance stabilizing transformation for \( \chi^2_d \)-distribution (Section 3.2. [24]) to achieve approximate confidence intervals for moderate \( n \). Since these inference procedures have been well developed, we shall not go into further theoretical details of the specific inference problems.

The major advantage of this ReX inference is that it is finite sample inference. Thus, it allows fast detection of the low-rank. In practice, we may face datasets with very large \( p \) and completely unknown correlation structure. In this case, we may assume that the correlations are i.i.d. uniformly distributed. By quickly checking the maximal entry in each sample and applying Theorem 3.3, we may get a good sense of the fixed rank. In particular, the ReX inference can be made with \( n < d \), when the eigenvalue thresholding procedures will fail. In the extreme, the inference can be made with only one sample. Thus, fast detection of low-rank is possible through this approach.

It is also noteworthy that the asymptotics in \( p \) indicates that the higher \( p \) is, the more accurate the inference is. Thus, instead of the “curse of
dimensionality,” the ReX inference can be viewed as a “blessing of dimensionality.”

4 Observation with Small Errors

In this section we consider the case when our observations may be subject to measurement errors, i.e., when we do not observe $X = L^T Z$ directly but instead observe $Y = X + \epsilon$, where $\epsilon = (\epsilon_1, \ldots, \epsilon_p)^T$ with independent $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$. In this case, although the rank of the correlation matrix of $X$ is $d$, the covariance matrix of $Y$ is of rank $p$. Nevertheless, we show that as long as the noise is small, we can still draw information about $d$ by utilizing the extremes.

To see the argument above, we notice that if we let $\sigma_{max} = \max_j \sigma_j^2$, then $\max_j \epsilon_j = O_p(\sigma_{max} \sqrt{\log p})$. Thus, as long as $\sigma_{max} = o(\sqrt{d(1 - p^{-2/(d-1)})/\log p})$ as $p \to \infty$, then $\max_j \epsilon_j = o_p(\sqrt{d(1 - p^{-2/(d-1)})})$. In this case, the asymptotic results in $Y$ are identical to those in $X$:

**Corollary 4.1** For any vector of $p$ standard Gaussian variables $X \sim \mathcal{N}(0, \Sigma)$ with rank$(\Sigma) = d$, and for $Y = X + \epsilon$ where $\epsilon = (\epsilon_1, \ldots, \epsilon_p)^T$ with independent $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$, if $\sigma_{max} = o(\sqrt{d(1 - p^{-2/(d-1)})/\log p})$, then for any fixed $\delta > 0$,

$$\lim_{p,d \to \infty} P\left(\frac{\|Y\|_\infty}{\sqrt{d(1 - p^{-2/(d-1)})}} > 1 + \delta\right) = 0. \tag{4.1}$$

In particular, if $\lim_{p \to \infty} d(\log \log p)^2/(\log p)^2 \to \infty$, then

$$\lim_{p \to \infty} P\left(\frac{\|Y\|_\infty}{\sqrt{d(1 - p^{-2/(d-1)})}} \leq 1\right) = 1. \tag{4.2}$$

We also have the corollary on the sharpness of the bound.

**Corollary 4.2** If $\sigma_{max} = o(\sqrt{d(1 - p^{-2/(d-1)})/\log p})$, and if $l_j$'s are i.i.d. uniformly distributed over the $(d - 1)$-sphere $S^{d-1}$, $\forall j$, then as $d \to \infty$ and $p \to \infty$,

$$\|Y\|_\infty/\sqrt{d(1 - p^{-2/(d-1)})} \xrightarrow{\text{prob.}} 1. \tag{4.3}$$

The ReX inference is also valid in $Y$ for a finite $d$:

**Corollary 4.3** If $l_j$'s are i.i.d. uniformly distributed over the $(d - 1)$-sphere $S^{d-1}$, $\forall j$, if $d$ is fixed as $p \to \infty$, and if $\sigma_{max} = o(\sqrt{d(1 - p^{-2/(d-1)})/\log p})$, then

$$\|Y\|_\infty/\sqrt{1 - p^{-2/(d-1)}} \xrightarrow{\text{dist.}} \chi_d. \tag{4.4}$$

14
5 Empirical Studies

5.1 Study of the ReX Trichotomy

We study 60,000 simulated datasets. We set \( p = 3000 \), and consider four different deficient rank scenarios: \( d_1 = 3 \), \( d_2 = 10 \), \( d_3 = 100 \) and \( d_4 = 300 \). The data are generated as in Section 3. We also consider the case when the \( X_j \)'s are i.i.d., so that \( d_5 = p = 3000 \). Note here \( \beta^* \log p = 10.048 \approx 10 \). So \( d_1 \) is about a fixed rank or super-low rank case, \( d_2 \) is about an exact-low rank case, while \( d_3 \) and \( d_4 \) are moderately-low rank cases. To emphasize the dependence of the distribution of extremes on the rank, we shall denote \( K_{p,d} = \| X \|_\infty \) if the rank of the covariance matrix of \( X \) is \( d \).

We first show the plot of kernel density estimates (using R default bandwidth) of the \( K_{p,d} \)'s for different \( d \)'s in Figure 1 below.

![Figure 1: Comparison of the distributions of \( K_{p,d} \) for \( p = 3000 \). The brown vertical line at the threshold of \( \sqrt{\log p} \) almost separate the distribution \( K_{p,d_1} \) in the super-low rank case from the distributions \( K_{p,d_3} \) and \( K_{p,d_4} \) in the moderately-low rank case.](image)

From the picture, we can clearly see the ReX trichotomy. If we set a threshold at \( \sqrt{\log p} \) (shown as the brown vertical line), then

1. for the super-low rank case \( d_1 \), the distribution of \( K_{p,d_1} \) is almost all on the left hand side of the threshold of \( \sqrt{\log p} \). It can be also easily checked that the distribution of \( K_{p,d_1} \) matches the \( \chi_3 \) density very well;
2. for the exact-low rank case \( d_2 \), the distribution of \( K_{p,d_2} \) is around \( \sqrt{\log p} \);
3. for the moderately-low rank case \( d_3 \) and \( d_4 \), the distributions of \( K_{p,d} \) are almost on the right hand side of the threshold of \( \sqrt{\log p} \), and they have
similar distributions even though the ranks substantially differ. Their distributions are also close to the distribution of $\max_j |X_j|$ when $X_j$’s are i.i.d. Gaussian.

From Figure 1 we see that there is a small right tail of the distribution $K_{p,d_1}$ that is on the right hand side of the threshold of $\sqrt{\log p}$ and there are tiny left tails of the distributions $K_{p,d_3}$ and $K_{p,d_4}$ that are on the left hand side of the threshold. Thus, the probability of correct classification here is close to 1 but not exactly 1. However, note only one sample ($n = 1$) has been considered by far. With a few more samples our classification can be much more accurate. Now consider we observe $n$ samples $X_{ij}$, $i = 1, \ldots, n$, and $j = 1, \ldots, p$. If we let $K_{i,p,d} = \|X_i\|_\infty = \max_j |X_{ij}|$ and simply look at the averages, $\bar{K}_{p,d} = \frac{1}{n} \sum_{i=1}^n K_{i,p,d}$, then we can do much better.

Figure 2 below illustrates the case $n = 30$. When we compare the distributions of $\bar{K}_{p,d}$ with 2,000 simulated datasets, we see that we have probability one for complete separation of the super-low rank case and moderately-low rank case. These categories can be classified by $K_{p,d}$ for large $p$ as described above.

![Figure 2: Comparison of the distributions of $\bar{K}_{p,d}$ for $p = 3000$ and $n = 30$. The distributions are more convergent, and the brown vertical line at the threshold of $\sqrt{\log p}$ separates the distribution $\bar{K}_{p,d_1}$ in the super-low rank case from the distributions $\bar{K}_{p,d_3}$ and $\bar{K}_{p,d_4}$ in the moderately-low rank case.](image)

We note here that since in the low-dimensional case the distribution of $K_{p,d}$ is approximately $\chi_d$, for which the sufficient statistics are $\sum_{i=1}^n \log K_{i,p,d}$, the simple average $\bar{K}_{p,d}$ is not even admissible as a point estimator. Therefore, although $\bar{K}_{p,d}$ is doing a great job, even better procedures should be available. Since our focus is on showing only the possibility of this asymptotic classification, we will not go into further details of this problem.
5.2 Study of the Coverage of the ReX Confidence interval

In this section we study the coverage probability of the ReX confidence interval proposed in Section 3.2. We set $p = 3000$ or $15000$ for which $\beta^\dagger \log p = 10.05$ or $12.06$ respectively. We then study the cases when $d$ ranges from 5 to 12 respectively. We make $n < d$ by setting $n$ to be the rounded number of $0.8d$. We simulated 1,000 datasets for each scenario. The confidence intervals are built through a slightly conservative variant of the square-root transformation of $K_{i,p,d}^2$: Since the asymptotic variance of $K_{p,d}^2 = \sum_{i=1}^n K_{i,p,d}^2/n$. However, we match the mean of $K_{p,d}^2$ at $d(1 - p^{-2/(d-1)})$. Thus, the lower and upper ends of the $(1 - \alpha)$-confidence intervals, $d_l$ and $d_u$, are obtained by solving equations

$$\sqrt{d_l(1 - p^{-2/(d_l-1)})} = \sqrt{K_{p,d}^2 + z_{1-\alpha/2}/\sqrt{2n}}$$

and

$$\sqrt{d_u(1 - p^{-2/(d_u-1)})} = \sqrt{K_{p,d}^2 - z_{1-\alpha/2}/\sqrt{2n}}$$

respectively. The coverage probabilities of these confidence intervals are summarized in Table 1.

Table 1: Coverage of 95% ReX confidence intervals for different $d$’s when $p_1 = 3000$ and $p_2 = 15000$.

| $d$ | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $p_1$ | 0.949 | 0.954 | 0.956 | 0.959 | 0.951 | 0.952 | 0.964 | 0.963 |
| $p_2$ | 0.954 | 0.955 | 0.963 | 0.964 | 0.963 | 0.963 | 0.972 | 0.964 |

Despite the difficult fact that the sample size $n < d$ when PCA based methods would fail, the coverage probabilities of these confidence intervals are satisfactory. This simulation result of the ReX method shows the feasibility of the finite sample inference for low-dimensional structures in high-dimensional datasets.

6 Discussions

This article provides an asymptotic upper bound, called the rank-extreme (ReX) bound, for arbitrary Gaussian distributions as $\sqrt{d(1 - p^{-2/(d-1)})}$. These results describe the relationship between $\ell_\infty$ norms of Gaussian vectors and the ranks of their covariance matrices. We then use the ReX association reversely towards the inference problem of low-dimensional linear dependency in high-dimensional data when the correlations are generated by inner products of i.i.d. uniformly distributed unit vectors. We show that
if $d$ is fixed, it is possible to make inference about $d$. We also show that if $d$ grows along with $p$, it is possible to compare $d$’s rate of growth with $\log p$. All the asymptotic results are in $p$ but not in $n$, so that fast detection and finite sample inference of a low-rank structure is possible by extracting the information contained in the maximum norm of Gaussian vectors.

It is important to note that extreme values as statistics are well-known to be not resistant to outliers. The asymptotic results in this article rely on the Gaussian assumption too. Therefore, the robustness of the inference needs to be carefully studied. We would also like to generalize the Gaussian assumptions to more general cases, for example, spherically symmetric distributions.

We should note that in this paper we found the ReX bound for the $\ell_\infty$ norm of Gaussian vectors with arbitrary correlation. We are interested in more accurate upper and lower bound for Gaussian vectors with special correlation structures. A related direction is that in this paper we consider ReX inference under uniformly distributed $l_j$’s. It would be interesting to study the ReX inference under a more general distribution of $l_j$’s, or under the framework of PCA models. The optimality of this approach needs to be studied as well. We hope to borrow from Bayesian methods in these studies.

One indication of the asymptotic rate $\sqrt{d(1 - p^{-2/(d-1)})}$ is that when $l_j$’s are i.i.d. uniformly distributed over the $(d-1)$-sphere $S^{d-1}$, we may estimate the rank of the correlation matrix $d$. This estimator $\hat{d}$ is easily defined as the rounded number of the solution to the equation

$$d(1 - p^{-2/(d-1)}) = \|X\|_\infty^2. \quad (6.1)$$

Since the function $d(1 - p^{-2/(d-1)})$ is monotone increasing in $d$, $\hat{d}$ is unique. We will study the asymptotic performance of this estimator in more details in a future paper.

There is an interesting fact that the case $d \sim \beta \log p$ covers the entire scope of that $\|X\|_\infty$ can take for arbitrary correlation matrices. Indeed, the function $\beta(1 - e^{-2/\beta})$ is monotone increasing in $\beta$ and its range is from 0 to 2. Therefore, it seems that we could concentrate on the case $d \sim \beta \log p$ and look at the inference about $\beta$, for example, the testing problem of $H_0 : \beta \leq \beta^\dagger$ or $H_1 : \beta > \beta^\dagger$.

The above fact also seems to be remotely related to the Johnson-Lindenstrauss Theorem [25]. In the J-L Theorem, it is shown that a random projection from $\mathbb{R}^p$ to a subspace of dimension $O(\log p)$ can well preserve the $\ell_2$ norm of any vector in $\mathbb{R}^p$. It would be interesting to see how well such a random projection can preserve the $\ell_\infty$ norm of the vector.
Acknowledgment

The author appreciates the great suggestions from L. D. Brown, A. Buja, T. Cai, J. Fan, J. S. Marron, and H. Shen. The author also thanks J. Berger, R. Berk, A. Budhiraja, E. Candes, L. de Haan, J. Galambos, E. George, J. Hannig, T. Jiang, I. Johnstone, A. Krieger, R. Leadbetter, D. Lin, J. Liu, W. Liu, Y. Liu, Z. Ma, E. Pitkin, S. Provan, D. Small, L. Wang, D. Zeng, L. Zhao, and C. Zong for very helpful discussions. The author is particularly grateful for L. A. Shepp for his inspiring introduction of the random packing literature.

A Proof of Theorem 2.1

To show (2.2), it is enough to show that for any fixed \( \epsilon \) that \( 0 < \epsilon < 1/(p^{2/(d-1)} - 1) \),

\[
\mathbb{P}\left( \max_j |l_j^T U| > \sqrt{(1 + \epsilon)(1 - p^{-2/(d-1)})} \right) \to 0. \tag{A.1}
\]

We follow the proof of Theorem 6.3 in [15].

For (A.1), we consider the Bonferroni bound that

\[
\mathbb{P}\left( \max_j |l_j^T U| > \sqrt{(1 + \epsilon)(1 - p^{-2/(d-1)})} \right)
\]

\[
= \mathbb{P}\left( \bigcup_j \{ |l_j^T U| > \sqrt{(1 + \epsilon)(1 - p^{-2/(d-1)})} \} \right)
\]

\[
\leq \sum_{j=1}^p \mathbb{P}\left( |l_j^T U| > \sqrt{(1 + \epsilon)(1 - p^{-2/(d-1)})} \right)
\]

\[
= p \mathbb{P}\left( |U| > \sqrt{(1 + \epsilon)(1 - p^{-2/(d-1)})} \right) \tag{A.2}
\]

where \( U \) is any coordinate of \( U \) or projection of \( U \) onto a unit vector. Now we use the fact that \( U^2 \sim \text{Beta}(1/2, (d - 1)/2) \), i.e., for any \( 0 < a < 1 \),

\[
\mathbb{P}[|U| > a] = \frac{1}{B(1/2, (d - 1)/2)} \int_{a^2}^1 x^{-1/2}(1 - x)^{(d-3)/2} dx. \tag{A.3}
\]

Due to the bound that \( \Gamma(x + 1/2)/\Gamma(x) < \sqrt{x} \), for any \( 0 < a < 1 \), we have

\[
\int_{a^2}^1 x^{-1/2}(1 - x)^{(d-3)/2} dx \leq \frac{2}{a(d - 1)}(1 - a^2)^{(d-1)/2}. \tag{A.4}
\]
Now we further bound (A.2) by

\[
pP\left( |U| > \sqrt{(1 + \epsilon)(1 - p^{-2/(d-1)})} \right) \leq p \frac{2}{(d-1)\pi} \frac{1}{\sqrt{(1 + \epsilon)(1 - p^{-2/(d-1)})}} (1 - (1 + \epsilon)(1 - p^{-2/(d-1)}))^{(d-1)/2}
\]

\[
= \frac{2}{\pi(1 + \epsilon) \sqrt{(d-1)(p^{2/(d-1)} - 1)}} \left( 1 - \epsilon(p^{2/(d-1)} - 1) \right)^{(d-1)/2}
\]

\[
\leq \frac{2}{\pi(1 + \epsilon) \sqrt{(d-1)(p^{2/(d-1)} - 1)}} \exp\left( -\frac{1}{2} \epsilon(d - 1)(p^{2/(d-1)} - 1) \right)
\]

(A.5)

Note that uniformly for any \( d \geq 2 \), \((d-1)(p^{2/(d-1)} - 1) \to \infty \) as \( p \to \infty \).

Thus, \( P\left( \max_j |l_j^T U| \right) > \sqrt{(1 + \epsilon)(1 - p^{-2/(d-1)})} \to 0 \) as \( p \to \infty \) regardless of the rate of \( d = d(p) \).

To see (2.3), note that if \( \lim_{p \to \infty} d/\log p = \beta > 0 \), then \( p^{1/(d-1)} \to e^{1/\beta} < \infty \). Thus we may let \( \epsilon \to 0 \) in (A.5) to get (2.3). Also, if \( d \to \infty \) but \( d/\log p \to 0 \), then (A.5) is further bounded by \( \frac{2}{\pi(d-1)}(1 + o(1)) \). Thus we have (2.3).

B Proof of Theorem 2.2

It is easy to show (4.1). To show (4.2), note that for any \( 0 < \epsilon < 1 \),

\[
P\left( \|X\|_\infty/\sqrt{d(1 - p^{-2/(d-1)})} > 1 \right) = P\left( \|Z\| \max_j |l_j^T U|/\sqrt{d(1 - p^{-2/(d-1)})} > 1 \right)
\]

\[
\leq P\left( \max_j |l_j^T U| > \sqrt{(1 - \epsilon)(1 - p^{-2/(d-1)})} \right) + P\left( \|Z\| > \sqrt{(1 + \epsilon)d} \right)
\]

(B.1)

We will show each of the two summands in the last line can be made small with a proper choice of \( \epsilon = \epsilon(p) \).
By the proof of Theorem 2.1, we see that
\[
P \left( \max_j |l_j^T U| > \sqrt{(1 - \varepsilon)(1 - p^{-2/(d-1)})} \right)
\leq \sqrt{\frac{2}{\pi (1 - \varepsilon)}} \frac{p^{1/(d-1)}}{\sqrt{(d - 1)(p^{2/(d-1)} - 1)}} \exp \left( \frac{1}{2} \varepsilon (d - 1) \left( \frac{p^{2/(d-1)}}{2} - 1 \right) \right) \tag{B.2}
\]
Note also that \( \|Z\|^2 \sim \chi_d^2 \). Thus by the Chernoff bound for \( \chi_d^2 \) distribution,
\[
P(\|Z\| > \sqrt{(1 + \varepsilon)d}) = \frac{((1 + \varepsilon)e^{-\varepsilon} - \varepsilon d/2)}{e^{\varepsilon(d - 1)/2}} \leq e^{-d\varepsilon^2/6} \tag{B.3}
\]
Due to (B.2) and (B.3), we let \( \varepsilon = \varepsilon(p) = \log \log p / (4 \log p) \). In the case when \( \lim_{p \to \infty} (\log \log p)^2 d / (\log p)^2 \to \infty \), both (B.2) and (B.3) converge to 0.

### C Proof of Theorem 3.1

Since we already have the upper bound, it is enough to show that for any fixed \( \varepsilon \) such that \( 0 < \varepsilon < 1/2 \),
\[
P \left( \max_j |l_j^T U| < \sqrt{(1 - \varepsilon)(1 - p^{-2/(d-1)})} \right) \to 0. \tag{C.1}
\]
We use the independence of \( l_j \)'s as in [21] and [15] to bound
\[
P \left( \max_j |l_j^T U| < \sqrt{(1 - \varepsilon)(1 - p^{-2/(d-1)})} \right) = \prod_{j=1}^p \left( \frac{2}{\pi (1 - \varepsilon)} \frac{p^{1/(d-1)}}{\sqrt{(d - 1)(p^{2/(d-1)} - 1)}} \exp \left( \frac{1}{2} \varepsilon (d - 1) \left( \frac{p^{2/(d-1)}}{2} - 1 \right) \right) \right). \tag{C.2}
\]
We will lower-bound the absolute value of the exponent in (C.2) by recalling (A.3). We first note that by integrating by parts, we have
\[
\int_{a^2}^1 x^{-1/2} (1 - x)^{(d-3)/2} dx \geq \frac{2}{d^2 - 1} a^{-3} (1 - a^2)^{(d-1)/2} ((d + 2)a^2 - 1) \tag{C.3}
\]
for any $0 < a < 1$. Thus, by the important fact that $\Gamma(x + 1)/\Gamma(x + 1/2) > \sqrt{x + 1/4}$, we have that for $p \to \infty$,

$$p \mathbb{P}\left(|U| > \sqrt{(1 - \epsilon)(1 - p^{-2/(d-1)})}\right)$$

$$= \frac{p}{B(1/2, (d - 1)/2)} \int_{(1-\epsilon)(1-p^{-2/(d-1)})}^{1} x^{-1/2}(1 - x)^{(d-3)/2} dx$$

$$\geq \sqrt{\frac{1}{\pi(1 - \epsilon)^3}} \frac{\sqrt{2d - 3}}{d^2 - 1} \frac{p^{1/(d-1)}}{\sqrt{(p^{2/(d-1)} - 1)^3}}.$$

\begin{equation}
\tag{C.4}
\end{equation}

In the last step of (C.4), we used again the fact that uniformly for any $d \geq 2$, $d(p^{2/(d-1)} - 1) \to \infty$ as $p \to \infty$. It is now easy to see that

$$p \mathbb{P}\left(|U| > \sqrt{(1 - \epsilon)(1 - p^{-2/(d-1)})}\right) \to \infty$$

\begin{equation}
\tag{C.5}
\end{equation}

as $p \to \infty$ regardless of the rate of $d = d(p)$, which completes the proof of Theorem 3.1.

References

[1] I. Markovsky, *Low-Rank Approximation: Algorithms, Implementation, Applications*. New York: Springer, 2012.

[2] B. Yang, *Projection approximation subspace tracking*, IEEE Trans. Signal Process., vol. 43, no. 1, pp. 95-107, Jan. 1995.

[3] R. Badeau, G. Richard, and B. David, *Fast and stable YAST algorithm for principal and minor subspace tracking*, IEEE Trans. Signal Process., vol. 56, no. 8, pp. 3437-3446, Aug. 2008.

[4] S. Bartelmaos and K. Abed-Meraim, *Fast principal component extraction using givens rotations*, IEEE Signal Process. Lett., vol. 15, pp. 369-372, 2008.

[5] X. G. Doukopoulos and G. V. Moustakides, *Fast and stable subspace tracking*, IEEE Trans. Signal Process., vol. 56, no. 4, pp. 1452-1465, Apr. 2008.
[6] A. Kavčič and B. Yang, *Adaptive rank estimation for spherical subspace trackers*, IEEE Trans. Signal Process., vol. 44, no. 6, pp. 1573-1579, Jun. 1996.

[7] D. J. Rabideau, *Fast, rank adaptive subspace tracking and applications*, IEEE Trans. Signal Process., vol. 44, no. 9, pp. 2229-2244, Sep. 1996.

[8] E. C. Real, D. W. Tufts, and J. W. Cooley, *Two algorithms for fast approximate subspace tracking*, IEEE Trans. Signal Process., vol. 47, no. 7, pp. 1936-1945, Jul. 1999.

[9] M. Shi, Y. Bar-Ness, and W. Su, *Adaptive estimation of the number of transmit antennas*, in Proc. IEEE Milit. Commun. Conf., 2007, pp. 1-5.

[10] S. Kritchman and B. Nadler, *Determining the number of components in a factor model from limited noisy data*, Chemomet. Intell. Lab. Syst., vol. 94, pp. 19-32, 2008.

[11] A. B. Owen and P. O. Perry, *Bi-cross-validation of the SVD and the non-negative matrix factorization*, Ann. Appl. Statist., vol. 3, pp. 564-594, 2009.

[12] P. O. Perry and P. J. Wolfe, *Minimax Rank Estimation for Subspace Tracking*, IEEE J. Sel. Topics Signal Process., vol. 4, no. 3, pp. 504–513, Jun. 2010.

[13] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*. New York: Wiley, 1978.

[14] M. R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, N.Y., 1983.

[15] R. Berk, L. D. Brown, A. Buja, K. Zhang, and L. Zhao, *Valid Post-Selection Inference*. Ann. Statist., vol. 41, 802–837, 2013.

[16] D. L. Donoho and I. M. Johnstone, *Ideal Spatial Adaptation by Wavelet Shrinkage*. Biometrika, vol. 81, 425-455, 1994.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning — Data Mining, Inference, and Prediction*. 2nd ed. Corr. 3rd printing. Springer Series in Statistics, New York: Springer, 2009.
[18] H. Leeb and B. M. Pötscher, *Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?* Ann. Statist., vol. 34, 2554–2591, 2006.

[19] H. Leeb and B. M. Pötscher, *Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?* Econometric Theory, vol. 24, 338–376, 2008.

[20] R. Berk, L. D. Brown, and L. Zhao, *Statistical Inference after Model Selection.* J. Quantitative Criminology, vol. 26, 217–236, 2010.

[21] A. D. Wyner, *Random Packings and Coverings of the Unit n-Sphere.* Bell System Technical Journal, vol. 46, 2111–2118, 1967.

[22] T. Cai and T. Jiang, *Phase transition in limiting distributions of coherence of high-dimensional random matrices.* J. Multivariate Analysis, vol.107, 24–39, 2012.

[23] T. Cai, J. Fan, and T. Jiang, *Distribution of Angles in Random Packing on Spheres.* J. Machine Learning Research, vol. 14, 1801–1828, 2013.

[24] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, 1998.

[25] W. B. Johnson and J. Lindenstrauss. *Extensions of Lipschitz Mappings into a Hilbert Space,* Conference in Modern Analysis and Probability (New Haven, Conn. 1982), Contemporary Mathematics, vol. 26, Providence, RI: American Mathematical Society, 189–206, 1984.