Submitted to the Annals of Statistics

IMPACTS OF HIGH DIMENSIONALITY IN FINITE SAMPLES*

By Jinchi Lv

University of Southern California

High-dimensional data sets are commonly collected in many contemporary applications arising in various fields of scientific research. We present two views of finite samples in high dimensions: a probabilistic one and a non-probabilistic one. With the probabilistic view, we establish the concentration property and robust spark bound for large random design matrix generated from elliptical distributions, with the former related to the sure screening property and the latter related to sparse model identifiability. An interesting concentration phenomenon in high dimensions is revealed. With the non-probabilistic view, we derive general bounds on dimensionality with some distance constraint on sparse models. These results provide new insights into the impacts of high dimensionality in finite samples.

1. Introduction. Thanks to the advances of information technologies, large-scale data sets with a large number of variables or dimensions are commonly collected in many contemporary applications that arise in different fields of sciences, engineering, and humanities. Examples include marketing data in business, panel data in economics and finance, genomics data in heath sciences, and brain imaging data in neuroscience, among many others. The emergence of a large amount of information contained in high-dimensional data sets provides opportunities, as well as unprecedented challenges, for developing new statistical methods and theory. See, for example, Hall (2006) and Fan and Li (2006) for insights and discussions on the statistical challenges associated with high dimensionality, and Fan and Lv (2010) for a brief review of some recent developments in high-dimensional sparse modeling with variable selection. The approach of variable selection aims to effectively identify important variables and efficiently estimate their effects on a response variable of interest.

For the purpose of prediction and variable selection, it is important to understand and characterize the impacts of high dimensionality in finite

*This research was supported by NSF CAREER Award DMS-0955316 and Grant DMS-0806030.

AMS 2000 subject classifications: Primary 62H99; secondary 60D99.

Keywords and phrases: High dimensionality, finite sample, sure independence screening, concentration phenomenon, geometric representation.

samples. Hall, Marron and Neeman (2005) investigated this problem under the asymptotic framework of fixed sample size n and diverging dimensionality p, and revealed an interesting geometric representation of high dimension, low sample size data. When viewed in the diverging p-dimensional Euclidean space, the randomness in the data vectors can be asymptotically squeezed into random rotation, with the shape of the rescaled n-polyhedron approaching deterministic, modulo the orientation. Such concentration phenomenon of random design matrix in high dimensions is also shared by the concentration property in Fan and Lv (2008) (see Definition 1), in the asymptotic setting of both n and p diverging. Geometrically, this property means that the configuration of the n sub-data vectors, modulo the orientation, becomes close to regular asymptotically. Such a property is key to establishing the sure screening property, which means that all important variables are retained in the reduced feature space with asymptotic probability one, of the sure independence screening (SIS) method introduced in Fan and Lv (2008).

The SIS uses the idea of independence learning by applying componentwise regression. Techniques of independence learning have been widely used for variable ranking and screening. Recent work on variable screening includes Fan and Fan (2008), Hall, Titterington and Xue (2009), Fan, Feng and Song (2011), Xue and Zou (2011), Zhu et al. (2011), Delaigle and Hall (2012), Li, Zhong and Zhu (2012), Mai and Zou (2012), and Bühlmann and Mandozzi (2012), among others. The utility of these methods is characterized by the sure screening property. In particular, Fan and Lv (2008) proved that the concentration property holds when the design matrix is generated from Gaussian distribution, and conjectured that it may well hold for a wide class of elliptical distributions. Samworth (2008) presented some simulation studies investigating such a property for non-Gaussian distributions. The first major contribution of our paper is to provide an affirmative answer to the conjecture posed in Fan and Lv (2008).

To ensure model identifiability and stability for reliable prediction and variable selection, it is practically important to control the collinearity for sparse models. Since it is well known that the level of collinearity among covariates typically increases with the model dimensionality, bounding the sparse model size can be effective in controlling model collinearity. Such a bound is characterized by the concept of robust spark (see Definition 2). Another contribution of the paper is to establish a lower bound on the robust spark in the setting of large random design matrix generated from the family of elliptical distributions.

In addition to the above probabilistic view of finite samples in high dimensions, we also present a non-probabilistic high-dimensional geometric

view. Both views are concerned with how much information finite sample contains. A fundamental question is what the impact of high dimensionality on differentiating the subspaces spanned by different sets of predictors is. Such a question is tied to the issue of model identifiability. In this paper, we intend to derive general bounds on dimensionality with some distance constraint on sparse models.

The rest of the paper is organized as follows. Section 2 establishes the concentration property and robust spark bound for large random design matrix generated from elliptical distributions. We investigate general bounds on dimensionality with distance constraint from a non-probabilistic point of view in Section 3. Section 4 presents two numerical examples to illustrate the theoretical results. We provide some discussions of our results and their implications in Section 5. All technical details are relegated to the Appendix.

2. Concentration property and robust spark bound of large random design matrix. In this section, we focus on the case of large random design matrix observed in a high-dimensional problem, in which each column vector contains the information of a particular covariate. In high-dimensional sparse modeling, a common practice is to assume that only a faction of all covariates, the so-called true or important covariates, contribute to the regression or classification problem, whereas the other covariates, the so-called noise covariates, are simply noise information. The inclusion of noise covariates can deteriorate the performance of the estimated model due to the well-known phenomenon of noise accumulation in high dimensional prediction (Fan and Fan, 2008; Fan and Lv, 2010). A crucial issue behind high-dimensional inference is to characterize the distance between the true underlying sparse model and other sparse models, under some discrepancy measure. Intuitively, such a distance can become smaller as the dimensionality increases, making it more difficult to distinguish the true model from the others. Therefore, it is a fundamental problem to characterize the impacts of high dimensionality in finite samples.

2.1. *Concentration property.* We start with the task of dimensionality reduction, particularly variable screening, which is useful in analyzing ultra-high dimensional data sets. With the idea of independence learning, Fan and Lv (2008) introduced the SIS method to reduce the dimensionality of the feature space from the ultra-high scale to a moderate scale, such as below sample size. They introduced an asymptotic framework under which the SIS enjoys the sure screening property even when the dimensionality can grow exponentially with the sample size; see their Theorem 1. The sure screening property means that the true model is contained in the much re-

duced model after variable screening with asymptotic probability one. In particular, a key ingredient of their asymptotic analysis is the so-called concentration property in Condition 2 of Fan and Lv (2008). They verified such property for random design matrix generated from Gaussian distribution, and conjectured that it may also hold for a wide class of elliptical distributions. To show that SIS is widely applicable, it is crucial to establish the concentration property for classes of non-Gaussian distributions.

The class of elliptical distributions, which is a wide family of distributions generalizing the multivariate normal distribution, has been broadly used in real applications. Examples of non-normal elliptical distributions include the Laplace distribution, t-distribution, Cauchy distribution, logistic distribution, and symmetric stable distribution. In particular, an important subclass of elliptical distributions is the family of mixtures of normal distributions. Mixture distributions provide a useful tool for describing heterogeneous populations. Elliptical distributions also play an important role in the theory of portfolio choice (Chamberlain, 1983; Owen and Rabinovitch, 1983). This is due to important properties that any affine transformation of elliptically distributed random variables still has an elliptical distribution and each elliptical distribution is uniquely determined by its location and scale parameters. An implication in portfolio theory is that if all asset returns jointly follow an elliptical distribution, then all portfolios are characterized fully by their location and scale parameters. We refer to Fang, Kotz and Ng (1990) for a comprehensive account of elliptical distributions.

Assume that $\mathbf{x} = (X_1, \cdots, X_p)^T$ is a p-dimensional random covariate vector having an elliptical distribution ν_p with mean $\mathbf{0}$ and nonsingular covariance matrix $\mathbf{\Sigma}$, and that we have a sample $(\mathbf{x}_i)_{i=1}^n$ of i.i.d. covariate vectors from this distribution. Then we have an $n \times p$ random design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$. By the definition of elliptical distribution (see Muirhead, 1982 or Fang, Kotz and Ng, 1990), the transformed p-dimensional random vector $\mathbf{z} = (Z_1, \cdots, Z_p)^T = \mathbf{\Sigma}^{-1/2}\mathbf{x}$ has a spherical distribution μ_p with mean $\mathbf{0}$ and covariance matrix I_p. Similarly, we define the transformed covariate vectors and transformed random design matrix as

$$\mathbf{z}_i = \mathbf{\Sigma}^{-1/2}\mathbf{x}_i \quad \text{and} \quad \mathbf{Z} = \mathbf{X}\mathbf{\Sigma}^{-1/2}, \tag{1}$$

where $i = 1, \cdots, n$. Clearly, $\mathbf{z}_1, \cdots, \mathbf{z}_n$ are n i.i.d. copies of the transformed random covariate vector $\mathbf{z}$. We denote by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ the largest and smallest eigenvalue of a given matrix, respectively. In high-dimensional problems, we often face the situation of $p \gg n$, so it is desirable to reduce the dimensionality of the feature space from p to a moderate one such as below sample size n. The SIS is capable of doing so when the random design matrix

$\mathbf{X}$ satisfies the following property, as introduced in Fan and Lv (2008).

DEFINITION 1 (Concentration property). *The random design matrix* $\mathbf{X}$ *is said to satisfy the concentration property if there exist some positive constants* c_1, C_1 *such that the deviation probability bound*

$$P\left\{\lambda_{\max}(\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T) > c_1 \quad or \quad \lambda_{\min}(\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T) < c_1^{-1}\right\} \leq \exp(-C_1 n) \tag{2}$$

holds for each $n \times \widetilde{p}$ *submatrix* $\widetilde{\mathbf{Z}}$ *of* $\mathbf{Z}$ *with* $cn < \widetilde{p} \leq p$ *and* $c > 1$ *some positive constant.*

As mentioned in the Introduction, the above concentration property shows a similar concentration phenomenon of large random design matrix to that in Hall, Marron and Neeman (2005). When the distribution ν_p of the covariate vector $\mathbf{x}$ is p-variate Gaussian, Fan and Lv (2008) proved that the random design matrix $\mathbf{X}$ satisfies the concentration property. We now consider a more general class of distributions including Gaussian distributions, the family of elliptical distributions. Assume that $P(\mathbf{z} = \mathbf{0}) = 0$. Then it follows from Theorem 1.5.6 in Muirhead (1982) that the p-variate spherical distribution μ_p has a density function with respect to the Lebesgue measure that is spherically symmetric on $\mathbb{R}^p$. We will work with the family of log-concave spherical distributions on $\mathbb{R}^p$ that satisfy the following two conditions.

CONDITION 1. *The density function* $\exp\{-U_p(\mathbf{v})\}$ *of the* p*-variate log-concave spherical distribution* μ_p *satisfies that for some positive constant* c_2,

$$\nabla^2 U_p(\mathbf{v}) \geq c_2 I_p \quad \textit{uniformly in } \mathbf{v} \in \mathbb{R}^p, \tag{3}$$

where $\nabla^2 U_p$ *denotes the Hessian matrix and* $\mathbf{A} \geq \mathbf{B}$ *means that* $\mathbf{A} - \mathbf{B}$ *is positive semidefinite for any symmetric matrices* $\mathbf{A}$ *and* $\mathbf{B}$.

CONDITION 2. *There exists some positive constant* $c_3 \leq 1$ *such that* $E|Z_1| \geq c_3$.

Condition 1 puts a constraint on the curvature of the log-density of distribution μ_p, and Condition 2 requires that the mean $E|Z_1|$ needs to be bounded from below. Clearly, log-concave spherical distributions satisfying Conditions 1–2 comprise a wide class containing Gaussian distributions. As seen in Lemma 2 later, Condition 1 entails that the corresponding spherical distribution is light-tailed, which is important for variable screening.

For heavy-tailed data sets, Delaigle and Hall (2012) showed that effective variable selection with untransformed data requires slower growth of dimensionality. In particular, they exploited variable transformation methods to transform the original data into light-tailed data and demonstrated their effectiveness and advantages. So in the presence of heavy-tailed data, one may work with the assumption of elliptical distributions on the transformed data.

The assumption of elliptical distributions is commonly used in dimension reduction and has also been used for variable screening. See, for example, Zhu et al. (2011). This assumption facilitates our technical analysis. Similar results may hold for more general family of distributions by resorting to techniques in random matrix theory. Some other variable screening methods such as in Mai and Zou (2012) require no such an assumption.

THEOREM 1 (Concentration property). *Under Conditions 1–2, the random design matrix* $\mathbf{X}$ *satisfies the concentration property (2).*

Theorem 1 shows that the concentration property holds not only for Gaussian distributions, but also for a wide class of elliptical distributions, as conjectured by Fan and Lv (2008) (see their Section 5.1). This provides an affirmative answer to their conjecture, showing that the SIS indeed enjoys the sure screening property for the random design matrix generated from a wide class of elliptical distributions. The proof of Theorem 1 relies on the following three lemmas that are of independent interest.

LEMMA 1. *Under Condition 1, each* q*-variate marginal distribution* $\widetilde{\mu}_q$ *of* μ_p *with* $1 \le q \le p$ *satisfies the logarithmic Sobolev inequality*

$$E_{\widetilde{\mu}_q}\left\{f^2 \log f^2\right\} \le 2C_2 E_{\widetilde{\mu}_q}\|\nabla f\|_2^2 \tag{4}$$

for any smooth function f *on* $\mathbb{R}^q$ *with* $E_{\widetilde{\mu}_q} f^2 = 1$*, where* $C_2 = c_2^{-1}$ *and* ∇f *denotes the gradient of function* f*.*

LEMMA 2. *Let* $\mathbf{z}_q$ *be an arbitrary* q*-dimensional subvector of* $\mathbf{z}$ *with* $1 \le q \le p$*. Then we have*

a) Under Condition 1, it holds for any $r \in (0, \infty)$ *that*

$$P\left\{\left|\|\mathbf{z}_q\|_2 - E\|\mathbf{z}_q\|_2\right| > r\right\} \le 2\exp\{-C_2^{-1} r^2/2\}; \tag{5}$$

b) It holds that

$$\sqrt{q} E|Z_1| \le E\|\mathbf{z}_q\|_2 \le \sqrt{q}. \tag{6}$$

LEMMA 3. *Assume that Conditions 1–2 hold,* $\mathbf{z}_q$ *is a* q*-dimensional subvector of* $\mathbf{z}$ *with* $n \le q \le p$*, and* $\mathbf{w} \sim N(\mathbf{0}, I_q)$*. Then there exist some positive constants* $c_4 < 1$*,* $c_5 > 1$*, and* C_3 *such that*

$$P\left\{\frac{\|\mathbf{z}_q\|_2}{\|\mathbf{w}\|_2} < c_4 \quad or \quad \frac{\|\mathbf{z}_q\|_2}{\|\mathbf{w}\|_2} > c_5\right\} \le 4\exp\{-C_3 n\}. \tag{7}$$

Lemma 1 shows that each marginal distribution of μ_p satisfies the logarithmic Sobolev inequality, which is an important tool for proving the concentration probability inequality for measures. Lemma 2 establishes that for each q-dimensional subvector $\mathbf{z}_q$ of $\mathbf{z}$, its L_2-norm $\|\mathbf{z}_q\|_2$ concentrates around the mean $E\|\mathbf{z}_q\|_2$ with significant probability, which is in turn sandwiched between two quantities $\sqrt{q}E|Z_1|$ and $\sqrt{q}$. Lemma 3 demonstrates an interesting phenomenon of measure concentration in high dimensions.

2.2. *Robust spark bound.* As is well-known in high-dimensional sparse modeling, controlling the level of collinearity for sparse models is essential for model identifiability and stable estimation. For a given $n \times p$ design matrix $\mathbf{X}$, there may exist another p-vector $\boldsymbol{\beta}_1$ that is different from the true regression coefficient vector $\boldsymbol{\beta}_0$ such that $\mathbf{X}\boldsymbol{\beta}_1$ is (nearly) identical to $\mathbf{X}\boldsymbol{\beta}_0$, when the dimensionality p is large compared with the sample size n. This indicates that model identifiability is generally not guaranteed in high dimensions when no additional constraint is imposed on the model parameter. In addition, the subdesign matrix corresponding to a sparse model should be well-conditioned to ensure reliable estimation of model parameters and nice convergence rates as in such as the least-squares or maximum likelihood estimation. As an example, the covariance matrix of the least-squares estimator is proportional to the inverse Gram matrix given by the design matrix.

Since the collinearity among the covariates increases with the dimensionality as evident from the geometric point of view, a natural and effective way to ensure model identifiability and reduce model instability is to control the size of sparse models. Such an idea has been adopted in Donoho and Elad (2003) for the problem of sparse recovery, which is the noiseless case of linear regression. In particular, they introduced the concept of spark as a bound on sparse model size to characterize model identifiability. The spark κ of the design matrix $\mathbf{X}$ is defined as the smallest possible positive integer such that there exists a singular $n \times \kappa$ submatrix of $\mathbf{X}$. This concept plays an important role in the problem of sparse recovery; see also Lv and Fan (2009). An implication is that the true model parameter vector $\boldsymbol{\beta}_0$ is uniquely defined as long as $\|\boldsymbol{\beta}_0\|_0 < \kappa/2$, which provides a basic condition

for model identifiability. For the problem of variable selection in the presence of noise, a stronger condition than provided by the spark is generally needed. For this purpose, the concept of spark was generalized in Zheng, Fan and Lv (2012) by introducing the concept of robust spark, as follows.

DEFINITION 2 (Robust spark). *The robust spark κ_c of the $n \times p$ design matrix $\mathbf{X}$ is defined as the smallest possible positive integer such that there exists an $n \times \kappa_c$ submatrix of $n^{-1/2}\mathbf{X}$ having a singular value less than a given positive constant c.*

It is easy to see that the robust spark κ_c approaches the spark of $\mathbf{X}$ as $c \to 0+$. The robust spark provides a natural bound on model size for effectively controlling the collinearity level of sparse models, which is referred to as the robust spark condition. For each sparse model with size $d < \kappa_c$, the corresponding $n \times d$ submatrix of $n^{-1/2}\mathbf{X}$ have all singular values bounded from below by c. The robust spark κ_c is always a positive integer no larger than $n + 1$. It is practically important in high-dimensional sparse modeling to show that the robust spark can be some large number diverging with the sample size n. We intend to build a lower bound on the robust spark for the case of random design matrix, following the setting in Section 2.1.

THEOREM 2 (Robust spark bound). *Assume that the rows of the $n \times p$ random design matrix $\mathbf{X}$ are i.i.d. as ν_p having mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$ and satisfying Conditions 1–2, with $\lambda_{\min}(\mathbf{\Sigma})$ bounded from below by some positive constant. Then with asymptotic probability one, $\kappa_c \geq \tilde{c}n/(\log p)$ for sufficiently small constant c and some positive constant $\tilde{c}$ depending only on c.*

Theorem 2 formally characterizes the order of the robust spark κ_c when the design matrix $\mathbf{X}$ is generated from the family of elliptical distributions. We see that sparse linear models of size as large as of order $O\{n/(\log p)\}$ can still be well separated from each other. On the other hand, when the true model size is beyond such an order, the true underlying sparse model may be indistinguishable from others in finite sample. Theorem 2 also justifies the range of the true sparse model size under which the problem of variable selection is meaningful. The deflation factor of $\log p$ represents the general price one has to pay for the search of important covariates in high dimensions.

The concept of robust spark shares a similar spirit as the restricted eigenvalue condition on the design matrix in Bickel, Ritov and Tsybakov (2009),

in the sense that both are sparse eigenvalue type conditions. Instead of constraining the sparse model size, the restricted eigenvalue condition uses an L_1-norm constraint on the parameter vector. As discussed in Zheng, Fan and Lv (2012), the robust spark condition can be weaker than the restricted eigenvalue condition, since the L_0-norm constraint can define a smaller subset than the L_1-norm constraint. Many other conditions have also been introduced to characterize the properties of variable selection methods such as the Lasso. See, for example, van de Geer and Bühlmann (2009) for a comprehensive comparison and discussions on these conditions.

In particular, the robust spark condition is weaker than the partial orthogonality condition, which requires that true covariates and noise covariates are essentially uncorrelated, with absolute correlation of the order $O(n^{-1/2})$. In contrast, the robust spark condition can allow for much stronger correlation between true covariates and noise covariates. The robust spark condition can also be weaker than the irrepresentable condition. To see this, let us consider the simple example constructed in Lv and Fan (2009). In their Example 1, the irrepresentable condition becomes the constraint that the maximum absolute correlation between the response and all noise covariates is bounded from above by $s^{-1/2}$, where s denotes the true model size. Since the response is a linear combination of true covariates in that example, this indicates that the irrepresentable condition can be stronger than the robust spark condition when the true model size grows.

3. General bounds on dimensionality with distance constraint. We have provided in Section 2 a probabilistic view of finite samples in high dimensions, with focus on large random design matrix generated from the family of elliptical distributions. It is also important to understand how the dimensionality plays an role in deterministic finite samples. For such a purpose, we take a high-dimensional geometric view of finite samples and derive general bounds on dimensionality using non-probabilistic arguments. With a slight abuse of notation, we now denote by $\mathbf{x}_j$ an n-dimensional vector of observations from the j-th covariate, and consider a collection of p covariates $\{\mathbf{x}_j : j = 1, \cdots, p\}$. Assume that each covariate vector $\mathbf{x}_j$ is rescaled to have L_2-norm $n^{1/2}$. Then all vectors $n^{-1/2}\mathbf{x}_j$, $j = 1, \cdots, p$, lie on the unit sphere S^{n-1} in the n-dimensional Euclidean space $\mathbb{R}^n$. We are interested in a natural question that how many variables there can be if the maximum collinearity of sparse models is controlled.

For each positive integer s, denote by $\mathcal{A}_s$ the set of all subspaces spanned by s of covariates $\mathbf{x}_j$'s. Assume that s is less than half of the spark κ of the $n \times p$ design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)$. Then each subspace in $\mathcal{A}_s$ is s-dimensional

and $|\mathcal{A}_s| = \binom{p}{s}$. To control the collinearity among the variables, it is desirable to bound the distances between s-dimensional subspaces in $\mathcal{A}_s$ away from zero, under some discrepancy measure. When each pair of subspaces in $\mathcal{A}_s$ has a positive distance, intuitively there cannot be too many of them. The geometry of the space of all s-dimensional subspaces of $\mathbb{R}^n$ is characterized by the Grassmann manifold $G_{n,s}$. To facilitate our presentation, we list in Appendix B some necessary background and terminology on the geometry and invariant measure of Grassmann manifold. In particular, $G_{n,s}$ admits an invariant measure which under a change of variable and symmetrization can be represented as a probability measure ν on $[0,1]^s$ with density given in (43).

With the aid of the measure ν, we can calculate the volumes of various shapes of neighborhoods in the Grassmann manifold, which are typically given in terms of the principal angles θ_i between an s-dimensional subspace of $\mathbb{R}^n$ and a fixed s-dimensional subspace with generator matrix $(I_s\ 0)$. The principal angles between subspaces are natural extensions of the concept of angle between lines. Let V_1, V_2 be two subspaces in $G_{n,s}$ having a set of principal angles $(\theta_1, \cdots, \theta_s)$, with $\pi/2 \geq \theta_1 \geq \cdots \geq \theta_s \geq 0$ and corresponding s pairs of unit vectors (v_{1i}, v_{2i}). If V_i is spanned by s of $\mathbf{x}_j$'s, then putting $r_i = \cos\theta_i$ and reversing the order give the canonical correlations $(r_s, \cdots, r_1)$ and corresponding pairs of canonical variables (v_{1i}, v_{2i}), for the two groups of variables.

There are three frequently used distances between subspaces V_1 and V_2 on the Grassmann manifold $G_{n,s}$: the geodesic distance $d_g(V_1, V_2) = (\sum_{i=1}^s \theta_i^2)^{1/2}$ (Wong, 1967), the chordal distance $d_c(V_1, V_2) = (\sum_{i=1}^s \sin^2\theta_i)^{1/2}$ (Conway, Hardin and Sloane, 1996), and the maximum chordal distance (Edelman, Arias and Smith, 1998)

$$d_m(V_1, V_2) = \sin\theta_1 = \max_{i=1}^s \sin\theta_i. \tag{8}$$

In view of the probability measure ν in (43), it seems natural to consider the latter two distances, which is indeed the case. To see this, let B_i be an $s \times n$ orthonormal generator matrix for V_i. Then V_i is uniquely determined by the projection matrix $P_i = B_i^T B_i$, which corresponds to the projection onto the s-dimensional subspace V_i. It is known that

$$d_c(V_1, V_2) = 2^{-1/2}\|P_1 - P_2\|_F \quad \text{and} \quad d_m(V_1, V_2) = \|P_1 - P_2\|_2,$$

where $\|\cdot\|_F$ and $\|\cdot\|_2$ denote the Frobenius norm and spectral norm (or operator norm) of a given matrix, respectively. These two matrix norms are commonly used in large covariance matrix estimation and other multivariate analysis problems.

We now bound the size of the set $\mathcal{A}_s \subset G_{n,s}$ of all subspaces spanned by s of covariates $\mathbf{x}_j$'s under some distance constraint, which in turn gives bounds on the dimensionality p. The probability measure ν on $[0,1]^s$ defined in (43) is a key ingredient in our analysis. When all the subspaces in $\mathcal{A}_s$ have distance at least $2\delta > 0$ under any distance d, it is easy to see that $\binom{p}{s} = |\mathcal{A}_s| \leq 1/\nu(B_{\delta,d})$, where $B_{\delta,d}$ denotes a ball of radius δ in Grassmann manifold $G_{n,s}$ under distance d. In particular, we focus on the maximum chordal distance defined in (8). Equivalently, the maximum chordal distance constraint gives the maximum principal angle constraint. Since the sample size n is usually small or moderate in many contemporary applications, we adopt the asymptotic framework of $s/n \to \gamma \in (0,1)$ as $n \to \infty$ for deriving asymptotic bounds on the dimensionality p.

THEOREM 3. *Assume that all subspaces spanned by s of covariates $\mathbf{x}_j$'s have maximum chordal distance at least a fixed constant $2\delta \in (0,1)$, and $s/n \to \gamma \in (0,1/2)$ as $n \to \infty$. Then we have*

$$\log p \lesssim (\log \delta^{-1})(1-\gamma)n + 2\log n + O(1), \tag{9}$$

where $\lesssim$ denotes asymptotic dominance.

Theorem 3 gives a general asymptotic bound on the dimensionality p under the maximum chordal distance constraint, or equivalently, the maximum principal angle constraint. We see that finite sample can allow for a large number of variables, in which sparse models with size much smaller than sample size n can still be distinguishable from each other. The leading order in the bound for $\log p$ is proportional to sample size n, with factors $\log \delta^{-1}$ and $1-\gamma$. This result is reasonable because larger δ means bigger separation of all s-dimensional subspaces spanned by covariates $\mathbf{x}_j$'s, and large γ means more such subspaces separated from each other, both cases leading to tighter constraint on the growth of dimensionality p. It is interesting that there are only two terms $O(\log n)$ and $O(1)$ following the leading order in the above bound on dimensionality.

The general bound on dimensionality with distance constraint in Theorem 3 also shares some similarity with the lower bound $O\{n/(\log p)\}$ on the robust spark in Theorem 2, although the former uses non-probabilistic arguments with no distributional assumption and the latter applies probabilistic arguments. The robust spark provides a natural bound on sparse model size to control collinearity for sparse models. Intuitively, when the dimensionality p grows with the sample size n, one expects tighter control on the robust spark through a deflation factor of $\log p$. Similarly, the upper

bound on the logarithmic dimensionality $\log p$ in Theorem 3 decreases with the minimum maximum chordal distance 2δ between sparse models through the factor $\log \delta^{-1}$. As mentioned in Section 2.2, these sparse eigenvalue type conditions play an important role in characterizing the variable selection properties including the model selection consistency for various regularization methods. Although the result in Theorem 3 can be viewed as the bound for the worst case scenario, it provides us caution and guidance on the growth of dimensionality in real applications, particularly when variable selection is an important goal in the studies.

In general, the robust spark κ_c provides a stronger measure on collinearity than the maximum chordal distance. To see this, assume that $s < \kappa_c/2$ and let V_1, V_2 be two subspaces spanned by two different sets of s of covariates $\mathbf{x}_j$'s. Then the maximum principal angle θ_1 between V_1 and V_2 is the angle between two vectors $v_i \in V_i$, where v_i is a linear combination of the corresponding set of covariate vectors $n^{-1/2}\mathbf{x}_j$ for each $i = 1, 2$. Since the union of these two sets of covariates has cardinality bounded from above by $2s < \kappa_c$, it follows from the definition of the robust spark that the angle θ_1 between v_1 and v_2 is bounded from zero, which entails that the maximum chordal distance between V_1 and V_2 is also bounded from zero. Conversely, when two s-dimensional subspaces V_1 and V_2 has the maximum chordal distance bounded from zero, the subdesign matrix corresponding to covariates in the sets can still be singular.

We next consider a stronger distance constraint than in Theorem 3, where in addition, all disjoint subspaces in $\mathcal{A}_s$ have minimum principal angles at least $\arcsin \delta_1$ for some $\delta_1 \in (0, \delta]$, with δ given in Theorem 3. Such disjoint subspaces are spanned by disjoint sets of s of covariates $\mathbf{x}_j$'s. In this case, it is natural to expect a tighter bound on the dimensionality p.

THEOREM 4. *Assume that conditions of Theorem 3 hold and all disjoint subspaces have minimum principal angles at least a fixed constant* $\arcsin \delta_1$ *with* $\delta_1 \in (0, \delta]$. *Then we have*

$$\log(p-s) \lesssim \left[(\log \delta^{-1})(1-\gamma) - c_{\delta_1} - \gamma - 2^{-1}\log(1-2\gamma)\right] n + 2\log n + O(1), \tag{10}$$

where $c_{\delta_1} = 2^{-1}\left[\log(1-\delta_1^2)^{-1}\right](1-\gamma) - 2^{-1}(1-\delta_1^2)^{-1}\delta_1^2(1-2\gamma)$.

Compared to the bound in Theorem 3, Theorem 4 indeed provides a tighter bound on the dimensionality p due to the additional distance constraint involving δ_1. We are interested in the asymptotic bound on the dimensionality when δ_1 is near zero. In this case, we have $c_{\delta_1} \sim \delta_1^2\gamma/2$. Observe that $\gamma + 2^{-1}\log(1-2\gamma) \leq 0$ and is of order $O(\gamma^2)$. It is generally difficult to

derive tight bounds over the whole ranges of δ_1 and γ. This is essentially due to the challenge of obtaining a globally tight function bounding the function f_1 defined in (32) from above, while retaining analytical tractability of evaluating the resulting integral.

We finally revisit the marginal correlation ranking, a widely used technique for analyzing large-scale data sets, from a non-probabilistic point of view. Given a sample of size n, the maximum correlation of noise covariates with the response variable can exceed the maximum correlation of true covariates with the response variable when the dimensionality p is high. Here the correlation between two n-vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is referred to as $\cos\theta = \mathbf{v}_1^T\mathbf{v}_2/(\|\mathbf{v}_1\|_2\|\mathbf{v}_2\|_2)$, where θ is the angle between them. It is important to understand the limit on the dimensionality p under which the above undesired phenomenon can happen.

THEOREM 5. *Let $r \in (0,1)$ be the maximum absolute correlation between s true predictors $\mathbf{x}_j$ and response vector $\mathbf{y}$ in $\mathbb{R}^n$ and assume that all $p-s$ noise predictors $\mathbf{x}_j$ have absolute correlations bounded by $\delta \in (0,1)$. Then there exists a noise predictor having absolute correlation with $\mathbf{y}$ larger than r if $\log(p-s) \geq 2^{-1}\left\{\log[4/(1-\delta^2)]\right\}(n-1) + 2^{-1}\log n + O(1)$.*

It is an interesting result that the above asymptotic bound on the dimensionality p depends only on $\delta \in (0,1)$, and is independent of the specific value of $r \in (0,1)$. The condition on the dimensionality is sufficient but not necessary in general, since one can always add an additional noise predictor having absolute correlation with $\mathbf{y}$ larger than r. Nevertheless, Theorem 5 gives us a general limit on dimensionality even when one believes that a majority of noise predictors have weak correlation with the response variable.

Meanwhile, we also see from Theorem 5 that the dimensionality p generally needs to be large compared to the sample size n such that a noise predictor may have the highest correlation with the response variable. This result is reflected in a common feature of many variable selection procedures including commonly used greedy algorithms, that is, initially selecting one predictor with the highest correlation with the response variable. See, for example, the LARS algorithm in Efron et al. (2004) and the LLA algorithm in Zou and Li (2008). Such a variable, which gives a sparse model with size one, commonly appears on the solution paths of many regularization methods for high-dimensional variable selection.

4. Numerical examples. In this section we provide two simulation examples to illustrate the theoretical results in Section 2, obtained through

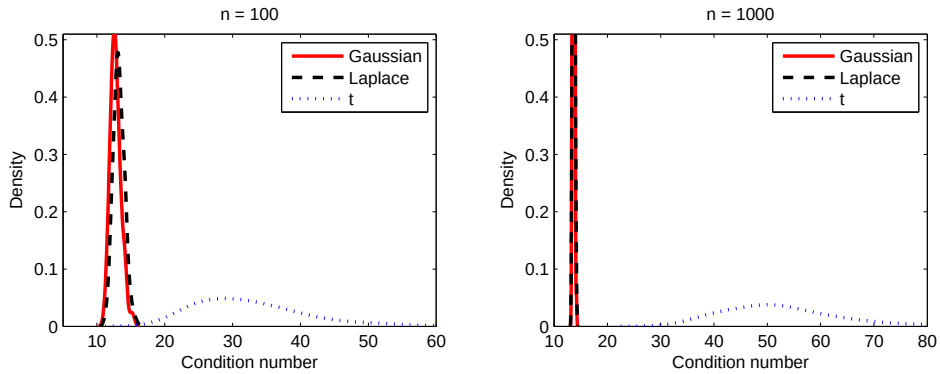

FIG 1. *Distributions of the condition number of* $\widetilde{p}^{-1}\mathbf{X}\mathbf{X}^T$ *in different scenarios of distributions of* $\mathbf{X}$ *with* $\widetilde{p} = 3n$ *and* $n = 100$ *and* 1000*, based on 100 simulations.*

probabilistic arguments. The first simulation example examines the concentration property for large random design matrix. Let $\mathbf{X}$ be an $n \times \widetilde{p}$ random design matrix with $\widetilde{p} = cn$ for some constant $c > 1$. We set $n = 100$ and 1000, and $c = 3$. We considered three scenarios of distributions: 1) each entry of $\mathbf{X}$ is sampled independently from $N(0, 1)$, 2) each entry of $\mathbf{X}$ is sampled independently from the Laplace distribution with mean 0 and variance 1, and 3) each row of $\mathbf{X}$ is sampled independently from the multivariate t-distribution with 10 degrees of freedom and then rescaled to have unit variances. In view of Definition 1, the concentration property of $\mathbf{X}$ is characterized by the distribution of the condition number of $\widetilde{p}^{-1}\mathbf{X}\mathbf{X}^T$. In each case, 100 Monte Carlo simulations were used to obtain the distribution of such condition number. Figure 1 depicts these distributions in different scenarios. We see that in scenarios 1 and 2, the condition number concentrates in the range of relatively small numbers, indicating the associated concentration property as shown in Theorem 1. In scenario 3 with multivariate t-distribution, one still observes the concentration phenomenon. However, since this distribution is relatively more heavy-tailed, we see that the distribution of the condition number becomes more spread out and shifts toward the range of large numbers.

The second simulation example investigates the robust spark bound for large $n \times p$ random design matrix $\mathbf{X}$. We adopted the same three scenarios of distributions as in the first simulation example, except that $n = 100$, and $p = 1000$ and 5000. In light of Theorem 2, we sampled randomly 1000 $n \times k$ submatrices of $n^{-1/2}\mathbf{X}$ each with $k = \lceil 2n/(\log p) \rceil$ columns and calculated the minimum of those 1000 smallest singular values. Similarly, in each case

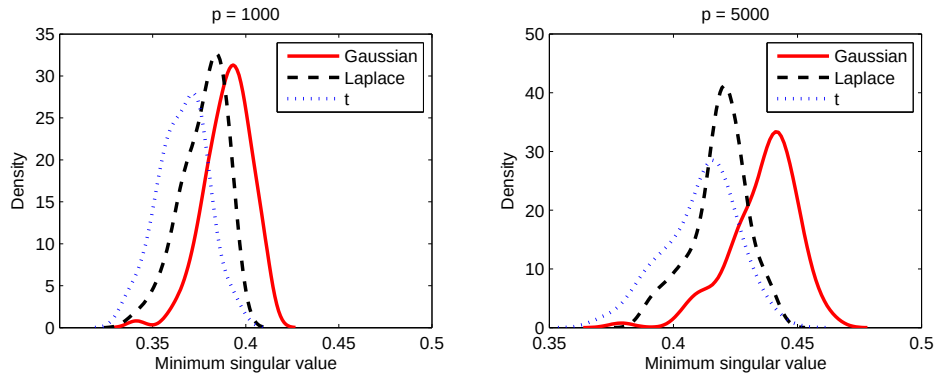

FIG 2. *Distributions of the minimum singular value over 1000 submatrices of* $n^{-1/2}\mathbf{X}$ *each with* $\lceil 2n/(\log p)\rceil$ *columns, in different scenarios of distributions of* $\mathbf{X}$ *with* $n = 100$*, and* $p = 1000$ *and* 5000*, based on 100 simulations.*

100 Monte Carlo simulations were used to obtain the distribution of such minimum singular value which is tied to the robust spark bound of $\mathbf{X}$. These distributions are shown in Figure 2. In particular, we see that the distribution of the minimum singular value concentrates clearly away from zero in each of the three scenarios of distributions. These numerical results indicate that the robust spark of random design matrix can indeed be at least of order $O\{n/(\log p)\}$, as shown in Theorem 2.

5. Discussions. We have investigated the impacts of high dimensionality in finite samples from two different perspectives: a probabilistic one and a non-probabilistic one. An interesting concentration phenomenon for large random design matrix has been revealed, as shown previously in Hall, Marron and Neeman (2005). We have shown that the concentration property, which is important in characterizing the sure screening property of the SIS, holds for a wide class of elliptical distributions, as conjectured by Fan and Lv (2008). We have also established a lower bound on the robust spark which is important in ensuring model identifiability and stable estimation. The high-dimensional geometric view of finite samples has lead to general bounds on dimensionality with distance constraint on sparse models, using non-probabilistic arguments.

Both probabilistic and non-probabilistic views provide understandings on how the dimensionality interacts with the sample size for large-scale data sets. Characterizing the limit of the dimensionality with respect to the sample size is key to the success of high-dimensional inference goals such as prediction and variable selection. We have focused on the family of ellipti-

cal distributions. It would be interesting to consider a more general class of distributions for future research.

Acknowledgments. The author sincerely thanks the Co-Editor, Associate Editor, and two referees for their valuable comments that improved significantly the paper.

APPENDIX A: PROOFS OF MAIN RESULTS

For notational simplicity, we use C to denote a generic positive constant, whose value may change from line to line.

A.1. Proof of Lemma 1. By Theorem 5.2 in Ledoux (2001), we know that Condition 1 entails the logarithmic Sobolev inequality (4) when $\widetilde{\mu}_q = \mu_p$. It remains to prove the logarithmic Sobolev inequality for any marginal distribution of μ_p. Let $1 \le q < p$ and $\widetilde{\mu}_q$ be a q-variate marginal distribution of μ_p. By the spherical symmetry of μ_p, without loss of generality we can assume that $\widetilde{\mu}_q$ is concentrated on $\mathbb{R}^q = \{\mathbf{v} = (v_1, \cdots, v_p)^T \in \mathbb{R}^p : v_{q+1} = \cdots = v_p = 0\}$. For any smooth function $\widetilde{f}$ on $\mathbb{R}^q$ with $E_{\widetilde{\mu}_q}\widetilde{f}^2 = 1$, define $f : \mathbb{R}^p \to \mathbb{R}$ by

$$f(v_1, \cdots, v_q, v_{q+1}, \cdots, v_p) = \widetilde{f}(v_1, \cdots, v_q). \tag{11}$$

Clearly f is a smooth function on $\mathbb{R}^p$ and

$$\nabla f(v_1, \cdots, v_q, v_{q+1}, \cdots, v_p) = \begin{bmatrix} \nabla \widetilde{f}(v_1, \cdots, v_q) \\ \mathbf{0} \end{bmatrix},$$

which shows that

$$\|\nabla f(v_1, \cdots, v_q, v_{q+1}, \cdots, v_p)\|_2^2 = \left\|\nabla \widetilde{f}(v_1, \cdots, v_q)\right\|_2^2. \tag{12}$$

In view of (11), it follows from Fubini's theorem that

$$E_{\mu_p} f^2 = E_{\widetilde{\mu}_q} \widetilde{f}^2 = 1.$$

Thus by (11), (12), and Fubini's theorem, applying the logarithmic Sobolev inequality (4) for μ_p to the smooth function f yields

$$E_{\widetilde{\mu}_q}\left\{\widetilde{f}^2 \log \widetilde{f}^2\right\} = E_{\mu_p}\left\{f^2 \log f^2\right\} \le 2C_2 E_{\mu_p}\|\nabla f\|_2^2 = 2C_2 E_{\widetilde{\mu}_q}\|\nabla \widetilde{f}\|_2^2,$$

which completes the proof.

A.2. Proof of Lemma 2. We first prove part a). By Lemma 1, the distribution of $\mathbf{z}_q$ satisfies the logarithmic Sobolev inequality (4). Observe that by the triangle inequality, the Euclidean norm $\|\cdot\|_2$ is 1-Lipschitz with respect to the metric induced by itself. Therefore the classical Herbst argument applies to prove the concentration inequality (5) (see, e.g., Theorem 5.3 in Ledoux, 2001). It remains to show part b). Note that $EZ_1^2 = 1$. By the spherical symmetry of μ_p, Hölder's inequality, and the Cauchy-Schwarz inequality, we have

$$E\|\mathbf{z}_q\|_2 \leq \sqrt{E\|\mathbf{z}_q\|_2^2} = \sqrt{qEZ_1^2} = \sqrt{q}$$

and

$$E\|\mathbf{z}_q\|_2 \geq E\left(q^{-1/2}\|\mathbf{z}_q\|_1\right) = \sqrt{q}E|Z_1|.$$

This concludes the proof.

A.3. Proof of Lemma 3. We first make a simple observation. The standard Gaussian distributions are special cases of spherical distributions. Recall that the q-variate standard Gaussian distribution γ_q has density function $\frac{1}{(2\pi)^{q/2}}e^{-\|\mathbf{v}\|_2^2/2}$, $\mathbf{v} \in \mathbb{R}^q$. Thus it is easy to check that γ_q satisfies Condition 1 with $c_2 = 1$. Let $\mathbf{w} = (W_1, \cdots, W_q)^T \sim N(\mathbf{0}, I_q)$. Then it follows immediately from Lemma 2 that for any $r \in (0, \infty)$,

$$P\left(|\|\mathbf{w}\|_2 - E\|\mathbf{w}\|_2| > r\right) \leq 2e^{-r^2/2}. \tag{13}$$

Note that

$$E|W_1| = 2\int_0^\infty u\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du = \sqrt{\frac{2}{\pi}}\int_0^\infty e^{-u^2/2}d\left(\frac{u^2}{2}\right) = \sqrt{\frac{2}{\pi}}. \tag{14}$$

By (6), (13), and (14), we have for any $r_1 \in (0, \sqrt{\frac{2}{\pi}})$,

$$P\left(q^{-1/2}\|\mathbf{w}\|_2 > 1 + r_1 \quad \text{or} \quad q^{-1/2}\|\mathbf{w}\|_2 < \sqrt{\frac{2}{\pi}} - r_1\right) \leq 2e^{-qr_1^2/2} \leq 2e^{-nr_1^2/2} \tag{15}$$

since $q \geq n$.

Now we get back to $\mathbf{z}_q$. It follows from (5) and (6) in Lemma 2 and Condition 2 that for any $r_2 \in (0, c_3)$,

$$P\left(q^{-1/2}\|\mathbf{z}_q\|_2 > 1 + r_2 \quad \text{or} \quad q^{-1/2}\|\mathbf{z}_q\|_2 < c_3 - r_2\right) \leq 2e^{-C_2^{-1}qr_2^2/2} \leq 2e^{-C_2^{-1}nr_2^2/2} \tag{16}$$

since $q \geq n$. Let

$$c_4 = \frac{c_3 - r_2}{1 + r_1} \quad \text{and} \quad c_5 = \frac{1 + r_2}{\sqrt{\frac{2}{\pi}} - r_1}.$$

Then combining (15) and (16) along with Bonferroni's inequality yields

$$\begin{aligned} P\left(\frac{\|\mathbf{z}_q\|_2}{\|\mathbf{w}\|_2} < c_4 \ \text{ or } \ \frac{\|\mathbf{z}_q\|_2}{\|\mathbf{w}\|_2} > c_5\right) &\leq 2e^{-nr_1^2/2} + 2e^{-C_2^{-1}nr_2^2/2} \\ &\leq 4e^{-C_3 n}, \end{aligned}$$

where $C_3 = \min(r_1^2/2, C_2^{-1}r_2^2/2)$. This completes the proof.

A.4. Proof of Theorem 1. In Section A.7, Fan and Lv (2008) proved that Gaussian distributions satisfy the concentration property (2), that is, for $\mathbf{Z} \sim N(\mathbf{0}, I_n \otimes I_p) = N(\mathbf{0}, I_{n\times p})$. We now consider the general situation where n rows of the $n \times p$ random matrix $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_n)^T$ are i.i.d. copies from the spherical distribution μ_p. Fix an arbitrary $n \times \widetilde{p}$ submatrix $\widetilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $cn < \widetilde{p} \leq p$, where $c \in (1, \infty)$. We aim to prove deviation inequality in (2) with different constants $c_1 \in (1, \infty)$ and $C_1 \in (0, \infty)$.

By the spherical symmetry, without loss of generality we can assume that $\widetilde{\mathbf{Z}}$ consists of the first $\widetilde{p}$ columns of $\mathbf{Z}$. Let

$$\widetilde{\mathbf{z}} = (Z_1, \cdots, Z_{\widetilde{p}})^T \quad \text{and} \quad \widetilde{\mathbf{Z}} = (\widetilde{\mathbf{z}}_1, \cdots, \widetilde{\mathbf{z}}_n)^T.$$

Clearly $\widetilde{\mathbf{z}}_1, \cdots, \widetilde{\mathbf{z}}_n$ are n i.i.d. copies of $\widetilde{\mathbf{z}}$. Take an $n \times \widetilde{p}$ random matrix

$$\mathbf{W} = (\mathbf{w}_1, \cdots, \mathbf{w}_n)^T \sim N(\mathbf{0}, I_n \otimes I_{\widetilde{p}}),$$

which is independent of $\widetilde{\mathbf{Z}}$. Then for each $i = 1, \cdots, n$, $\mathbf{w}_i$ has the $\widetilde{p}$-variate standard Gaussian distribution. It is well-known that $\mathbf{w}_i/\|\mathbf{w}_i\|_2$ has the Haar distribution on the unit sphere $S^{\widetilde{p}-1}$ in $\widetilde{p}$-dimensional Euclidean space $\mathbb{R}^{\widetilde{p}}$, i.e., the uniform distribution on $S^{\widetilde{p}-1}$.

Since the distribution of $\widetilde{\mathbf{z}}$ is a marginal distribution of μ_p, the spherical symmetry of μ_p entails that of the distribution of $\widetilde{\mathbf{z}}$. It follows easily from the assumption of $P(\mathbf{z} = \mathbf{0}) = 0$ that $P(\widetilde{\mathbf{z}} = \mathbf{0}) = 0$. Thus by Theorem 1.5.6 in Muirhead (1982), $\widetilde{\mathbf{z}}/\|\widetilde{\mathbf{z}}\|_2$ is uniformly distributed on $S^{\widetilde{p}-1}$ and is independent of $\|\widetilde{\mathbf{z}}\|_2$. This along with the above fact shows that for each $i = 1, \cdots, n$,

$$\widetilde{\mathbf{z}}_i \overset{(\mathrm{d})}{=\!=} \|\widetilde{\mathbf{z}}_i\|_2 \left(\frac{\mathbf{w}_i}{\|\mathbf{w}_i\|_2}\right) = \left(\frac{\|\widetilde{\mathbf{z}}_i\|_2}{\|\mathbf{w}_i\|_2}\right)\mathbf{w}_i, \tag{17}$$

where we use the symbol $\overset{(d)}{=\!=}$ to denote being identical in distribution. Hereafter, for notational simplicity we do not distinguish $\widetilde{\mathbf{z}}_i$ and $\left(\frac{\|\widetilde{\mathbf{z}}_i\|_2}{\|\mathbf{w}_i\|_2}\right)\mathbf{w}_i$.

Define the $n \times n$ diagonal matrix

$$\mathbf{Q} = \operatorname{diag}\left\{\frac{\|\widetilde{\mathbf{z}}_1\|_2}{\|\mathbf{w}_1\|_2}, \cdots, \frac{\|\widetilde{\mathbf{z}}_n\|_2}{\|\mathbf{w}_n\|_2}\right\}.$$

Then we have

$$\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T = \mathbf{Q}\left(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T\right)\mathbf{Q},$$

which entails

$$\begin{aligned}\min_{i=1}^n \frac{\|\widetilde{\mathbf{z}}_i\|_2^2}{\|\mathbf{w}_i\|_2^2}\lambda_{\min}(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T)I_n &\le \min_{i=1}^n \frac{\|\widetilde{\mathbf{z}}_i\|_2^2}{\|\mathbf{w}_i\|_2^2}\left(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T\right) \le \mathbf{Q}\left(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T\right)\mathbf{Q} \\ &= \widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T \le \max_{i=1}^n \frac{\|\widetilde{\mathbf{z}}_i\|_2^2}{\|\mathbf{w}_i\|_2^2}\left(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T\right) \\ &\le \max_{i=1}^n \frac{\|\widetilde{\mathbf{z}}_i\|_2^2}{\|\mathbf{w}_i\|_2^2}\lambda_{\max}(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T)I_n.\end{aligned}$$

This shows that

$$\lambda_{\min}(\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T) \ge \min_{i=1}^n \frac{\|\widetilde{\mathbf{z}}_i\|_2^2}{\|\mathbf{w}_i\|_2^2}\lambda_{\min}(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T) \tag{18}$$

and

$$\lambda_{\max}(\widetilde{p}^{-1}\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T) \le \max_{i=1}^n \frac{\|\widetilde{\mathbf{z}}_i\|_2^2}{\|\mathbf{w}_i\|_2^2}\lambda_{\max}(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T). \tag{19}$$

As mentioned before, we have for some $c_1 \in (1, \infty)$ and $C_1 \in (0, \infty)$,

$$P\left(\lambda_{\max}(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T) > c_1 \text{ or } \lambda_{\min}(\widetilde{p}^{-1}\mathbf{W}\mathbf{W}^T) < 1/c_1\right) \le e^{-C_1 n}. \tag{20}$$

Note that $\widetilde{p} > n$. Thus by (7) in Lemma 3, an application of Bonferroni's inequality gives

$$P\left(\min_{i=1}^n \frac{\|\widetilde{\mathbf{z}}_i\|_2^2}{\|\mathbf{w}_i\|_2^2} < c_4 \quad \text{or} \quad \max_{i=1}^n \frac{\|\widetilde{\mathbf{z}}_i\|_2^2}{\|\mathbf{w}_i\|_2^2} > c_5\right) \le 4ne^{-C_3 n}, \tag{21}$$

where $c_4 \in (0, 1)$, $c_5 \in (1, \infty)$, and $C_3 \in (0, \infty)$. Therefore by Bonferroni's inequality, combining (20) and (21) proves the deviation inequality in (2) by appropriately changing the constants $c_1 \in (1, \infty)$ and $C_1 \in (0, \infty)$. This concludes the proof.

A.5. Proof of Theorem 2. Using the similar arguments as in the proof of Theorem 1, we can prove that there exist some universal positive constants c_6, C such that the deviation probability bound

$$P\left\{\lambda_{\min}(n^{-1}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}}) < c_6\right\} \le \exp(-Cn) \tag{22}$$

holds for each $n \times \widetilde{p}$ submatrix $\widetilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $\widetilde{p} = c_7 n$ and $c_7 < 1$ some positive constant. This is because Lemmas 1-2 are free of the dimension q of the marginal distribution, and Lemma 3 still holds with the choice of $q = \widetilde{p}$. We should also note that the deviation probability bound (22) holds when $\widetilde{\mathbf{Z}} \sim N(\mathbf{0}, I_n \otimes I_{\widetilde{p}})$, which is entailed by the concentration property (2) proved in Fan and Lv (2008) for Gaussian distributions.

For each set $\alpha \subset \{1, \cdots, p\}$ with $|\alpha| = \widetilde{p}$, denote by $\mathbf{\Sigma}_{\alpha,\alpha}$ the principal submatrix of $\mathbf{\Sigma}$ corresponding to variables in α, and $\mathbf{X}_\alpha$ a submatrix of the design matrix $\mathbf{X}$ consisting of columns with indices in α. It follows easily from the representation of elliptical distributions that $\widetilde{\mathbf{X}}_\alpha = \mathbf{X}_\alpha \mathbf{\Sigma}_{\alpha,\alpha}^{-1/2}$ has the same distribution as $\mathbf{Z}_\alpha$. Since $\lambda_{\min}(\mathbf{\Sigma})$ is bounded from below by some positive constant, we have

$$\lambda_{\min}\big(n^{-1}\mathbf{X}_\alpha^T\mathbf{X}_\alpha\big) \ge \lambda_{\min}\big(n^{-1}\widetilde{\mathbf{X}}_\alpha^T\widetilde{\mathbf{X}}_\alpha\big)\lambda_{\min}(\mathbf{\Sigma}_{\alpha,\alpha}) \ge C\lambda_{\min}\big(n^{-1}\widetilde{\mathbf{X}}_\alpha^T\widetilde{\mathbf{X}}_\alpha\big),$$

where C is some positive constant. Therefore, combining the above results yields

$$P\left\{\lambda_{\min}\big(n^{-1}\mathbf{X}_\alpha^T\mathbf{X}_\alpha\big) < c_6\right\} \le \exp(-Cn) \tag{23}$$

with a possibly different positive constant c_6. Note that the positive constants involved are universal ones. We choose a positive integer $K = 2^{-1}Cn/(\log p) \le \widetilde{p}$. Then an application of the Bonferroni inequality together with (23) gives

$$P\left\{\min_{|\alpha|=K} \lambda_{\min}\big(n^{-1}\mathbf{X}_\alpha^T\mathbf{X}_\alpha\big) < c_6\right\} \le \sum_{|\alpha|=K} \exp(-Cn) \le p^K \exp(-Cn) \to 0$$

as $n \to \infty$. This shows that with asymptotic probability one, the robust spark $\kappa_c \ge K$ for any $c \le c_6$, which completes the proof.

A.6. Proof of Theorem 3. For the maximum chordal distance d_m, by noting that $x_i = \sin^2\theta_i$, we have a simple representation of the neighborhood $B_{\delta,d_m} = \{(x_1, \cdots, x_s) \in [0,1]^s : \max_{i=1}^s x_i \le \delta^2\}$ for $\delta \in (0, 1/2)$. We need to calculate its volume under the probability measure ν given in (43). In light of (43), a change of variable $y_i = \delta^{-2}x_i$ gives

$$d\nu = \delta^{s(n-s)}K_{n,s}f(y_1, \cdots, y_s) \prod_{1\le i<j\le s} |y_i - y_j| \prod_{i=1}^s y_i^{\alpha-1} dy_1 \cdots dy_s, \tag{24}$$

where $f(y_1, \cdots, y_s) = \prod_{i=1}^{s}(1-\delta^2 y_i)^{-1/2}$ over $[0,1]^s$. Observe that without the term f in (24), $\nu(B_{\delta,d_m})$ would become Selberg's integral which is a generalization of the beta integral (Mehta, 2004). We will evaluate this integral by sandwiching the function f between two functions of the same form. Since the function $(1-\delta^2 y)^{-1/2}$ is increasing and convex on $[0,1]$, it follows that $1 + c_1 y \le (1-\delta^2 y)^{-1/2} \le 1 + c_2 y$, where $c_1 = \delta^2/2$ and $c_2 = \delta^2(1-\delta^2)^{-3/2}/2$. This shows that

$$\prod_{i=1}^{s}(1+c_1 y_i) \le f(y_1, \cdots, y_s) \le \prod_{i=1}^{s}(1+c_2 y_i).$$

Thus, we obtain a useful representation of the volume of the neighborhood

$$\nu(B_{\delta,d_m}) = \delta^{s(n-s)} K_{n,s} I(c), \tag{25}$$

where $I(c)$ with some $c \in [c_1, c_2]$ is an integral given in the following lemma.

LEMMA 4. *For each $c > 0$, we have*

$$\begin{aligned} I(c) &\equiv \int_{[0,1]^s} \prod_{i=1}^{s}(1+cy_i) \prod_{1\le i<j\le s} |y_i - y_j| \prod_{i=1}^{s} y_i^{\alpha-1} dy_1 \cdots dy_s \\ &= 2^s \pi^{-s/2} \sum_{m=0}^{s} \binom{s}{m} c^m \prod_{i=s-m}^{s-1} \frac{\alpha + 2^{-1} i}{\alpha + 2^{-1}(s+i+1)} \\ &\quad \cdot \prod_{i=0}^{s-1} \frac{\Gamma(\alpha + 2^{-1} i)\Gamma(1 + 2^{-1}(i+1))\Gamma(1+2^{-1} i)}{\Gamma(\alpha + 2^{-1}(s+i+1))}, \end{aligned} \tag{26}$$

where the factor containing m equals 1 when $m = 0$.

Proof of Lemma 4. Observe that the integrand in (26) is symmetric in $y_1, \cdots, y_s$. Thus an expansion of $\prod_{i=1}^{s}(1+cy_i)$ gives

$$\begin{aligned} &\int_{[0,1]^s} \prod_{i=1}^{s}(1+cy_i) \prod_{1\le i<j\le s} |y_i - y_j| \prod_{i=1}^{s} y_i^{\alpha-1} dy_1 \cdots dy_s \\ &= \sum_{m=0}^{s} \binom{s}{m} c^m \int_{[0,1]^s} \prod_{i=1}^{m} y_i \prod_{1\le i<j\le s} |y_i - y_j| \prod_{i=1}^{s} y_i^{\alpha-1} dy_1 \cdots dy_s, \end{aligned}$$

where $\prod_{i=1}^{m} y_i = 1$ when $m = 0$. The above integrals are exactly Aomoto's

extension of Selberg's integral (Mehta, 2004) and can be calculated as

$$\int_{[0,1]^s} \prod_{i=1}^{m} y_i \prod_{1\le i<j\le s} |y_i - y_j| \prod_{i=1}^{s} y_i^{\alpha-1} dy_1 \cdots dy_s$$
$$= 2^s \pi^{-s/2} \prod_{i=s-m}^{s-1} \frac{\alpha + 2^{-1} i}{\alpha + 2^{-1}(s+i+1)} \prod_{i=0}^{s-1} \frac{\Gamma(\alpha + 2^{-1} i)\Gamma(1 + 2^{-1}(i+1))\Gamma(1 + 2^{-1} i)}{\Gamma(\alpha + 2^{-1}(s+i+1))},$$

where the factor containing m equals 1 when $m = 0$. This completes the proof of Lemma 4.

Let us continue with the proof of Theorem 3. By assumption, $s \sim \gamma n$ as $n \to \infty$, so $\delta^{s(n-s)}\pi^{-s/2} \sim \delta^{\gamma(1-\gamma)n^2}\pi^{-\gamma n/2}$. Applying Stirling's formula for large factorials gives $s! \sim (2\pi\gamma n)^{1/2}(\gamma/e)^{\gamma n} n^{\gamma n}$. Thus by omitting $O(n)$ and smaller order terms,

$$\log[\delta^{s(n-s)}\pi^{-s/2}/s!] \sim (\log\delta)\gamma(1-\gamma)n^2 - \gamma n \log n. \tag{27}$$

Using Stirling's formula for the Gamma function $\Gamma(t+1) \sim (2\pi t)^{1/2}(t/e)^t$ as $t \to \infty$ and noting that $A_j = 2\pi^{j/2}/\Gamma(j/2)$ and $\alpha = (n - 2s + 1)/2$, we derive

$$\prod_{i=0}^{s-1} \frac{A_{n-s-i}}{A_{n-i}} \prod_{i=0}^{s-1} \frac{\Gamma(\alpha + 2^{-1} i)}{\Gamma(\alpha + 2^{-1}(s+i+1))} \sim \pi^{-s^2/2}(2e)^{s/2}(n-s-1)^{(n-s-1)/2}(n-1)^{-(n-1)/2},$$

which entails that

$$\log\left\{\prod_{i=0}^{s-1} \frac{A_{n-s-i}}{A_{n-i}} \prod_{i=0}^{s-1} \frac{\Gamma(\alpha + 2^{-1} i)}{\Gamma(\alpha + 2^{-1}(s+i+1))}\right\} \sim -(\log\pi)\gamma^2 n^2/2 - \gamma n(\log n)/2. \tag{28}$$

Similarly, it follows from the identities $\Gamma(t+1) = t\Gamma(t)$ and $\Gamma(1) = 1$ that

$$2^{-s} \prod_{i=0}^{s-1} A_{s-i}^2 \prod_{i=0}^{s-1} \Gamma(1 + 2^{-1}(i+1))\Gamma(1 + 2^{-1} i) = s!\pi^{s(s+1)/2}/\Gamma(s/2).$$

This shows that

$$\log\left\{2^{-s} \prod_{i=0}^{s-1} A_{s-i}^2 \prod_{i=0}^{s-1} \Gamma(1 + 2^{-1}(i+1))\Gamma(1 + 2^{-1} i)\right\} \sim (\log\pi)\gamma^2 n^2/2 + \gamma n(\log n)/2. \tag{29}$$

It remains to consider the last term. Note that

$$\prod_{i=s-m}^{s-1} \frac{\alpha + 2^{-1}i}{\alpha + 2^{-1}(s+i+1)} = \frac{(n-s)!}{(n+1)!}\frac{(n-m+1)!}{(n-m-s)!}$$
$$\sim \left(\frac{n-m+1}{n+1}\right)^{s+1}\left(\frac{n-s}{n+1}\right)^{n-s+1/2}\left(\frac{n-m+1}{n-m-s}\right)^{n-m-s+1/2} \sim e^{O(n)},$$

which entails that

(30)
$$\log\left\{\sum_{m=0}^{s}\binom{s}{m}c^m \prod_{i=s-m}^{s-1}\frac{\alpha+2^{-1}i}{\alpha+2^{-1}(s+i+1)}\right\} \sim \log[e^{O(n)}(1+c)^s] = O(n).$$

Thus combining (25) and (27)–(30) yields

$$\log(\nu(B_{\delta,d_m})) \sim (\log\delta)\gamma(1-\gamma)n^2 - \gamma n\log n + O(n). \tag{31}$$

Since all the subspaces in $\mathcal{A}_s$ have maximum chordal distance at least 2δ, it holds that $\binom{p}{s} = |\mathcal{A}_s| \le 1/\nu(B_{\delta,d_m})$. Hence by (31),

$$\log\binom{p}{s} \lesssim (\log\delta^{-1})\gamma(1-\gamma)n^2 + \gamma n\log n + O(n),$$

where $\lesssim$ denotes asymptotic dominance. It is easy to derive $\log\binom{p}{s} \gtrsim \gamma n\log p - \gamma n\log n$. These two results lead to the claimed bound on $\log p$, which concludes the proof.

A.7. Proof of Theorem 4. Let us fix an arbitrary subset $\{\mathbf{x}_{j_1},\cdots,\mathbf{x}_{j_s}\}$ and denote by $\mathcal{A}_s^{p-s}$ the set of s-subspaces spanned by s of the remaining $p-s$ $\mathbf{x}_j$'s. By assumption, $\mathcal{A}_s^{p-s}$ lies in a neighborhood in the Grassmann manifold $G_{n,s}$ that is characterized by the set $R_{\delta_1} = \{(x_1,\cdots,x_s)\in[0,1]^s : \min_{i=1}^s x_i \ge \delta_1^2\}$, since $x_i = \sin^2\theta_i$. In view of (43), a change of variable $y_i = (1-\delta_1^2)^{-1}(1-x_i)$ gives

$$d\nu = (1-\delta_1^2)^{s^2/2}K_{n,s}f_1(y_1,\cdots,y_s)\prod_{1\le i<j\le s}|y_i-y_j|\prod_{i=1}^{s}y_i^{-1/2}dy_1\cdots dy_s, \tag{32}$$

where $f_1(y_1,\cdots,y_s) = \prod_{i=1}^s[1-(1-\delta_1^2)y_i]^{\alpha-1}$ over $[0,1]^s$. Clearly $[1-(1-\delta_1^2)y_i]^{\alpha-1} \ge (1-y_i)^{\alpha-1}$ for $\alpha\ge 1$, which together with (43) and (32) entails that $(1-\delta_1^2)^{s^2/2}$ is a lower bound on the integral $\nu(R_{\delta_1})$. However, we need an upper bound on it. The idea is to bound the function f_1 by an exponential function.

We are more interested in the asymptotic behavior of $\nu(R_{\delta_1})$ when δ_1 is near zero. Using the inequality $\log(1+t) \le t$, we derive

$$1-(1-\delta_1^2)y = (1-\delta_1^2)[1+(1-\delta_1^2)^{-1}\delta_1^2 - y] \le (1-\delta_1^2)e^{(1-\delta_1^2)^{-1}\delta_1^2}e^{-y}.$$

This leads to

$$f_1(y_1,\cdots,y_s) \le [(1-\delta_1^2)e^{(1-\delta_1^2)^{-1}\delta_1^2}]^{s(\alpha-1)}\prod_{i=1}^{s} e^{-(\alpha-1)y_i}.$$

Thus we have

$$\nu(R_{\delta_1}) \le (1-\delta_1^2)^{2^{-1}s^2+s(\alpha-1)}e^{(1-\delta_1^2)^{-1}\delta_1^2 s(\alpha-1)}K_{n,s}B_\alpha, \tag{33}$$

where $B_\alpha = \int_{[0,1]^s}\prod_{1\le i<j\le s}|y_i-y_j|\prod_{i=1}^{s} y_i^{-1/2}e^{-(\alpha-1)y_i}dy_1\cdots dy_s$. Since $\alpha-1=(n-2s-1)/2\to\infty$ as $n\to\infty$, a change of variable $z_i=(\alpha-1)y_i$ gives

$$\begin{aligned} B_\alpha &= (\alpha-1)^{-s^2/2}\int_{[0,\alpha-1]^s}\prod_{1\le i<j\le s}|z_i-z_j|\prod_{i=1}^{s} z_i^{-1/2}e^{-z_i}dz_1\cdots dz_s \\ &\le (\alpha-1)^{-s^2/2}\int_{[0,\infty)^s}\prod_{1\le i<j\le s}|z_i-z_j|\prod_{i=1}^{s} z_i^{-1/2}e^{-z_i}dz_1\cdots dz_s. \end{aligned}$$

Note that the last integral is a Selberg type integral related to the Laguerre polynomials (Mehta, 2004), which can be calculated exactly. This along with the identities $\Gamma(t+1)=t\Gamma(t)$ and $\Gamma(3/2)=\pi^{1/2}/2$ yields

$$B_\alpha \lesssim (\alpha-1)^{-s^2/2}s!\pi^{-s/2}\prod_{i=1}^{s}\Gamma^2(i/2).$$

By assumption, $s\sim\gamma n$ as $n\to\infty$. It is easy to show that

$$\log[(1-\delta_1^2)^{2^{-1}s^2+s(\alpha-1)}e^{(1-\delta_1^2)^{-1}\delta_1^2 s(\alpha-1)}] \sim -c_{\delta_1}\gamma n^2 + O(n), \tag{34}$$

where $c_{\delta_1} = 2^{-1}\left[\log(1-\delta_1^2)^{-1}\right](1-\gamma) - 2^{-1}(1-\delta_1^2)^{-1}\delta_1^2(1-2\gamma)$. It remains to consider the term $K_{n,s}B_\alpha$. By (42), we have

$$\begin{aligned} 2^{-s}\pi^{-s/2}\widetilde{K}_{n,s}\prod_{i=1}^{s}\Gamma^2(i/2) &= \prod_{i=0}^{s-1}\frac{\Gamma((n-i)/2)}{\Gamma((n-s-i)/2)} \\ &\sim \prod_{i=0}^{s-1}\left(\frac{n-i-2}{n-s-i-2}\right)^{(n-s-i-1)/2}\left[\frac{(n-2)!}{(n-s-2)!}\right]^{s/2}(2e)^{-s^2/2}, \end{aligned}$$

where we used Stirling's formula for the Gamma function in the last step. It follows from $s \sim \gamma n$ that

$$\prod_{i=0}^{s-1} \left(\frac{n-i-2}{n-s-i-2} \right)^{(n-s-i-1)/2} \lesssim \left(\frac{1-\gamma}{1-2\gamma} \right)^{\gamma(1-\gamma)n^2/2},$$

and an application of Stirling's formula for large factorials gives

$$\begin{aligned} \left[\frac{(n-2)!}{(n-s-2)!} \right]^{s/2} (2e)^{-s^2/2} &\sim \left(\frac{n-2}{n-s-2} \right)^{s(n-s-3/2)/2} [(n-2)/(2e^2)]^{s^2/2} \\ &\lesssim (1-\gamma)^{-\gamma(1-\gamma)n^2/2} [n/(2e^2)]^{\gamma^2 n^2/2}. \end{aligned}$$

Note that $(\alpha-1)^{-s^2/2} \sim [(1-2\gamma)n/2]^{-\gamma^2 n^2/2}$. Combing these results together yields

$$\begin{aligned} \log(K_{n,s} B_\alpha) &= \log[(\alpha-1)^{-s^2/2} 2^{-s} \pi^{-s/2} \widetilde{K}_{n,s} \prod_{i=1}^{s} \Gamma^2(i/2)] \\ &\lesssim -[2\gamma + \log(1-2\gamma)]\gamma n^2/2. \end{aligned} \tag{35}$$

It follows from (33)–(35) that

$$\log(\nu(R_{\delta_1})) \lesssim -c_{\delta_1} \gamma n^2 - [2\gamma + \log(1-2\gamma)]\gamma n^2/2 + O(n). \tag{36}$$

Finally we are ready to derive a bound on the dimensionality p. Since $\mathcal{A}_s^{p-s}$ lies in a neighborhood in $G_{n,s}$ characterized by the set R_{δ_1} and all the subspaces in $\mathcal{A}_s^{p-s}$ have maximum chordal distance at least δ, it holds that $\binom{p-s}{s} = |\mathcal{A}_s^{p-s}| \leq \nu(R_{\delta_1})/\nu(B_{\delta,d_m})$. Aided by (31) and (36), a similar argument as in the proof of Theorem 3 gives the claimed bound on $\log(p-s)$. This completes the proof.

A.8. Proof of Theorem 5. To prove the conclusion, we use the terminology introduced in Section 3. Note that the n-vectors $\mathbf{x}_j$ and $\mathbf{y}$ can be viewed as elements of Grassmannian manifold $G_{n,1}$, which consists of all one-dimensional subspaces of $\mathbb{R}^n$. The absolute correlation between two n-vectors is given by $\cos\theta_1$, where $\theta_1 \in [0, \pi/2]$ is the principal angle between the two corresponding one-dimensional subspaces. We use the parametrization with local coordinate θ_1 at the one-dimensional subspace $L = \{t\mathbf{y} : t \in \mathbb{R}\}$ spanned by $\mathbf{y}$. Then the uniform distribution on the Grassmann manifold $G_{n,1}$ can be expressed in local coordinate θ_1 and gives a probability measure ν in (43) with $s = 1$ on $[0, 1]$ through a change of variable $x_1 = \sin^2\theta_1$, where $\prod_{1 \leq i < j \leq s} |x_i - x_j| = 1$ in this case. Consider the maximum chordal

distance on $G_{n,1}$, which is defined as $\sin\theta_1$. For any $t > 0$, denote by B_{t,d_m} a ball of radius t centered at L in $G_{n,1}$ under the maximum chordal distance, i.e., $B_{t,d_m} = \{x_1 \in [0,1] : x_1 = \sin^2\theta_1 \le t^2\}$ in local coordinate. We need to calculate the volumes of B_{t_1,d_m} with $t_1 = 2^{-1}\sin\cos^{-1}\delta = (1-\delta^2)^{1/2}/2$ and $B^c_{t_2,d_m}$, the complement of B_{t_2,d_m} with $t_2 = \sin\cos^{-1} r = (1-r^2)^{1/2}$, under the measure ν.

In view of (43), we have

$$\nu(B_{t,d_m}) = \nu([0,t^2]) = K_{n,1}\int_0^{t^2} x_1^{(n-3)/2}(1-x_1)^{-1/2}dx_1, \tag{37}$$

where $K_{n,1} = \widetilde{K}_{n,1}/2 = A_1^2 A_{n-1}/(4A_n)$ with $A_j = 2\pi^{j/2}/\Gamma(j/2)$ the area of the unit sphere $S^{j-1} \subset \mathbb{R}^j$. It follows from Stirling's formula for the Gamma function that

$$\begin{aligned} K_{n,1} &= \pi^{-1/2}\Gamma(n/2)/\Gamma((n-1)/2) \sim \pi^{-1/2}\left(\frac{n-2}{2e}\right)^{1/2}\left(\frac{n-2}{n-3}\right)^{(n-2)/2} \\ &\sim (2\pi)^{-1/2}n^{1/2}. \end{aligned} \tag{38}$$

It remains to evaluate the integral in (37). Note that $(1-x_1)^{-1/2}$ is bounded between 1 and ∞ on $[0,t^2]$ for t bounded away from 1. Thus we have

$$\int_0^{t^2} x_1^{(n-3)/2}(1-x_1)^{-1/2}dx_1 > \int_0^{t^2} x_1^{(n-3)/2}dx_1 = 2(n-1)^{-1}t^{n-1}, \tag{39}$$

where both sides have the same asymptotic order. Combining (37)–(39) yields $\nu(B_{t,d_m}) > c_n t^{n-1}$ with $c_n \sim (2/\pi)^{1/2}n^{-1/2}$, and $\nu(B_{t,d_m}) \sim (2/\pi)^{1/2}n^{-1/2}t^{n-1}$ for t bounded away from 1. Since all the $p-s$ noise predictors $\mathbf{x}_j$ have absolute correlations bounded by $\delta \in (0,1)$, we have

$$p - s \le \nu(B^c_{t_2,d_m})/\nu(B_{t_1,d_m}) \tag{40}$$

if there exists no noise predictor that has absolute correlation with $\mathbf{y}$ larger than $r \in (0,1)$. The right hand side of (40) is $[1-\nu(B_{t_2,d_m})]/\nu(B_{t_1,d_m})$, which is less than and has the same asymptotic order as $1/\nu(B_{t_1,d_m}) < c_n^{-1}[4/(1-\delta^2)]^{(n-1)/2} \sim (\pi/2)^{1/2}n^{1/2}[4/(1-\delta^2)]^{(n-1)/2}$. This together with (40) concludes the proof.

APPENDIX B: GEOMETRY AND INVARIANT MEASURE OF GRASSMANN MANIFOLD

We briefly introduce some necessary background and terminology on the geometry and invariant measure of Grassmann manifold. Let V_1 and V_2

be two s-dimensional subspaces of $\mathbb{R}^n$ and $d_g(\cdot,\cdot) = \arccos|\langle\cdot,\cdot\rangle|$ be the geodesic distance on S^{n-1}, that is, the distance induced by the Euclidean metric on $\mathbb{R}^n$. It was shown by James (1954) that as v_1 and v_2 vary over $V_1 \cap S^{n-1}$ and $V_2 \cap S^{n-1}$ respectively, $d_g(v_1, v_2)$ has a set of s critical values $\angle(V_1, V_2) = (\theta_1, \cdots, \theta_s)$ with $\pi/2 \geq \theta_1 \geq \cdots \geq \theta_s \geq 0$, corresponding to s pairs of unit vectors (v_{1i}, v_{2i}), $i = 1, \cdots, s$. Each critical value θ_i is exactly the angle between v_{1i} and v_{2i}, and v_{1i} is orthogonal to v_{1j} and v_{2j} if $j \neq i$. The principal angles θ_i are unique and if none of them are equal, the principle vectors (v_{1i}, v_{2i}) are unique up to a simultaneous direction reversal. In general, the dimensions of V_1 and V_2 can be different, in which case s should be their minimum.

All s-dimensional subspaces of $\mathbb{R}^n$ form a space, the so-called Grassmann manifold $G_{n,s}$. It is a compact Riemannian homogeneous space, of dimension $s(n-s)$, isomorphic to $O(n)/(O(s) \times O(n-s))$, where $O(j)$ denotes the orthogonal group of order j. It is well known that $G_{n,s}$ admits an invariant measure μ. It can be constructed by viewing $G_{n,s}$ as $V_{n,s}/O(s)$, where $V_{n,s} \cong O(n)/O(n-s)$ denotes the Stiefel manifold of all orthonormal s-frames (that is, sets of s orthonormal vectors) in $\mathbb{R}^n$. By deriving the exterior differential forms on those manifolds (James, 1954), $d\mu(V)$ can be expressed in local coordinates, at the s-dimensional subspace with generator matrix $(I_s\ 0)$, as a product of three independent densities $\prod_{i=1}^3 d\mu_i$, where

$$d\mu_1 = \widetilde{K}_{n,s} \prod_{i=1}^{s} (\sin\theta_i)^{n-2s} \prod_{1\leq i<j\leq s} (\sin^2\theta_i - \sin^2\theta_j) d\theta_1 \cdots d\theta_s \tag{41}$$

over $\Theta = \{(\theta_1, \cdots, \theta_s) : \pi/2 > \theta_1 > \cdots > \theta_s > 0\}$, and $d\mu_2$ and $d\mu_3$ are independent of parameters $(\theta_1, \cdots, \theta_s)$. The normalization constant is given by

$$\widetilde{K}_{n,s} = \prod_{i=0}^{s-1} \frac{A_{s-i}^2 A_{n-s-i}}{2A_{n-i}}, \tag{42}$$

where $A_j = 2\pi^{j/2}/\Gamma(j/2)$ is the area of the unit sphere S^{j-1}. A change of variable $x_i = \sin^2\theta_i$ and symmetrization in (41) yield a probability measure ν on $[0,1]^s$ with density

$$d\nu = K_{n,s} \prod_{1\leq i<j\leq s} |x_i - x_j| \prod_{i=1}^{s} x_i^{\alpha-1} \prod_{i=1}^{s} (1-x_i)^{-1/2} dx_1 \cdots dx_s, \tag{43}$$

where $K_{n,s} = \widetilde{K}_{n,s}/(2^s s!)$ and $\alpha = (n-2s+1)/2$.

REFERENCES

[1] Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–1732.

[2] Bühlmann, P. and Mandozzi, J. (2012). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Manuscript*.

[3] Chamberlain, G. (1983). A characterization of the distributions that imply mean-variance utility functions. *Journal of Economic Theory* **29**, 185–201.

[4] Conway, J. H., Hardin, R. H. and Sloane, N. J. A. (1996). Packing lines, planes, etc.: Packings in Grassmannian spaces. *Experiment. Math.* **5**, 139–159.

[5] Delaigle, A. and Hall, P. (2012). Effect of heavy tails on ultra high dimensional variable ranking methods. *Statistica Sinica* **22**, 909–932.

[6] Donoho, D. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proc. Natl. Acad. Sci. USA* **100**, 2197–2202.

[7] Edelman, A., Arias, T. and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**, 303–353.

[8] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407–499.

[9] Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36**, 2605–2637.

[10] Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in ultra-high dimensional additive models. *J. Amer. Statist. Assoc.* **106**, 544–557.

[11] Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians* (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. **III**, 595–622.

[12] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849–911.

[13] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica* **20**, 101–148.

[14] Fang, K.-T., Kotz, S. and Ng, K.-W. (1990). *Symmetric Multivariate and Related Distributions*. London: Chapman and Hall.

[15] Hall, P. (2006). Some contemporary problems in statistical science. In *Madrid Intelligencer* (Edited by F. Chamizo and A. Quirós), 38–41. Springer, New York.

[16] Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. Roy. Statist. Soc. Ser. B* **67**, 427–444.

[17] Hall, P., Titterington, D. M. and Xue, J.-H. (2009). Tilting methods for assessing the influence of components in a classifier. *J. Roy. Statist. Soc. Ser. B* **71**, 783–803.

[18] James, A. T. (1954). Normal multivariate analysis and the orthogonal group. *Ann. Math. Statist.* **25**, 40–75.

[19] Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. Cambridge: American Mathematical Society.

[20] Li, R., Zhong, W. and Zhu, L.P. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107**, 1129–1139.

[21] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37**, 3498–3528.

[22] Mai, Q. and Zou, H. (2012). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, to appear.

[23] Mehta, M. L. (2004). *Random Matrices* (3rd edition). Academic Press, Amsterdam.

[24] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York: John

Wiley.

[25] Owen, J. and Rabinovitch, R. (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *Journal of Finance* **38**, 745–752.

[26] Samworth, R. (2008). Discussion of "Sure independence screening for ultrahigh dimensional feature space." *J. Roy. Statist. Soc. Ser. B* **70**, 888–889.

[27] van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.* **3**, 1360–1392.

[28] Wong, Y.-C. (1967). Differential geometry of Grassmann manifolds. *Proc. Nat. Acad. Sci. U.S.A.* **57**, 589–594.

[29] Xue, L. and Zou, H. (2011). Sure independence screening and compressed random sensing. *Biometrika* **98**, 371–380.

[30] Zheng, Z., Fan, Y. and Lv, J. (2012). High-dimensional thresholded regression and shrinkage effect. *Manuscript*.

[31] Zhu, L.P., Li, L., Li, R. and Zhu, L.X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106**, 1464–1475.

[32] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509–1566.

Information and Operations
Management Department
Marshall School of Business
University of Southern California
Los Angeles, CA 90089
USA
E-mail: jinchilv@marshall.usc.edu