Regularity Properties of High-dimensional Covariate Matrices*

Edgar Dobriban

Department of Statistics, Stanford University
e-mail: dobriban@stanford.edu

and

Jianqing Fan

Department of Operations Research and Financial Engineering, Princeton University
e-mail: jqfan@princeton.edu

Abstract: Regularity properties such as the incoherence condition, the restricted isometry property, compatibility, restricted eigenvalue and ℓ_q sensitivity of covariate matrices play a pivotal role in high-dimensional regression and compressed sensing. Yet, like computing the spark of a matrix, we first show that it is NP-hard to check the conditions involving all submatrices of a given size.

This motivates us to investigate what classes of design matrices satisfy these conditions. We demonstrate that the most general property, ℓ_q sensitivity, holds with high probability for covariate matrices sampled from populations with a suitably regular covariance matrix. The probability lower bound and sample size required depend on the tail behavior of the random noise. We examine this for three important cases, bounded, sub-Gaussian, and finite moment random noises.

We further show that ℓ_q sensitivity is preserved under natural operations on the data. Our work is particularly important for many statistical applications, in which the covariates are observational and correlated and can be thought of as fixed in advance.

AMS 2000 subject classifications: Primary 62J05; secondary 68Q17, 62H12.
Keywords and phrases: high-dimensional regression, instrumental variables, sparse estimation, compressed sensing, random matrix, restricted eigenvalue, compatibility, ℓ_q sensitivity, computational complexity, NP-hardness.

*Fan's research was partially supported by NIH grants R01GM100474-01 and NIH R01-GM072611 and NSF grants DMS-1206464 and DMS-0704337. The bulk of the research was carried out while Edgar Dobriban was an undergraduate student at Princeton University.

1. Introduction

The analysis of high-dimensional data is a central topic of statistics, motivated by advances in science, technology and engineering. Recent research revealed that estimation in high dimensions may be possible if the problems are suitably sparse. As a typical example, consider linear regression where most of the coefficients of the parameter vector are vanishing. In this setting, popular estimators include the Lasso (Chen, Donoho and Saunders, 2001; Tibshirani, 1996), folded concave penalized least-squares such as SCAD (Fan and Li, 2001), and the Dantzig selector (Candès and Tao, 2007). Sparsity has been exploited in a number of other questions, for instance instrumental variables regression in the presence of endogeneity (Gautier and Tsybakov, 2011).

The Lasso and Dantzig selector have small estimation error as long as the matrix of covariates obeys one of a variety of conditions. The incoherence condition of Donoho and Huo (2001) provides the earliest and simplest example. Later Candès and Tao (2005) introduced the restricted isometry property and showed its application to the Dantzig selector (Candès and Tao, 2007). In subsequent work Bickel, Ritov and Tsybakov (2009) analyzed the estimators under the weaker and more general restricted eigenvalue (RE) condition. The compatibility conditions of van de Geer (2007) are closely related. See van de Geer and Bühlmann (2009) for the relationship between these properties. Gautier and Tsybakov (2011) have recently introduced an estimator for instrumental variables regression, along with the ℓ_q sensitivity properties that guarantee small estimation error. ℓ_q sensitivity is the weakest and most general of the above properties, and also applies to linear regression. It is closely related to the cone invertibility factors of Ye and Zhang (2010).

We investigate in depth the conditions of the design matrices needed for high-dimensional sparse estimation. We first deal with the computational complexity of checking the properties on general design matrices. The locations of the non-vanishing coefficients of the regression parameter are unknown, so we must make a non-degeneracy assumption uniformly over all subsets of a given size. This suggests that the conditions may be hard to check. We confirm this by showing that checking any of the restricted eigenvalue, compatibility, and ℓ_q sensitivity properties for general data matrices is NP-hard. This implies that there is no efficient way to check them, under the widely believed conjecture that $\mathsf{P} \neq \mathsf{NP}$. Our result builds on the recent proof that computing the spark and checking the restricted isometry property is NP-hard (Bandeira et al., 2013; Tillmann and Pfetsch, 2012).

Verifying the needed matrix properties is an important problem, recognized in a number of places in the literature. Tao (2007); Raskutti, Wainwright and Yu (2010); d'Aspremont and El Ghaoui (2011) discuss it as a problem of interest. Verification leads to guarantees that the inference procedure was successful. From a statistical point of view, the numerical values of the regularity conditions yield confidence sets for the regression parameter. The difficulty of their computation has already motivated several research works. For instance convex relaxations have been proposed for approximating the restricted isometry contant (d'Aspremont, Bach and Ghaoui, 2008; Lee and Bresler, 2008), and linear relaxations for the ℓ_q sensitivity (Gautier and Tsybakov, 2011).

Incoherence conditions are easy to check in polynomial time, in contrast to the hardness for the other properties. However they do not provide optimal rate of convergence: they require the sample size n of quadratic order s^2 (Bunea, 2007; Bandeira et al., 2012), while other conditions allow linear order (see e.g. Candès and Tao, 2005; Raskutti, Wainwright and Yu, 2010).

As a way to address the problem of non-verifiability, we show that our conditions hold with high probability if the covariate matrix is randomly sampled from a suitably well-behaved distribution. This extends the well-understood results for matrices with independent entries to correlated observation vectors, generated independently from a regular population. Previous results have been obtained for RIP (e.g. Rauhut, Schnass and Vandergheynst, 2008; Vershynin, 2010), and RE (Raskutti, Wainwright and Yu, 2010; Rudelson and Zhou, 2012). We establish new results for the more general ℓ_q sensitivity, under three probability models: observations that are (1) sub-gaussian, (2) bounded, and (3) have bounded moments. These results are useful if a population covariance model is known and easier to analyze. This is often the case, as illustrated by our examples, and those in van de Geer and Bühlmann (2009); Raskutti, Wainwright and Yu (2010).

Finally, we show that the ℓ_q sensitivity property is preserved under several natural operations on the data matrix. It is initially hard to ascertain whether this crucial property holds, but then there is a range of transformations one can apply to the data while preserving it.

In Section 2, we give definitions and the setup of our problem. In Section 3 we present our results, which are proven in Section 5. We finish with discussion in Section 4.

2. Definitions and Setup

We start with some basic notation, and then introduce the problems and notions we study: regression, associated estimators, regularity properties, sub-gaussian variables, and computational complexity.

2.1. Some notation

We denote by $|v|_q$ the vector ℓ_q norm. An s-sparse vector has at most s non-vanishing coordinates. For a set $S \subset \{1, \dots, p\}$ we denote by $|S|$ its cardinality and S^c its complement. For a vector $v = (v_1, \dots, v_p)^T$ and a subset S, we denote $v_S = (v_1 \mathbf{1}_{\{1 \in S\}}, \dots, v_p \mathbf{1}_{\{p \in S\}})^T$, where $\mathbf{1}_{\{.\}}$ is the indicator function. We denote by $\|M\|_{\max}$ the maximum absolute value of the entries of matrix M. For two sequences a_n and b_n of scalars, $a_n = O(b_n)$ means that there is a constant $c > 0$, such that $a_n \leq c b_n$ for all sufficiently large n. $a_n \asymp b_n$ means that $a_n = O(b_n)$ and $b_n = O(a_n)$. For random variables X_n, we write $X_n = O_P(1)$ if the collection X_n is bounded in probability, sometimes called uniformly tight. For two sequences of random variables X_n, Y_n, the notation $X_n = O_P(Y_n)$ means that there is a sequence of random variables $R_n = O_P(1)$, such that $X_n = R_n Y_n$.

2.2. Regression problems and estimators

In linear regression we want to explain a response variable y as a linear function of p covariates $x_1, \dots, x_p$, up to a noise term ε, via the model $y = \sum_{i=1}^p x_i \beta_i + \varepsilon$. To estimate β, we observe n independent samples: the $n \times 1$ response vector Y and the covariate vectors $X_1, X_2, \cdots, X_p$ of dimension n, forming the columns of an $n \times p$ matrix X. Hence now with a noise vector ε with independent $N(0, \sigma^2)$ entries, we have the model

$$Y = X\beta + \varepsilon.$$

We wish to estimate the p-dimensional parameter vector β in the case $n \ll p$.

We assume that most of the coordinates of β are vanishing, and that the design matrix X is regular, as specified in the next section. The locations of nonzero coordinates are unknown to us. In this setting the Lasso, or ℓ_1-penalized least squares, is a popular estimator (Tibshirani, 1996; Chen, Donoho and Saunders, 2001):

$$\hat{\beta}_{Lasso} = \arg\min_{\beta} \frac{1}{2n} |Y - X\beta|_2^2 + \lambda \sum_{i=1}^p |\beta_i|,$$

for a given regularization parameter.

The Dantzig selector is another estimator for this problem, which for a known noise level σ takes the form (Candès and Tao, 2007):

$$\hat{\beta}_{Dantzig} = \arg\min |\beta|_1, \text{ subject to } \left|\frac{1}{n} X^T (Y - X\beta)\right|_\infty \leq \sigma A \sqrt{\frac{2\log(p)}{n}},$$

where A is a tuning parameter.

In instrumental variables regression we also start with the model $y = \sum_{i=1}^p x_i \beta_i + \varepsilon$. Now some x_i may be correlated with the noise, in which case they are called endogeneous. Further, we have additional variables z_i, $i = 1, \dots L$, called instruments, that are uncorrelated with the noise. In addition to X, we observe n independent samples of z_i, which are arranged in the $n \times L$ matrix Z. In this setting, Gautier and Tsybakov (2011) propose the Self-Tuning Instrumental Variables (STIV) estimator, a generalization of the Dantzig selector. In the case where the noise level σ is known, STIV takes the form:

$$\min_{\beta \in \mathcal{I}} |D_X^{-1} \beta|_1 \tag{1}$$

with the minimum over the polytope

$$\mathcal{I} = \{\beta \in \mathbb{R}^p : \left|\frac{1}{n} D_Z Z^T (Y - X\beta)\right|_\infty \leq \sigma A \sqrt{\frac{2\log(L)}{n}}\}.$$

Here D_X and D_Z are the diagonal matrices with $(D_X)_{ii}^{-1} = \max_{k=1,\dots,n} |x_{ki}|$, $(D_Z)_{ii}^{-1} = \max_{k=1,\dots,n} |z_{ki}|$.

2.3. Regularity properties

The cone (rather, union of cones) $C(s, \alpha)$ is the set of vectors such that the ℓ_1 norm is concentrated on some s coordinates:

$$C(s, \alpha) = \{v \in \mathbb{R}^p : \exists S \subset \{1, \dots, p\}, |S| = s, \alpha |v_S|_1 \geq |v_{S^c}|_1\}.$$

The regularity properties discussed depend on a triplet of parameters (s, α, γ). In all cases s is the sparsity size of the problem, α is the cone opening parameter in $C(s, \alpha)$, and γ is the lower bound. They are all positive numbers. The first matrix property is the Restricted Eigenvalue condition from Bickel, Ritov and Tsybakov (2009); Koltchinskii (2009).

Definition 2.1. *A matrix X obeys the* **Restricted Eigenvalue** condition $RE(s, \alpha, \gamma)$, if

$$\frac{|Xv|_2}{|v_S|_2} \geq \gamma, \text{ for all } v \in C(s, \alpha), \alpha|v_S|_1 \geq |v_{S^c}|_1.$$

Bickel, Ritov and Tsybakov (2009) show that if the normalized data matrix $1/\sqrt{n}X$ obeys $RE(s, \alpha, \gamma)$ and β is s-sparse, then the estimation error is small:

$$|\hat{\beta} - \beta|_2 = O_P\left(\frac{1}{\gamma^2}\sqrt{\frac{s \log p}{n}}\right)$$

for both the Dantzig and Lasso selectors. The 'cone opening' α required in the restricted eigenvalue property equals 1 for Dantzig; 3 for Lasso. Next, we describe the compatibility condition from van de Geer (2007).

Definition 2.2. *A matrix X obeys the* **compatibility** condition with positive parameters (s, α, γ) if

$$\frac{\sqrt{s}|Xv|_2}{|v_S|_1} \geq \gamma, \text{ for all } v \in C(s, \alpha), \alpha|v_S|_1 \geq |v_{S^c}|_1.$$

The two conditions are very similar. The only difference is that the ℓ_1 versus ℓ_2 norm in the denominator. The inequality $|v_S|_1 \leq \sqrt{s}|v_S|_2$ immediately implies that the compatibility conditions are formally weaker than the RE assumptions. van de Geer (2007) provides an ℓ_1 oracle inequality for the Lasso under the compatibility condition. See also van de Geer and Bühlmann (2009); Bühlmann and van de Geer (2011).

The third and last assumption analyzed in this paper is the ℓ_q sensitivity property from Gautier and Tsybakov (2011).

Definition 2.3. *Let $q \geq 1$. The $n \times p$ matrix X and $n \times L$ matrix Z satisfy the ℓ_q* **sensitivity** property with parameters (s, α, γ), if

$$\frac{s^{1/q}|n^{-1}Z^TXv|_\infty}{|v|_q} \geq \gamma, \text{ for all } v \in C(s, \alpha).$$

Gautier and Tsybakov (2011) show that ℓ_q sensitivity is weaker than the restricted eigenvalue and compatibility conditions. In the case $Z = X$ the definition reduces to the cone invertibility factors of Ye and Zhang (2010). We note that the definition in Gautier and Tsybakov (2011) differs in normalization. We do not normalize for simplicity, to avoid the dependencies introduced by this process. As shown in Theorem 2.4, our definition works for

an un-normalized version of the STIV estimator. The argument is classical, but more general than Candès and Tao (2007) due to the use of instruments and ℓ_q sensitivity. Consider

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{I}} |\beta|_1 \quad \text{with} \quad \mathcal{I} = \{\beta \in \mathbb{R}^p : \left|\frac{1}{n} Z^T (Y - X\beta)\right|_\infty \leq \sigma\lambda\}. \tag{2}$$

Theorem 2.4. *Assume that z_j, $j = 1, \ldots, L$, and ε are mean zero sub-gaussian variables with sub-gaussian norm at most σ, β is s-sparse, and X, Z obey the ℓ_q sensitivity property with parameters $(s, 1, \gamma)$. Then, with n independent samples of data, taking $\lambda = A\sqrt{\frac{2\log(L)}{n}}$, the unnormalized STIV estimator (2) obeys*

$$\|\hat{\beta} - \beta\|_q = O_P\left(\frac{A\sigma s^{1/q}}{\gamma}\sqrt{\frac{\log(L)}{n}}\right).$$

Finally, we introduce the incoherence condition and restricted isometry property, which serve as contrasts with the above conditions. For an $n \times p$ matrix X whose columns $\{X_j\}_{j=1}^p$ are normalized to length $\sqrt{n}$, the mutual incoherence condition holds if $X_i^T X_j \leq \gamma/s$ for some positive γ. Such a notion was defined in Donoho and Huo (2001), and later used by Bunea (2007) to derive oracle inequalities for the Lasso.

A matrix X obeys the restricted isometry property with parameters s and δ if $(1-\delta)|v|_2^2 \leq |Xv|_2^2 \leq (1+\delta)|v|_2^2$ for all s-sparse vectors v (Candès and Tao, 2005).

2.4. Sub-gaussian vectors

The L_p norm of a random variable is $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$. A random variable X satisfying $\sup_{p \geq 1} p^{-1/2}\|X\|_p < \infty$ is called sub-gaussian, and its sub-gaussian norm is defined as $\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2}\|X\|_p$ (Vershynin, 2010). The random vector $\underline{X}$ is sub-gaussian if all one-dimensional marginals are sub-gaussian. The sub-gaussian norm of p-dimensional random vector $\underline{X}$ is then defined as

$$\|\underline{X}\|_{\psi_2} = \sup_{x \in S^{p-1}} \|\langle \underline{X}, x\rangle\|_{\psi_2}.$$

Here S^{p-1} is the Euclidean unit sphere in $\mathbb{R}^p$.

2.5. Notions from computational complexity

In complexity theory, problems are classified according to the computational resources - time and memory - needed to solve them on a Turing machine, a model for the computer (Arora and Barak, 2009).

A well-known example of a complexity class is P, consisting of the problems decidable in polynomial time in the size of the input. For input encoded in n bits, a yes or no answer must be found in time $O(n^k)$ for some fixed k. Another important class is NP, the decision problems for which already existing solutions can be verified in polynomial time. This is usually much easier than solving the question itself in polynomial-time. For instance, the subset-sum problem: 'Given a set of integers, does there exist a subset with zero sum?' is in NP, since one can easily check any purported solution - a subset of the given integers - to see if it indeed solves the problem. However, finding this subset seems harder: simply enumerating all subsets is not a polynomial-time algorithm.

Formally, the definition of NP requires that if the answer is *yes*, then there exists an easily verifiable proof. We have $\mathsf{P} \subset \mathsf{NP}$, since a polynomial-time solution is a certificate verifiable in polynomial time. However, it is a famous open problem to decide if P equals NP (Cook, 2000). It is widely believed in the complexity community that $\mathsf{P} \neq \mathsf{NP}$.

To compare the computational hardness of various problems, one can reduce known hard problems to the novel questions of interest, thereby demonstrating the difficulty of the novel problems. Specifically, a problem A is polynomial-time reducible to a problem B, if an oracle solving B - that is an immediate solver for an instance of B - can be queried once to give a polynomial-time algorithm to solve A. This is also variously known as a polynomial-time many-one reduction, strong reduction or Karp reduction. A problem is NP-*hard* if every problem in NP reduces to it, namely it is at least as difficult as all other problems in NP. If one reduces a known NP-hard problem to a new question, this demonstrates the NP-hardness of the new problem.

3. Results

3.1. Computational Complexity

We first show that the common conditions needed for sparse estimation are unfortunately NP-hard to verify. This builds on the recent results that com-

puting the spark and checking restricted isometry are NP-hard (Bandeira et al., 2013; Tillmann and Pfetsch, 2012).

Theorem 3.1. *Let X be an $n \times p$ matrix, Z an $n \times L$ matrix, $0 < s < n$, and $\alpha, \gamma > 0$. It is NP-hard to decide any of the following problems:*

1. *Does X obey the restricted eigenvalue condition with parameters (s, α, γ)?*
2. *Does X satisfy the compatibility conditions with parameters (s, α, γ) ?*
3. *Do X, Z obey the ℓ_q sensitivity property with parameters (s, α, γ)?*

The proof of Theorem 3.1 is found in Section 5.2. The theorem implies that there is no efficient way to check if a matrix is regular, provided $\mathsf{P} \neq \mathsf{NP}$.

Conditions like restricted isometry and restricted eigenvalue are central to both high-dimensional statistics and compressed sensing. Our result has more important implications for statistics. In signal processing and compressed sensing, one has a choice of a suitable random matrix - for instance with iid normal entries. Various random matrix ensembles appropriate for signal processing applications are regular with high probability, obeying even the restricted isometry property (Candès and Tao, 2005). Thus there may not be an urgent need for verification.

In statistical applications, however, the data matrix is often observational and correlated. The correlation between predictors is in many cases unknown, and may be substantial. It can be hard to judge if the matrix is regular. Therefore checking regularity conditions is a more important issue for statistics than for signal processing.

3.2. ℓ_q sensitivity for correlated designs

Due to the hardness of verification of regularity conditions, it is of paramount importance to provide sufficient conditions for ℓ_q sensitivity to hold for random matrices sampled from a high-dimensional correlated random vector. To this end, we first define a population version of ℓ_q sensitivity. Let $\underline{X}$ and $\underline{Z}$ be p and L-dimensional zero-mean random vectors and denote by $\Psi = \mathbb{E}\underline{Z}\underline{X}^T$ the $L \times p$ matrix of covariances with $\Psi_{ij} = \mathbb{E}(Z_i X_j)$.

Definition 3.2. *The $L \times p$ matrix of covariances Ψ satisfies the ℓ_q sensitivity property for $q \geq 1$ with parameters (s, α, γ) if*

$$\min_{v \in C(s,\alpha)} \frac{s^{1/q} |\Psi v|_\infty}{|v|_q} \geq \gamma.$$

Note that when $\underline{Z} = \underline{X}$, Ψ is the covariance matrix of $\underline{X}$. In particular, when $\Psi = I_p$, as in many designs of compressed sensing, it possesses the ℓ_q-sensitivity. See also Example 3.7 below and its proof.

Population ℓ_q sensitivity corresponds to the sample version with $n = \infty$. It is a necessary and natural condition to impose. Together with tail conditions it is sufficient to guarantee the regularity condition of random matrices sampled from such a population. This is indeed shown in the following theorem, in three different models: sub-gaussian vectors, bounded coordinates, and finite moments.

Theorem 3.3. *Let $\underline{X}$ and $\underline{Z}$ be zero-mean random vectors, such that the matrix of population covariances Ψ satisfies the ℓ_q sensitivity property, $q \geq 1$, with parameters (s, α, γ). Let $a > 0$ be fixed. Given n iid samples and any $\delta > 0$, the matrix $\hat{\Psi} = \frac{1}{n} Z^T X$ obeys ℓ_q sensitivity with parameters $(s, \alpha, \gamma - \delta)$, with high probability under each of the following settings:*

1. *If $\underline{X}$ and $\underline{Z}$ are sub-gaussian with fixed constants, then sample ℓ_q sensitivity holds with probability at least $1 - (2pL)^{-a}$, provided that the sample size is at least $n \geq cs^2 \log(2pL)$.*
2. *If the entries of the vectors are bounded by fixed constants, the property also holds with probability at least $1 - (2pL)^{-a}$, whenever $n \geq cs^2 \log(2pL)$.*
3. *If the entries have bounded moments: $\mathbb{E}|X_i|^{4r} < C_x < \infty$, $\mathbb{E}|Z_j|^{4r} < C_z < \infty$ for some positive integer r and all i, j, then ℓ_q sensitivity holds with probability at least $1 - 1/n^a$, assuming the sample size is at least $n^{1-a/r} \geq cs^2 (pL)^{1/r}$.*

The constant c does not depend on n, L, p and s, only on the other parameters of each case. It is given explicitly in the proofs in Section 5.3. The statements require $n \asymp s^2$ within a logarithmic order for the first two cases, and it would be interesting to know if the rate can be improved. Further, note that bounded random vectors are formally also sub-gaussian, but the sub-gaussian norm scales as $\sqrt{p}$. We get better results for bounded vectors if we treat them directly.

Related results have been obtained for the restricted isometry property (Rauhut, Schnass and Vandergheynst, 2008; Rudelson and Zhou, 2012) and restricted eigenvalue condition (Raskutti, Wainwright and Yu, 2010; Rudelson and Zhou, 2012). We investigate the ℓ_q sensitivity property since it's weaker and more general, also applicable to instrumental variables regression.

Theorem 3.3 can be extended to the case of the mixture distributions with different tail properties, i.e. (X, Z) is sampled from population $\mathbb{P}_1$ with probability p_1, and $\mathbb{P}_2$ with probability $1 - p_1$. We show this in the simplest case, a mixture of bounded and sub-gaussian random vectors. For $k = 1, 2$ let

$\Psi_k = \mathbb{E}_k Z X^T$ denote the matrix of covariances of X and Z under population $\mathbb{P}_k$

Theorem 3.4. *Suppose the distribution of random vectors X, Z is a mixture of a sub-gaussian distribution $\mathbb{P}_1$ and a coordinate-wise bounded distribution $\mathbb{P}_2$, with fixed mixture probability. Suppose further that either of the two matrices of covariances Ψ_1 or Ψ_2 obeys the ℓ_q sensitivity with lower bound γ and that $\|\Psi_1 - \Psi_2\|_{\max} \leq \delta/s$. Then for each $\nu > 0$, the matrix of sample covariances of n independent samples of (X, Z) obeys ℓ_q sensitivity with sparsity size s and lower bound $\gamma - (\delta + \nu)(1 + \alpha)$, with probability $1 - 4(2Lp)^{-\rho}$, if $n \geq cs^2 \log(2pL)$, for some constants ρ, c.*

Again, ρ and c are constants that do not depend on n, L, p, s. We prove Theorem 3.4 in Section 5.4. From the proof, one can see that the condition on Ψ-matrices can be relaxed to the ℓ_q sensitivity of the matrix $p_1\Psi_1 + (1 - p_1)\Psi_2$, where p_1 is the probability of getting the sample from $\mathbb{P}_1$.

In addition to the uncorrelated covariance matrices that satisfy the ℓ_q sensitivity (See Example 3.7 below and its proof), we introduce a more general class of covariance matrices that possess such a property.

Definition 3.5. *The $L \times p$ matrix Ψ is called* s-comprehensive *if for any subset $S \subset \{1, \ldots, p\}$ of size s, and for each pattern of signs $\varepsilon \in \{-1, 1\}^S$, there exists either a row w of Ψ such that $\operatorname{sgn}(w_i) = \varepsilon_i$ for $i \in S$, and $w_i = 0$ otherwise, or a row with $\operatorname{sgn}(w_i) = -\varepsilon_i$ for $i \in S$, and $w_i = 0$ otherwise.*

Note that when $L = p$, diagonal matrices are 1-comprehensive. However, when $L \neq p$, none of the other conditions are applicable. This illustrates that ℓ_q sensitivity is the most general property. By simple counting, $L \geq 2^{s-1}\binom{p}{s}$. We show that an s-comprehensive covariance matrix obeys the ℓ_1 sensitivity property.

Theorem 3.6. *Suppose the $L \times p$ matrix of covariances Ψ is s-comprehensive, and that all non-vanishing entries in Ψ have absolute value at least $c > 0$. Then Ψ obeys the ℓ_1 sensitivity property with parameters s, α and $\gamma = sc/(1 + \alpha)$.*

The proof of Theorem 3.6 is found in Section 5.5. The theorem shows that the larger the value s and hence the value L, the smaller c is required. It presents an interesting tradeoff between the number of instruments L and the strength of non-vanishing components of Ψ.

Finally, we give several examples to demonstrate that the ℓ_q sensitivity is indeed weaker than other regularity conditions. The technical proofs of the results in Examples 3.7 and 3.9 can be found in Section 5.6.

Example 3.7. *If Σ is a diagonal matrix with entries $d_1, d_2, \ldots, d_p$, then restricted isometry property holds if $1+\delta \geq d_i \geq 1-\delta$ for all i. Restricted eigenvalue only requires $d_i \geq \gamma$. The same condition is required for compatibility. This example shows why restricted isometry is the most stringent property. Further, ℓ_1 sensitivity holds even if a finite number of d_i go to zero at rate $1/s$ (shown in Section 5.6). In this latter case, all other regularity conditions fail. This example shows that l_q regularity is much weaker than other regularity conditions.*

The next examples further delineate between the various properties.

Example 3.8. *For the equal correlations model $\Sigma = (1-\rho)I_p + \rho ee^T$, restricted isometry requires $\rho < 1/(s-1)$. In contrast, restricted eigenvalue, compatibility, and ℓ_q sensitivity hold with lower bound $1-\rho$ (see van de Geer and Bühlmann (2009); Raskutti, Wainwright and Yu (2010)).*

Example 3.9. *If Σ has diagonal entries equal to 1, $\sigma_{12} = \sigma_{21} = \rho$, and all other entries are equal to zero, then compatibility and ℓ_1 sensitivity hold as long as $1-\rho \asymp 1/s$ (proven in Section 5.6). In such a case, however, the restricted eigenvalues are of order $1/s$. This is an example where compatibility and ℓ_1 sensitivity hold but the restricted eigenvalue condition fails.*

3.3. Operations preserving regularity

While it is difficult to check that a covariate matrix is regular, this property is preserved under natural operations that do not change the covariance structure by much. We show this for the ℓ_q sensitivity, in analogy to results on restricted isometry property (eg. Bandeira et al. (2012) and references therein).

We provide two theorems, both proven in Section 5.7. First we have a theorem about linear transformations of the data matrix that preserve regularity. Let X and Z be covariate matrices as in the rest of the paper.

Theorem 3.10. *1. Perform the orthogonal transformation M on each covariate: let $X' = MX$, $Z' = MZ$. Then (X', Z') obey the same ℓ_q sensitivity properties as (X, M).*

2. Let M be a cone-preserving *linear transformation $\mathbb{R}^L \to \mathbb{R}^L$, such that for all $v \in C(s, \alpha)$ we have $Mv \in C(s', \alpha')$ and let $X' = XM$. Suppose further that $|Mv|_q \geq c|v|_q$ for all v in $C(s, \alpha)$. If (X, Z) obeys the ℓ_q sensitivity property with parameters (s', α', γ), then (X', Z) has ℓ_q sensitivity with parameters $(s, \alpha, c\gamma)$.*

3. *Let M be a linear transformation $\mathbb{R}^p \to \mathbb{R}^p$ such that for all v, $|Mv|_\infty \geq c|v|_\infty$. If we transform $Z' = ZM$, and (X, Z) obeys the ℓ_q sensitivity property with lower bound γ, then (X, Z') obeys the same property with lower bound $c\gamma$.*

Our second result is about the additive operations on the covariate matrix that preserve regularity. We use the induced matrix norms $|M|_{a,b} = \sup_v |Mv|_b/|v|_a$. In particular, note that $|M|_{1,1}$ is the maximum ℓ_1 column sum and $|M|_{1,\infty}$ is the maximum absolute entry denoted by $\|M\|_{\max}$ elsewhere in the paper.

Theorem 3.11. *1. If Σ obeys ℓ_q sensitivity with lower bound γ, and $|\Delta|_{q,\infty} \leq \delta/s^{1/q}$, then $\Sigma + \Delta$ obeys ℓ_q sensitivity with lower bound $\gamma - \delta$.*

2. *Let (X_1, Z_1) and (X_2, Z_2) be such that (X_1, Z_1) has ℓ_q sensitivity with lower bound γ. Further suppose $|Z_1^T(X_2 - X_1)/n|_{q,\infty} \leq \varepsilon$, $|(Z_2 - Z_1)^T X_1/n|_{q,\infty} \leq \delta$ and $|(Z_2 - Z_1)^T(X_2 - X_1)/n|_{q,\infty} \leq c$. Then $[(X_1 + X_2)/2, (Z_1 + Z_2)/2]$ has ℓ_q sensitivity with lower bound $\gamma - s^{1/q}(2\varepsilon + 2\delta + c)/4$.*

4. Discussion

This paper presented an in-depth study of the matrix properties required for high-dimensional sparse estimation. We considered the restricted eigenvalue and compatibility properties, and the more general ℓ_q sensitivity condition, also applicable to instrumental variables regression. First we showed that they are unfortunately NP-hard to check. The results are important because in statistical applications the data is typically observational, and one cannot rely on the known regularity of iid random matrices.

For problems where a model of the covariance matrix is available, we have formulated high probability sufficient conditions for ℓ_q sensitivity. Finally we have established that several natural matrix operations preserve the ℓ_q sensitivity.

Our work raises further questions about the interplay of estimation and computation, specifically for sparse regression models. It would be interesting to study if there are statistically efficient estimators relying on computationally verifiable conditions. Specifically, can one devise an estimation method for sparse linear regression with mean squared error of minimax optimal order $s\log(p)/n$, relying on a condition that is also efficiently verifiable? The current theory falls short: incoherence requires $n \asymp s^2\log(p)$ samples to hold, and restricted eigenvalues are NP-hard to check. This is an

important research area, as illustrated by the recent work Chandrasekaran and Jordan (2013).

Finally, accurate and efficiently computable approximations to the values of regularity constants could provide efficient confidence intervals. Previous work on this problem has relied on convex relaxations, unfortunately leading to confidence intervals that are wider by a factor s than those theoretically possible (d'Aspremont, Bach and Ghaoui, 2008; Lee and Bresler, 2008; Gautier and Tsybakov, 2011). Some recent progress involves significance testing for adaptive linear models (Lockhart et al., 2013). Improvements in this direction would be of significant theoretical and practical value.

5. Proofs

5.1. Proof of Theorem 2.4

Proof. From a classical argument (e.g. Candès and Tao, 2007) it follows that

$$\max_{i=1,\dots,L} n^{-1}\langle z_i, \varepsilon\rangle \leq \sigma A\sqrt{\frac{2\log(L)}{n}}$$

with high probability, which is equivalent to $\beta \in \mathcal{I}$. From now on, assume that this event holds. Then, as $\hat{\beta}$ minimizes the ℓ_1 norm over $\mathcal{I}$, we have $|\hat{\beta}|_1 \leq |\beta|_1$ or $|\delta_{S^c}|_1 \leq |\delta_S|_1$ with $\delta = \hat{\beta} - \beta$. Hence δ is in the cone $C(s, 1)$. Further,

$$|\frac{1}{n}Z^T X\delta|_\infty \leq |\frac{1}{n}Z^T(Y - X\beta)|_\infty + |\frac{1}{n}Z^T(Y - X\hat{\beta})|_\infty \leq 2\sigma\lambda.$$

Therefore using the ℓ_q sensitivity we conclude

$$|\delta|_q \leq \frac{s^{1/q}}{\gamma}|\frac{1}{n}Z^T X\delta|_\infty \leq \frac{2\sigma\lambda s^{1/q}}{\gamma}.$$

This is the desired claim. □

5.2. Proof of Theorem 3.1

The spark of a matrix X, denoted $spark(X)$, is the smallest number of linearly dependent columns. The proof of our complexity result, Theorem 3.1, consists of a polynomial-time reduction from the NP-hard problem of computing the spark of a matrix (see Bandeira et al. (2013); Tillmann and Pfetsch (2012) and references therein).

Lemma 5.1. *Given an $n \times p$ matrix with integer entries X, and a sparsity size $0 < s < p$, it is* NP*-hard to decide if the spark of X is at most s.*

We also need the following technical lemma, which provides bounds on the singular values of matrices with bounded integer entries. For a matrix X, we denote by $\|X\|_2$ or $\|X\|$ its operator norm. Furthermore, we denote by X_S the submatrix of X obtained by taking the columns with indices in S.

Lemma 5.2. *Let X be an $n \times p$ matrix with integer entries. Let $M = \max_{i,j} |X_{ij}|$. Then,*

$$\|X\|_2 \leq 2^{\lceil \log_2(\sqrt{np}M) \rceil}.$$

Further, let $0 < s < n$. If $spark(X) > s$, then for any $S \subset \{1, \ldots, p\}$, $|S| = s$, we have:

$$\lambda_{\min}(X_S^T X_S) \geq 2^{-2n\lceil \log_2(nM) \rceil}.$$

Proof. The first claim follows from:

$$\|X\|_2 \leq \sqrt{np}\|X\|_{\max} \leq 2^{\lceil \log_2(\sqrt{np}M) \rceil}.$$

For the second claim, let X_S denote a submatrix of X with an arbitrary index set S of size s. Then $spark(X) > s$ implies that X_S is non-singular. Since the absolute values of the entries of X lie in $\{0, \ldots, M\}$, the entries of $X_S^T X_S$ are integers with absolute values between 0 and nM^2, namely $\|X_S^T X_S\|_{\max} \leq nM^2$. Moreover, since the non-negative and nonzero determinant of $X_S^T X_S$ is integer, it must be at least 1. Hence,

$$\begin{aligned} 1 \leq \prod_{i=1}^{s} \lambda_i(X_S^T X_S) &\leq \lambda_{\min}(X_S^T X_S)\lambda_{\max}(X_S^T X_S)^{s-1} \\ &\leq \lambda_{\min}(X_S^T X_S)(s\|X_S^T X_S\|_{\max})^{s-1}. \end{aligned}$$

Rearranging, we get

$$\lambda_{\min}(X_S^T X_S) \geq (snM^2)^{-s+1} \geq (nM)^{-2n} \geq 2^{-2n\lceil \log_2(nM) \rceil}.$$

In the middle inequality we have used $s \leq n$. This is the desired bound. □

For the proof we need the notion of *encoding length*, which is the size in bits of an object. Thus, an integer M has size $\lceil \log_2(M) \rceil$ bits. Hence the size of the matrix X is at least $np + \lceil \log_2(M) \rceil$: at least one bit for each entry, and $\lceil \log_2(M) \rceil$ bits to represent the largest entry. To ensure that the reduction is polynomial-time, we must make sure in particular that the size

in bits of the parameters involved is polynomial in the size of the input X. As in standard treatments of computational complexity, the numbers here are rational (Arora and Barak, 2009).

Proof of Theorem 3.1. It is enough to consider X with integer entries. For each property and given sparsity size s, we will exhibit parameters (α, γ) of a size in bits polynomial in that of the input X, such that:

1. $spark(X) \leq s \implies X$ does not obey the regularity property with parameters (α, γ),
2. $spark(X) > s \implies X$ obeys the regularity property with parameters (α, γ).

Hence, any polynomial-time algorithm for deciding if the regularity property holds for (X, s, α, γ), can, with just one call, in polynomial time decide if $spark(X) \leq s$. Here it is crucial that (α, γ) are polynomial in the size of X, so that the whole reduction is polynomial in X.

Since deciding $spark(X) \leq s$ is NP-hard by Theorem 5.1, this shows the desired NP-hardness of checking the conditions. For ℓ_q sensitivity, we in fact show that the subproblem where $Z = X$ is NP-hard, thus the full problem is also clearly NP-hard.

Now we provide the required parameters (α, γ) for each regularity condition. Similar ideas are used when comparing the conditions.

For the **restricted eigenvalue** condition, the first claim follows any $\gamma > 0$, and any $\alpha > 0$. To see this, if the spark of X at most s, there is a nonzero s-sparse vector v in the kernel of X, and $|Xv|_2 = 0 < \gamma |v_S|_2$, where S is any set containing the nonzero coordinates. This v is clearly also in the cone $C(s, \alpha)$, and so X does not obey RE with parameters (s, α, γ).

We now prove the second claim for the **restricted eigenvalue**. If $spark(X) > s$, then for each index set S of size s, the submatrix X_S is non-singular. We now show that this implies a non-vanishing lower bound on the RE constant of X. Indeed, consider a vector v in the cone $C(s, \alpha)$, and assume specifically that $\alpha |v_S|_1 \geq |v_{S^c}|_1$. Using the simple identity $Xv = X_S v_S + X_{S^c} v_{S^c}$, we have

$$\begin{aligned} |Xv|_2 &= |X_S v_S + X_{S^c} v_{S^c}|_2 \geq |X_S v_S|_2 - |X_{S^c} v_{S^c}|_2 \\ &\geq \sqrt{\lambda_{\min}(X_S^T X_S)} |v_S|_2 - \|X_{S^c}\|_2 |v_{S^c}|_2. \end{aligned}$$

Further, since v is in the cone, we have

$$|v_{S^c}|_2 \leq |v_{S^c}|_1 \leq \alpha |v_S|_1 \leq \alpha \sqrt{s} |v_S|_2. \tag{3}$$

Since X_S is non-degenerate and integer-valued, we can use the bounds from Lemma 5.2. Consequently, with $M = \|X\|_{\max}$, we obtain

$$\begin{aligned} |Xv|_2 &\geq |v_S|_2 \left(\sqrt{\lambda_{\min}(X_S^T X_S)} - \|X_{S^c}\| \alpha \sqrt{s} \right) \\ &\geq |v_S|_2 \left(2^{-n\lceil \log_2(nM) \rceil} - 2^{\lceil \log_2(\sqrt{np}M) \rceil} \alpha \sqrt{s} \right). \end{aligned}$$

By choosing, say, $\alpha = 2^{-2n\lceil \log_2(npM) \rceil}$, $\gamma = 2^{-2n\lceil \log_2(npM) \rceil}$, we easily conclude after some computations that $|Xv|_2 \geq \gamma |v_S|_2$.

Moreover, the size in bits, or encoding length, of the parameters is polynomially related to that of X. Indeed, the size in bits of both parameters is $2n\lceil \log_2(npM) \rceil$, and the size of X is at least $np + \lceil \log_2(M) \rceil$, as discussed before the proof. Note that $2n\lceil \log_2(npM) \rceil \leq (np + \lceil \log_2(M) \rceil)^2$. Hence we have showed both required conditions.

The argument for the **compatibility** conditions is nearly identical. Indeed, the first claim is satisfied for any $\gamma > 0$: for any nonzero s-sparse vector v in the kernel of X, with S containing the support of v, we have $\sqrt{s}|Xv| = 0 < \gamma |v_S|_1$.

For the second claim we argue as above, and obtain

$$\frac{\sqrt{s}|Xv|_2}{|v_S|_1} \geq \frac{|Xv|_2}{|v_S|_2} \geq 2^{-n\lceil \log_2(nM) \rceil} - 2^{\lceil \log_2(\sqrt{np}M) \rceil} \alpha \sqrt{s}.$$

Therefore the same choices of α, γ work in this case as well.

Finally, we deal with the ℓ_q **sensitivity** property. The first condition is again satisfied for all $\alpha > 0$ and $\gamma > 0$. Indeed, If the spark of X is at most s, there is a nonzero s-sparse vector v in its kernel, and thus $|X^T X v|_\infty = 0$.

For the second condition, we note that

$$|Xv|_2^2 = v^T X^T X v \leq |v|_1 |X^T X v|_\infty$$

For v in the cone, $\alpha |v_S|_1 \geq |v_{S^c}|_1$ and hence

$$|v|_2 \geq |v_S|_2 \geq \frac{1}{\sqrt{s}} |v_S|_1 \geq \frac{1}{\sqrt{s}(1+\alpha)} |v|_1.$$

Combination of the last two results gives

$$\frac{s|X^T X v|_\infty}{n|v|_1} \geq \frac{s|Xv|_2^2}{n|v|_1^2} \geq \frac{1}{n(1+\alpha)^2} \frac{|Xv|_2^2}{|v|_2^2}.$$

Finally, since $q \geq 1$, we have $|v|_1 \geq |v|_q$, and as v is in the cone, $|v|_2^2 = |v_S|_2^2 + |v_{S^c}|_2^2 \leq (1+\alpha^2 s)|v_S|_2^2$, by inequality (3). Therefore,

$$\frac{s^{1/q}|X^T X v|_\infty}{n|v|_q} \geq \frac{s^{1/q-1}}{n(1+\alpha)^2(1+\alpha^2 s)} \frac{|Xv|_2^2}{|v_S|_2^2}.$$

Hence we essentially reduced to restricted eigenvalues. From the proof of that case, the choice $\alpha = 2^{-2n\lceil \log_2(npM)\rceil}$ gives

$$\frac{|Xv|_2}{|v_S|_2} \geq 2^{-2n\lceil \log_2(npM)\rceil}.$$

Hence for this α we also have

$$\frac{s^{1/q}|X^T X v|_\infty}{n|v|_2} \geq 2^{-5(n+1)\lceil \log_2(npM)\rceil},$$

where we have applied a number of coarse bounds. Thus X obeys the ℓ_q sensitivity property with the parameters $\alpha = 2^{-2n\lceil \log_2(npM)\rceil}$ and $\gamma = 2^{-5n\lceil \log_2(npM)\rceil}$. As in the previous case, the size in bits of these parameters are polynomial in the size in bits of X. This shows that both conditions hold, and thus proves the correctness of the reduction for ℓ_q sensitivity. This completes the proof of Theorem 3.1.

The values of (α, γ) used in the proof are outside the range where the regularity properties lead to effective bounds for the estimation error. This choice is essential for the current proof, but it is a question of interest to extend it to the regime where α and γ are independent of n and p.

5.3. Proof of Theorem 3.3

The ℓ_q sensitivity property of random matrices relies on large deviation inequalities for random inner products. After establishing such inequalities, we finish the proofs quite directly, essentially by a union bound. We discuss the three probabilistic settings one by one, proving the required lemmas along the way. Thus, the proof of Theorem 3.3 is split into three parts: 5.3.1, 5.3.2, 5.3.3.

5.3.1. Sub-gaussian variables

A random variable X is sub-exponential if $\sup_{p\geq 1} p^{-1}\|X\|_p < \infty$, and the sub-exponential norm (or constant) is then $\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}\|X\|_p$. We use the following Bernstein-type inequality; see Corollary 5.17 in Vershynin (2010).

Lemma 5.3 (Bernstein for sub-exponential)**.** *If $X_1, \dots, X_N$ are independent centered sub-exponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$, then for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i\right| \geq t\right) \leq 2\exp\left(-cN\min\left(\frac{t^2}{K^2}, \frac{t}{K}\right)\right),$$

where $c \geq 1/(8e^2)$ is a constant independent of N.

Bernstein's lemma immediately implies a deviation inequality for inner products. We state it separately for clarity. It is also an extension of a lemma used in covariance matrix estimation (Bickel and Levina, 2008; Ravikumar, 2011).

Lemma 5.4 (Deviation of Inner Products for Sub-gaussians)**.** *Let X and Z be zero-mean sub-gaussian random variables, with sub-gaussian norms $\|X\|_{\psi_2}$, $\|Z\|_{\psi_2}$ respectively. Then, given n iid samples of X and Z, the sample covariance satisfies the tail bound:*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i Z_i - \mathbb{E}(XZ)\right| \geq t\right) \leq 2\exp(-cn\min(t/K, t^2/K^2)).$$

where $K := 4\|X\|_{\psi_2}\|Z\|_{\psi_2}$.

Proof. The proof consists of a direct application of Bernstein's inequality. We only need to bound the sub-exponential norms of $U_i = X_i Z_i - \mathbb{E}(X_i Z_i)$. In general if X, Z are sub-gaussian, then XZ is sub-exponential and moreover

$$\|XZ\|_{\psi_1} \leq 2\|X\|_{\psi_2}\|Z\|_{\psi_2}. \tag{4}$$

Indeed by the Cauchy-Schwartz inequality $\left(\mathbb{E}|XZ|^p\right)^2 \leq \mathbb{E}|X|^{2p}\mathbb{E}|Z|^{2p}$. Hence also

$$p^{-1}\left(\mathbb{E}|XZ|^p\right)^{1/p} \leq 2(2p)^{-1/2}\left(\mathbb{E}|X|^{2p}\right)^{1/2p}(2p)^{-1/2}\left(\mathbb{E}|Z|^{2p}\right)^{1/2p}.$$

Taking the supremum over $p \geq 1/2$ of both sides leads to the desired inequality (4).

The U_i are iid random variables, and their sub-exponential norm is by the triangle inequality, the norm inequality (4), and Cauchy-Schwartz, at most

$$\|U_i\|_{\psi_1} \leq \|X_i Z_i\|_{\psi_1} + |\mathbb{E}XZ| \leq 2\|X\|_{\psi_2}\|Z\|_{\psi_2} + \left(\mathbb{E}X^2\mathbb{E}Z^2\right)^{1/2}.$$

Further by definition $\left(\mathbb{E}X^2\right)^{1/2} \leq \sqrt{2}\|X\|_{\psi_2}$, hence the sub-exponential norm is at most

$$\|U_i\|_{\psi_1} \leq 4\|X\|_{\psi_2}\|Z\|_{\psi_2}.$$

Thus the result follows by a direct application of Bernstein's inequality. □

With these preparations, we now prove Theorem 3.3 for the sub-gaussian case. By a union bound over the Lp entries of the matrix $\Psi - \hat{\Psi}$

$$P(\|\Psi - \hat{\Psi}\|_{\max} \geq t) \leq \sum_{i,j} P(|\Psi_{i,j} - \hat{\Psi}_{i,j}| \geq t) \leq Lp \max_{i,j} P(|\Psi_{i,j} - \hat{\Psi}_{i,j}| \geq t).$$

By Lemma 5.4 each probability is upper bounded by a term of the form $2\exp(-cn\min(t/K, t^2/K^2))$, where K varies with i, j. The largest of these bounds corresponds to the largest of the K-s. Hence the K in the largest term is $4\max_{i,j}\|X_i\|_{\Psi_2}\|Z_j\|_{\Psi_2}$. By the definition of sub-gaussian norm, this is at most $4\|\underline{X}\|_{\Psi_2}\|\underline{Z}\|_{\Psi_2}$, where the $\underline{X}$ and $\underline{Z}$ are now p and L-dimensional vectors, respectively.

Therefore we have the uniform bound

$$P(\|\Psi - \hat{\Psi}\|_{\max} \geq t) \leq 2Lp\exp(-cn\min(t/K, t^2/K^2)) \tag{5}$$

with $K = 4\|\underline{X}\|_{\Psi_2}\|\underline{Z}\|_{\Psi_2}$.

We choose t such that $(a+1)\log(2Lp) = cnt^2/K^2$, that is $t = \sqrt{\frac{K^2(a+1)\log(2Lp)}{cn}}$. Since we can assume $(a+1)\log(2Lp) \leq cn$ by the scaling in the statement, the relevant term is the one quadratic in t: the total probability of error is $(2Lp)^{-a}$. From now on, we will work on the high-probability event that $\|\Psi - \hat{\Psi}\|_{\max} \leq t$.

For any vector v

$$|\Psi v|_\infty - \left|\hat{\Psi}v\right|_\infty \leq \left|(\Psi - \hat{\Psi})v\right|_\infty \leq \|\Psi - \hat{\Psi}\|_{\max}|v|_1 \leq t|v|_1.$$

That is, with high probability it holds uniformly for all v that:

$$\left|\hat{\Psi}v\right|_\infty \geq |\Psi v|_\infty - R\sqrt{\frac{\log(2pL)}{n}}|v|_1 \tag{6}$$

for the constant $R = \sqrt{\frac{K^2(a+1)}{c}}$.

For vectors v in $C(s, \alpha)$, we bound the ℓ_1 norm by the ℓ_q norm, $q \geq 1$, in the usual way, to get a term depending on s rather than on all p coordinates:

$$|v|_1 \leq (1+\alpha)|v_S|_1 \leq (1+\alpha)s^{1-1/q}|v_S|_q \leq (1+\alpha)s^{1-1/q}|v|_q. \tag{7}$$

Introducing this into (6) gives with high probability over all $v \in C(s, \alpha)$:

$$\frac{s^{1/q}\left|\hat{\Psi}v\right|_\infty}{|v|_q} \geq \frac{s^{1/q}\,|\Psi v|_\infty}{|v|_q} - R(1+\alpha)s\sqrt{\frac{\log(2pL)}{n}}.$$

If we choose n such that

$$n \geq \frac{K^2(1+a)(1+\alpha)^2}{c\delta^2}s^2\log(2pL),$$

then the second term will be at most δ. Further since Ψ obeys the ℓ_q sensitivity assumption, the first term will be at least γ. This shows that $\hat{\Psi}$ satisfies the the ℓ_q sensitivity assumption with constant $\gamma - \delta$ with high probability, and finishes the proof.

To summarize, it suffices if the sample size is at least

$$n \geq \frac{\log(2pL)(a+1)}{c}\max\left(1, \frac{K^2(1+\alpha)^2}{\delta^2}s^2\right). \tag{8}$$

The key to the proof, inequality (6), is similar in spirit to the one used in Raskutti, Wainwright and Yu (2010) to establish the Restricted Eigenvalue condition for correlated designs. However, our argument also easily allows a two-sided high-probability bound

$$\left||\Psi v|_\infty - \left|\hat{\Psi}v\right|_\infty\right| \leq R\sqrt{\frac{\log(pL)}{n}}|v|_1.$$

Hence, the population ℓ_q sensitivity property is both necessary and sufficient for the sample version. This is not necessarily clear from the proofs for the Restricted Eigenvalue condition (Raskutti, Wainwright and Yu, 2010; Rudelson and Zhou, 2012).

5.3.2. Bounded variables

If the components of the vectors X, Z are bounded, then essentially the same proof goes through. The sub-exponential norm of $X_iZ_j - \mathbb{E}(X_iZ_j)$ is bounded - by a different argument - because $|X_iZ_j - \mathbb{E}(X_iZ_j)| \leq 2C_xC_z$, hence $\|X_iZ_j - \mathbb{E}(X_iZ_j)\|_{\Psi_1} \leq 2C_xC_z$. Hence Lemma 5.4 holds with the same proof, where now the value of $K := 2C_xC_z$ is different. The rest of the proof only relies on Lemma 5.4, so it goes through unchanged. Therefore, with the same sample requirement (8), the matrix of sample covariances obeys the ℓ_q sensitivity with high probability.

5.3.3. Variables with bounded moments

For variates with bounded moments, we also need a large deviation inequality for inner products. We were unable to find a reference for this specific instance of a large deviation inequality, so we give a proof below. The general flow of the argument is classical, and relies on the Markov inequality and a moment-of-sum computation (e.g. Petrov (1995)). The closest result we are aware of is a lemma used in covariance matrix estimation (Ravikumar, 2011). Our result can be viewed as an extension of theirs, and the proof is shorter.

Lemma 5.5 (Deviation for Bounded Moments - Khintchine-Rosenthal)**.** *Let X and Z be zero-mean random variables, and r a positive integer, such that $\mathbb{E}X^{4r} = C_x < \infty$, $\mathbb{E}Z^{4r} = C_z < \infty$. Then, given n iid samples from X and Z, the sample covariance satisfies the tail bound:*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i Z_i - \mathbb{E}(XZ)\right| \geq t\right) \leq \frac{2^{2r} r^{2r} \sqrt{C_x C_z}}{t^{2r} n^r}.$$

Proof. Let $Y_i = X_i Z_i - \mathbb{E}XZ$, and $k = 2r$. By the Markov inequality, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i\right| \geq t\right) \leq \frac{\mathbb{E}\left|\sum_{i=1}^{n} Y_i\right|^k}{(tn)^k}.$$

We now bound the k-th moment of the sum $\sum_{i=1}^{n} Y_i$ using a type of classical argument, often referred to as Kintchine's or Rosenthal's inequality. We can write, recalling that $k = 2r$ is even,

$$\mathbb{E}\left|\sum_{i=1}^{n} Y_i\right|^k = \sum_{i_1, i_2, \ldots, i_k \in \{1,\ldots,n\}} \mathbb{E}(Y_{i_1} Y_{i_2} \ldots Y_{i_k}) \tag{9}$$

By the mutual independence of Y_i

$$\mathbb{E}(Y_1^{a_1} Y_2^{a_2} \ldots Y_n^{a_n}) = \mathbb{E}Y_1^{a_1} \mathbb{E}Y_2^{a_2} \ldots \mathbb{E}Y_n^{a_n}.$$

As $\mathbb{E}Y_i = 0$, the summands for which there is a Y_i singleton vanish. For the remaining terms, we bound by Jensen's inequality $(\mathbb{E}|Y|^{r_1})^{1/r_1} \leq (\mathbb{E}|Y|^{r_2})^{1/r_2}$ for $0 \leq r_1 \leq r_2$. So a generic term is at most

$$(\mathbb{E}|Y|^k)^{a_1/k} (\mathbb{E}|Y|^k)^{a_2/k} \ldots (\mathbb{E}|Y|^k)^{a_n/k} = \mathbb{E}|Y|^k.$$

Above we have used that $a_1 + \ldots + a_n = k$. Hence, each non-vanishing term in the summation (9) was upper bounded by the same constant. To estimate

the sum, we are left with the combinatorial problem of counting the number of sequences of non-negative integers $(a_1, \ldots, a_n)$ that sum to k, and such that no term is 1. Here, if some $a_i > 0$, then $a_i \geq 2$. Thus, there are at most $k/2 = r$ nonzero elements. Therefore, the number of such sequences is not more than the number of ways to choose r places out of n, multiplied by the number of ways to distribute $2r$ elements among those places:

$$\binom{n}{r} r^{2r} \leq n^r r^{2r}.$$

Thus, we have proved that

$$\mathbb{E}\left|\sum_{i=1}^n Y_i\right|^{2r} \leq n^r r^{2r} \mathbb{E}|Y|^{2r}.$$

Further, we make an explicit bound in terms of the moments of X, Z. By the Minkowski and Jensen inequalities

$$\mathbb{E}|Y|^k = \mathbb{E}|X_i Z_i - \mathbb{E}X_i Z_i|^k \leq \left((\mathbb{E}|X_i Z_i|^k)^{1/k} + \mathbb{E}|X_i Z_i|\right)^k \leq 2^k \mathbb{E}|X_i Z_i|^k.$$

Further by Cauchy-Schwartz $\mathbb{E}|X_i Z_i|^k \leq \sqrt{\mathbb{E}|X_i|^{2k}\mathbb{E}|Z_i|^{2k}} = \sqrt{C_x C_z}$. Introducing this bound for the moment of sum in the Markov inequality leads to the desired bound

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Y_i\right| \geq t\right) \leq \frac{2^{2r} r^{2r} \sqrt{C_x C_z}}{t^{2r} n^r}.$$

□

We are ready to prove Thm 3.3. By a union bound, the probability that $\|\Psi - \hat{\Psi}\|_{\max} \geq t$ is at most

$$Lp \frac{2^{2r} r^{2r} \sqrt{C_x C_z}}{t^{2r} n^r}.$$

Since r is fixed, for simplicity of notation, we can denote $C_0^{2r} = 2^{2r} r^{2r} \sqrt{C_x C_z}$. Choosing $t = C_0 (Lp)^{1/2r} n^{-1/2 + a/(2r)}$, the above probability is at most $1/n^a$.

The bound

$$|\Psi v|_\infty - \left|\hat{\Psi} v\right|_\infty \leq \left|(\Psi - \hat{\Psi})v\right|_\infty \leq \|\Psi - \hat{\Psi}\|_{\max} |v|_1.$$

holds as before, so we conclude that with probability $1 - 1/n^a$, for all $v \in C(s, \alpha)$:

$$\frac{s^{1/q} \left|\hat{\Psi} v\right|_\infty}{|v|_q} \geq \frac{s^{1/q} |\Psi v|_\infty}{|v|_q} - (1+\alpha) st.$$

From the choice of t, for sample size at least

$$n^{1-a/r} \geq \frac{C_0^2 (1+\alpha)^2}{\delta^2} (Lp)^{1/r} s^2$$

the error term on the left hand side is at most δ. In this case $\hat{\Psi}$ satisfies the the ℓ_q sensitivity assumption with constant $\gamma - \delta$ with high probability.

5.4. Proof of Theorem 3.4

This result and Theorem 3.3 have closely related proofs, relying in essence on the same large deviation inequalities. Let p_1 and $1 - p_1$ denote the mixture probabilities. Then, the outcome of a sample from $\mathbb{P}_1$ corresponds to a Bernoulli trial with success probability p_1. The number of samples n_1 from $\mathbb{P}_1$ is a realization from a Binomial(n, p_1) random variable. The first part of the analysis is conditional on $N_1 = n_1$.

Let X and Z be the two matrices of observations, and let (X_i, Z_i) denote the matrices of the samples from distribution $\mathbb{P}_i$. Without loss of generality, assume that Ψ_1 satisfies the ℓ_q sensitivity. Then we can write the matrix of sample covariances as

$$\hat{\Psi} = \frac{1}{n} Z^T X = \frac{1}{n} \left(Z_1^T X_1 + Z_2^T X_2 \right),$$

which can be further decomposed as

$$\frac{n_1}{n} \left(\frac{1}{n_1} Z_1^T X_1 - \Psi_1 \right) + \frac{n_2}{n} \left(\frac{1}{n_2} Z_2^T X_2 - \Psi_2 \right) + \frac{n_2}{n} (\Psi_2 - \Psi_1) + \Psi_1.$$

The main term is Ψ_1, and the first three terms are error terms. The first two are stochastic (call them M_1, M_2), and are bounded as in Theorem 3.3, while the third term (call it N) is small because the Ψ_i are close to one another.

In more detail, note that

$$|\hat{\Psi} v|_\infty \geq |\Psi v|_\infty - (\|M_1\|_{\max} + \|M_2\|_{\max} + \|N\|_{\max}) |v|_1$$

We bound the ratios $n_1/n \leq 1$, $n_2/n \leq 1$ by the constant 1. The uniform large deviation inequality (5) from the same theorem can be applied to both samples, yielding bounds for $\|M_i\|_\infty$:

$$P(\|\Psi_i - \hat{\Psi}_i\|_{\max} \geq t) \leq 2Lp \exp(-cn_i \min(t/K, t^2/K^2))$$

with $K = \max(4\|X_1\|_{\Psi_2}\|Z_1\|_{\Psi_2}, 2C(X_2)C(Z_2))$. Here we have assumed without loss of generality that the first sample is sub-gaussian, and the second one is bounded. K is the maximum of two expressions depending on these norms, the same expressions as in Theorem 3.3.

We choose t_i as in the proof of Theorem 3.3: $t_i = \sqrt{\frac{K^2(a+1)\log(2Lp)}{cn_i}}$. As long as $(a+1)\log(2Lp) \leq cn_i$, the total probability of error is $2(2Lp)^{-a}$.

Now, for the first time in this proof, we consider the number of samples n_1 from the first distribution as random. Since it's a Binomial(n, p_1) random variable, Hoeffding's inequality holds:

$$\mathbb{P}_{p_1,p_2}\left\{\left|\frac{n_1}{n} - p_1\right| \geq \varepsilon\right\} \leq 2\exp(-2\varepsilon^2 n).$$

So with probability at least $1 - 2\exp(-2\varepsilon^2 n)$, the deviation $|n_1/n - p_1| \leq \varepsilon$, and thus also $|n_2/n - p_2| \leq \varepsilon$. We work on this event in what follows.

Let $q = \min(p_1, p_2)$, and choose $\varepsilon = q/2$. Then $n_i \geq n(p_i - \varepsilon) \geq n(q - \varepsilon) = nq/2$. Hence the bounds simplify if we substitute for n_i, and become

$$\|M_1\|_{\max} + \|M_2\|_{\max} \leq 2\sqrt{\frac{2K^2(a+1)\log(2Lp)}{qcn}}$$

with probability at least $1 - 2(2Lp)^{-a}$, as long as

$$2(a+1)\log(2Lp) \leq qcn. \tag{10}$$

Now we combine these in the main bound. By assumption $\|(\Psi_2 - \Psi_1)\|_{\max} \leq \delta/s$. Let t denote the bound obtained for $\|M_1\|_{\max} + \|M_2\|_{\max}$. We thus have

$$|\hat{\Psi}v|_\infty \geq |\Psi v|_\infty - (t + \delta/s)|v|_1$$

which together with (7) leads to

$$\frac{s^{1/q}|\hat{\Psi}v|_\infty}{|v|_q} \geq \frac{s^{1/q}|\Psi v|_\infty}{|v|_q} - (ts + \delta)(1 + \alpha).$$

In order for the term ts to be at most ν, we need

$$n \geq \frac{8K^2(1+a)}{qc\nu^2} s^2 \log(2pL).$$

If this and (10) happens, then the previous display implies the statement of the theorem, by the ℓ_q sensitivity of Ψ_1. The probability of error is at most $2(2Lp)^{-a} + 2\exp(-2q^2 n)$. Because of (10), the second probability is also of the form $2(2Lp)^{-\theta}$, where now $\theta = 4q(a+1)/c$. So we can choose $\rho = \min(a, \theta)$ to get the simpler form of the probability bound claimed in the statement. This proves the theorem.

5.5. Proof of Theorem 3.6

To bound the term $|\Psi v|_\infty$ in the ℓ_1 sensitivity, we use the s-comprehensive property. Indeed, let $v \in C(s, \alpha)$. By the symmetry of the s-comprehensive property, we can assume without loss of generality that $|v_1| \geq |v_2| \geq \ldots \geq |v_p|$. Then if S denotes the first s components, $\alpha |v_S|_1 \geq |v_{S^c}|_1$.

Consider the sign pattern of the top s components of v: $\varepsilon = (\mathrm{sgn}(v_1), \ldots, \mathrm{sgn}(v_s))$. Since Ψ is s-comprehensive, it has a row w with matching sign pattern. Then we can compute

$$\langle w, v \rangle = \sum_{i \in S} |w_i| \mathrm{sgn}(w_i) v_i = \sum_{i \in S} |w_i| \mathrm{sgn}(v_i) v_i = \sum_{i \in S} |w_i||v_i|.$$

Hence the inner product is lower bounded by

$$\min_{i \in S} |w_i| \sum_{i \in S} |v_i| \geq c \sum_{i \in S} |v_i|.$$

Combining the above, the ℓ_1 sensitivity is at least:

$$\frac{s|\langle w, v \rangle|}{|v|_1} \geq \frac{sc|v_S|_1}{(1+\alpha)|v_S|_1} = \frac{cs}{(1+\alpha)}.$$

This proves the stated thesis.

5.6. Proof of claims in Examples 3.7, 3.9

We must bound the ℓ_1 sensitivity for the two specific covariance matrices Σ. We first verify the claim in Example 3.7. For the diagonal matrix with entries $d_1, \ldots, d_p > 0$, we have

$$m = |\Sigma v|_\infty = \max(|d_1 v_1|, \ldots, |d_p v_p|).$$

Then summing $|v_i| \leq m/d_i$ for i in any set S with size s:

$$|v_S|_1 \leq m \sum_{i \in S} 1/d_i.$$

We want to bound this for $v \in C(s,\alpha)$, so let S be the subset of dominating coordinates for which $|v_{S^c}|_1 \le \alpha |v_S|_1$. It follows that

$$|v|_1 \le (1+\alpha)|v_S|_1 \le (1+\alpha)m \sum_{i\in S} 1/d_i.$$

Therefore

$$\frac{s|\Sigma v|_\infty}{|v|_1} \ge \frac{s}{(1+\alpha)\sum_{i\in S} 1/d_i} \ge \frac{1}{(1+\alpha)s^{-1}\sum_{i=1}^s 1/d_{(i)}},$$

where $\{d_{(i)}\}_{i=1}^p$ is the order of $\{d_i\}_{i=1}^p$, arranged from the smallest to the largest. The harmonic average in the lower bound can be bounded away from zero even several d_i-s are of order $O(1/s)$. For instance if $d_{(1)} = \cdots = d_{(k)} = 1/s$ and $d_{(k+1)} > 1/c$ for some constant c and integer $k < s$, then the ℓ_1 sensitivity is at least

$$\frac{s|\Sigma v|_\infty}{|v|_1} \ge \frac{1}{(1+\alpha)(k+(1-k/s)c)},$$

which is bounded away from zero whenever k is bounded. In this setting the smallest eigenvalue of Σ is $1/s$, so only the ℓ_1 sensitivity holds out of all regularity properties.

We now consider Example 3.9. For this specific covariance matrix,

$$m = |\Sigma v|_\infty = \max(|v_1 + \rho v_2|, |v_2 + \rho v_1|, |v_3|, \ldots, |v_p|).$$

The coordinate v_1 can be bounded as follows:

$$|v_1| = \left|\frac{1}{1-\rho^2}(v_1+\rho v_2) - \frac{\rho}{1-\rho^2}(\rho v_1 + v_2)\right| \le m\left(\frac{1}{1-\rho^2} + \frac{\rho}{1-\rho^2}\right)$$

leading to $|v_1| \le m/(1-\rho)$. Similarly $|v_2| \le m/(1-\rho)$. Furthermore, For each $i \notin \{1,2\}$, we have $|v_i| \le m$. Thus, for any set S,

$$|v_S|_1 \le m\left(\frac{2}{1-\rho} + s - 2\right).$$

For any $v \in C(s,\alpha)$,

$$|v|_1 \le (1+\alpha)|v_S|_1 \le (1+\alpha)m\left(\frac{2}{1-\rho} + s - 2\right)$$

Hence we obtain the lower bound on the ℓ_1 sensitivity

$$\frac{s|\Sigma v|_\infty}{|v|_1} \ge \frac{s}{(1+\alpha)(2/(1-\rho)+s-2)}.$$

If $1-\rho = 1/s$, this bound is at least $1/3(1+\alpha)$, showing that ℓ_1 sensitivity holds. However, the smallest eigenvalue is also $1-\rho = 1/s$, so the other regularity properties (restricted eigenvalue, compatibility), fail to hold as $s \to \infty$.

5.7. Proofs from Section 3.3

For the first claim of Theorem 3.10, note that $(Z')^T X' = (MZ)^T MX = Z^T X$ since M is orthonormal. ℓ_q sensitivity of the pair of matrices (X, Z) only depends on the matrix $Z^T X$, which is preserved under the orthonormal transformation. Hence, the transformed matrices inherit the regularity property.

For the second claim, note $(Z')^T X' v = Z^T X(Mv)$. If v is any vector in the cone $C(s, \alpha)$, we have $Mv \in C(s', \alpha')$ by the cone-preserving property. Hence by the ℓ_q sensitivity of X, Z

$$\frac{s^{1/q}|1/nZ^T X(Mv)|_\infty}{|Mv|_q} \geq \gamma.$$

Further by the condition on M: $|Mv|_q \geq c|v|_q$. Multiplying these two inequalities yields the ℓ_q sensitivity for X', Z.

For the last claim, we write $(Z')^T X' v = MZ^T Xv$. By the ℓ_q sensitivity of X, Z, for all $v \in C(s, \alpha)$,

$$\frac{s^{1/q}|1/nZ^T Xv|_\infty}{|v|_q} \geq \gamma.$$

However, $|M(1/nZ^T Xv)|_\infty \geq c|1/nZ^T Xv|_\infty$ by the assumption on M. Multiplying these inequalities gives the desired ℓ_q sensitivity of X, Z', completing the proof of Theorem 3.10.

Finally, for the proof of Theorem 3.11, note the following inequality, which has already been used in the paper:

$$|(\Sigma+\Delta)v|_\infty \geq |\Sigma v|_\infty - |\Delta v|_\infty \geq |\Sigma v|_\infty - |\Delta|_{q,\infty}|v|_q.$$

Since $|\Delta|_{q,\infty} \leq \delta/s^{1/q}$, we have, using the assumed ℓ_q sensitivity of Σ, that $s|(\Sigma+\Delta)v|_\infty/|v|_q \geq \gamma - \delta$, as required.

For the second part, we start by letting $U = X_2 - X_1$, $V = Z_2 - Z_1$, so that $(X_1 + X_2, Z_1 + Z_2) = (2X_1 + U, 2Z_1 + V)$, and

$$(Z_1+Z_2)^T(X_1+X_2) = 4Z_1^T X_1 + 2V^T X_1 + 2Z_1^T U + V^T U.$$

Using the triangular inequality $|(Z_1+Z_2)^T(X_1+X_2)v|_\infty \geq |4Z_1^TX_1|_\infty - (|2V^TX_1v|_\infty + |2Z_1^TUv|_\infty + |V^TUv|_\infty)$. We can bound the three error terms as follows:

$$|V^TX_1v|_\infty \leq |V^TX_1|_{q,\infty}|v|_q,$$

and similarly for the other two terms. Combining them yields

$$\frac{s^{1/q}|1/n(Z_1+Z_2)^T(X_1+X_2)v|_\infty}{|v|_q} \geq 4\frac{s^{1/q}|1/nZ_1^TX_1v|_\infty}{|v|_q} - c,$$

where $c = s^{1/q}/n(|V^TX_1|_{q,\infty} + 2|Z_1^TU|_{q,\infty} + |V^TU|_{q,\infty})$. Dividing by 4 gives the desired ℓ_q sensitivity for $(Z_1+Z_2)^T(X_1+X_2)/4$.

References

Arora, S. and Barak, B. (2009). *Computational Complexity: A Modern Approach*. Cambridge University Press.

Bandeira, A. S., Fickus, M., Mixon, D. G. and Wong, P. (2012). The road to deterministic matrices with the restricted isometry property. *arXiv:1202.1234*.

Bandeira, A., Dobriban, E., Mixon, D. and Sawin, W. (2013). Certifying the Restricted Isometry Property is Hard. *Information Theory, IEEE Transactions on, to appear*.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics* **36** 2577–2604.

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37** 1705–1732.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*, 1st ed. *Springer Series in Statistics*. Springer.

Bunea, F. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1** 169–194. Mathematical Reviews number (MathSciNet): MR2312149; Zentralblatt MATH identifier: 1146.62028.

Candès, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory* **51** 4203– 4215.

Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35** 2313–2351.

Chandrasekaran, V. and Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences* **110** E1181–E1190.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (2001). Atomic Decomposition by Basis Pursuit. *SIAM Review* **43** 129–159.

Cook, S. (2000). The P versus NP Problem. http://www.claymath.org/millennium/P_vs_NP/Official_Problem_Description.pdf.

d'Aspremont, A., Bach, F. and Ghaoui, L. E. (2008). Optimal Solutions for Sparse Principal Component Analysis. *J. Mach. Learn. Res.* **9** 1269–1294.

d'Aspremont, A. and El Ghaoui, L. (2011). Testing the nullspace property using semidefinite programming. *Math. Program.* **127** 123–144.

Donoho, D. L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on* **47** 2845–2862.

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

Gautier, E. and Tsybakov, A. B. (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv:1105.2454*.

Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828. Mathematical Reviews number (MathSciNet): MR2555200; Zentralblatt MATH identifier: 05815956.

Lee, K. and Bresler, Y. (2008). Computing performance guarantees for compressed sensing. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* 5129 –5132.

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2013). A significance test for the lasso. *ArXiv e-prints*.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Clarendon PresS.

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted Eigenvalue Properties for Correlated Gaussian Designs. *Journal of Machine Learning Research* **11** 2241–2259.

Rauhut, H., Schnass, K. and Vandergheynst, P. (2008). Compressed Sensing and Redundant Dictionaries. *IEEE Transactions on Information Theory* **54** 2210–2219.

Ravikumar, P. (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980.

Rudelson, M. and Zhou, S. (2012). Reconstruction from Anisotropic Random Measurements. In *Proceedings of the 25th Annual Conference on Learning Theory*.

Tao, T. (2007). Open question: deterministic UUP matrices. http://terrytao.wordpress.com/2007/07/02/open-question-deterministic-

uup-matrices/.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 267–288.

Tillmann, A. M. and Pfetsch, M. E. (2012). The Computational Complexity of RIP, NSP, and Related Concepts in Compressed Sensing. *arXiv:1205.2081*.

van de Geer, S. (2007). The deterministic Lasso. In *JSM Proceedings*. Americal Statistical Association.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3** 1360–1392.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*. Chapter 5 of: Compressed Sensing, Theory and Applications. Cambridge, 2012.

Ye, F. and Zhang, C. (2010). Rate Minimaxity of the Lasso and Dantzig Selector for the lq Loss in lr Balls. *Journal of Machine Learning Research* **11** 3519–3540.