Geometric Inference for General High-Dimensional Linear Inverse Problems

Jiapeng Liu

Department of Statistics
Purdue University

June 23, 2016

(Work by Tony Cai, Tengyuan Liang and Alexander Rakhlin)

Introduction

Model Setting

High dimensional linear inverse model:

$$Y = \mathcal{X}(M) + Z \tag{1}$$

- $M \in \mathbb{R}^p$ is the vectorized version of the parameter of interest.
- $\mathcal{X} : \mathbb{R}^p \to \mathbb{R}^n$ is a linear operator, and $Z \in \mathbb{R}^n$ is a noise vector.
- We observe $(\mathcal{X}, Y)$ and wish to recover M.
- In the high-dimensional setting, p is much larger than n.

Introduction

Examples of linear inverse problems.

High Dimension Linear Regression.

$$Y = XM + Z$$

- $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ with $p \gg n$, $M \in \mathbb{R}^p$ is a sparse signal, and $Z \in \mathbb{R}^n$ is a noise vector. The goal is to recover the unknown sparse signal $M \in \mathbb{R}^p$ through an efficient algorithm.
- Many estimation methods including l_1-regularized procedures such as the Lasso and Dantzig Selector have been developed and analyzed.
- A common approach for inference is to first construct a de-biased Lasso or de-biased scaled-Lasso estimator and then make inference based on the asymptotic normality of low-dimensional functionals of the de-biased estimator.

Introduction

Two examples of linear inverse problems.

Trace Regression.

One observe $(X_i, Y_i), i = 1, ..., n$ with

$$Y_i = \mathrm{Tr}(X_i^T M) + Z_i,$$

- $Y \in \mathbb{R}$, $X \in \mathbb{R}^{p_1 \times p_2}$ are measurement matrices, and Z_i are noise. Here the dimension of the parameter M is $p = p_1 p_2 \gg n$. The goal is to recover the unknown low rank matrix $M \in \mathbb{R}^{p_1 \times p_2}$.
- A number of constrained and penalized nuclear minimization methods have been introduced and studied in both the noiseless and noisy settings.

Introduction

Contributions of this paper.

Unified convex algorithms.

- A general computationally feasible convex program that provides near optimal rate of convergence simultaneously for a collection of high-dimensional linear inverse problems.
- A general efficient convex program that leads to statistical inference for linear contrasts of M.

Local geometric theory.

- A unified theoretical framework is provided for analyzing high-dimensional linear inverse problems based on the local conic geometry and duality.

Basic Convex Geometry

- M is commonly assumed to have a low complexity structure, with respect to a given atom set $\mathcal{A}$.
- Suppose $\mathcal{A}$ is a collection of some basic atoms. A parameter M is of **complexity** k in terms of the atoms in $\mathcal{A}$ if M can be expressed as a linear combination of at most k atoms in $\mathcal{A}$, i.e., there exists a decomposition

$$M = \Sigma_{a \in \mathcal{A}} c_a(M) \cdot a, \text{ where } \Sigma_{a \in \mathcal{A}} \mathbb{I}_{\{c_a(M) \neq 0\}} \leq k$$

Basic Convex Geometry

- The **atomic norm** $\|x\|_{\mathcal{A}}$ for any $x \in \mathbb{R}^p$ is defined as the gauge (Minkowski functional) of conv($\mathcal{A}$):

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 : x \in t \ \mathrm{conv}(\mathcal{A})\}$$

  It can also be written as

$$\|x\|_{\mathcal{A}} = \inf\Big\{\Sigma_{a \in \mathcal{A}} c_a : x = \Sigma_{a \in \mathcal{A}} c_a(M) \cdot a, c_a \geq 0\Big\}$$

- The **dual norm** of this atomic norm is defined as:

$$\|x\|_{\mathcal{A}}^* = \sup\{\langle x, a\rangle : a \in \mathcal{A}\} = \sup\{\langle x, a\rangle : \|a\|_{\mathcal{A}} \leq 1\}$$

Basic Convex Geometry

- The unit ball with respect to the atomic norm $\|\cdot\|_{\mathcal{A}}$ is the convex hull of the set of atoms $\mathcal{A}$.
- The **tangent cone** (recession cone) at x with respect to the scaled unit ball $\|x\|_{\mathcal{A}}\text{conv}(\mathcal{A})$ is defined to be

$$T_{\mathcal{A}}(x) = \text{cone}\{h : \|x + h\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}}\}.$$

- $T_{\mathcal{A}}(x)$ determines the geometric property of the neighborhood around the true parameter M.

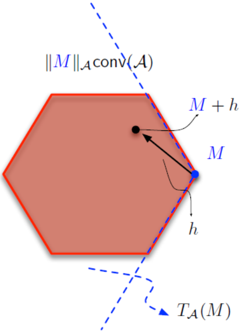

Figure 1: Tangent cone general illustration 2D. The red shaped area is the scaled convex hull of atom set. The blue dashed line forms the tangent cone at M. Black arrow denotes the possible directions inside the cone.

High Dimension Linear Regression.

- The atom set $\mathcal{A}$ consists of the unit basis vectors $\{\pm e_i\}$,
- The atomic norm $\|\cdot\|_{\mathcal{A}}$ is the vector l_1 norm,
- The dual norm $\|\cdot\|_{\mathcal{A}}^*$ is the vector l_∞ norm.

Trace Regression

- The atom set $\mathcal{A}$ consists of the rank one matrices (matrix manifold) $\mathcal{A} = \{uv^* : \|u\|_{l_2} = 1, \|v\|_{l_2} = 1\}$
- The atomic norm is the nuclear norm (sum of *singular values*).
- The dual norm is the spectral norm (the largest *singular values*).

Basic Convex Geometry

Definition (Gaussian Width)

For a compact set $K \in \mathbb{R}^p$, the **Gaussian width** is defined as

$$w(K) := \mathbb{E}_g \left[\sup_{v \in K} \langle g, v \rangle \right].$$

where $g \sim N(0, I_p)$ is the standard multivariate Gaussian vector.

- When $M \in \mathbb{R}^p$ is a s-sparse vector,

$$w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{s \ \mathrm{log} p/s}.$$

- When $M \in \mathbb{R}^{p \times q}$ is a rank-r matrix,

$$w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{r(p+q-r)}.$$

Basic Convex Geometry

Definition (Sudakov Minoration Estimate)

The **Sudakov estimate** of a compact set $K \in \mathbb{R}^p$ is defined as

$$e(K) := \sup_{\epsilon} \epsilon \sqrt{\log \mathcal{N}(K, \epsilon)}.$$

where $\mathcal{N}(K, \epsilon)$ denotes the ϵ-covering number of set K with respect to the Euclidean norm.

- Sudakov estimate determines the complexity of the cone $T_{\mathcal{A}}(M)$. It is useful for the minimax lower bound analysis.

Basic Convex Geometry

Definition (Volume Ratio)

$$v(K) := \sqrt{p}\left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(B_2^p)}\right)^{\frac{1}{p}}.$$

Definition (local asphericity ratio)

$$\gamma_{\mathcal{A}}(M) := \sup\left\{\frac{\|h\|_{\mathcal{A}}}{\|h\|_{l_2}} : h \in T_{\mathcal{A}}(M), h \neq 0\right\}.$$

- This measures how extreme the atomic norm is relative to the l_2 norm within the local tangent cone.

Point Estimation via Convex Relaxation

- Suppose we observe $(\mathcal{X}, Y)$ in (1) where M is assumed to have low complexity w.r.t. a given atom set $\mathcal{A}$. We use a generic convex constrained minimization procedure to estimate M:

$$\hat{M} = \arg\min_{M} \{\|M\|_{\mathcal{A}} : \|\mathcal{X}^*(Y - \mathcal{X}(M))\|_{\mathcal{A}}^* \leq \lambda\} \qquad (2)$$

  where λ is a tuning parameter (localization radius) that depends on the sample size, noise level, and geometry of the atom set $\mathcal{A}$.
- $\mathcal{X}^* : \mathbb{R}^n \to \mathbb{R}^p$ is the adjoint operator of $\mathcal{X}$ such that

$$\langle \mathcal{X}(V_1), V_2 \rangle = \langle V_1, \mathcal{X}^*(V_2) \rangle$$

Point Estimation via Convex Relaxation

- $$\begin{aligned}\|\mathcal{X}^*(Y-\mathcal{X}(M))\|_{\mathcal{A}}^* &= \sup\{\langle \mathcal{X}^*(Y-\mathcal{X}(M)), a\rangle : a\in\mathcal{A}\}\\ &= \sup\{\langle Y-\mathcal{X}(M), \mathcal{X}(a)\rangle : a\in\mathcal{A}\}\end{aligned}$$
- The Dantzig selector for high-dimensional sparse regression (Candes et al. (2007)) and the constrained nuclear norm minimization (Candes et al. (2011)) for trace regression are particular examples of (2).

Statistical Inference via Convex Feasibility Program.

- In model (1), let $M \in \mathbb{R}^p$ be the vectorized parameter of interest, and $\{e_i, 1 \le i \le p\}$ are the corresponding basis vectors. Consider the following convex feasibility problem for matrix $\Omega \in \mathbb{R}^{p \times p}$, where each row $\Omega_{i\cdot}$ satisfies

$$\|\mathcal{X}^* \mathcal{X} \Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \le \eta, \quad \forall\ 1 \le i \le p \tag{3}$$

  where η is some tuning parameter that depends on the sample size and geometry of the atom set $\mathcal{A}$.

- One can also solve a stronger version of the above convex program for $\eta \in \mathbb{R}, \Omega \in \mathbb{R}^{p \times p}$ simultaneously

$$(\Omega, \eta_n) = \arg\min_{\Omega, \eta} \{\eta : \|\mathcal{X}^* \mathcal{X} \Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \le \eta, \quad \forall\ 1 \le i \le p\}.$$

Statistical Inference via Convex Feasibility Program.

- This convex feasibility program can be viewed as a unified treatment for general linear inverse models.
- Built upon the constrained minimization estimator $\hat{M}$ in (2) and feasible matrix in (3), the de-biased estimator for inference on parameter M is defined as

$$\tilde{M} := \hat{M} + \Omega \mathcal{X}^*(Y - \mathcal{X}(\hat{M})). \tag{4}$$

Local Geometric Theory: Gaussian Setting.

Local Geometric Upper Bound.

Definition (Gaussian Ensemble Design)

Let $\mathcal{X} \in \mathbb{R}^{n \times p}$ be the matrix form of the linear operator $\mathcal{X} : \mathbb{R}^p \to \mathbb{R}^n$. $\mathcal{X}$ is $\mathbf{Gaussian\ ensemble}$ if each element is $i.i.d$ Gaussian random variable with mean 0 and variance $\frac{1}{n}$.

- In this part, we assume the noise vector Z is Gaussian and the linear operator $\mathcal{X}$ is the Gaussian ensemble design.

Local Geometric Theory: Gaussian Setting.

Local Geometric Upper Bound.

- We need to choose a suitable λ (in the convex program (2)) to guarantee that the true parameter M is in the feasible set with high probability.
- **The choice of λ.** In the case of Gaussian noise the tuning parameter is chosen as

$$\lambda_{\mathcal{A}}(\mathcal{X}, \sigma, n) = \frac{\sigma}{\sqrt{n}} \left\{ \omega(\mathcal{X}\mathcal{A}) + \delta \cdot \sup_{v \in \mathcal{A}} \|\mathcal{X}v\|_{l_2} \right\} \asymp \frac{\sigma}{\sqrt{n}} \omega(\mathcal{X}\mathcal{A}) \quad (5)$$

  where $\mathcal{X}\mathcal{A}$ is the image of the atom set under the linear operator $\mathcal{X}$, and $\delta > 0$ can be chosen arbitrarily according to the probability of success we would like to attain.
- $\lambda_{\mathcal{A}}(\mathcal{X}, \sigma, n)$ is a global parameter that depends on $\mathcal{X}$ and $\mathcal{A}$, but, not on the complexity of M.

Local Geometric Theory: Gaussian Setting.

Local Geometric Upper Bound.

Theorem 1 (Gaussian Ensemble: Convergence Rate)

Suppose we observe $(\mathcal{X}, Y)$ as in (1) with the Gaussian ensemble design and $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. Let $\hat{M}$ be the solution of (2) with λ chosen as in (5). Let $0 < c < 1$ be a constant. For any $\delta > 0$, if

$$n \geq \frac{4[\omega(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c},$$

then with probability at least $1 - 3exp(-\delta^2/2)$

$$\|\hat{M} - M\|_{l_2} \leq \frac{2\sigma}{(1-c)^2} \cdot \frac{\gamma_{\mathcal{A}}(M)\omega(\mathcal{X}\mathcal{A})}{\sqrt{n}},$$

$$\|\hat{M} - M\|_{\mathcal{A}} \leq \frac{2\sigma}{(1-c)^2} \cdot \frac{\gamma_{\mathcal{A}}^2(M)\omega(\mathcal{X}\mathcal{A})}{\sqrt{n}},$$

$$\|\mathcal{X}(\hat{M} - M)\|_{l_2} \leq \frac{2\sigma}{(1-c)} \cdot \frac{\gamma_{\mathcal{A}}(M)\omega(\mathcal{X}\mathcal{A})}{\sqrt{n}}.$$

Local Geometric Theory: Gaussian Setting.

Local Geometric Upper Bound.

- Theorem 1 gives bounds for the estimation error under both the l_2 *norm* loss and the *atomic norm* loss as well as for the in sample prediction error.
- The upper bounds are determined by the geometric quantities $\omega(\mathfrak{X}\mathcal{A}), \gamma_{\mathcal{A}}(M)$ and $\omega(B_2^p \cap T_{\mathcal{A}}(M))$.
- Given any $\epsilon > 0$, the smallest sample size n to ensure the recovery error $\|\hat{M} - M\|_{l_2} \leq \epsilon$ with probability at least $1 - 3exp(-\delta^2/2)$ is

$$n \geq \max\left\{ \frac{4\sigma^2}{(1-c)^4} \cdot \frac{\gamma_{\mathcal{A}}^2(M)\omega^2(\mathfrak{X}\mathcal{A})}{\epsilon^2}, \; \frac{4\omega^2(B_2^p \cap T_{\mathcal{A}}(M))}{c^2} \right\}.$$

Local Geometric Inference.

Theorem 2 (Geometric Inference)

Suppose we observe $(\mathcal{X}, Y)$ as in (1) with the Gaussian ensemble design and $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. Assume $p \geq n \gtrsim \omega^2(B_2^p \cap T_{\mathcal{A}}(M))$. If the tuning parameters λ, η are chosen with

$$\lambda \asymp \frac{\sigma}{\sqrt{n}} \omega(\mathcal{X}\mathcal{A}), \quad \eta \asymp \frac{1}{\sqrt{n}} \omega(\mathcal{X}\mathcal{A}),$$

convex programs (2) and (3) have non-empty feasibility set for Ω with high probability, and the following decomposition

$$\tilde{M} - M = \Delta + \frac{\sigma}{\sqrt{n}} \Omega \mathcal{X}^* W$$

holds, where $W \sim N(0, I_n)$ and $\Delta \in \mathbb{R}^p$ satisfies

$$\|\Delta\|_{l_\infty} \lesssim \gamma_{\mathcal{A}}^2(M) \cdot \lambda\eta \asymp \sigma \frac{\gamma_{\mathcal{A}}^2(M) \omega^2(\mathcal{X}\mathcal{A})}{n}.$$

Local Geometric Inference.

Theorem 2 (Geometric Inference)(Cont.)

Suppose $(n, p(n))$ as a sequence satisfies

$$\limsup_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}^2(M)\omega^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

then for any $v \in \mathbb{R}^p, \|v\|_{l_1} \leq \rho$ with ρ finite, we have the asymptotic normality for the functional $\langle v, \tilde{M} \rangle$,

$$\frac{\sqrt{n}}{\delta}\Big(\langle v, \tilde{M}\rangle - \langle v, M\rangle\Big) = \sqrt{v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v} \cdot Z_0 + o_p(1)$$

where $Z_0 \sim N(0,1)$ and $\lim_{n,p(n)\to\infty} o_p(1) = 0$ means convergence in probability.

Remark:

- It follows from Theorem 2 that a valid asymptotic $(1-\alpha)$-level confidence intervals for M_i, $1 \le i \le p$ (when v is taken as e_i in Theorem 2) is

$$\left[\tilde{M}_i + \Phi^{-1}(\frac{\alpha}{2})\sigma\sqrt{\frac{[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]_{ii}}{n}},\ \tilde{M}_i + \Phi^{-1}(1-\frac{\alpha}{2})\sigma\sqrt{\frac{[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]_{ii}}{n}}\right]$$

- For hypothesis testing problem

$$H_0 : \Sigma_{i=1}^p v_i M_i = v_0 \quad \text{v.s.} \quad H_\alpha : \Sigma_{i=1}^p v_i M_i \neq v_0.$$

  The test statistic is $\frac{\sqrt{n}(\langle v, \tilde{M}\rangle - v_0)}{\sigma(v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v)^{1/2}}$ and under the null, it follows an asymptotic standard normal distribution as $n \to \infty$.
- Since Ω is an estimate of the inverse of $\mathcal{X}^*\mathcal{X}$, the asymptotic variance nearly achieves the Fisher information lower bound.

Minimax Lower Bound for Local Tangent Cone.

Theorem 3 (Lower bound Based on Local Tangent Cone)

Suppose we observe $(\mathcal{X}, Y)$ as in (1) with the Gaussian ensemble design and $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. Let $0 < c < 1$ be a constant. For any $\delta > 0$, if

$$n \geq \frac{4[\omega(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c}.$$

Then with probability at least $1 - 2exp(-\delta^2/2)$

$$\inf_{\hat{M}} \sup_{M' \in T_{\mathcal{A}}(M)} \mathbb{E}_{\cdot|\mathcal{X}} \|\hat{M} - M'\|_{l_2}^2 \geq \frac{c_0 \sigma^2}{(1+c)^2} \cdot \left(\frac{e(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}} \right)^2$$

for some universal constant $c_0 > 0$. Here $\mathbb{E}_{\cdot|\mathcal{X}}$ stands for the conditional expectation given the design matrix $\mathcal{X}$, and the probability statement is with respect to the distribution of $\mathcal{X}$ under the Gaussian ensemble design.

Minimax Lower Bound for Local Tangent Cone.

- There exists a universal constant $c > 0$ such that

$$
\begin{aligned}
c \cdot e(B_2^p \cap T_{\mathcal{A}}(M)) &\leq \omega(B_2^p \cap T_{\mathcal{A}}(M)) \\
&\leq 24 \int_0^\infty \sqrt{\log \mathcal{N}(B_2^p \cap T_{\mathcal{A}}(M), \epsilon)}\, d\epsilon
\end{aligned}
$$

- Under the Gaussian setting, both in terms of the upper bound and lower bound, geometric complexity measures govern the diffculty of the estimation problem, through closely related quantities Gaussian width and Sudakov estimate.

Local Geometric Theory: General Setting.

Definition (local isometry constants)

$$\phi_{\mathcal{A}}(M, \mathcal{X}) := \inf\left\{ \frac{\|\mathcal{X}(h)\|_{\mathcal{A}}}{\|h\|_{l_2}} : h \in T_{\mathcal{A}}(M), h \neq 0 \right\}.$$

$$\psi_{\mathcal{A}}(M, \mathcal{X}) := \sup\left\{ \frac{\|\mathcal{X}(h)\|_{\mathcal{A}}}{\|h\|_{l_2}} : h \in T_{\mathcal{A}}(M), h \neq 0 \right\}.$$

- The local isometry constants measure how well the linear operator preserves the l_2 norm within the local tangent cone.
- Intuitively, the larger the ψ or the smaller the ϕ is, the harder the recovery is.

Local Geometric Theory: General Setting.

Local Geometric Upper Bound.

Theorem 4 (Geometrizing Local Convergence)

Suppose we observe $(\mathcal{X}, Y)$ as in (1). Assume $\mathcal{X}$ is fixed. Condition on the event that the noise vector Z satisfies, for some given λ_n

$$\|\mathcal{X}^*(Z)\|_{\mathcal{A}}^* \leq \lambda_n$$

Let $\hat{M}$ be the solution to the convex program (2) with λ_n being the tuning parameter. Then

$$\|\hat{M} - M\|_{l_2} \leq \frac{2\gamma_{\mathcal{A}}(M)}{\phi_{\mathcal{A}}^2(M, \mathcal{X})} \lambda_n,$$

$$\|\hat{M} - M\|_{\mathcal{A}} \leq \frac{2\gamma_{\mathcal{A}}^2(M)}{\phi_{\mathcal{A}}^2(M, \mathcal{X})} \lambda_n,$$

$$\|\mathcal{X}(\hat{M} - M)\|_{l_2} \leq \frac{2\gamma_{\mathcal{A}}(M)}{\phi_{\mathcal{A}}(M, \mathcal{X})} \lambda_n.$$

Local Geometric Theory: General Setting.

Local Geometric Upper Bound.

- λ_n quantifies the uncertainty in estimation for a given sample size. It is a global parameter which does not depend on the local geometry.
- $\gamma_{\mathcal{A}}(M)$ and $\phi_{\mathcal{A}}(M, \mathfrak{X})$ depend on the local tangent cone geometry.
- Theorem 4 holds deterministically under the conditions on $\|\mathfrak{X}^*(Z)\|_{\mathcal{A}}^*$ and $\phi_{\mathcal{A}}(M, \mathfrak{X})$. It does not require distributional assumptions on noise, nor does it impose conditions on the design matrix.

Local Geometric Theory: General Setting.

Theorem 5 (General Local Minimax Lower Bound)

Let $T \in \mathbb{R}^p$ be a compact convex cone. The minimax lower bound for the linear inverse model (1), if restricted to the cone T, is

$$\inf_{\hat{M}} \sup_{M \in T} \mathbb{E}_{\cdot|\mathcal{X}} \|\hat{M} - M\|_{l_2}^2 \geq \frac{c_0 \sigma^2}{\psi^2} \cdot \Big(\frac{e(B_2^p \cap T)}{\sqrt{n}} \vee \frac{v(B_2^p \cap T)}{\sqrt{n}} \Big)^2$$

where $\hat{M}$ is any measurable estimator, $\psi = \sup_{v \in B_2^p \cap T} \|\mathcal{X}(v)\|_{l_2}$ and c_0 is a universal constant. Here $\mathbb{E}_{\cdot|\mathcal{X}}$ is conditioned on the design matrix. $\mathcal{X}$ $e(\cdot)$ and $v(\cdot)$ denote the Sudakov estimate and volume ratio. Then

$$\begin{aligned} &\inf_{\hat{M}} \sup_{M' \in T_{\mathcal{A}}(M)} \mathbb{E}_{\cdot|\mathcal{X}} \|\hat{M} - M'\|_{l_2}^2 \\ \geq &\frac{c_0 \sigma^2}{\psi_{\mathcal{A}}^2(M, \mathcal{X})} \cdot \Big(\frac{e(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}} \vee \frac{v(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}} \Big)^2 \end{aligned}$$

Application for High-Dimensional Linear Regression.

Corollary 1

Consider the linear regression model $Y = XM + Z$. Assume that $X \in \mathbb{R}^{n\times p}$ is the Gaussian ensemble design and the parameter of interest $M \in \mathbb{R}^p$ is of sparsity s. Let $\hat{M}$ be the solution to the constrained l_1 minimization (2) with $\lambda = C_1\sigma\sqrt{\frac{\log p}{n}}$. If $n \geq C_2 s \log p$, then

$$\|\hat{M} - M\|_{l_2} \lesssim \sigma\sqrt{\frac{s\log p}{n}},$$

$$\|\hat{M} - M\|_{l_1} \lesssim \sigma s\sqrt{\frac{\log p}{n}},$$

$$\|X(\hat{M} - M)\|_{l_2} \lesssim \sigma\sqrt{\frac{\log p}{n}}.$$

with high probability, where $C_1, C_2 > 0$ are some universal constants.

References

- Candes, E. J. and Plan, Y. (2011), Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on,*, **57(4)**, 2342-2359.
- Candes, E. J. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* , **35**, 2313-2351.