Estimation of High-Dimensional Low Rank Matrices

Alexandre Tsybakov (joint work with Angelika Rohde)

Laboratoire de Statistique, CREST

Marne-la-Vallée, May 18, 2010

Trace Regression Model

- We observe $(X_i, Y_i), i = 1, \ldots, N$ such that

$$Y_i \;=\; \mathrm{trace}\left(X_i' A^*\right) \;+\; \xi_i, \quad i = 1, ..., N,$$

ξ_i i.i.d. random errors, $X_i \in \mathbb{R}^{m \times T}$ known, $A^* \in \mathbb{R}^{m \times T}$ unknown

- Problems:
  - estimation of A^*;
  - prediction $=$ estimation of $X \mapsto \mathrm{trace}\left(X' A^*\right)$.
- Focus on:
  - High-dimensional setting: $mT \gg N$.
  - A^* is a matrix of small rank, $r = \mathrm{rank}(A^*) \ll \min(m, T)$.
  - Sparse matrices X_i (**masks**): few non-zero entries.

Examples: 1. Point masks.

$$X_i \in \left\{ \sum_{j=1}^{d} e_{k_j}(m) e'_{l_j}(T) : \ 1 \le k_j \le m, 1 \le l_j \le T \right\},$$

$e_k(m)$'s the canonical basis vectors in $\mathbb{R}^m$.

- $d = 1$: Matrix Completion Problem. Suppose we observe only $N \ll mT$ entries of matrix $A^* \in \mathbb{R}^{m \times T}$ with/without noise

$$\longrightarrow \text{ can we guess the many other entries?}$$

- Applications: Recommendation systems, e.g., Netflix; dimension $mT \sim 10^9$, $N \sim 10^7$.
- Role of the rank: Let $m = T$ $\Rightarrow$ completion impossible if $N < (2m - r)r$, where $r = \mathrm{rank}(A^*)$

- Two cases of matrix completion:
  - USR matrix completion = Uniform Sampling at Random; masks X_i i.i.d. uniformly distributed on the set

$$\left\{ e_k(m)e_l'(T) : \ 1 \le k \le m, 1 \le l \le T \right\} .$$

    Non-noisy case: Candès/Recht (2008), Candès/Tao (2009).
  - Collaborative filtering. Random or deterministic masks X_i, which are all **distinct**.

Examples: 2. Column or row masks

- Multi-task learning = longitudinal (or panel, or cross-section) data analysis
- $N = nT$ where T number of tasks; n number of observations per task.
- Vectors of parameters $a_t^* \in \mathbb{R}^m$, $t = 1, \dots, T$ for tasks,

$$A^* = (a_1^* \cdots a_T^*).$$

- X_i's are **column masks**, only one non-zero column $\mathbf{x}^{(t,s)} \in \mathbb{R}^m$:

$$X_i \in \{(\mathbf{0} \cdots \mathbf{0} \underbrace{\mathbf{x}^{(t,s)}}_{t} \mathbf{0} \cdots \mathbf{0}),\ t = 1, \dots, T,\ s = 1, \dots, n\}.$$

- Column $\mathbf{x}^{(t,s)}$ = the vector of predictor variables corresponding to sth observation for the tth task.

Thus, for each $i = 1, \ldots, N$ there exists a pair (t, s) with $t = 1, \ldots, T, \; s = 1, \ldots, n$, such that

$$\mathrm{trace}(X_i' A^*) = (a_t^*)' \mathbf{x}^{(t,s)}.$$

If we denote by $Y^{(t,s)}$ and $\xi^{(t,s)}$ the corresponding values Y_i and ξ_i, our trace regression model can be written as a collection of T standard vector regression models:

$$Y^{(t,s)} = (a_t^*)' \mathbf{x}^{(t,s)} + \xi^{(t,s)}, \quad t = 1, \ldots, T, \; s = 1, \ldots, n.$$

(Usual formulation of multi-task learning model.)

- Suppose $A^* = (a_1^* \cdots a_T^*)$ has small rank $\equiv$ "tasks are related".
- Problems: estimation of A^*, prediction.

Examples: 3. "Complete" matrices X_i

- All the entries of X_i are i.i.d. Rademacher or Gaussian $\mathcal{N}(0,1)$.
- X_i are no longer **masks**.
- Computationally hard when mT is large, e.g., $mT \sim 10^9$.
- Our results cover this case but it is not of our main interest.
- Parallel work on this case: Negahban/Wainwright (2009) with $N \gg mT$; Candès/Plan (2010). Without noise: Recht/al. (2007).

Our aim is to construct estimators $\widehat{A}$ of matrix A^* such that the following distance measures are small with probability close to 1:

- Prediction loss $\quad d^2(\widehat{A}, A^*) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{trace}^2((\widehat{A} - A^*)' X_i)$
- Schatten-q loss $\quad \| \widehat{A} - A^* \|_{S_q}^q$

$\| \cdot \|_{S_q}$ denotes Schatten-q (quasi-)norm

$$\|A\|_{S_q} = \left(\sum_{j=1}^{m \wedge T} \sigma_j(A)^q \right)^{1/q}, \quad q > 0,$$

with $\sigma_i(A)$'s singular values of matrix $A \in \mathbb{R}^{m \times T}$.

Prototype reference: Vector estimation and Lasso

- We observe $(X_i, Y_i), i = 1, \dots, N$, such that

$$Y_i \;=\; X_i'\beta \;+\xi_i, \quad i = 1, \dots, N,$$

$X_i \in \mathbb{R}^p,\ \beta \in \mathbb{R}^p,\ \xi_i \overset{iid}{\sim} \mathcal{N}(0,1)$

- High-dimensional setting: $p \gg N$.

- Sparsity index s of β = number of non-zero components of β is small;

$$s = |\beta|_0 = \sum_{j=1}^{p} I\{i : \beta_j \neq 0\} \ll p.$$

- vector case: LASSO estimator

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - X_i' \beta \right)^2 \; + \; \lambda |\beta|_1 \right\},$$

$|\beta|_1 = \ell_1$-norm of β, $\lambda > 0$ tuning parameter.

- matrix case: Schatten-1 estimator

$$\widehat{A} \in \underset{A \in \mathbb{R}^{m \times T}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i - \mathrm{trace}\left(X_i' A \right) \Big)^2 \; + \; \lambda \left\| A \right\|_{S_1} \right\}.$$

- penalized least squares with Schatten (quasi-)norm penalty

motivation: shrinkage towards low-rank representations

- vector case: LASSO estimator

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - X_i' \beta \right)^2 \ + \ \lambda |\beta|_1 \right\},$$

$|\beta|_1 = \ell_1$-norm of β, $\lambda > 0$ tuning parameter.

- matrix case: Schatten-p estimator

$$\widehat{A} \in \underset{A \in \mathbb{R}^{m \times T}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i - \mathrm{trace}\left(X_i' A \right) \Big)^2 \ + \ \lambda \| A \|_{S_p}^p \right\}, \ 0 < p \leq 1.$$

- penalized least squares with Schatten (quasi-)norm penalty

motivation: shrinkage towards low-rank representations

Sparsity Oracle Inequalities – Vector Case

Prediction loss: $d^2(\widehat{\beta}, \beta) = \frac{1}{N}|\mathbf{X}(\widehat{\beta} - \beta)|_2^2$,

$\mathbf{X} = (X_{ji})_{1\le i\le N; 1\le j\le p}$, and $|\cdot|_q$, $q \ge 1$, is the ℓ_q norm.

Theorem (Bickel, Ritov and T., 2009, Rigollet and T., 2010)

Consider the Lasso estimator $\widehat{\beta}$ with $\lambda = A\sqrt{\frac{\log p}{N}}$,$A > 2\sqrt{2}$. Then with probability at least $1 - p^{1-A^2/8}$, under the RI condition,

$$d^2(\widehat{\beta}, \beta) \le C\left(\frac{s \log p}{N}\right), s = |\beta|_0, \quad \text{"FAST" rate},$$

and, under NO assumption on $\mathbf{X}$,

$$d^2(\widehat{\beta}, \beta) \le C|\beta|_1\sqrt{\frac{\log p}{N}} \quad \text{"SLOW" rate}.$$

Sparsity Oracle Inequalities – Matrix Case???

- Investigate two possibilities:
  (i) "Fast" rates scheme. Here we need some strong conditions, such as matrix analogs of RI assumption.
  (ii) "Slow" rates scheme. We need essentially no assumption on the masks but some mild assumptions on the Schatten norm of A^*.

- **The outcome is surprising**:
  (i) "Fast" rates scheme (i.e., using RI) essentially fails when we deal with very sparse masks X_i.
  (ii) "Slow" rates scheme leads to the rates which are NOT slow if matrices X_i are very sparse!

- Schatten-p estimator:

$$\widehat{A} \in \underset{A \in \mathbb{R}^{m \times T}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{j=1}^{N} \Big(Y_j - \mathrm{trace}\,(X_j' A) \Big)^2 + \lambda \| A \|_{S_p}^{p} \right\}, \quad p \leq 1$$

- Prediction loss:

$$d^2(\widehat{A}, A^*) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{trace}^2((\widehat{A} - A^*)' X_i)$$

- Basic inequality

$$d^2(\widehat{A}, A^*) \leq \underbrace{2 \frac{1}{N} \sum_{i=1}^{N} \xi_i \,\mathrm{trace}\Big((\widehat{A} - A^*)' X_i \Big)}_{\text{"stochastic term"}} + \lambda \Big(\| A^* \|_{S_p}^{p} - \| \widehat{A} \|_{S_p}^{p} \Big)$$

Lemma

Under appropriate assumptions, with probability $\geq 1 - \exp(-C(m+T))$,

$$\left| \frac{1}{N} \sum_{i=1}^{N} \xi_i \mathrm{trace}\big((\widehat{A} - A^*)' X_i\big) \right| \;\leq\; \frac{\delta}{2} I_{\{0<p<1\}} d^2\big(\widehat{A}, A^*\big) + \tau \delta^{p-1} \|\widehat{A} - A^*\|_{S_p}^p,$$

for all $\delta > 0$, where $0 < \tau < \infty$ is an explicitly given parameter(m, T, N).

Difficulty: requires some new tools, e.g., ϵ-entropy of the (non-convex) Schatten-p ball $\{A \in \mathbb{R}^{m\times m} : \|A\|_{S_p} \leq 1\}$, $p < 1$, in the Frobenius norm, with explicit dependence on p

$\tau =$ "EFFECTIVE NOISE LEVEL";

Choose $\lambda = 4\tau$

Examples of "noise levels" τ (Gaussian ξ_i)

Assumptions on X_i	Assumptions on N, m, T, p	"Noise levels" τ
Unif. bounded $\mathcal{L}$	$p = 1$	$c\left(\frac{m+T}{N}\right)^{1/2}$
Unif. bounded $\mathcal{L}$	$0 < p < 1,\ m = T$	$c(p)\left(\frac{m}{N}\right)^{1-p/2}$
USR matrix compl.	$p = 1,\ (m+T)mT > N$	$c\frac{m+T}{N}$
Collab. filtering	$p = 1$	$c\frac{(m+T)^{1/2}}{N}$

The sampling operator $\mathcal{L} : A \mapsto \left(\mathrm{trace}(X_1'A), ..., \mathrm{trace}(X_N'A)\right)/\sqrt{N}$ is uniformly bounded if there exists a constant $c_0 < \infty$ such that

$$|\mathcal{L}(A)|_2^2 \;\le\; c_0 \|A\|_{S_2}^2$$

for all matrices $A \in \mathbb{R}^{m\times T}$ where $|\cdot|_2$ is the Euclidean norm in $\mathbb{R}^N$.

We first explore the "Slow rates" scheme: without Restricted Isometry

"Slow rates" scheme

- Basic inequality + Lemma, setting $\delta = 1/2$ and $\lambda = 4\tau$:

$$d^2(\widehat{A}, A^*) \;\le\; 8\tau\Big(\|\widehat{A} - A^*\|_{S_p}^p + \|A^*\|_{S_p}^p - \|\widehat{A}\|_{S_p}^p\Big) \;\le\; 16\tau\|A^*\|_{S_p}^p$$

since $\|A + B\|_{S_p}^p \le \|A\|_{S_p}^p + \|B\|_{S_p}^p$, $p \le 1$.

Theorem (Sparsity Oracle Inequality – "Slow rates" scheme)

Let $0 < p \le 1$, $\lambda = 4\tau$. Then, for cases listed in the table above,

$$d^2\big(\widehat{A}, A^*\big) \;\le\; 16\tau\|A^*\|_{S_p}^p,$$

with probability $\ge 1 - \exp(-C(m+T))$ where $C > 0$ is independent of N, m, T.

Remarks

- The rate is faster for smaller p in the penalty.
- If $\sigma_1(A^*) \le C$ we have the bound

$$d^2\left(\widehat{A}, A^*\right) \;\le\; Cr\tau.$$

So, the rates are FAST or VERY FAST:

- for uniformly bounded sampling operator with $m = T$, $p = (\log(N/m))^{-1}$:

$$d^2\left(\widehat{A}, A^*\right) \sim \frac{rm}{N}\log\left(\frac{N}{m}\right),$$

- for USR matrix completion with $p = 1$:

$$d^2\left(\widehat{A}, A^*\right) \sim \frac{r(m+T)}{N},$$

- for collaborative filtering with $p = 1$:

$$d^2\left(\widehat{A}, A^*\right) \sim \frac{r(m+T)^{1/2}}{N}.$$

Rate heuristics for prediction loss: Square matrix case

- $A^* \in \mathbb{R}^{m\times m}$ and $\mathrm{rank}(A^*) = r$
  $\Rightarrow \quad (2m-r)r$ free parameters

$$r \ll m \quad \Rightarrow \quad \text{intrinsic dimension} \sim rm$$

$$\textbf{Rate} = \quad \frac{\textit{intrinsic dimension}}{\textit{sample size}} \sim \frac{rm}{N} \left(\ll \frac{m^2}{N} \right)$$

- For USR matrix completion setting we achieve the optimal rate heuristics using the "slow rate" argument if the maximal singular value of A^* is uniformly bounded.

- Collaborative filtering leads to even faster convergence rates as compared to USR matrix completion.

- On the difference from the vector problems, the log-factor is can be avoided in the rates if the maximal singular value is uniformly bounded.

- Another difference is that the concentration is exponential and not polynomial in the dimension.

We now turn to "Fast rates" scheme:

with Restricted Isometry

Restricted Isometry: Vector versus Matrix

- Vector case. Restricted Isometry: $\exists\ 0 < \delta_s < 1$ such that

$$(1-\delta_s)|\beta|_2 \le \frac{1}{\sqrt{N}}|\mathbf{X}\beta|_2 \le (1+\delta_s)|\beta|_2$$

for all $\beta \in \mathbb{R}^p$ with sparsity index $|\beta|_0 \le s$.

- Matrix case. Restricted Isometry RI(r,ν) condition:
$\exists\ 0 < \delta_r < 1$ such that

$$(1-\delta_r)\|A\|_{S_2} \ \le\ \nu\left(\frac{1}{N}\sum_{i=1}^{N}\mathrm{trace}^2(A'X_i)\right)^{1/2} \ \le\ (1+\delta_r)\|A\|_{S_2}$$

for all $A \in \mathbb{R}^{m\times T}$ with $\mathrm{rank}(A) \le r$. Scaling factor ν.

Examples.

1. USR matrix completion. Point masks. The scaling constant in matrix version of Restricted Isometry is

$$\nu \sim \sqrt{mT}.$$

But we can only achieve it if $N > mT$
$\Rightarrow$ "matrix completion catastrophe", see below...

2. Multi-task learning. Column masks. The scaling constant is

$$\nu \sim \sqrt{T}.$$

3. "Complete" matrices X_i. All Gaussian or Rademacher entries. Restricted isometry with scaling constant

$$\nu = 1,$$

cf. Recht et al. (2007).

Assumptions on X_i	Assumptions on N, m, T, p	"Noise levels" τ
Unif. bounded $\mathcal{L}$	$p = 1$	$c\left(\frac{m+T}{N}\right)^{1/2}$
Unif. bounded $\mathcal{L}$	$0 < p < 1$, $m = T$	$c(p)\left(\frac{m}{N}\right)^{1-p/2}$

Theorem (Sparsity Oracle Inequality – "Fast" scheme: with RI)

Let $\mathrm{rank}(A^*) \le r$. *Assume the RI (br,ν) condition with a sufficiently large* $b = b(p)$ *and some* $0 < \nu < \infty$. *Let the sampling operator* $\mathcal{L}$ *be uniformly bounded. Then, for the Schatten-p estimator* $\widehat{A}$ *with* $\lambda = 4\tau$, *with* τ *as in the table above we have*

$$
\begin{aligned}
d^2\left(\widehat{A}, A^*\right) \;&\le\; C r \tau^{\frac{2}{2-p}} \nu^{\frac{2p}{2-p}}, \\
\|\widehat{A} - A^*\|_{S_q}^q \;&\le\; C r \tau^{\frac{q}{2-p}} \nu^{\frac{2q}{2-p}}, \quad \forall\, q \in [p, 2],
\end{aligned}
$$

with probability $\ge 1 - \exp(-C(m+T))$ *where* $C > 0$ *is independent of* N, m, T.

Remarks

- "Complete" matrices X_i. Then $\nu = 1$. If also $p = 1$, we have the bound

$$d^2(\widehat{A}, A^*) \;\leq\; Cr\tau^2 \sim \frac{r(m+T)}{N}.$$

Same for the Frobenius norm. This is the optimal rate.

- USR matrix completion: no Restricted Isometry if $mT \gg N$. The RI scheme does not apply.

Example: USR matrix completion

X_i point masks which are i.i.d. and uniformly distributed on

$$\left\{ e_k(m)e_l'(T) : 1 \le k \le m,\, 1 \le l \le T \right\}.$$

Set $\delta_{kl}^{(i)} = I_{\{X_i = e_k(m)e_l'(T)\}}$. Then $\forall\ A \in \mathbb{R}^{m\times T}$:

$$\frac{mT}{N}\sum_{i=1}^{N} \mathrm{tr}^2(X_i'A) \;=\; \frac{mT}{N}\sum_{i=1}^{N}\sum_{k,l} a_{kl}^2\,\delta_{kl}^{(i)} \;=\; \sum_{k,l} a_{kl}^2 \left(\frac{mT}{N}\sum_{i=1}^{N}\delta_{kl}^{(i)}\right).$$

But $\mathbf{E}\left(\frac{mT}{N}\sum_{i=1}^{N}\delta_{kl}^{(i)}\right) = 1$ for all k, l, and $\sum_{k,l} a_{kl}^2 = \|A\|_{S_2}^2$.

- $\Rightarrow$ the RI condition, if it holds, should be naturally scaled by $\nu = \sqrt{mT}$, a very large value.

Matrix completion: the RI catastrophe

$$\frac{mT}{N}\sum_{i=1}^{N}\mathrm{trace}^2(X_i'A) \;=\; \sum_{k,l} a_{kl}^2\Big(\frac{mT}{N}\sum_{i=1}^{N}\delta_{kl}^{(i)}\Big).$$

$\mathbf{E}\Big(\frac{mT}{N}\sum_{i=1}^N \delta_{kl}^{(i)}\Big) = 1$ for all k, l, and $\sum_{k,l} a_{kl}^2 = \|A\|_{S_2}^2$.

- Since $\delta_{kl}^{(i)}$ are i.i.d. Bernoulli(1/(mT)),

$\mathrm{Var}\Big(\frac{mT}{N}\sum_{i=1}^N \delta_{kl}^{(i)}\Big) \sim \frac{mT}{N}$ $\Rightarrow$ RI condition requires $mT < N$!

$\Rightarrow$ nothing can be done under the requirement $mT \gg N$ which is intrinsic for matrix completion problems.

$\Rightarrow$ Restricted isometry not adapted to problems with sparse masks

Theorem (Matrix completion, I)

Let $\xi_1, \dots, \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables, and assume that $m = T > 1$, $N > \mathrm{e}\, m$ and that X_i are i.i.d. uniformly distributed on

$$\Big\{ e_k(m) e_l'(T) : \, 1 \le k \le m, \, 1 \le l \le T \Big\}.$$

Let $A^* \in \mathbb{R}^{m \times m}$ with $\mathrm{rank}(A^*) \le r$ and the maximal singular value $\sigma_1(A^*) \le (N/m)^{C^*}$ for some $0 < C^* < \infty$. Set

$$p = (\log(N/m))^{-1}.$$

Then, $\forall\ \vartheta \ge c^2$ with a universal constant $c > 0$, for a proper choice of $\lambda = \lambda(\vartheta)$, the Schatten-p estimator $\widehat{A}$ satisfies:

$$d(\widehat{A}, A^*)^2 \le C \vartheta \, \frac{rm}{N} \log\left(\frac{N}{m}\right)$$

with probability $\ge 1 - c \, \exp(-\vartheta m / c^2)$ for some $c > 0$.

Theorem (Matrix completion, II)

Let ξ_i, $i = 1, ..., N$, *with*

$$\boldsymbol{E}|\xi_i|^l \ \le \ \frac{1}{2} l! \sigma^2 H^{l-2}, \quad l = 2, 3, ...,$$

with some finite constants σ *and* H*. Assume that* $mT(m + T) > N$ *and that the* X_i *are point masks, which are iid uniformly distributed on*

$$\left\{ e_k(m) e_l'(T) : \ 1 \le k \le m, \ 1 \le l \le T \right\}$$

and independent from $\xi_1, ..., \xi_N$*. Then with an appropriate choice of* $\lambda = \lambda(m, T, N, \sigma, H)$*, the Schatten-*1 *estimator* $\widehat{A}$ *satisfies*

$$d(\widehat{A}, A^*)^2 \ \le \ 16 \bar{C} \|A^*\|_{S_1} \frac{m + T}{N}$$

with probability at least $1 - 4 \exp\{-(2 - \log 5)(m + T)\}$*, where* $\bar{C} = \bar{C}(\sigma, H)$*.*

Theorem (Matrix completion, III)

Let ξ_i, $i = 1, ..., N$, iid $\mathcal{N}(0, \sigma^2)$. Consider the problem of collaborative filtering (i.e. N different point masks). Then the Schatten-1 estimator $\widehat{A}$ with $\lambda = \lambda(m, T, N, \sigma)$ satisfies

$$d(\widehat{A}, A^*)^2 \ \leq \ 256\|A^*\|_{S_1} \frac{\sqrt{m+T}}{N}$$

with probability at least $1 - 2\exp\{-(4 - \log 5)(m + T)\}$.

- collaborative filtering leads to faster convergence rates as compared to USR matrix completion setting
- the log-factor is avoidable for uniformly bounded maximal singular value

Theorem (Multi-task learning)

Let $\xi_1, \dots, \xi_N$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables, and assume that $m = T > 1$, $n > \mathrm{e} \log m$. Consider the multi-task learning problem with $A^* \in \mathbb{R}^{m \times m}$, $\mathrm{rank}(A^*) \le r$ and the maximal singular value $\sigma_1(A^*) \le (n/\log m)^{C^*}$ for some $0 < C^* < \infty$. Assume that the spectra of the task Gram matrices Ψ_t are uniformly in t bounded from above by a $c_0 T$ where $c_0 < \infty$. Set

$$p = (\log n - \log\log m)^{-1}.$$

Then, $\forall\ \vartheta \ge 1$, for a proper choice of $\lambda = \lambda(\vartheta)$, the Schatten-$p$ estimator $\widehat{A}$ satisfies:

$$d(\widehat{A}, A^*)^2 \le C\vartheta \, \frac{r}{n} \log\left(\frac{n}{\log m}\right) \log m$$

with probability $\ge 1 - C\, m^{-\vartheta/C^2}$ for some $C > 0$.

Matrix versus Vector Sparsity

- linear dependence on $\mathrm{rank}(A^*)$

  $\sim$ linear dependence on sparsity index s

- (at least) linear dependence on m

  $\not\sim$ logarithmic dependence on p

- impossible to recover all low-rank matrices

  (counter-) example: $e_i e_j'$, with e_i's the canonical unit vectors

- possible to recover most of them?

Theorem (Candès & Tao 2009)

In the non-noisy setting ($\xi_i \equiv 0$), under the strong incoherence condition (SIC), *exact recovery is possible with high probability for*

$$N > C\, rm(\log m)^6, \qquad r = \mathrm{rank}(A^*),$$

observed entries with locations uniformly sampled at random.

Heuristics:

SIC ensures that the singular vectors of A^* are sufficiently "spread out" or "incoherent"

Matrix completion is possible by convex programming:

$$\text{minimize } \|A\|_{S_1}$$
$$\text{subject to } Y_i = \mathrm{trace}\left(X_i' A\right), \ i = 1, \dots, N$$

- $\|.\|_{S_p}$ denotes Schatten-p (quasi-)norm

$$\|A\|_{S_p} = \left(\sum_{j=1}^{m} \sigma_j(A)^p \right)^{1/p}, \quad p > 0,$$

$\sigma_i(A)$'s singular values of A

- Equivalent: y_{ij}, $(i,j) \in \Omega \subset \{1, 2, ..., m\}^2$ observed entries

$$\text{minimize } \|A\|_{S_1}$$
$$\text{subject to } a_{ij} = y_{ij}, \ (i,j) \in \Omega$$

- Candès and Recht (2008), Candès and Tao (2009)
  $\longrightarrow$ focus on exact recovery

- Candès and Plan (2009)

  $\longrightarrow$ same setting in the presence of noise,
  proposed estimators $\widehat{A}$ of A^* and evaluated $\|\widehat{A} - A^*\|_{S_2}$

  $\longrightarrow$ establish bounds on $\|\widehat{A} - A^*\|_{S_2}^2$ of order m^3
  when $A^* \in \mathbb{R}^{m \times m}$ and the noise is Gaussian

  $\longrightarrow$ argued that even the oracle cannot achieve better rate in
  the Frobenius norm than rm^3/N, which is rather pessimistic

$\Rightarrow$ Nothing reasonable can be achieved for the Frobenius norm in the matrix completion problem

Sparsity for Matrices (Two notions of matrix sparsity)

- small number of non-zero entries
  - $\rightarrow$ Meinshausen and Bühlmann (2006) (in view of inverse covariance matrices and graphical models)
  - $\rightarrow$ Bickel and Lewina (2008) (banded covariance matrices)
  - $\rightarrow$ Wainwright, Yu (2008), ...
- newly introduced in the framework of matrix completion:
  - $\rightarrow$ sparsity quantified by the rank (Recht et al. 2007)

    $$\text{sparse matrix} = \text{small rank matrix}$$
  - $\rightarrow$ Negahban and Wainwright (2009), Candès and Plan (2010) (using restricted isometry of sampling operator)
- We assume:
  - masks X_i have small number of non-zero entries
  - A^* is of small rank, $\mathrm{rank}(A^*) \ll m$