

Convex relaxation and high-dimensional matrices

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Based on joint work with:

Alekh Agarwahl (UC Berkeley)
Sahand Negahban (UC Berkeley)
Pradeep Ravikumar (UT Austin)
Bin Yu (UC Berkeley)

Introduction

High-dimensional data sets are everywhere:

- social networks
- computer vision
- recommender systems and collaborative filtering
- astronomy datasets
- and so on....

Introduction

High-dimensional data sets are everywhere:

- social networks
- computer vision
- recommender systems and collaborative filtering
- astronomy datasets
- and so on....

Question:

Suppose that $n = 100$ and $d = 1000$. Do we expect theory requiring $n \rightarrow +\infty$ and $d = \mathcal{O}(1)$ to be useful?

Introduction

High-dimensional data sets are everywhere:

- social networks
- computer vision
- recommender systems and collaborative filtering
- astronomy datasets
- and so on....

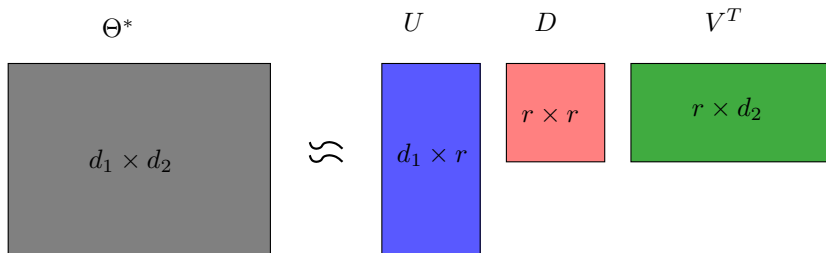
Question:

Suppose that $n = 100$ and $d = 1000$. Do we expect theory requiring $n \rightarrow +\infty$ and $d = \mathcal{O}(1)$ to be useful?

Modern viewpoint:

- non-asymptotic results (valid for all (n, d))
- allow for $n \ll d$ or $n \asymp d$
- investigate various types of **low-dimensional structure**

(Nearly) low-rank matrices



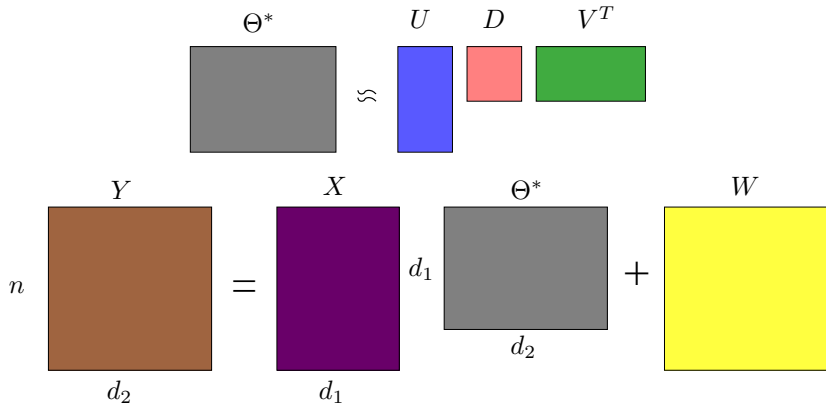
Matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ with $\text{rank } r \ll \min\{d_1, d_2\}$.

Example: Multiview imaging



Low-rank multitask regression


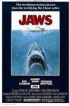



- d_2 tasks in d_1 dimensions
- unknown matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ of approximate rank r



Observations:

- predictor matrix $X \in \mathbb{R}^{n \times d_1}$
- output matrix $Y \in \mathbb{R}^{n \times d_2}$

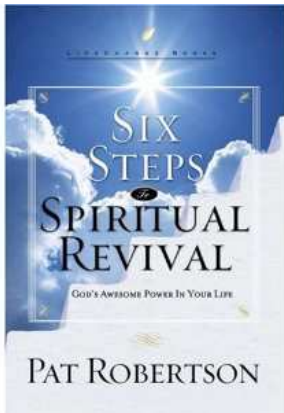
Example: Collaborative filtering

| | | | | | | |
|---|---|---|---|-----|-----|---|
| |  |  |  | ... | ... |  |
|  | 4 | * | 3 | ... | ... | * |
|  | 3 | 5 | * | ... | ... | 2 |
|  | 5 | 4 | 3 | ... | ... | 3 |
|  | 2 | * | * | ... | ... | 1 |

Universe of d_1 individuals and d_2 films Observe $n \ll d_1 d_2$ ratings

(e.g., Srebro, Alon & Jaakkola, 2004)

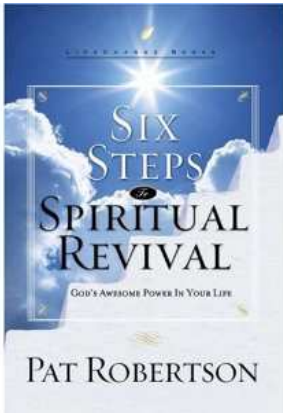
Security and robustness issues



Spiritual guide

Break-down of Amazon recommendation system, 2002.

Security and robustness issues



Spiritual guide

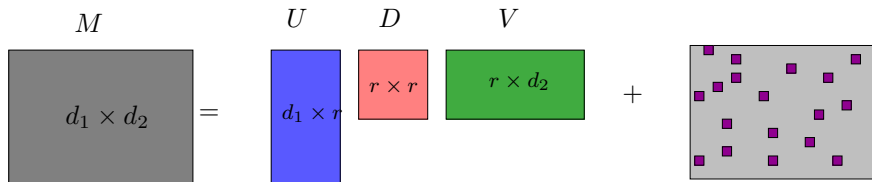


Sex manual

Break-down of Amazon recommendation system, 2002.

Example: Matrix decomposition

Unknown matrix M decomposed into sum:



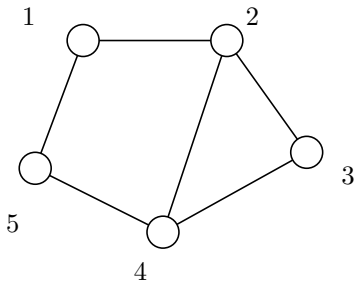
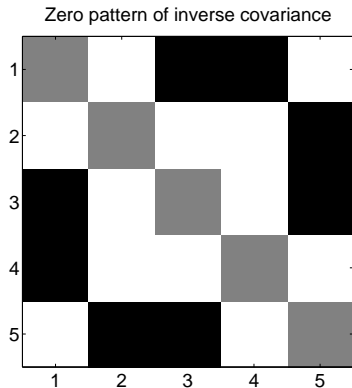
$$M = \underbrace{\Theta^*}_{\text{Low-rank component}} + \underbrace{\Gamma^*}_{\text{Sparse component}}$$

In collaborative filtering:

- low rank component Θ^* represents true user information
- sparse component Γ^* represents adversarial noise

(Chandrasekaran, Sanghavi, Parillo & Willsky, 2009; Candes et al., 2010; Xu et al., 2010; Hsu et al., 2010; Agarwal et al., 2011)

Example: Learning graphical models

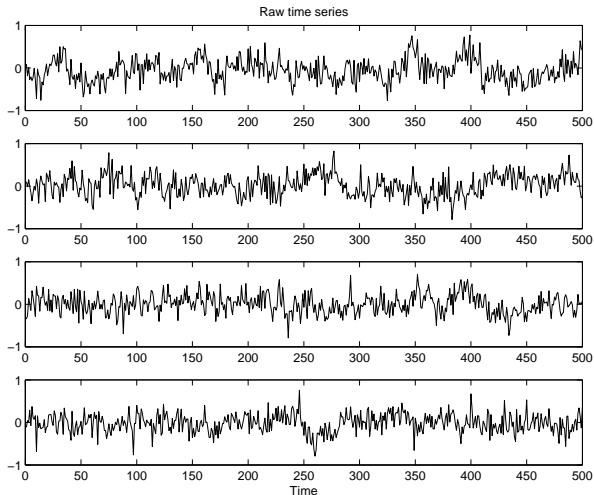


$$\mathbb{P}(x_1, x_2, \dots, x_d) \propto \exp\left(-\frac{1}{2}x^T \Gamma^* x\right).$$

Problems with hidden/latent variables lead to sparse/low-rank decompositions.

(Chandrasekaran, Parillo & Willsky, 2010)

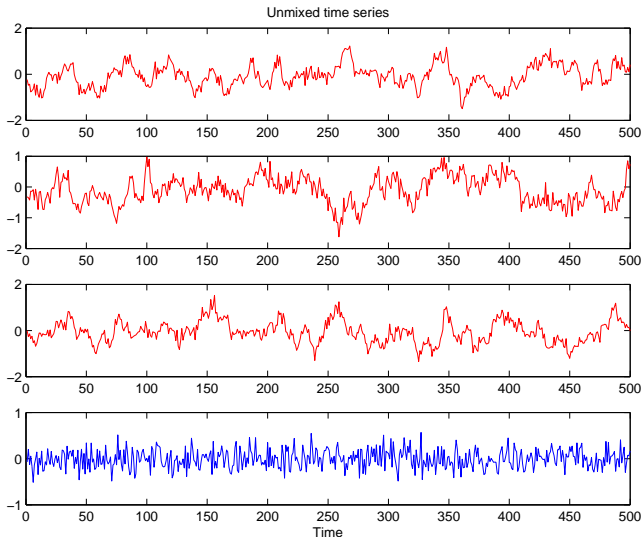
Example: Constrained system identification



Sample paths of first-order time series in $d = 100$ dimensions.

$$X(t+1) = \Theta^* X(t) + W(t), \quad t = 1, 2, \dots$$

Example: Constrained system identification



State within a **3-dimensional subspace**, remaining **97 dimensions of noise**

Remainder of talk

- 1 Matrix regression problems
 - ▶ Regularization with nuclear norm
 - ▶ Restricted strong convexity
 - ▶ A general theorem

- 2 Various examples
 - ▶ Matrix sketching
 - ▶ Matrix completion
 - ▶ Matrix decomposition

Matrix regression problems

For sample size n , define an observation operator $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^n$:

$$\underbrace{\Theta^*}_{d_1 \times d_2 \text{ matrix}} \quad \mapsto \quad \underbrace{\mathfrak{X}(\Theta^*)}_{n\text{-vector of observations}}$$

Operator \mathfrak{X} and output $y \in \mathbb{R}^n$ linked via noisy linear model:

$$y = \mathfrak{X}(\Theta^*) + w.$$

Matrix regression problems

For sample size n , define an observation operator $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^n$:

$$\underbrace{\Theta^*}_{d_1 \times d_2 \text{ matrix}} \quad \mapsto \quad \underbrace{\mathfrak{X}(\Theta^*)}_{n\text{-vector of observations}}$$

Operator \mathfrak{X} and output $y \in \mathbb{R}^n$ linked via noisy linear model:

$$y = \mathfrak{X}(\Theta^*) + w.$$

Estimate unknown Θ^* by minimizing loss function

$$\hat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \{ \mathcal{L}(\Theta; y, \mathfrak{X}) + \lambda_n \|\Theta\|_{\text{nuc}} \},$$

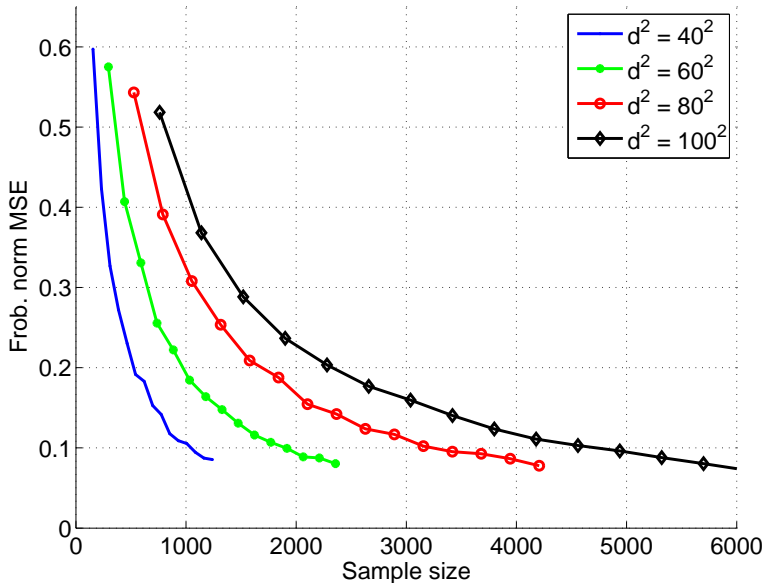
regularized with nuclear norm $\|\Theta\|_{\text{nuc}} = \sum_{j=1}^{\min\{d_1, d_2\}} \sigma_j(\Theta)$

Least-squares loss is commonly used:

$$\mathcal{L}(\Theta; y, \mathfrak{X}) = \frac{1}{2n} \|y - \mathfrak{X}(\Theta)\|_2^2.$$

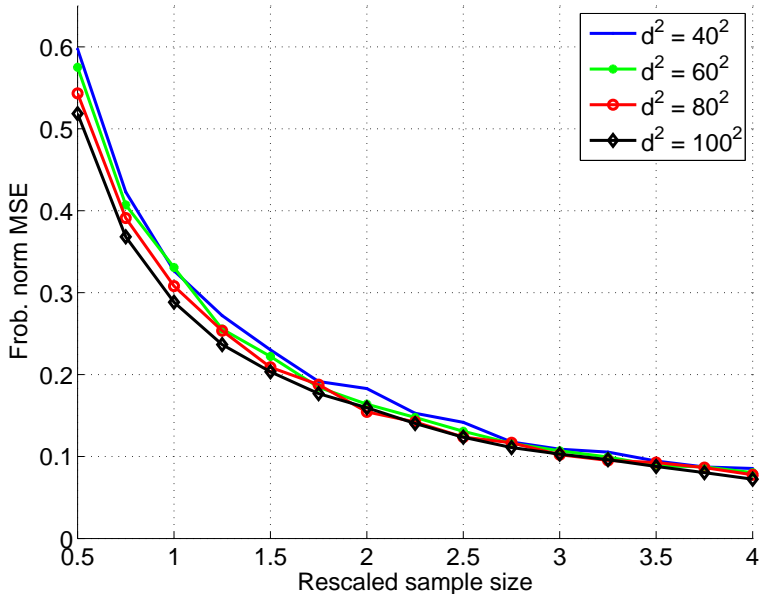
Noisy matrix completion (unrescaled)

MSE versus raw sample size ($q = 0$)

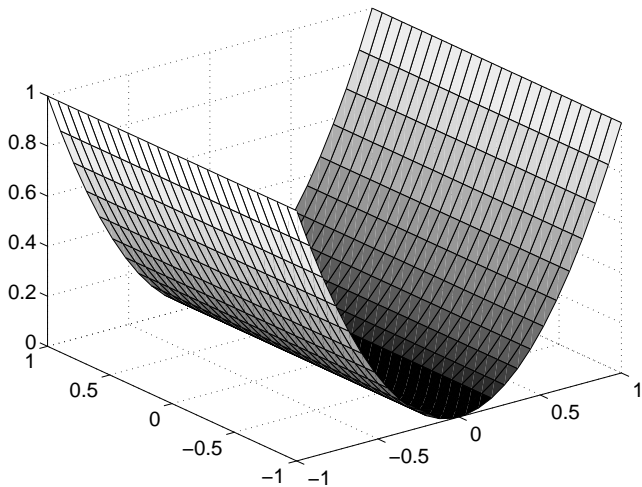


Noisy matrix completion (rescaled)

MSE versus rescaled sample size ($q = 0$)



Strong convexity never holds



When $n \ll d_1 d_2$, the Hessian $\nabla^2 \mathcal{L}(\Theta; y, \mathfrak{X}) = \frac{1}{n} \mathfrak{X}^* \mathfrak{X}$ has nullspace of dimension $(d_1 d_2) - n$.

Restricted strong convexity (RSC)

Definition (Negahban et al., 2009)

The operator $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^n$ satisfies RSC (w.r.t. nuclear norm) with curvature $\gamma > 0$ and tolerance $\kappa > 0$

$$\frac{\|\mathfrak{X}(\Theta)\|_2^2}{n} \geq \gamma(\mathfrak{X}) \|\Theta\|_F^2 - \kappa(\mathfrak{X}) \|\Theta\|_{\text{nuc}}^2 \quad \text{for all } \Theta \in \mathbb{R}^{d_1 \times d_2}.$$

Restricted strong convexity (RSC)

Definition (Negahban et al., 2009)

The operator $\mathfrak{X} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^n$ satisfies RSC (w.r.t. nuclear norm) with curvature $\gamma > 0$ and tolerance $\kappa > 0$

$$\frac{\|\mathfrak{X}(\Theta)\|_2^2}{n} \geq \gamma(\mathfrak{X}) \|\Theta\|_F^2 - \kappa(\mathfrak{X}) \|\Theta\|_{\text{nuc}}^2 \quad \text{for all } \Theta \in \mathbb{R}^{d_1 \times d_2}.$$

- 1 Reduces to ordinary strong convexity if $\kappa = 0$, but this **never** holds when $n \ll d_1 d_2$.
- 2 Guarantees that least-squares loss $\mathcal{L}(\Theta) = \frac{1}{2n} \|y - \mathfrak{X}(\Theta)\|_2^2$ is strongly convex in a restricted sense.
- 3 Generalizes to other loss functions and regularizers.
- 4 Substantially milder requirement than restricted **isometry**.

General guarantee for regression with nuclear norm

Given $y = \mathfrak{X}(\Theta^*) + w$, estimate Θ^* by solving the SDP:

$$\hat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \|y - \mathfrak{X}(\Theta)\|_2^2 + \lambda_n \|\Theta\|_{\text{nuc}} \right\}.$$

General guarantee for regression with nuclear norm

Given $y = \mathfrak{X}(\Theta^*) + w$, estimate Θ^* by solving the SDP:

$$\hat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \|y - \mathfrak{X}(\Theta)\|_2^2 + \lambda_n \|\Theta\|_{\text{nuc}} \right\}.$$

Conditions:

- operator \mathfrak{X} satisfies RSC with curvature $\gamma(\mathfrak{X})$ and tolerance $\kappa(\mathfrak{X})$.
- regularization parameter satisfies $\lambda_n \geq 2\|\mathfrak{X}^*(w)\|_{\text{op}}/n$.

General guarantee for regression with nuclear norm

Given $y = \mathfrak{X}(\Theta^*) + w$, estimate Θ^* by solving the SDP:

$$\hat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \|y - \mathfrak{X}(\Theta)\|_2^2 + \lambda_n \|\Theta\|_{\text{nuc}} \right\}.$$

Conditions:

- operator \mathfrak{X} satisfies RSC with curvature $\gamma(\mathfrak{X})$ and tolerance $\kappa(\mathfrak{X})$.
- regularization parameter satisfies $\lambda_n \geq 2\|\mathfrak{X}^*(w)\|_{\text{op}}/n$.

Theorem (Negahban & W., 2009)

For any matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, any solution $\hat{\Theta}$ to the SDP satisfies the bound

$$\|\hat{\Theta} - \Theta^*\|_F^2 \lesssim \min_{r \in \{1, 2, \dots, \min\{d_1, d_2\}\}} \frac{\bar{\lambda}_n}{\gamma(\mathfrak{X})} \left\{ \underbrace{\frac{\bar{\lambda}_n r}{\gamma(\mathfrak{X})}}_{\text{Estim. error}} + \underbrace{\sum_{j=r+1}^{\min\{d_1, d_2\}} \sigma_j(\Theta^*)}_{\text{Approximation error}} \right\},$$

where $\bar{\lambda}_n = \max\{\lambda_n, \kappa(\mathfrak{X})\}$.

Example: Matrix completion

Random operator $\mathfrak{X} : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^n$ with

$$[\mathfrak{X}(\Theta^*)]_i = \Theta_{a(i)b(i)}^* = \langle\langle E_{a(i)b(i)}, \Theta^* \rangle\rangle,$$

where $(a(i), b(i))$ is a matrix index sampled u.a.r.

Even in noiseless setting, model is **unidentifiable**:

Consider a rank one matrix:

$$\Theta^* = e_1 e_1^T = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

Past work has imposed **eigen-incoherence** conditions.

(Recht & Candes, 2008;

Chandrasekaran et al., 2009 Gross, 2009; Keshavan et al., 2009)

A milder “spikiness” condition

Consider the “poisoned” low-rank matrix:

$$\Theta^* = \Gamma^* + \delta e_1 e_1^T = \Gamma^* + \delta \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

where Γ^* is rank $r - 1$, all eigenvectors perpendicular to e_1 .

Excluded by **eigen-incoherence** for all $\delta > 0$.

A milder “spikiness” condition

Consider the “poisoned” low-rank matrix:

$$\Theta^* = \Gamma^* + \delta e_1 e_1^T = \Gamma^* + \delta \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

where Γ^* is rank $r - 1$, all eigenvectors perpendicular to e_1 .

Excluded by **eigen-incoherence** for all $\delta > 0$.

Control by **spikiness ratio**:

(Negahban & W., 2010)

$$1 \leq \frac{d \|\Theta^*\|_\infty}{\|\Theta\|_F} \leq d.$$

Noisy matrix completion (general ℓ_q -balls)

Suppose that Θ^* lies in the ℓ_q -ball:

$$\mathbb{B}_q(R_q) := \left\{ \Theta \in \mathbb{R}^{d \times d} \mid \sum_{j=1}^d |\sigma_j(\Theta)|^q \leq R_q \right\}.$$

Special case $q = 0$ means Θ^* has rank $r = R_0$.

Corollary (Negahban & W., 2010)

If noise is zero-mean with ν -sub-exponential tails, and Θ^ has spikiness at most α , then*

$$\|\hat{\Theta} - \Theta^*\|_F^2 \lesssim R_q \left((\nu^2 \vee 1) (\alpha)^2 \frac{d \log d}{n} \right)^{1 - \frac{q}{2}}$$

with high probability.

Other work for exactly low rank matrices

In this [special case](#), our result gives:

$$\|\hat{\Theta} - \Theta^*\|_F \lesssim \max\{\nu, \alpha\} \sqrt{\frac{rd \log d}{n}}.$$

Other work for exactly low rank matrices

In this [special case](#), our result gives:

$$\|\hat{\Theta} - \Theta^*\|_F \lesssim \max\{\nu, \alpha\} \sqrt{\frac{rd \log d}{n}}.$$

Candes & Plan, 2009:

- analyzed nuclear norm relaxation
- under **eigen-incoherence conditions** with parameter μ , sufficient for exact recovery
- based on extrapolation from exact recovery:

$$\|\hat{\Theta} - \Theta^*\|_F \lesssim \nu \mu \left\{ \sqrt{d} + \frac{\sqrt{n}}{d} \right\}.$$

- for fixed noise variance ν^2 , diverges as $d \rightarrow +\infty$; also diverges as $n \rightarrow +\infty$ for fixed d

Other work for exactly low rank matrices

In this [special case](#), our result gives:

$$\|\hat{\Theta} - \Theta^*\|_F \lesssim \max\{\nu, \alpha\} \sqrt{\frac{rd \log d}{n}}.$$

Keshavan, Montanari & Oh, 2009:

- analyzed alternative method based on trimmed SVD
- established bound

$$\|\hat{\Theta} - \Theta^*\|_F \lesssim \nu \mu \kappa(\Theta^*) \sqrt{\frac{rd}{n}},$$

- bound grows with **matrix condition number** $\kappa(\Theta^*) = \frac{\sigma_{\max}(\Theta^*)}{\sigma_{\min}(\Theta^*)}$
- eigen-incoherence conditions are imposed

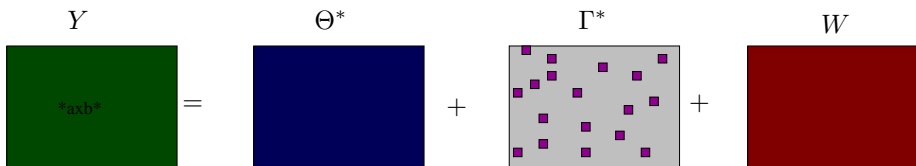
Example: Noisy matrix decomposition

$$Y = \Theta^* + \Gamma^* + W$$

The diagram shows the decomposition of matrix Y into three components: Θ^* , Γ^* , and W . Matrix Y is a dark green square with dimensions $a \times b$. Matrix Θ^* is a dark blue square. Matrix Γ^* is a light gray square with scattered small purple squares representing noise. Matrix W is a dark red square.

$$Y = \Theta^* + \Gamma^* + W$$

Example: Noisy matrix decomposition



$$Y = \Theta^* + \Gamma^* + W$$

Method with two regularizers plus “spikiness” control on Θ :

$$(\hat{\Theta}, \hat{\Gamma}) \in \arg \min_{(\Theta, \Gamma)} \left\{ \frac{1}{2n} \|y - (\Theta + \Gamma)\|_2^2 + \lambda_n \|\Theta\|_{\text{nuc}} + \mu_n \|\Gamma^*\|_1 \right\}.$$

- Noiseless version: Chandrasekaran et al., 2009; Candes et al. 2010; Xu et al., 2010.
- Noisy version: Hu et al., 2010.

Consequences for noisy matrix decomposition

Theorem (Agarwal, Negahban & W., 2011)

With appropriate choice of regularization parameters (λ_n, μ_n) , the squared Frob. error $e^2(\hat{\Theta}, \hat{\Gamma})$ of any SDP solution satisfies

$$e^2 \leq \underbrace{c_1 \nu^2 \left(\frac{r(d_1 + d_2)}{d_1 d_2} \right)}_{\text{Low-rank component}} + \underbrace{c_1 \nu^2 \left(\frac{k \log\left(\frac{d_1 d_2}{k}\right)}{d_1 d_2} \right)}_{\text{Sparse component}} + \underbrace{c_1 \frac{\alpha_d^2 k}{d_1 d_2}}_{\text{Unidentifiable component}}$$

with high probability.

Consequences for noisy matrix decomposition

Theorem (Agarwal, Negahban & W., 2011)

With appropriate choice of regularization parameters (λ_n, μ_n) , the squared Frob. error $e^2(\hat{\Theta}, \hat{\Gamma})$ of any SDP solution satisfies

$$e^2 \leq \underbrace{c_1 \nu^2 \left(\frac{r(d_1 + d_2)}{d_1 d_2} \right)}_{\text{Low-rank component}} + \underbrace{c_1 \nu^2 \left(\frac{k \log\left(\frac{d_1 d_2}{k}\right)}{d_1 d_2} \right)}_{\text{Sparse component}} + \underbrace{c_1 \frac{\alpha_d^2 k}{d_1 d_2}}_{\text{Unidentifiable component}}$$

with high probability.

Intuition:

- effective sample size $n = d_1 d_2$
- **low-rank component** has $\approx r(d_1 + d_2)$ degrees of freedom
- **sparse component** has k non-zeros hidden in $d_1 d_2$, and hence $\approx k \log\left(\frac{d_1 d_2}{k}\right)$ degrees of freedom
- term $\alpha_d^2 \frac{k}{d_1 d_2}$ is unavoidable due to **unidentifiability**

Minimax-optimality

- minimax error over a matrix family:

$$\mathfrak{M}(\mathcal{F}) := \inf_{(\tilde{\Theta}, \tilde{\Gamma})} \sup_{(\Theta^*, \Gamma^*) \in \mathcal{F}} \mathbb{E}[\|\tilde{\Theta} - \Theta^*\|_F^2 + \|\tilde{\Gamma} - \Gamma^*\|_F^2],$$

Minimax-optimality

- minimax error over a matrix family:

$$\mathfrak{M}(\mathcal{F}) := \inf_{(\tilde{\Theta}, \tilde{\Gamma})} \sup_{(\Theta^*, \Gamma^*) \in \mathcal{F}} \mathbb{E}[\|\tilde{\Theta} - \Theta^*\|_F^2 + \|\tilde{\Gamma} - \Gamma^*\|_F^2],$$

- low-rank plus sparse family

$$\mathcal{F}_{\text{sp}} := \left\{ (\Theta^*, \Gamma^*) \mid \text{rank}(\Theta^*) \leq r, |\text{supp}(\Gamma^*)| \leq k, \|\Theta^*\|_\infty \leq \frac{\alpha_d}{\sqrt{d_1 d_2}} \right\}.$$

Minimax-optimality

- minimax error over a matrix family:

$$\mathfrak{M}(\mathcal{F}) := \inf_{(\tilde{\Theta}, \tilde{\Gamma})} \sup_{(\Theta^*, \Gamma^*) \in \mathcal{F}} \mathbb{E}[\|\tilde{\Theta} - \Theta^*\|_F^2 + \|\tilde{\Gamma} - \Gamma^*\|_F^2],$$

- low-rank plus sparse family

$$\mathcal{F}_{\text{sp}} := \left\{ (\Theta^*, \Gamma^*) \mid \text{rank}(\Theta^*) \leq r, |\text{supp}(\Gamma^*)| \leq k, \|\Theta^*\|_\infty \leq \frac{\alpha_d}{\sqrt{d_1 d_2}} \right\}.$$

Theorem (Agarwal, Negahban & W, 2011)

There is a universal constant $c_0 > 0$ such that for all $\alpha_d \geq 32\sqrt{\log(d_1 d_2)}$, we have

$$\mathfrak{M}(\mathcal{F}_{\text{sp}}(r, k, \alpha_d)) \geq c_0 \nu^2 \left\{ \frac{r(d_1 + d_2)}{d_1 d_2} + \frac{k \log\left(\frac{d_1 d_2 - k}{k/2}\right)}{d_1 d_2} \right\} + c_0 \frac{\alpha_d^2 k}{d_1 d_2}.$$

Summary

- high-dimensional matrix problems occur in many settings
 - estimators based on nuclear norm and other convex matrix regularizers are popular
 - a single theoretical result:
 - ▶ provides guarantees for many models
 - ▶ resulting bounds are minimax-optimal (over all algorithms) in many cases
-

Some references:

- S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu (2009). A unified framework for high-dimensional analysis of M -estimators with decomposable regularizers, NIPS Conference.
- S. Negahban and M. J. Wainwright (2009). Estimation rates of (near) low-rank matrices with noise and high-dimensional scaling. arxiv.org/abs/0912.5100. To appear in *Annals of Statistics*.
- S. Negahban and M. J. Wainwright (2010). Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. arxiv.org/abs/0112.5100, September 2010.