Adaptive one-bit matrix completion

Joseph Salmon
Télécom Paristech, Institut Mines-Télécom

Joint work with Jean Lafond (Télécom Paristech)
Olga Klopp (Crest / MODAL'X, Université Paris Ouest)
Éric Moulines (Télécom Paristech)

Motivation : recommender systems

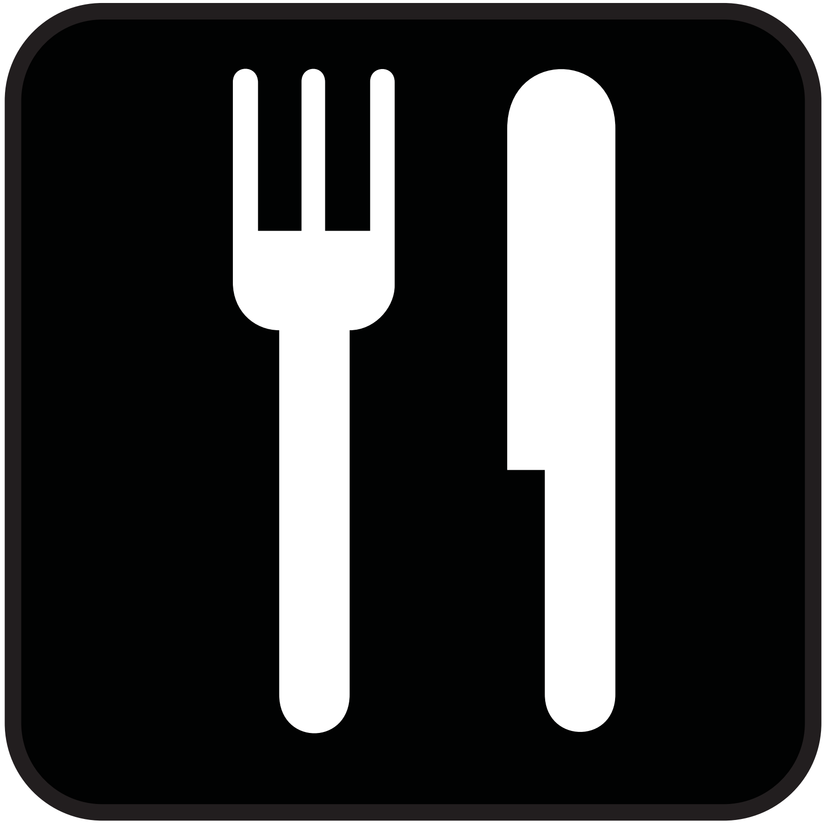
movies (Netflix, Itunes, Allociné)

Motivation : recommender systems

songs (Pandora, Itunes)

Motivation : recommender systems

restaurants (Yelp, la Fourchette)



Motivation : recommender systems

trips, hotels (TripAdvisor, Voyages-SNCF)

Motivation : recommender systems

...

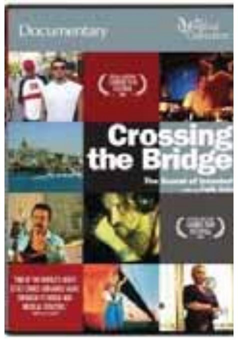
Motivation : recommender systems

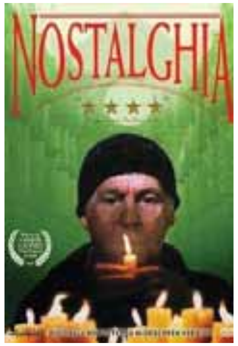
In all those cases matrix completion is **a** crucial ingredient (not the only one) for improving recommender systems Koren *et al.* [2009]

Recommendation systems : movie example

Medhi	★★★★★	★	?	★	★★★★★
Lorne	★	★★★★★	?	★★★	★★★
Raquel	?	?	?	★★★★	★
Maria	?	★★	★★★★	★★★★★	★★★
Robert	★★	★★★★	?	?	?

Recommendation systems : movie example





Medhi	★★★★★	★	★★★★★	★	★★★★★
Lorne	★	★★★★★	★★	★★★	★★★
Raquel	★	★★★★	★★	★★★★	★
Maria	★★★★★	★★	★★★★	★★★★★	★★★
Robert	★★	★★★★	★	★★★★★	★★★★

Other aspect of matrix completion

- Quantum physics

Other aspect of matrix completion

- Quantum physics
- Image/signal processing with missing pixels

Other aspect of matrix completion

- Quantum physics
- Image/signal processing with missing pixels
- Communications

Other aspect of matrix completion

- Quantum physics
- Image/signal processing with missing pixels
- Communications
- Analysis of survey data

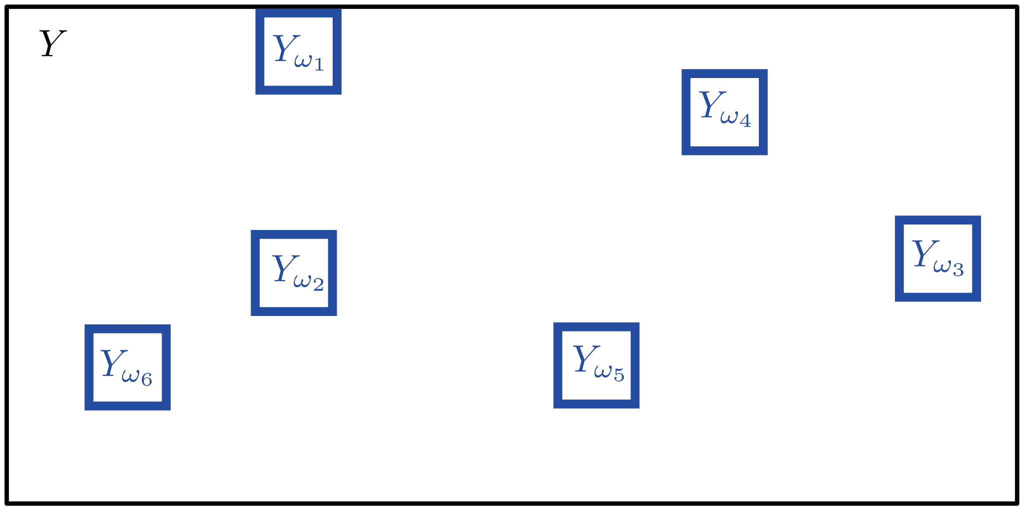
Other aspect of matrix completion

- Quantum physics
- Image/signal processing with missing pixels
- Communications
- Analysis of survey data
- ...

Classical theoretical model : partial observation and Gaussian noise

Observation model

- Matrix of true ratings : $X^* \in \mathbb{R}^{m_1 \times m_2}$ (to recover)
- Indexes observed : $(\omega_i)_{1 \le i \le n} \overset{i.i.d.}{\sim}$ Unif over $[m_1] \times [m_2]$,
- Noisy observations : $Y_{\omega_i} = X^*_{\omega_i} + \sigma \varepsilon_{\omega_i}$ for $1 \le i \le n$
- σ : noise level, ε : centered standard Gaussian random vector



$n = 6$

<u>Rem</u>: potentially $n \ll m_1 m_2$

<u>Rem</u>: randomness sources : 1) index picking 2) degraded answer

Some dataset sizes

Parameter Size	m_1	m_2	n
MovieLens	$70 \cdot 10^3$	$10 \cdot 10^3$	$10 \cdot 10^6$
NetFlix	$2.5 \cdot 10^6$	$17 \cdot 10^3$	$100 \cdot 10^6$
Yahoo	$1 \cdot 10^6$	$600 \cdot 10^3$	$250 \cdot 10^6$

Low rank and matrix factorization

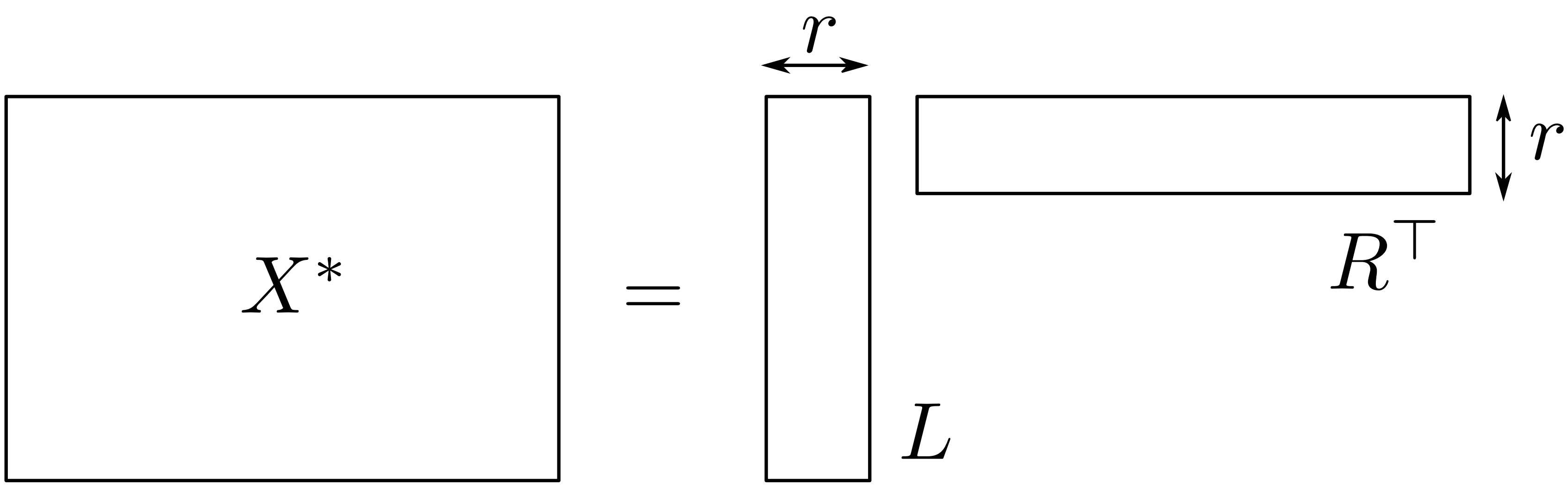
Underlying simplifying assumption : $\boxed{r^* = \mathrm{rank}(X^*)}$ is small

Consequence :

- pass from $m_1 m_2$ to $r^*(m_1 + m_2)$ degrees of freedom

Interpretation :

- a combination of few items can represent all of them
- a combination of few users can represent all of them



$$X^* = L \cdot R^\top$$

Popular estimator

Least square penalized by trace/nuclear norm

$$\hat{X} = \operatorname*{arg\,min}_{X \in \mathbb{R}^{m_1 \times m_2}} \frac{1}{2} \sum_{i=1}^{n} (Y_{\omega_i} - X_{\omega_i})^2 + \lambda \|X\|_{\sigma,1}$$

- $\|X\|_{\sigma,1}$: trace/nuclear norm (ℓ_1 norm of the singular values)
- $\lambda > 0$: regularization parameter controlling data-fitting / low rank trade-off

Rem:

- vector case : $\|\cdot\|_1$ regularization $\Rightarrow$ sparsity (LASSO)
- matrix case : $\|\cdot\|_{\sigma,1}$ regularization $\Rightarrow$ low rank

Previous theoretical work on matrix completion with

- noise-free scenario : Recht, Fazel and Parrilo [2010] Candès and Recht [2009] Candès and Tao [2010]
- additive noise scenario : Candès and Plan [2010] Koltchinskii, Tsybakov and Lounici [2011] Negahban and Wainwrigh [2012] Klopp [2014]

Typical results :

Klopp [2014]

For $\lambda = C\sigma\sqrt{\frac{\log(m_1+m_2)}{\min(m_1,m_2)n}}$, w.h.p.

$$\frac{\|\hat{X} - X^*\|_F^2}{m_1 m_2} \leq c \max(\sigma^2, \|X^*\|_\infty^2) \frac{r^* \max(m_1, m_2) \log(m_1 + m_2)}{n}$$

Rem: can be extended to non uniform sampling provided each coefficient is sampled sufficiently often

Limits of the previous model

- Generally ratings are discrete (0-1, 1-5 stars, etc.)
- In surveys, answers are naturally discrete (yes/no, classes, etc.)
- Variance of the noise model (implicitly) assumed identical for all entries. Cases with picky distribution *e.g.*, movies with agreement (only 5's) / disagreement among the audience (lots of 1's and lots of 5's).

Binary model

Observation model Davenport *et al.* [2012]

- Matrix of true ratings to recover : $X^* \in \mathbb{R}^{m_1 \times m_2}$
- Indexes observed : $(\omega_i)_{1 \le i \le n} \overset{i.i.d.}{\sim}$ Unif over $[m_1] \times [m_2]$,
- Indirect observations :

$$\mathbb{P}(Y_{\omega_i} = 1) = f(X^*_{\omega_i}) \text{ and } \mathbb{P}(Y_{\omega_i} = -1) = 1 - f(X^*_{\omega_i}) ,$$

where f is a link function taking value in $[0, 1]$.

Rem: Uniform sampling only for the sake of simplicity
Rem: To obtain theoretical guarantees $\log(f(\cdot))$ and $\log(1 - f(\cdot))$ need to be concave (*e.g.*, logit, probit)

The estimator

The log-likelihood of the observations $X \to \mathrm{L}(X)$:

$$\mathrm{L}(X) = \sum_{i=1}^{n} \left[\mathbb{1}_{\{Y_{\omega_i}=1\}} \log(f(X_{\omega_i})) + \mathbb{1}_{\{Y_{\omega_i}=-1\}} \log(1 - f(X_{\omega_i})) \right] .$$

Penalized log-likelihood estimator

$$\hat{X} = \underset{X \in \mathbb{R}^{m_1 \times m_2}}{\arg\min}\ \mathrm{F}(X) , \quad \text{where} \quad \mathrm{F}(X) = -\frac{1}{n}\mathrm{L}(X) + \lambda \|X\|_{\sigma,1} ,$$

with $\lambda > 0$ a regularization parameter.

Results

Proposed result

For $\lambda = C\sqrt{\frac{\log(m_1+m_2)}{\min(m_1,m_2)n}}$ w.h.p.

$$\mathrm{KL}\left(f(X^*), f(\hat{X})\right) \leq c^* \frac{r^* \max(m_1, m_2) \log(m_1 + m_2)}{n}$$

where we define the Kullback-Liebler divergence :

$$\mathrm{KL}\left(P, Q\right) := \frac{1}{m_1 m_2} \sum_{\substack{1 \leq i \leq m_1 \\ 1 \leq j \leq m_2}} \left[P_{i,j} \log \frac{P_{i,j}}{Q_{i,j}} + (1 - P_{i,j}) \log \frac{1 - P_{i,j}}{1 - Q_{i,j}} \right].$$

Multinomial Coordinate Lifted Gradient Desc. : Dudik *et al.* [2012]

Data: Observations : Y

ini. param. : $\theta_0 \in \Theta_+$; tolerance : ϵ ; maximum iterations : K

Result: $\theta \in \Theta_+$

Initialization : $\theta \leftarrow \theta_0$, $\mathrm{conv} \leftarrow 0$, $k \leftarrow 0$

while $k \leq K$ ***and*** $\mathrm{conv} = 0$ **do**

  Compute top singular vectors pair of $(-\nabla \mathrm{F}(W_\theta))$: u, v Let

  $g = \lambda + \left\langle \nabla \mathrm{L}, uv^\top \right\rangle$

  if $g \leq -\epsilon/2$ **then**

    $\beta = \arg\min_{b\in\mathbb{R}} \mathrm{F}\left(\theta + (b\delta_{uv^\top})\right)$

    $\theta \leftarrow \theta + \beta\delta_{uv^\top}$; $k \leftarrow k + 1$

  end

  else

    if $g \leq \epsilon$ **then**

      $\mathrm{conv} \leftarrow 1$

    end

    else

      $\theta \leftarrow \arg\min_{\theta' \in \mathbb{R}+\mathbb{K}, \mathrm{supp}(\theta') \subset \mathrm{supp}(\theta)} \mathrm{F}(\theta')$; $k \leftarrow k + 1$

    end

  end

end

Main interests compared to other classical methods

- Does not require full SVD as proximal methods
- Convex formulation which offers strong theoretical guarantees
- Well adapted to sparse structure

Numerical experiments

- Simulate X^* for $m_1 \times m_2 = 100 \times 150,\ 300 \times 450,\ 900 \times 1350$ and $r^* = 5$

Numerical experiments

- Simulate X^* for $m_1 \times m_2 = 100 \times 150,\ 300 \times 450,\ 900 \times 1350$ and $r^* = 5$
- For each X^* simulate with logit distribution n observations, from $n = 10000$ to 500000

Numerical experiments

- Simulate X^* for $m_1 \times m_2 = 100 \times 150,\ 300 \times 450,\ 900 \times 1350$ and $r^* = 5$
- For each X^* simulate with logit distribution n observations, from $n = 10000$ to 500000
- For Gaussian and Binomial estimator, choose λ by cross validation

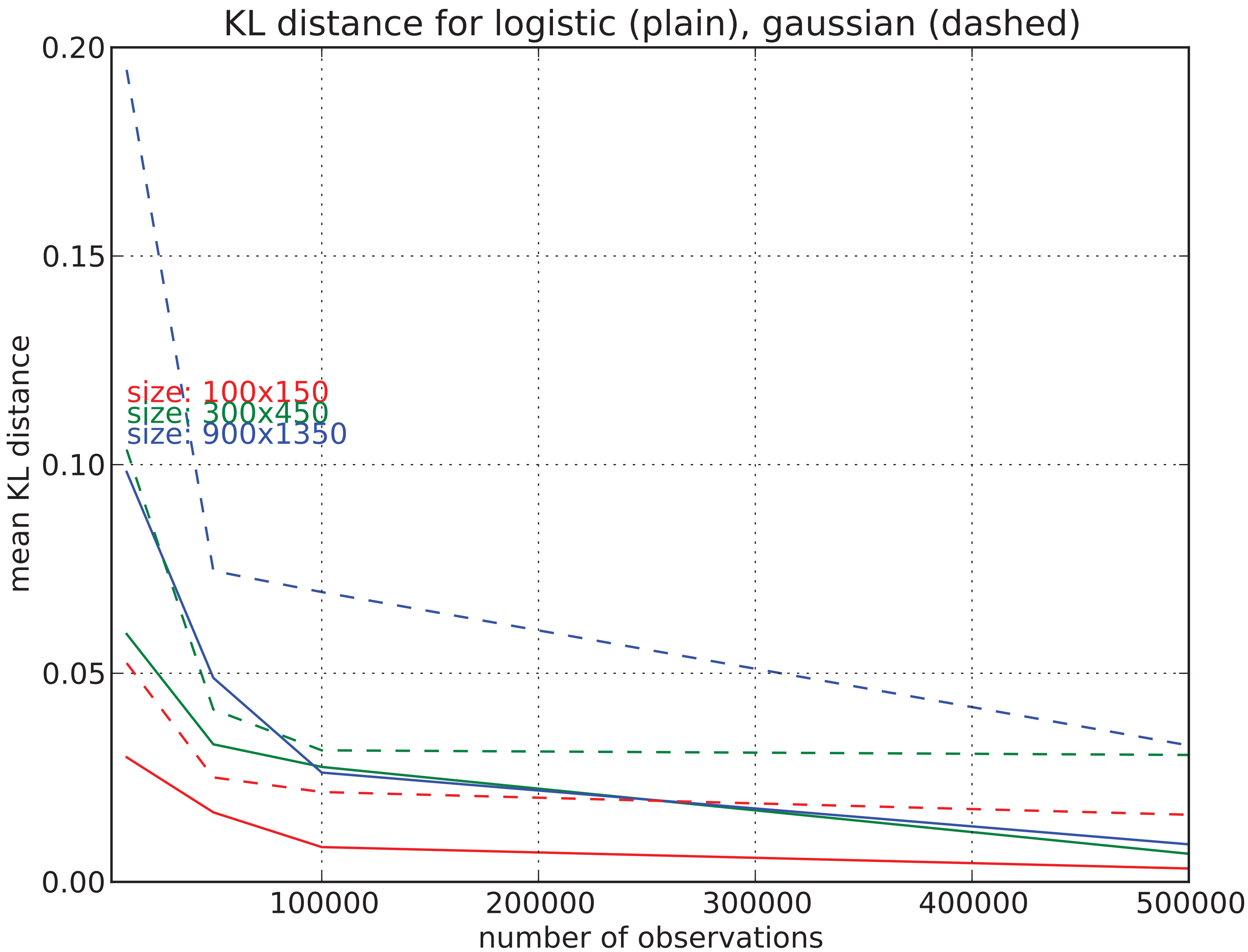
Numerical experiments

- Simulate X^* for $m_1 \times m_2 = 100 \times 150$, 300×450, 900×1350 and $r^* = 5$
- For each X^* simulate with logit distribution n observations, from $n = 10000$ to 500000
- For Gaussian and Binomial estimator, choose λ by cross validation
- For Gaussian and Binomial estimator, estimate X^*

Numerical experiments

- Simulate X^* for $m_1 \times m_2 = 100 \times 150,\ 300 \times 450,\ 900 \times 1350$ and $r^* = 5$
- For each X^* simulate with logit distribution n observations, from $n = 10000$ to 500000
- For Gaussian and Binomial estimator, choose λ by cross validation
- For Gaussian and Binomial estimator, estimate X^*
- For Gaussian and Binomial estimator, compute KL divergence

Illustration over simulation (with cross-validation choice for λ)



Conclusion

- New results for binary / logit matrix completion
- No need to know a bound on the rank or to make a "spikiness" assumption
- Fast algorithm based on Lifted Coordinate Descent Dudik *et al.* [2012]
- Extension to multinomial under some separability

Références I

- E. J. Candès and Y. Plan.
  Matrix completion with noise.
  Proceedings of the IEEE, 98(6) :925–936, 2010.

- E. J. Candès and B. Recht.
  Exact matrix completion via convex optimization.
  Found. Comput. Math., 9(6) :717–772, 2009.

- E. J. Candès and T. Tao.
  The power of convex relaxation : Near-optimal matrix completion.
  IEEE Trans. Inf. Theory, 56(5) :2053–2080, 2010.

- M. Dudík, Z. Harchaoui, and J. Malick.
  Lifted coordinate descent for learning with trace-norm regularization.
  In *AISTATS*, 2012.

- M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters.
  1-bit matrix completion.
  CoRR, abs/1209.3672, 2012.

Références II

- Y. Koren, R. Bell, and C. Volinsky.
  Matrix factorization techniques for recommender systems.
  Computer, 42(8) :30–37, 2009.
- O. Klopp.
  Noisy low-rank matrix completion with general sampling distribution.
  Bernoulli, 2(1) :282–303, 02 2014.
- V. Koltchinskii, A. B. Tsybakov, and K. Lounici.
  Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion.
  Ann. Statist., 39(5) :2302–2329, 2011.
- S. Negahban and M. J. Wainwright.
  Restricted strong convexity and weighted matrix completion : optimal bounds with noise.
  J. Mach. Learn. Res., 13 :1665–1697, 2012.

Références III

- B. Recht, M. Fazel, and P. A. Parrilo.

  Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization.

  SIAM review, 52(3) :471–501, 2010.