Probabilistic Low-Rank Matrix Completion with Adaptive Spectral Regularization Algorithms

Adrien Todeschini

Inria Bordeaux

JdS 2014, Rennes
Aug. 2014

Joint work with François Caron (Univ. Oxford), Marie Chavent (Inria, Univ. Bordeaux)

Disclaimer

- Not a fully Bayesian approach.
- Derivation of an EM algorithm for MAP estimation.
- But builds on a hierarchical prior construction.

Outline

Introduction

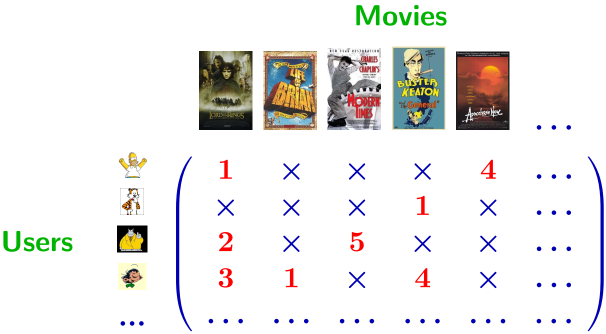
Hierarchical adaptive spectral penalty

EM algorithm for MAP estimation

Experiments

Matrix Completion

- Netflix prize
- 480k users and 18k movies providing 1-5 ratings
- 99% of the ratings are missing
- Objective: predict missing entries in order to make recommendations



Matrix Completion

Objective

Complete a matrix $\boldsymbol{X}$ of size $\boldsymbol{m} \times \boldsymbol{n}$ from a subset of its entries

Applications

- Recommender systems
- Image inpainting
- Imputation of missing data

$$\begin{pmatrix} \square & \times & \times & \times & \square & \dots \\ \times & \times & \times & \square & \times & \dots \\ \square & \times & \square & \times & \times & \dots \\ \square & \square & \times & \square & \times & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix}$$

Matrix Completion

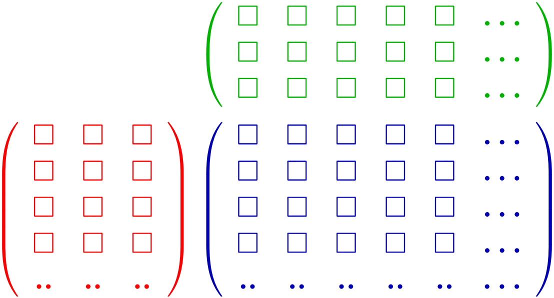
- Potentially large matrices (each dimension of order $\mathbf{10^4 - 10^6}$)
- Very sparsely observed (1%-10%)

Low rank Matrix Completion

- Assume that the complete matrix $\boldsymbol{Z}$ is of low rank

$$\underbrace{\boldsymbol{Z}}_{m\times n} \simeq \underbrace{\boldsymbol{A}}_{m\times k}\underbrace{\boldsymbol{B}^T}_{k\times n}$$

with $k \ll \min(m,n) = r$.

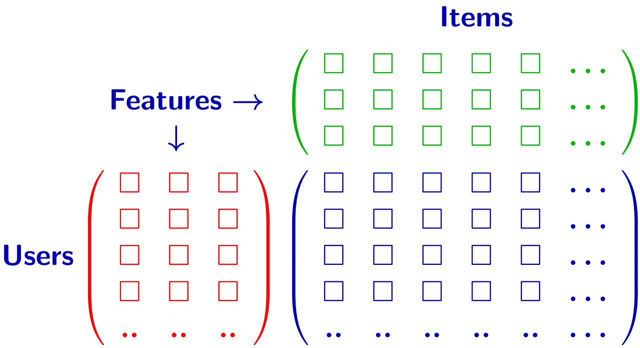


Low rank Matrix Completion

- Assume that the complete matrix Z is of low rank

$$\underbrace{Z}_{m\times n} \simeq \underbrace{A}_{m\times k}\underbrace{B^T}_{k\times n}$$

with $k \ll \min(m, n) = r$.



Low rank Matrix Completion

- Let $\Omega \subset \{1,\ldots,m\} \times \{1,\ldots,n\}$ be the subset of observed entries
- For $(i,j) \in \Omega$

$$X_{ij} = Z_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$$

where $\sigma^2 > 0$

Low rank Matrix Completion

- Optimization problem

$$\underset{Z}{\text{minimize}} \quad \underbrace{\frac{1}{2\sigma^2} \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2}_{-\text{ loglikelihood}} + \underbrace{\lambda \,\text{rank}(Z)}_{\text{penalty}}$$

  where $\lambda > 0$ is some regularization parameter.
- Non-convex
- Computationally hard for general subset Ω

Low rank Matrix Completion

- Matrix completion with nuclear norm penalty

$$\underset{Z}{\text{minimize}} \quad \underbrace{\frac{1}{2\sigma^2} \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2}_{-\text{ loglikelihood}} + \underbrace{\lambda \|Z\|_*}_{\text{penalty}}$$

  where $\|Z\|_*$ is the nuclear norm of Z, or the sum of the singular values of Z.

- Convex relaxation of the rank penalty optimization

[Fazel, 2002, Candès and Recht, 2009, Candès and Tao, 2010]

Low rank Matrix Completion

- Complete matrix $\boldsymbol{X}$
- Nuclear norm objective function

$$\underset{\boldsymbol{Z}}{\text{minimize}} \quad \frac{1}{2\sigma^2}||\boldsymbol{X} - \boldsymbol{Z}||_F^2 + \lambda\, ||\boldsymbol{Z}||_*$$

  where $||\cdot||_F^2$ is the Frobenius norm
- Global solution given by a soft-thresholded SVD

$$\widehat{\boldsymbol{Z}} = \mathbf{S}_{\lambda\sigma^2}(\boldsymbol{X})$$

  where $\mathbf{S}_\lambda(\boldsymbol{X}) = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{D}}_\lambda\widetilde{\boldsymbol{V}}^T$ with
  $\widetilde{\boldsymbol{D}}_\lambda = \text{diag}((\widetilde{d}_1 - \lambda)_+, \ldots, (\widetilde{d}_r - \lambda)_+)$
  and $t_+ = \max(t, 0)$.

[Cai et al., 2010, Mazumder et al., 2010]

Low rank Matrix Completion

Soft-Impute algorithm

- Start with an initial matrix $\boldsymbol{Z^{(0)}}$
- At each iteration $\boldsymbol{t = 1, 2, \ldots}$
  - Replace the missing elements in $\boldsymbol{X}$ with those in $\boldsymbol{Z^{(t-1)}}$
  - Perform a soft-thresholded SVD on the completed matrix, with shrinkage $\boldsymbol{\lambda}$ to obtain the low rank matrix $\boldsymbol{Z^{(t)}}$

[Mazumder et al., 2010]

Low rank Matrix Completion

- Thresholding rule

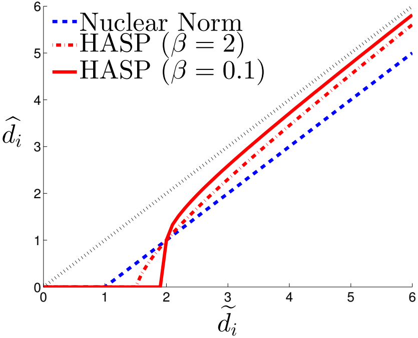


Figure : Thresholding rules on the singular values $\widetilde{\boldsymbol{d}}_{\boldsymbol{i}}$ of $\boldsymbol{X}$

Outline

Nuclear Norm penalty

- Maximum A Posteriori (MAP) estimate

$$\widehat{Z} = \arg\max_{Z} \left[\log p(X|Z) + \log p(Z)\right]$$

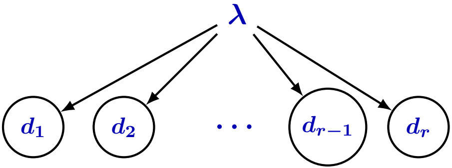
under the prior

$$p(Z) \propto \exp\left(-\lambda \|Z\|_*\right)$$

where $Z = UDV^T$ with $D = \text{diag}(d_1, d_2, \ldots, d_r)$, and

$$U, V \overset{\text{iid}}{\sim} \text{Haar uniform prior on unitary matrices}$$

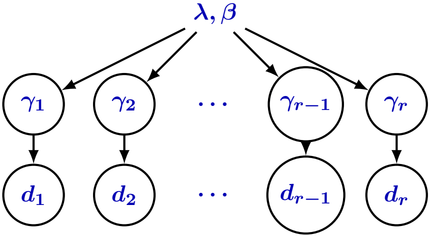
$$d_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$$



Hierarchical adaptive spectral penalty

- Each singular value has its own random shrinkage coefficient
- Hierarchical model, for each singular value $i = 1, \ldots, r$

$$
\begin{aligned}
d_i | \gamma_i &\sim \mathrm{Exp}(\gamma_i) \\
\gamma_i &\sim \mathrm{Gamma}(a, b)
\end{aligned}
$$



- We set $a = \lambda b$ and $b = \beta$

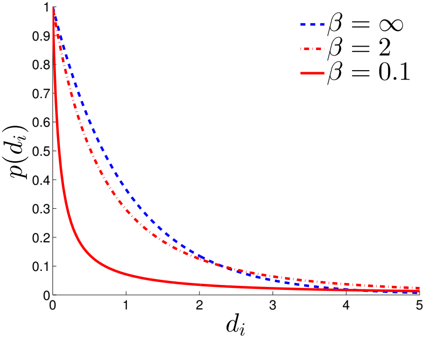
[Todeschini et al., 2013]

Hierarchical adaptive spectral penalty

- Marginal distribution over d_i:

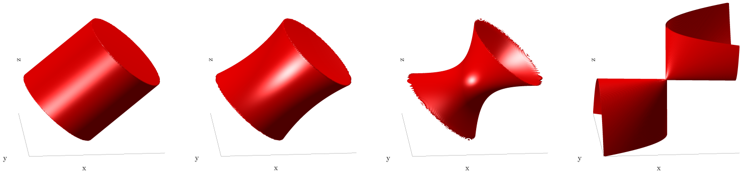
$$p(d_i) = \int_0^\infty \mathrm{Exp}(d_i;\gamma_i)\,\mathrm{Gamma}(\gamma_i;a,b)d\gamma_i = \frac{ab^a}{(d_i+b)^{a+1}}$$

Pareto distribution with heavier tails than exponential distribution

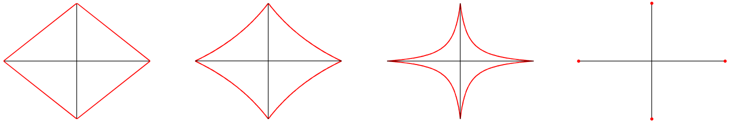


Hierarchical adaptive spectral penalty

$$\boldsymbol{pen(Z) = -\log p(Z) = \sum_{i=1}^{r} (a+1)\log(b+d_i)}$$



(a) Nuclear norm (b) HASP ($\boldsymbol{\beta = 1}$) (c) HASP ($\boldsymbol{\beta = 0.1}$) (d) Rank penalty



(e) $\boldsymbol{\ell_1}$ norm (f) HAL ($\boldsymbol{\beta = 1}$) (g) HAL ($\boldsymbol{\beta = 0.1}$) (h) $\boldsymbol{\ell_0}$ norm

- Admits as special case the nuclear norm penalty $\boldsymbol{\lambda||Z||_*}$ when $\boldsymbol{a = \lambda b}$ and $\boldsymbol{b \to \infty}$.

Outline

EM algorithm for MAP estimation

Expectation Maximization (EM) algorithm to obtain a MAP estimate

$$\widehat{Z} = \arg\max_{Z} \left[\log p(X|Z) + \log p(Z) \right]$$

i.e. to minimize

$$L(Z) = \frac{1}{2\sigma^2} \left\| P_\Omega(X) - P_\Omega(Z) \right\|_F^2 + (a+1) \sum_{i=1}^{r} \log(b + d_i)$$

where

$$P_\Omega(X)(i,j) = \begin{cases} X_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

$$P_\Omega^\perp(X)(i,j) = \begin{cases} 0 & \text{if } (i,j) \in \Omega \\ X_{ij} & \text{otherwise} \end{cases}$$

EM algorithm for MAP estimation

- Latent variables: $\gamma = (\gamma_1, \dots, \gamma_r)$ and $P_\Omega^\perp(X)$
- E step:

$$Q(Z, Z^\star) = \mathbb{E}\left[\log(p(P_\Omega^\perp(X), Z, \gamma)) | Z^\star, P_\Omega(X)\right]$$
$$= C - \frac{1}{2\sigma^2}\|X^\star - Z\|_F^2 - \sum_{i=1}^r \omega_i d_i$$

where $X^\star = P_\Omega(X) + P_{\Omega^\perp}(Z^\star)$ and $\omega_i = \mathbb{E}[\gamma_i | d_i^\star] = \frac{a+1}{b+d_i^\star}$.

EM algorithm for MAP estimation

- M step:

$$\underset{Z}{\text{minimize}} \quad \frac{1}{2\sigma^2} \|X^\star - Z\|_F^2 + \sum_{i=1}^{r} \omega_i d_i \tag{1}$$

(1) is an adaptive spectral penalty regularized optimization problem, with weights $\omega_i = \frac{a+1}{b+d_i^\star}$.

$$d_1^\star \geq d_2^\star \geq \ldots \geq d_r^\star$$

$$\Rightarrow 0 \leq \omega_1 \leq \omega_2 \leq \ldots \leq \omega_r \tag{2}$$

Given condition (2), the solution is given by a weighted soft-thresholded SVD

$$\widehat{Z} = \mathrm{S}_{\sigma^2\omega}(X^\star) \tag{3}$$

where $\mathrm{S}_\omega(X) = \widetilde{U}\widetilde{D}_\omega\widetilde{V}^T$ with
$\widetilde{D}_\omega = \mathrm{diag}((\widetilde{d}_1 - \omega_1)_+, \ldots, (\widetilde{d}_r - \omega_r)_+)$.

[Gaïffas and Lecué, 2011]

EM algorithm for MAP estimation

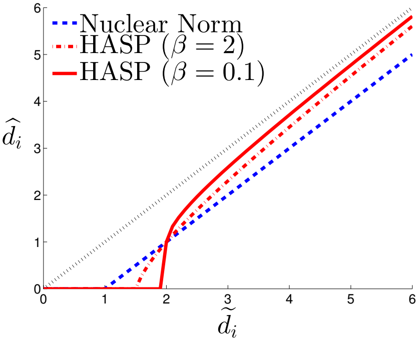


Figure : Thresholding rules on the singular values $\widetilde{d}_i$ of X

The weights will penalize less heavily higher singular values, hence reducing bias.

EM algorithm for MAP estimation

Hierarchical Adaptive Soft Impute (HASI) algorithm for matrix completion

Initialize $Z^{(0)}$ with Soft-Impute. At iteration $t \geq 1$

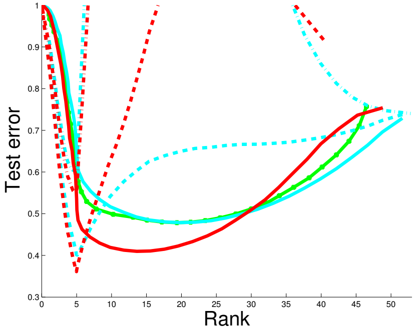
- For $i = 1, \dots, r$, compute the weights $\omega_i^{(t)} = \frac{a+1}{b+d_i^{(t-1)}}$
- Set $Z^{(t)} = \mathrm{S}_{\sigma^2 \omega^{(t)}} \left(P_\Omega(X) + P_\Omega^\perp(Z^{(t-1)}) \right)$

EM algorithm for MAP estimation

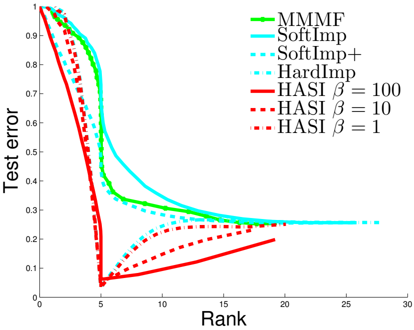
- HASI algorithm admits the Soft-Impute algorithm as a special case when $a = \lambda b$ and $b = \beta \to \infty$. In this case, $\omega_i^{(t)} = \lambda$ for all i.
- When $\beta < \infty$, the algorithm adaptively updates the weights so that to penalize less heavily higher singular values.

Outline

Simulated data



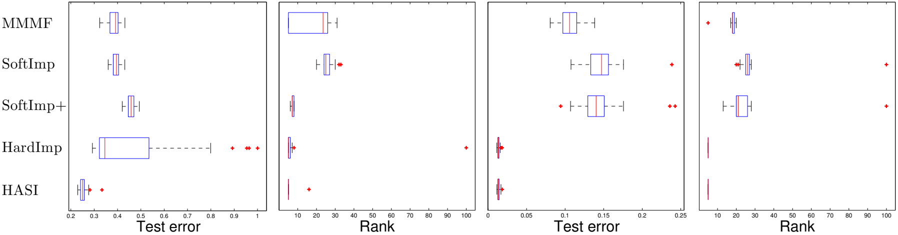
(a) SNR=1; 50% missing; rank=5



(b) SNR=10; 80% missing; rank=5

Simulated data

We then remove 20% of the observed entries as a validation set to estimate the regularization parameters. We use the unobserved entries as a test set.

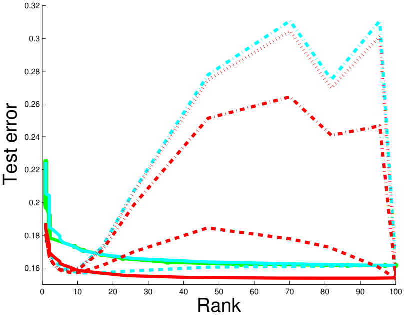


(c) SNR=1; 50% miss. (d) SNR=1; 50% miss. (e) SNR=10; 80% miss. (f) SNR=10; 80% miss.

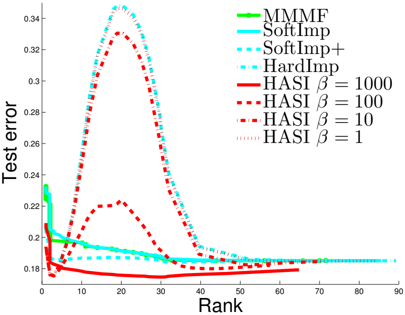
Collaborative filtering examples (Jester)

	Jester 1 **24983×100** **27.5%** miss.		Jester 2 **23500×100** **27.3%** miss.		Jester 3 **24938×100** **75.3%** miss.	
Method	NMAE	Rank	NMAE	Rank	NMAE	Rank
MMMF	0.161	95	0.162	96	0.183	58
Soft Imp	0.161	100	0.162	100	0.184	78
Soft Imp+	0.169	14	0.171	11	0.184	33
Hard Imp	0.158	7	0.159	6	0.181	4
HASI	**0.153**	100	**0.153**	100	**0.174**	30

Collaborative filtering examples (Jester)



(g) Jester 1



(h) Jester 3

Collaborative filtering examples (MovieLens)

	MovieLens 100k **943 × 1682** **93.7%** miss.		MovieLens 1M **6040 × 3952** **95.8%** miss.	
Method	NMAE	Rank	NMAE	Rank
MMMF	0.195	50	**0.169**	30
Soft Imp	0.197	156	0.176	30
Soft Imp+	0.197	108	0.189	30
Hard Imp	0.190	7	0.175	8
HASI	**0.187**	35	0.172	27

Conclusion and perspectives

- Conclusion:
  - Good results compared to several alternative low rank matrix completion methods.
  - Bridge between nuclear norm and rank regularization algorithms.
  - Can be extended to binary matrices
  - Non-convex optimization, but experiments show that initializing the algorithm with the Soft-Impute algorithm provides very satisfactory results.
  - Matlab code available online
- Perspectives:
  - Fully Bayesian approach
  - Tensor factorization
  - Online EM

Bibliography I

Cai, J., Candès, E., and Shen, Z. (2010).
A singular value thresholding algorithm for matrix completion.
SIAM Journal on Optimization, 20(4):1956–1982.

Candès, E. and Recht, B. (2009).
Exact matrix completion via convex optimization.
Foundations of Computational mathematics, 9(6):717–772.

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

Fazel, M. (2002).
Matrix rank minimization with applications.
PhD thesis, Stanford University.

Gaïffas, S. and Lecué, G. (2011).
Weighted algorithms for compressed sensing and matrix completion.
arXiv preprint arXiv:1107.1638.

Larsen, R. M. (2004).
Propack-software for large and sparse svd calculations.
Available online. URL http://sun. stanford. edu/rmunk/PROPACK.

Bibliography II

Mazumder, R., Hastie, T., and Tibshirani, R. (2010).
Spectral regularization algorithms for learning large incomplete matrices.
The Journal of Machine Learning Research, 11:2287–2322.

Todeschini, A., Caron, F., and Chavent, M. (2013).
Probabilistic low-rank matrix completion with adaptive spectral regularization algorithms.
In *Advances in Neural Information Processing Systems*, pages 845–853.

Thank you