Complétion de matrice de rang faible probabiliste à l'aide d'algorithmes de régularisation spectrale adaptatifs

Adrien Todeschini [1], François Caron [2] & Marie Chavent [3]

[1] *INRIA - IMB - Univ. Bordeaux, 33405 Talence*
Adrien.Todeschini@inria.fr
[2] *Univ. Oxford, Dept. of Statistics, Oxford, OX1 3TG, UK*
Caron@stats.ox.ac.uk
[3] *Univ. Bordeaux - IMB - INRIA, 33000 Bordeaux*
Marie.Chavent@u-bordeaux2.fr

Résumé. Nous proposons une nouvelle classe d'algorithmes pour la complétion de matrice de rang faible. Notre approche s'appuie sur de nouvelles fonctions de pénalité sur les valeurs singulières de la matrice de rang faible. En exploitant une représentation basée sur un modèle de mélange de cette pénalité, nous montrons qu'un ensemble de variables latentes convenablement choisi permet de dériver un algorithme EM pour obtenir une estimation du Maximum A Posteriori de la matrice de rang faible complétée. L'algorithme résultant est un algorithme à seuillage doux itératif qui adapte de manière itérative les coefficients de réduction associés aux valeurs singulières. [1]

Mots-clés. complétion de matrice, factorisation de matrice, modélisation de rang faible, algorithme EM, modèle hiérarchique bayésien, SVD, filtrage collaboratif

Abstract. We propose a novel class of algorithms for low rank matrix completion. Our approach builds on novel penalty functions on the singular values of the low rank matrix. By exploiting a mixture model representation of this penalty, we show that a suitably chosen set of latent variables enables to derive an EM algorithm to obtain a Maximum A Posteriori estimate of the completed low rank matrix. The resulting algorithm is an iterative soft-thresholded algorithm which iteratively adapts the shrinkage coefficients associated to the singular values.

Keywords. matrix completion, matrix factorization, low-rank modeling, EM algorithm, Bayesian hierarchical model, SVD, collaborative filtering

1 Introduction

La complétion de matrice a attiré beaucoup d'attention au cours des dernières années. L'objectif est de compléter une matrice de dimension potentiellement importante basée

1. Cet article est une version courte de l'article Todeschini *et al.* (2013) publié dans la conférence internationale NIPS.

sur un petit (et potentiellement bruité) sous-ensemble de ses entrées (Srebro *et al.* (2005); Candès et Plan (2010)). Une application populaire consiste à construire des systèmes de recommandation automatiques, où les lignes correspondent à des utilisateurs, les colonnes à des items et les entrées peuvent être des notes. L'objectif est alors de prévoir les préférences d'utilisateur à partir d'un sous-ensemble des entrées.

Dans de nombreux cas, il est raisonnable de supposer que la matrice $m \times n$ inconnue Z peut être approchée par une matrice de rang faible $Z \simeq AB^T$ où A et B sont respectivement de taille $m \times k$ et $n \times k$, avec $k \ll \min(m, n)$. Dans l'application aux systèmes de recommandation, l'hypothèse de rang faible est judicieuse car il est communément admis que seul un petit nombre de facteurs contribue aux préférences de l'utilisateur. La structure de rang faible implique donc une sorte de collaboration entre les différents utilisateurs/items.

On observe généralement une version bruitée X_{ij} de certaines entrées $(i, j) \in \Omega$ où $\Omega \subset \{1, \ldots, m\} \times \{1, \ldots, n\}$. Pour $(i, j) \in \Omega$

$$X_{ij} = Z_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

La complétion de matrice de rang faible peut être effectuée par la résolution du problème d'optimisation suivant

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

où $\lambda > 0$ est un paramètre de régularisation. Pour un sous-ensemble quelconque Ω, le problème d'optimisation (2) est très complexe et beaucoup d'auteurs ont préconisé l'utilisation d'une relaxation convexe de (2) (Fazel (2002); Candès et Plan (2010); Mazumder *et al.* (2010)), donnant le problème d'optimisation convexe suivant

$$\underset{Z}{\text{minimize}} \quad \frac{1}{2\sigma^2} \sum_{(i,j)\in\Omega} (X_{ij} - Z_{ij})^2 + \lambda \, \|Z\|_* \tag{3}$$

où $\|Z\|_*$ est la norme nucléaire de Z, soit la somme des valeurs singulières de Z. Mazumder *et al.* (2010) ont proposé un algorithme itératif, appelé Soft-Impute, pour résoudre le problème de minimisation régularisé par la norme nucléaire (3).

Nous montrons ici que la solution de la fonction objectif (3) peut être interprété comme l'estimateur Maximum A Posteriori (MAP) lorsque l'on suppose que les valeurs singulières de Z sont *i.i.d* selon une distribution exponentielle de taux λ. En utilisant cette interprétation bayésienne, nous proposons des pénalités concaves alternatives à la norme nucléaire, obtenues en considérant que les valeurs singulières sont issues d'un mélange de distributions exponentielles. Nous montrons que cette classe de pénalités comble le vide entre la norme nucléaire et la pénalité de rang, et qu'un algorithme Espérance-Maximisation (EM) simple peut être dérivé afin d'obtenir des estimations du MAP. L'algorithme résultant adapte de manière itérative les coefficients de réduction associés aux valeurs singulières. Il peut être considéré comme l'équivalent pour les matrices des algorithmes de repondération ℓ_1 (Candès *et al.* (2008); Lee *et al.* (2010)) pour la régression linéaire multivariée.

Nous montrons également que l'algorithme Soft-Impute de Mazumder *et al.* (2010) est obtenu comme un cas particulier.

Enfin, notre exposé fournira des preuves empiriques de l'intérêt de l'approche proposée sur des données réelles et simulées.

2 Pénalité spectrale adaptative hiérarchique

La solution $\widehat{Z}$ au problème d'optimisation (3) peut être interprétée comme le Maximum A Posteriori sous la vraisemblance (1) et l'*a priori* $p(Z) \propto \exp\left(-\lambda \left\|Z\right\|_*\right)$.

En supposant $Z = UDV^T$, avec $D = \mathrm{diag}(d_1, d_2, \ldots, d_r)$ on peut encore décomposer cet *a priori* en

$$p(Z) = p(U)p(V)p(D)$$

où on suppose la distribution *a priori* uniforme de Haar sur les matrices unitaires U et V, et l'*a priori* exponentiel sur les valeurs singulières d_i, d'où

$$p(d_1, \ldots, d_r) = \prod_{i=1}^{r} \mathrm{Exp}(d_i; \lambda) \tag{4}$$

où $\mathrm{Exp}(x; \lambda) = \lambda \exp(-\lambda x)$ est la densité de probabilité de la distribution exponentielle de paramètre λ évaluée en x. La distribution exponentielle a son mode en zéro, ce qui favorise ainsi des solutions parcimonieuses.

Nous proposons ici des pénalités/distributions *a priori* alternatives, qui comblent le vide entre les pénalités de rang et de norme nucléaire. Nos pénalités sont basées sur une construction hiérarchique bayésienne. Les problèmes d'optimisation pour obtenir le MAP associé peuvent être résolus par un algorithme EM.

On considère l'*a priori* hiérarchique suivant sur la matrice de rang faible Z. On suppose toujours que $Z = UDV^T$, où les matrices unitaires U et V suivent un *a priori* uniforme et $D = \mathrm{diag}(d_1, \ldots, d_r)$. On suppose maintenant que chaque valeur singulière d_i a son propre paramètre de régularisation γ_i.

$$p(d_1, \ldots, d_r | \gamma_1, \ldots \gamma_r) = \prod_{i=1}^{r} p(d_i|\gamma_i) = \prod_{i=1}^{r} \mathrm{Exp}(d_i; \gamma_i)$$

On suppose également que les paramètres de régularisation sont eux-mêmes *i.i.d* selon distribution gamma

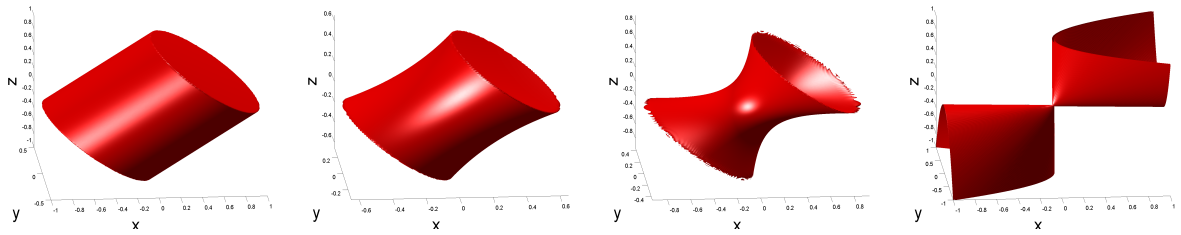
$$p(\gamma_1, \ldots, \gamma_r) = \prod_{i=1}^{r} p(\gamma_i) = \prod_{i=1}^{r} \mathrm{Gamma}(\gamma_i; \lambda\beta, \beta)$$

où $\mathrm{Gamma}(x; a, b)$ est la densité de probabilité de la distribution gamma de paramètres a et b évaluée en x.

La distribution marginale de d_i est donc un mélange continu de distributions exponentielles

$$p(d_i) = \int_0^\infty \mathrm{Exp}(d_i; \gamma_i)\, \mathrm{Gamma}(\gamma_i; \lambda\beta, \beta) d\gamma_i = \frac{\lambda\beta^{\lambda\beta+1}}{(d_i + \beta)^{\lambda\beta+1}} \tag{5}$$

Il s'agit d'une distribution de Pareto qui a des queues plus lourdes que la distribution exponentielle. Plus β est faible, plus les queues de la distribution sont lourdes. Lorsque $\beta \to \infty$, on retrouve la distribution exponentielle de taux λ.



(a) Norme nucléaire (b) HASP ($\beta = 1$) (c) HASP ($\beta = 0.1$) (d) Pénalité de rang

FIGURE 1 – Variétés de pénalité constante pour une matrice symétrique 2×2, $Z = [x, y; y, z]$ pour (a) la norme nucléaire, (b-c) HASP avec $\lambda = 1$ (b) $\beta = 1$ et (c) $\beta = 0.1$, et (d) la pénalité de rang.

Soit

$$pen(Z) = -\log p(Z) = -\sum_{i=1}^{r} \log(p(d_i)) = C_1 + \sum_{i=1}^{r} (\lambda\beta + 1) \log(\beta + d_i) \tag{6}$$

la pénalité induite par l'*a priori* $p(Z)$. Nous appelons la pénalité (6) *Hierarchical Adaptive Spectral Penalty* (HASP). Sur la figure 1 sont représentées les boules de pénalité constante pour une matrice symétrique 2×2 pour la norme nucléaire, HASP et la pénalité de rang.

La pénalité (6) admet comme cas particuliers la pénalité de norme nucléaire $\lambda||Z||_*$ lorsque $\beta \to \infty$.

3 Algorithme EM pour l'estimation du MAP

Soit l'opérateur $P_\Omega(X)$ et son complémentaire $P_\Omega^\perp(X)$

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

En utilisant la représentation (5) basée sur un mélange d'exponentielles, nous montrons maintenant comment dériver un algorithme EM pour obtenir une estimation du MAP

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

ce qui équivaut à minimiser la fonction objectif

$$L(Z) = \frac{1}{2\sigma^2} \left\| P_\Omega(X) - P_\Omega(Z) \right\|_F^2 + (\lambda\beta + 1) \sum_{i=1}^{r} \log(b + d_i) \tag{7}$$

Nous allons dériver l'algorithme EM en utilisant les variables latentes γ et $P_\Omega^\perp(X)$.

L'étape E est donnée par

$$\begin{aligned} Q(Z, Z^*) &= \mathbb{E}\left[\log(p(P_\Omega(X), P_\Omega^\perp(X), Z, \gamma))|Z^*, P_\Omega(X)\right] \\ &= C_2 - \frac{1}{2\sigma^2}\left\{\left\|P_\Omega(X) + P_\Omega^\perp(Z^*) - Z\right\|_F^2\right\} - \sum_{i=1}^r \mathbb{E}[\gamma_i|d_i^*]d_i \end{aligned}$$

Par conséquent, à chaque itération de l'algorithme EM, l'étape de M consiste à résoudre le problème d'optimisation

$$\underset{Z}{\text{minimize}} \quad \frac{1}{2\sigma^2}\|X^* - Z\|_F^2 + \sum_{i=1}^r \omega_i d_i \tag{8}$$

où $\omega_i = \mathbb{E}[\gamma_i|d_i^*] = \frac{\partial}{\partial d_i^*}\left[-\log p(d_i^*)\right] = \frac{\lambda\beta+1}{\beta+d_i^*}$ et $X^* = P_\Omega(X) + P_\Omega^\perp(Z^*)$ est la matrice observée, complétée par les entrées de Z^*.

On a maintenant un problème de matrice complète.

Cas particulier Lorsque $\beta \to \infty$, on obtient à chaque itération $\omega_i^{(t)} = \lambda$ pour tout i. Par conséquent, (8) est un problème d'optimisation de matrice complète régularisé par la norme nucléaire dont la solution est une décomposition en valeur singulière (SVD) à seuillage doux de X^* (Cai *et al.* (2010); Mazumder *et al.* (2010)), i.e.

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

où $\mathbf{S}_\lambda(X^*) = \widetilde{U}\widetilde{D}_\lambda\widetilde{V}^T$ avec $\widetilde{D}_\lambda = \text{diag}((\widetilde{d}_1 - \lambda)_+, \ldots, (\widetilde{d}_r - \lambda)_+)$ et $t_+ = \max(t, 0)$. $X^* = \widetilde{U}\widetilde{D}\widetilde{V}^T$ est la SVD de X^* avec $\widetilde{D} = \text{diag}(\widetilde{d}_1, \ldots, \widetilde{d}_r)$ et $\widetilde{d}_1 \geq \widetilde{d}_2 \ldots \geq \widetilde{d}_r$.

Cas général Lorsque $\beta < \infty$, (8) est un problème d'optimisation régularisé par une norme nucléaire pondérée adaptative, avec des poids ω_i.

Sans perte de généralité, on suppose que $d_1^* \geq d_2^* \geq \ldots \geq d_r^*$, ce qui implique

$$0 \leq \omega_1 \leq \omega_2 \leq \ldots \leq \omega_r \tag{9}$$

Les poids ci-dessus pénalisent donc moins lourdement les plus grandes valeurs singulières, réduisant ainsi le biais.

Tel que l'ont montré Gaïffas et Lecué (2011), une solution globale à (8) sous la contrainte d'ordre (9) est donnée par une SVD à seuillage doux pondéré

$$Z^{(t)} = \mathbf{S}_{\sigma^2\omega}(X^*) \tag{10}$$

où $\mathbf{S}_\omega(X^*) = \widetilde{U}\widetilde{D}_\omega\widetilde{V}^T$ avec $\widetilde{D}_\omega = \text{diag}((\widetilde{d}_1 - \omega_1)_+, \ldots, (\widetilde{d}_r - \omega_r)_+)$.

Algorithm 1 *Hierarchical Adaptive Soft Impute* (HASI)

Initialiser $Z^{(0)}$. A l'itération $t \geq 1$

- Pour $i = 1, \ldots, r$, calculer les poids $\omega_i^{(t)} = \frac{\lambda\beta+1}{\beta+d_i^{(t-1)}}$
- Définir $Z^{(t)} = \mathbf{S}_{\sigma^2\omega^{(t)}}\left(P_\Omega(X) + P_\Omega^\perp(Z^{(t-1)})\right)$
- Si $\frac{L(Z^{(t-1)}) - L(Z^{(t)})}{L(Z^{(t-1)})} < \varepsilon$ alors retourner $\widehat{Z} = Z^{(t)}$

L'algorithme 1 résume la procédure *Hierarchical Adaptive Soft Impute* (HASI) qui converge vers un minimum local de la fonction objectif (7).

Il admet l'algorithme Soft-Impute de Mazumder *et al.* (2010) comme cas particulier lorsque $\beta \to \infty$. Au contraire, lorsque $\beta < \infty$, notre algorithme met à jour de manière adaptative les poids afin de pénaliser moins lourdement les valeurs singulières élevées.

Bien que le problème d'optimisation ne soit pas convexe, nos expériences montrent qu'une initialisation avec l'algorithme Soft-Impute de Mazumder *et al.* (2010) fournit des résultats très satisfaisants.

L'exposé montrera des comparaisons numériques entre notre approche et de récentes alternatives, montrant l'intérêt de l'approche proposée pour la complétion de matrice de rang faible.

Bibliographie

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

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

Candès, E., Wakin, M. et Boyd, S. (2008). Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905.

Fazel, M. (2002). *Matrix rank minimization with applications*. Thèse de doctorat, Stanford University.

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

Lee, A., Caron, F., Doucet, A. et Holmes, C. (2010). A hierarchical Bayesian framework for constructing sparsity-inducing priors. *arXiv preprint arXiv :1009.1914*.

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

Srebro, N., Rennie, J. et Jaakkola, T. (2005). Maximum-Margin Matrix Factorization. *Dans Advances in neural information processing systems*, volume 17, pages 1329–1336. MIT Press.

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