Analyse en composantes principales globalement parcimonieuse

Pierre-Alexandre Mattei [1], Charles Bouveyron [1] & Pierre Latouche [2]

[1] *Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes,*
pierre-alexandre.mattei@parisdescartes.com,
charles.bouveyron@parisdescartes.fr
[2] *Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne*
pierre.latouche@univ-paris1.fr

Résumé. Avec l'essor actuel des données de grande dimension, les versions parcimonieuses de l'analyse en composantes principales (ACP) se sont imposées comme de simples mais efficaces manières de sélectionner des variables sans supervision. Cependant, lorsqu'il s'agit de trouver plusieurs axes principaux parcimonieux, ces approches sont délicates à interpréter car chaque axe a son propre support. Afin de pallier cette défaillance, nous proposons une approche bayésienne qui permet d'obtenir plusieurs composantes principales ayant le même support. Pour le praticien, il est ainsi possible de déterminer quelles variables permettent le mieux de résumer ses données. Pour ce faire, nous utilisons un modèle d'ACP probabiliste sans bruit associé à un *a priori* gaussien. Dans ce cadre, nous obtenons pour la première fois une expression explicite de la vraisemblance marginale. Afin d'obtenir une méthode compatible avec des données de grande dimension, nous proposons également une simple relaxation de ce modèle. Comme le support est identique pour tous les axes trouvés, nous appelons notre approche *analyse en composantes principales probabiliste globalement parcimonieuse* (GSPPCA).

Mots-clés. Régression linéaire, parcimonie, sélection bayésienne de variables, algorithme EM, données ouvertes

Abstract. With the flourishing development of high-dimensional data, sparse versions of principal component analysis (PCA) have imposed themselves as simple, yet powerful ways of selecting relevant features in an unsupervised manner. However, when several sparse principal components are computed, the interpretation of the selected variables may be difficult since each axis has its own sparsity pattern and has to be interpreted separately. To overcome this drawback, we propose a Bayesian procedure that allows to obtain several sparse components with the same sparsity pattern. This allows the practitioner to identify the original variables which are relevant to describe the data. To this end, using Roweis' probabilistic interpretation of PCA and an isotropic Gaussian prior on the loading matrix, we provide the first exact computation of the marginal likelihood of a Bayesian PCA model. In order to avoid the drawbacks of discrete model selection, we propose a simple relaxation of our framework. Since the sparsity pattern is common to all components, we call this approach globally sparse probabilistic PCA (GSPPCA).

Keywords. PCA, sparsity, Bayesian variable selection, VEM algorithm, high-dimensional data.

1 Introduction

Des premières études de résultats d'examens scolaires par Hotelling (1933) jusqu'à l'exploration des données de micromatrices d'ADN (Ringnér, 2008), l'analyse en composantes principales (ACP) s'est imposée comme un outil statistique de base pour la réduction de dimension de données. Cependant, dans le cadre de données de grande dimension, l'ACP présente deux principaux inconvénients :

— d'une part, les axes principaux empiriques deviennent inconsistants lorsque le nombre de variables est plus grand que le nombre d'observations (Johnstone et Lu, 2009)

— d'autre part, les composantes principales sont combinaisons linéaires de *toutes* les variables de départ, et ne permettent ainsi pas d'effectuer une sélection de variables bien souvent souhaitable dans ce cadre.

Afin de remédier à ces défauts, plusieurs approches parcimonieuses de l'ACP ont été proposées (voir par exemple Jolliffe *et al.* (2003), Zou *et al.* (2006), Archambeau et Bach (2009) ou Zhang *et al.* (2012)). Ces méthodes permettent effectivement d'obtenir des axes creux, mais qui n'ont pas forcément le même support. Par conséquent, elles ne permettent pas d'effectuer de réelle sélection de variable. Nous proposons ici d'utiliser la sélection de modèles Bayésienne afin d'obtenir des axes principaux ayant le même support, et donc de déterminer quelles variables sont pertinentes dans un jeu de données. Nous appelons cette procédure *analyse en composantes principales globalement paricmonieuse* par opposition aux approches *locales*, c'est à dire qui autorisent un support différent pour chaque axe pricnipal. Pour ce faire, nous utilisons le modèles d'ACP probabiliste de Roweis (1998) pour lequel nous proposons, pour la premiere fois, une expression explicite de la vraisemblance marginale. Le modèle finalement choisi est celui qui maximise cette vraisemblance. Pour plus de détails, le lecteur est renvoyé à la publication de Mattei *et al.* (2016).

2 Sélection de variables pour l'ACP probabiliste

Par la suite, on suppose donné un échantillon i.i.d. $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$ que l'on souhaite projeter sur un espace de dimension plus faible d. Les observations sont stockées dans la matrice $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$.

2.1 Le modèle

Le modèle d'ACP probabiliste (ACPP) de Tipping et Bishop (1999) s'écrit

$$\mathbf{x} = \mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon} \tag{1}$$

où $\mathbf{y} \sim \mathcal{N}(0, \mathbf{I}_d)$ est un vecteur gaussien latent de dimension faible, $\mathbf{W}$ est une matrice de paramètres de taille $p \times d$ et $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$ est un bruit gaussien. Tipping et Bishop (1999) prouvèrent que l'estimateur du maximum de vraisemblance $\mathbf{W}_{\text{ML}}$ de $\mathbf{W}$ permet de retrouver les axes principaux de la matrice de données.

Afin de procéder à une analyse parcimonieuse bayésienne, nous proposons d'utiliser des *a priori* gaussiens indépendants $w_{ij} \sim \mathcal{N}(0, 1/\alpha^2)$ ainsi qu'un vecteur binaire $\mathbf{v} \in \{0, 1\}^p$ séparant les variables pertinentes de celles peu utiles. Le modèle devient ainsi

$$\mathbf{x} = \mathbf{V}\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon} \tag{2}$$

où $\mathbf{V} = \text{diag}(\mathbf{v})$.

De nombreux modèles bayésiens similaires ont été introduits, sans jamais que les vraisemblances marginales soient calculées exactement (Bishop, 1999; Archambeau et Bach, 2009; Lázaro-Gredilla et Titsias, 2011). Ici, nous proposons de nous placer dans le cadre particulier d'ACPP introduit par Roweis (1998) qui remarqua que le modèle d'ACPP permet de retrouver les axes principaux même dans le cas limite $\sigma \to 0$. Pour ce faire, nous considérons la modification suivant du modèle d'ACPP

$$\mathbf{x} = \mathbf{V}\mathbf{W}\mathbf{y} + \bar{\mathbf{V}}\boldsymbol{\varepsilon_1} + \mathbf{V}\boldsymbol{\varepsilon_2} \tag{3}$$

où $\boldsymbol{\varepsilon_1} \sim \mathcal{N}(0, \sigma_1^2 \mathbf{I}_p)$ est le bruit des variables inactives et $\boldsymbol{\varepsilon_2} \sim \mathcal{N}(0, \sigma_2^2 \mathbf{I}_p)$ est le bruit des variables actives, tout en gardant à l'esprit que nous souhaitons nous placer dans le cas $\sigma_2 \to 0$. Comme prouvé dans Mattei *et al.* (2016), dans ce cas précis, la vraisemblance marginale peut être calculée explicitement.

Theorème 1. *Lorsque $\sigma_2 \to 0$, $\mathbf{x}$ converge en probabilité vers une variable aléatoire $\tilde{\mathbf{x}}$ dont la densité est*

$$p(\tilde{\mathbf{x}}|\mathbf{v}, \alpha, \sigma_1) = C e^{\frac{-||\bar{\mathbf{v}} \odot \tilde{\mathbf{x}}||_2^2}{2\sigma_1^2}} ||\mathbf{v} \odot \tilde{\mathbf{x}}||_2^{\frac{d-q}{2}} K_{\frac{q-d}{2}} \left(\alpha ||\mathbf{v} \odot \tilde{\mathbf{x}}||_2\right), \tag{4}$$

où

$$C = \frac{\alpha^{\frac{q+d}{2}} (2\pi)^{-q/2} 2^{1-d/2}}{\Gamma\left(\frac{d}{2}\right)}.$$

Ce théorème permet donc de calculer efficacement la vraisemblance marginale

$$\log p(\tilde{\mathbf{X}}|\mathbf{v}, \alpha, \sigma_1) = \sum_{i=1}^{n} \log p(\tilde{\mathbf{x}}_i|\mathbf{v}, \alpha, \sigma_1)$$

qui doit être maximisée dans le cadre de la sélection de variables bayésienne (Kass et Raftery, 1995).

TABLE 1 – F-score pour 50 répétitions

	$n = p/2$	$n = p$	$n = 2p$
SSPCA-CV	0.944 ± 0.061	0.985 ± 0.022	1 ± 0
GSPPCA	0.97 ± 0.071	0.985 ± 0.34	1 ± 0
SPCA	0.771 ± 0.11	0.972 ± 0.026	0.95 ± 0.028

2.2 Inférence en grande dimension à l'aide d'une relaxation simple

En dépit du résultat précédent, la maximisation de la vraisemblance marginale demeure délicate en raison de discrétion de $\mathbf{v}$ qui peut prendre 2^p valeurs potentielles. Similairement à Latouche *et al.* (2015), nous considérons donc la relaxation simple suivante : le vecteur $\mathbf{v}$ est remplacé par un vecteur continu $\mathbf{u} \in [0, 1]^p$. Dans le cadre de ce modèle relâché, un algorithme de type *variational expectation-maximization* (VEM) permet de maximiser rapidement une borne inférieure de la vraisemblance marginale. Pour plus de détails sur la mise en place de cet algorithme, le lecteur est renvoyé à Mattei *et al.* (2016). Après convergence, le vecteur $\mathbf{u}$ doit être binarisé afin d'obtenir un estimateur de $\mathbf{v}$. Afin d'effectuer cette binarisation, on calcule la valeur de la vraisemblance marginale exacte évaluée en les p valeurs de $\mathbf{v}$ obtenues en conservant uniquement les k plus grands coefficients de $\mathbf{u}$ pour $k \leq p$. Le vecteur $\mathbf{u}$ ayant la plus grande vraisemblance est alors choisi.

3 Comparaisons numériques

Nous avons comparé la selection de modèle de notre ACP globalement parcimonieuse à celle obtenue par validation croisée par Jenatton *et al.* (2009) ainsi qu'à un algorithme "localement parcimonieux", celui de Zou *et al.* (2006) (dans lequel la sélection est effectuée en conservant les variables expliquant 99% de la variance complète). En simulant selon le modèle avec $p = 100$, $d = 10$, $\sigma = 0.6$, nous obtenons les F-scores de la table 1.

Nous voyons que les méthodes globales permettent de retrouver la parcimonie des données de manière bien plus efficace. De nombreuses autres expériences sont disponibles dans Mattei *et al.* (2016)

4 Conclusion

La sélection non supervisée de variables est un problème complexe et bien souvent mal posé (en particulier lorsqu'aucune tâche prédictive telle que le *clustering* n'est envisagée). Nous avons montré qu'elle peut trouver une modélisation simple dans le cadre de l'ACP globalement parcimonieuse. Il serait intéressant d'explorer par la suite plus d'application à des données réelles et de voir dans quel mesure la dimension d de l'espace latent peut

être estimée dans le cadre d'un modèle proche du nôtre.

Références

Archambeau, C. et F. Bach. 2009, «Sparse probabilistic projections», dans *Advances in neural information processing systems*, p. 73–80.

Bishop, C. M. 1999, «Variational principal components», dans *Proceedings of the Ninth International Conference on Artificial Neural Networks*, p. 509–514.

Hotelling, H. 1933, «Analysis of a complex of statistical variables into principal components.», *Journal of educational psychology*, vol. 24, n° 6, p. 417.

Jenatton, R., G. Obozinski et F. Bach. 2009, «Structured sparse principal component analysis», dans *International Conference on Artificial Intelligence and Statistics*.

Johnstone, I. M. et A. Y. Lu. 2009, «On consistency and sparsity for principal components analysis in high dimensions», *Journal of the American Statistical Association*, vol. 104, n° 486.

Jolliffe, I. T., N. T. Trendafilov et M. Uddin. 2003, «A modified principal component technique based on the lasso», *Journal of Computational and Graphical Statistics*, vol. 12, n° 3, p. 531–547.

Kass, R. E. et A. E. Raftery. 1995, «Bayes factors», *Journal of the american statistical association*, vol. 90, n° 430, p. 773–795.

Latouche, P., P.-A. Mattei, C. Bouveyron et J. Chiquet. 2015, «Combining a relaxed EM algorithm with Occam's razor for bayesian variable selection in high-dimensional regression», *Journal of Multivariate Analysis*, vol. in press, doi :d10.1016/j.jmva.2015.09.004.

Lázaro-Gredilla, M. et M. K. Titsias. 2011, «Spike and slab variational inference for multitask and multiple kernel learning», dans *Advances in neural information processing systems*, p. 2339–2347.

Mattei, P.-A., C. Bouveyron et P. Latouche. 2016, «Globally sparse probabilistic pca», dans *International Conference on Artificial Intelligence and Statistics*.

Ringnér, M. 2008, «What is principal component analysis ?», *Nature biotechnology*, vol. 26, n° 3, p. 303–304.

Roweis, S. 1998, «EM algorithms for PCA and SPCA», *Advances in neural information processing systems*, p. 626–632.

Tipping, M. E. et C. M. Bishop. 1999, «Probabilistic principal component analysis», *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 61, n° 3, p. 611–622.

Zhang, Y., A. d'Aspremont et L. El Ghaoui. 2012, «Sparse PCA : Convex relaxations, algorithms and applications», dans *Handbook on Semidefinite, Conic and Polynomial Optimization*, Springer, p. 915–940.

Zou, H., T. Hastie et R. Tibshirani. 2006, «Sparse principal component analysis», *Journal of computational and graphical statistics*, vol. 15, n° 2, p. 265–286.