Un algorithme EM pour une version parcimonieuse de l'analyse en composantes principales probabiliste

Charles Bouveyron*, Julien Jacques**

*Université Paris Descartes - Laboratoire MAP5
charles.bouveyron@parisdescartes.fr,
http://w3.mi.parisdescartes.fr/∼cbouveyr/
**Université Lumière Lyon 2 - Laboratoire ERIC
julien.jacques@univ-lyon2.fr
http://eric.univ-lyon2.fr/∼jjacques/

Résumé. Nous considérons une version parcimonieuse de l'analyse en composantes principales probabiliste. La pénalité ℓ_1 imposée sur les composantes principales rend leur interprétation plus aisée en ne faisant dépendre ces dernières que d'un nombre restreint de variables initiales. Un algorithme EM, simple de mise en œuvre, est proposé pour l'estimation des paramètres du modèle. La méthode de l'heuristique de pente est finalement utilisée pour choisir le coefficient de pénalisation.

1 Introduction

L'analyse en composantes principales (ACP, Jolliffe (1986)) est une des méthodes, si ce n'est la méthode, d'analyse exploratoire les plus couramment utilisées. Elle a été ré-interprétée sous un formalisme probabiliste par Tipping et Bishop (1999), montrant que les composantes principales pouvaient être estimées par maximum de vraisemblance dans le cadre d'un modèle à variables latentes. Avec l'avénement des données de grande dimension, la problématique consistant à sélectionner un petit nombre de variables d'intérêt parmi l'ensemble des variables disponibles est devenue primordiale. Un des soucis majeurs de l'ACP dans cette optique est que les composantes principales sont définies comme une combinaison linéaire de l'ensemble des variables initiales. Des versions parcimonieuses de l'ACP (Zou et al., 2004) ainsi que de sa version probabiliste (Guan et Dy, 2009) ont été proposées récemment. La version parcimonieuse de Zou et al. (2004) repose sur l'ajout d'une pénalisation de type ℓ_1 au problème des moindres carrés, qui nécessite le choix du coefficient de pénalisation de façon heuristique. Dans Guan et Dy (2009), une version sparse bayésienne de l'ACP probabiliste est proposée. Nous proposons dans ce travail une alternative fréquentiste utilisant un algorithme EM pour l'inférence. La procédure d'estimation obtenue à l'avantage d'être particulièrement simple, et ne nécessite pas le choix de loi a priori. Elle offre en outre la possibilité de considérer le problème du choix de la pénalité comme un problème de choix de modèles.

2 Analyse en composantes principales probabiliste

Soit $\mathbf{y}$ un vecteur aléatoire observé de dimension p, et $\mathbf{x}$ un vecteur aléatoire latent (non observé), de dimension d, relié à $\mathbf{y}$ par l'équation suivante :

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

où $\mathbf{W}$ est une matrice $p \times d$, $\boldsymbol{\mu}$ est le vecteur moyenne (supposé nul dans la suite, $\boldsymbol{\mu} = 0$), et $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Conditionnellement au vecteur $\mathbf{x}$, la distribution des vecteurs observés est :

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{W}\mathbf{x}, \sigma^2\mathbf{I}). \tag{2}$$

En supposant $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, la distribution marginale du vecteur observé est :

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\mathbf{W}^t + \sigma^2\mathbf{I}). \tag{3}$$

Ce modèle est un modèle de type "*factor analysis*" (Bartholomew et al., 2011), popularisé par Tipping et Bishop (1999) sous le nom d'analyse en composantes principales probabiliste (*Probabilistic Principal Component Analysis (PPCA)*). En effet, une estimation par maximum de vraisemblance des paramètres du modèle à l'aide d'un algorithme EM (Dempster et al., 1977), considérant le vecteur latent $\mathbf{x}$ comme manquant, conduit à estimer les colonnes de $\mathbf{W}$ par les vecteurs propres de la matrice de covariance empirique, vecteurs qui ne sont rien d'autres que les axes principaux classiques.

Soit $\mathbf{y}_1, \ldots, \mathbf{y}_n$ un échantillon i.i.d de vecteurs observés. L'algorithme EM consiste à maximiser de façon itérative la log-vraisemblance complétée par les données non observées $\mathbf{x}_1, \ldots, \mathbf{x}_n$:

$$\ell_c = \sum_{i=1}^{n} \left(-\frac{p}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}_i - \mathbf{W}\mathbf{x}_i)^t(\mathbf{y}_i - \mathbf{W}\mathbf{x}_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\mathbf{x}_i^t\mathbf{x}_i \right). \tag{4}$$

3 Une version parcimonieuse de l'analyse en composantes principales probabiliste

Dans ce travail, nous considérons une version parcimonieuse de l'analyse en composantes principales probabiliste. L'objectif est d'obtenir des axes principaux déterminés uniquement grâce à un nombre restreint de variables initiales, et ainsi faciliter leur interprétation. De plus, comme nous le verrons par la suite, l'approche probabiliste de l'ACP permet de sélectionner le paramètre de pénalité par des méthodes classiques de sélection de modèles.

Dans l'optique d'introduire de la parcimonie au sein de axes principaux, nous considérons une pénalité ℓ_1 sur les colonnes de la matrice $\mathbf{W}$. La vraisemblance complétée à maximiser est alors la suivante :

$$\ell_c^{pen} = \ell_c - \lambda \sum_{\ell=1}^{d} ||\mathbf{w}_\ell||_1 \tag{5}$$

où $\mathbf{w}_\ell = (w_{1\ell}, \ldots, w_{p\ell})^t$ est la ℓ-ème colonne de $\mathbf{W}$ et $\lambda > 0$ est le paramètre de pénalisation. L'algorithme EM est un algorithme itératif qui alterne deux étapes (E et M), décrites ci-après.

La q-ème itération de l'étape E consiste à calculer l'espérance de ℓ_c^{pen} sous la loi $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{(q)})$, où $\boldsymbol{\theta}^{(q)} = (\mathbf{W}^{(q)}, \sigma^{2(q)})$ est la valeur courante de l'estimation des paramètres du modèles :

$$E[\ell_c^{pen}(\boldsymbol{\theta})|\boldsymbol{\theta}^{(q)}] = -\sum_{i=1}^{n}\left(\frac{p}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}\mathbf{y}_i^t\mathbf{y}_i - \frac{1}{\sigma^2}\mathbf{e}_i^t\mathbf{W}^t\mathbf{y}_i + \frac{1}{2}tr(\mathbf{S}_i) + \frac{1}{2\sigma^2}tr(\mathbf{W}^t\mathbf{W}\mathbf{S}_i)\right) - \lambda\sum_{j=1}^{p}\sum_{\ell=1}^{d}|w_{j\ell}| + c \tag{6}$$

où $c = -\frac{n}{2}(p+d)\ln 2\pi$ est une constante indépendante des paramètres du modèle et où

$$\mathbf{e}_i = \mathbf{M}^{-1}{\mathbf{W}^{(q)}}^t\mathbf{y}_i, \qquad \text{et} \qquad \mathbf{S}_i = \sigma^{2(q)}\mathbf{M}^{-1} + \mathbf{e}_i\mathbf{e}_i^t \tag{7}$$

avec $\mathbf{M} = {\mathbf{W}^{(q)}}^t\mathbf{W}^{(q)} + \sigma^{2(q)}\mathbf{I}$.

L'étape M consiste alors à maximiser $E[\ell_c^{pen}(\boldsymbol{\theta})|\boldsymbol{\theta}^{(q)}]$ en fonction de $\boldsymbol{\theta}$. De sorte à faciliter la maximisation, nous considérons l'approximation de la norme ℓ_1 par la forme quadratique suivante (Fan et Li, 2001) :

$$|w_{j\ell}| \simeq |w_{j\ell}^{(q)}| + \frac{1}{2}\frac{\text{sign}(w_{j\ell}^{(q)})}{|w_{j\ell}^{(q)}|}(w_{j\ell}^2 - {w_{j\ell}^{(q)}}^2) \tag{8}$$

qui est valide lorsque $w_{j\ell}^{(q)} \simeq w_{j\ell}$. La maximisation de $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ en fonction de $\mathbf{W}$ n'ayant pas de solution analytique, nous utilisons une approche alternative consistant à maximiser $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ en fonction de chaque éléments de la matrice $\mathbf{W}$ successivement. On obtient alors, en dérivant $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ par rapport à $w_{j\ell}$ et en égalant à 0 :

$$w_{j\ell}^{(q+1)} = \frac{\sum_{i=1}^{n}\left(e_{i\ell}y_{ij} - \frac{1}{2}\sum_{k\neq\ell}s_{i\ell k}w_{jk}^{(q)}\right)}{\sigma^{2(q)}\lambda\frac{\text{sign}(w_{j\ell}^{(q)})}{|w_{j\ell}^{(q)}|} + \sum_{i=1}^{n}s_{i\ell\ell}}. \tag{9}$$

Cette dérivation élément par élément ne conduit pas nécessairement au maximum de $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$, mais suffit pour faire augmenter la log-vraisemblance à chaque étape de l'algorithme. On obtient alors un algorithme GEM (*Generalized* EM) qui conserve les mêmes propriétés de convergence qu'un algorithme EM classique.

L'estimateur de la variance résiduelle est quant à lui :

$$\sigma^{2(q+1)} = \frac{1}{Np}\sum_{i=1}^{n}\left\{\mathbf{y}_i^t\mathbf{y}_i - 2\mathbf{e}_i^t{\mathbf{W}^{(q+1)}}^t\mathbf{y}_i + tr\left(\mathbf{S}_i{\mathbf{W}^{(q+1)}}^t\mathbf{W}^{(q+1)}\right)\right\}. \tag{10}$$

et s'avère être identique à celui de la version non sparse de l'analyse en composantes principales probabiliste (Tipping et Bishop, 1999).

4 Sélection de λ par l'heuristique de pente

Dans Zou et al. (2004), le choix de la pénalité λ est réalisé de manière heuristique en se basant sur l'éboulis des valeurs propres. L'idée de la stratégie que nous proposons est d'estimer le modèle sur une grille de valeurs de λ, et d'utiliser un outil de sélection de modèles pour choisir le meilleur modèle. Les outils classiques de sélection de modèles sont par exemple les critères AIC (Akaike, 1974) et BIC (Schwarz, 1978), qui pénalisent la log-vraisemblance $\ell(\hat{\theta})$ de la façon suivante : $\mathrm{AIC} = \ell(\hat{\boldsymbol{\theta}}) - \gamma$ et $\mathrm{BIC} = \ell(\hat{\boldsymbol{\theta}}) - \frac{\gamma}{2}\log(n)$, où γ est le nombre de paramètres libres du modèles et n le nombre d'observations. La valeur de γ dépend directement de la valeur de λ puisqu'elle est égale au nombre d'éléments non nuls dans $\mathbf{W}$ plus un (pour la variance résiduelle). Même si ces critères sont largement utilisés et asymptotiquement consistants, ils sont aussi connus pour être plus efficaces sur simulations que sur données réelles.

Pour surmonter ce problème, Birgé et Massart (2007) ont récemment proposé une méthode dirigée par les données pour calibrer la pénalité des critères pénalisés, connue sous le nom d'heuristique de pente. L'heuristique de pente a été proposé initialement dans un cadre d'un modèle de régression gaussien homoscédastique, mais a ensuite été étendue à d'autres situations. Birgé et Massart (2007) ont démontré qu'il existait une pénalité minimale et que de considérer une pénalité égale au double de la pénalité minimale permettait d'approcher le modèle oracle en terme de risque. La pénalité minimale est en pratique estimée par la pente de la partie linéaire de la log-vraisemblance $\ell_c^{pen}(\hat{\boldsymbol{\theta}})$ exprimée en fonction de la complexité du modèle. Le critère associé est alors défini par :

$$\mathrm{SHC} = \ell(\hat{\boldsymbol{\theta}}) - 2\hat{s}\gamma, \tag{11}$$

où $\hat{s}$ est l'estimation de la pente de la partie linéaire de $\ell_c^{pen}(\hat{\boldsymbol{\theta}})$. Une revue détaillée et des conseils d'implémentation sont donnés dans Baudry et al. (2012).

5 Illustrations numériques

Nous choisissons pour illustrer notre méthodologie un jeu de données classique issu de l'*UCI machine learning repository* : le jeux de données USPS. Le jeu original contient 7291 images représentant des chiffres manuscrits de 0 à 9. Chaque chiffre est une image en niveaux de gris de taille 16×16, représentée par un vecteur de dimension 256. Pour cette expérience, nous avons extrait un sous ensemble de 1756 images correspondant aux chiffres 3, 5 et 8. Nous réalisons sur ces données une ACP ainsi que l'ACP parcimonieuse que nous proposons. Pour cette dernière, nous fixons le nombre maximum d'itérations de l'algorithme EM à 500 et le seuil de convergence à 10^{-6}, et nous considérons une grille de valeurs de λ de 0 à 150, avec un pas de 1. La méthode de l'heuristique de pente (figure 1) conduit à choisir $\lambda = 126$. Dans un but illustratif, nous discutons ici les résultats concernant les deux premières composantes principales. La figure 2 représente la projection des 1756 images dans le premier plan principal de l'ACP ainsi que les deux premières composantes principales, tandis que la figure 3 propose la même représentation pour l'ACP parcimonieuse. Nous pouvons noter que les deux méthodes définissent un premier plan principal relativement discriminant vis-à-vis des trois types d'images. Tout l'intérêt de l'ACP parcimonieuse est que les composantes principales

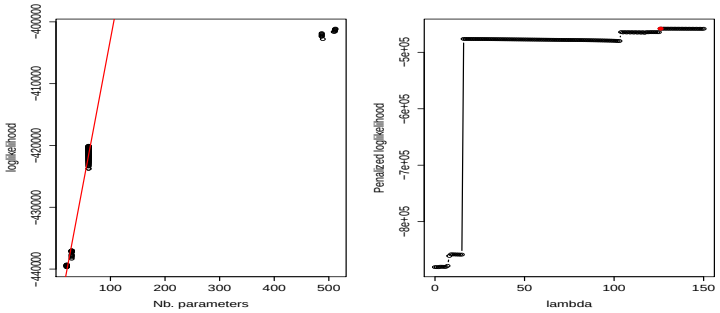

FIG. 1 – *Heuristique de pente : log-vraisemblance en fonction du nombre de paramètres (gauche) et log-vraisemblance pénalisée en fonction de* λ.

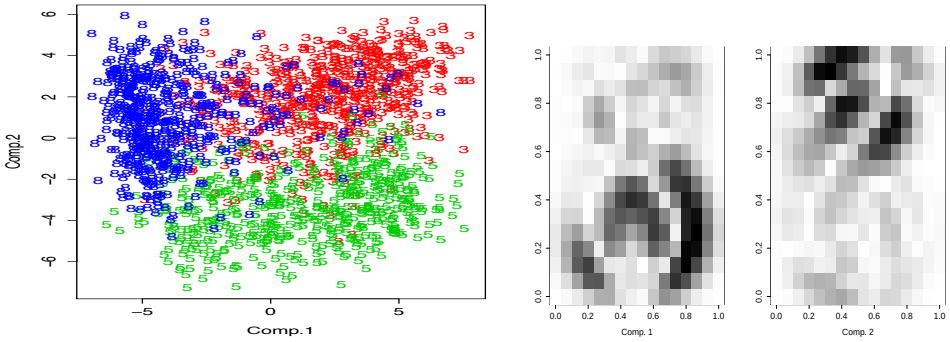

FIG. 2 – *Représentation des 1756 images dans le premier plan principal (gauche) et deux premières composantes principales (droite) obtenues par l'analyse en composantes principales.*

obtenues captent un signal semblable à celles de l'ACP, tout en étant très parcimonieuses puisqu'elles ne dépendent que de peu de variables initiales (21 pour la première composante et 19 pour la seconde sur les 256 initialement disponibles).

Références

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Bartholomew, D., M. Knott, et I. Moustaki (2011). *Latent Variable Models and Factor Analysis : A Unified Approach*. Wiley Series in Probability and Statistics. Wiley.

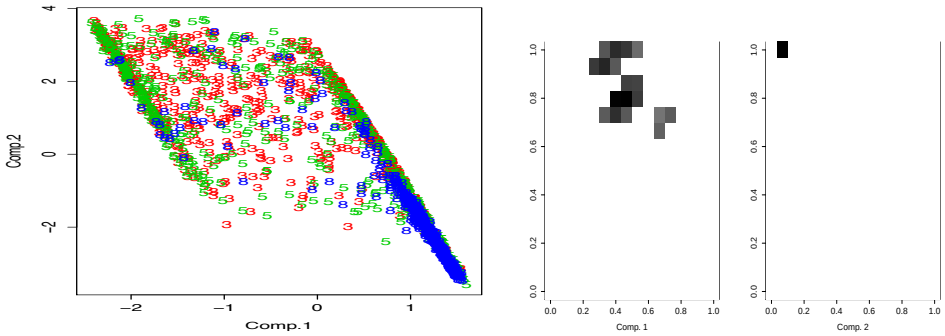

FIG. 3 – *Représentation des 1756 images dans le premier plan principal (gauche) et deux premières composantes principales (droite) obtenues par l'analyse en composantes principales parcimonieuse.*

Baudry, J.-P., C. Maugis, et B. Michel (2012). Slope heuristics : overview and implementation. *Statistics and Computing 22*(2), 455–470.

Birgé, L. et P. Massart (2007). Minimal penalties for gaussian model selection. *Probability theory and related fields 138*(1-2), 33–73.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

Fan, J. et R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Guan, Y. et J. Dy (2009). Sparse probabilistic principal component analysis. In *In Proc. AISTATS'2009, JMLR W&CP*, pp. 185–192.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer-Verlag.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*, 461–464.

Tipping, M. et C. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B 61*, 611–622.

Zou, H., T. Hastie, et R. Tibshirani (2004). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*, 265–286.

Summary

We consider a parsimonious version of probabilistic principal component analysis and we proposed an EM algorithm for model inference. The ℓ_1 penalty imposed on the principal components makes their interpretation easier, by linking them with only a limited number of original variables. The EM algorithm, easy to implement, is proposed to estimate the model parameters. The slope heuristic is used to select the intensity of the penalty.