Sélection de modèle bayésienne exacte pour l'analyse en composantes principales

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

[1] *Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes/Sorbonne Paris Cité*
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. Nous présentons une méthode de sélection de modèle bayésienne pour déterminer le nombre de composantes principales. Notre approche est basée sur un calcul explicite de la vraisemblance marginale dans le cadre d'un nouvel *a priori* de type normal-gamma. Ainsi, les probabilités *a posteriori* des modèles peuvent être déterminées de façon exacte et un nombre d'axes optimal choisi. Les hyperparamètres sont choisis à l'aide d'une méthode heuristique simple. Dans un cadre non-asymptotique, nous montrons à l'aide de simulations que cette méthode exacte est compétitive avec les procédés habituels de sélection de dimension, bayésiens ou non.

Mots-clés. ACP, parcimonie, réduction de dimension, sélection de modèle

Abstract. We present a Bayesian model selection framework to estimate the number of principal components. Our approach is based on a closed-form expression of the marginal likelihood obtained using a new normal-gamma prior. Consequently, posterior probabilities of models can be exactly computed and an optimal number of components can be inferred. Hyperparameters are chosen using simple heuristics. In a non-asymptotic framework, we show on simulated data that this exact method is competitive with both Bayesian and frequentist state-of-the-art methods.

Keywords. PCA, sparsity, dimension reduction, model selection

1 Introduction

En dépit de son ubiquité s'étendant à l'ensemble des domaines de la statistique (Jolliffe et Cadima, 2016), l'analyse en composantes principales (ACP) ne bénéficie pas d'une technique de détermination automatique du nombre d'axes largement acceptée. Généralement, des critères heuristiques sont utilisés par le praticien au regard des valeurs propres de la matrice de covariance empirique des données. Cependant, cette méthode ancienne, popularisée notamment par Cattell (1966), est en général surpassée par d'autres approches plus récentes, comme la validation croisée (Josse et Husson, 2012) ou des méthodes de maximum de vraisemblance (Zhu et Ghodsi, 2006; Bouveyron *et al.*, 2011).

La sélection de modèle bayésienne (Robert, 2006, chap. 7) permet théoriquement de proposer une réponse automatique à ce problème. Cependant, dans le cadre des *a priori* classiques, il n'existe pas d'expression explicite de la vraisemblance marginale et des approximations sont en général utilisées (Bishop, 1999; Minka, 2000; Archambeau et Bach, 2009). Nous présentons ici un *a priori* de type normal-gamma permettant de calculer exactement la vraisemblance marginale, et donc d'obtenir un choix de dimension optimal d'un point de vue bayésien.

2 Sélection de modèle 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. Les observations sont stockées dans la matrice $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$.

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

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

où $\mathbf{y}_i \sim \mathcal{N}(0, \mathbf{I}_d)$ est un vecteur gaussien latent de dimension faible, $\mathbf{W}$ est une matrice de paramètres de taille $p \times d$ et $\boldsymbol{\varepsilon}_i|\sigma \sim \mathcal{N}(0, \sigma^2\mathbf{I}_p)$ est un bruit gaussien d'écart-type $\sigma > 0$ (indépendant du vecteur latent) pour tout $i \in \{1, ..., n\}$. 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 classiques de la matrice de données.

Afin de procéder à une analyse bayésienne, nous proposons d'utiliser des *a priori* gaussiens indépendants $w_{jk} \sim \mathcal{N}(0, \phi^{-1})$ pour $j \in \{1, ..., p\}$ et $k \in \{1, ..., d\}$ où $\phi > 0$ représente un hyperparamètre de précision.

De nombreux modèles bayésiens similaires ont été introduits, sans jamais que les vraisemblances marginales associées soient calculées exactement (Bishop, 1999; Minka, 2000; Archambeau et Bach, 2009) excepté dans le cas d'un modèle sans bruit inadapté au problème de sélection de dimension (Bouveyron *et al.*, 2016). En effet, comme le paramètre de variance du bruit change considérablement lorsqu'on ajoute ou retire une dimension, choisir un *a priori* pour ce paramètre est crucial si l'on souhaite choisir d.

Ici, nous utilisons un *a priori* gamma sur la variance du bruit

$$\sigma^2 \sim \text{Gamma}(a, b), \tag{2}$$

où $a > 0$ et $b > 0$ sont des hyperparamètres positifs. En choisissant $b = \phi/2$, la loi marginale des données devient une loi de Laplace généralisée, pour des raisons techniques omises dans ce résumé – cf. Kozubowski *et al.* (2013) et Mattei (2017) pour plus de détails. Plus précisément, on peut montrer que

$$\forall i \in \{1, ..., n\}, \quad \mathbf{x}_i \sim GAL_p(2\phi^{-1}\mathbf{I}_p, 0, a + d/2), \tag{3}$$

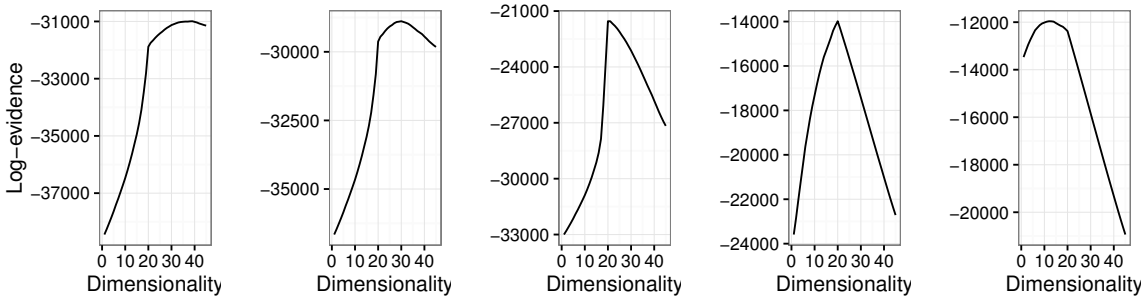

FIGURE 1 – Différente courbes de vraisemblance marginale pour différentes valeurs de ϕ. $\phi*$ correspond au maximum de notre critère heuristique.

où GAL_p représente la loi de Laplace généralisée de Kotz *et al.* (2001, p. 257). Cette expression constitue le premier calcul explicite de vraisemblance marginale pour un modèle bayésien bruité bâti sur l'ACPP. Notons cependant qu'Ando (2009) a également obtenu une expression exacte pour un modèle d'analyse factorielle à facteurs Student, modèle qui n'est pas sans lien avec l'ACPP.

La dimension choisie par notre méthode sera donc celle conduisant à la plus grande vraisemblance marginale. Notons tout de même que ce résultat dépend des hyperparamètres ϕ et a. Afin de choisir ces paramètres, nous utilisons deux heuristiques. D'une part, nous choisissons a afin que l'*a priori* sur σ mette de la masse près d'un estimateur $\hat{\sigma}$ (comme par exemple l'estimateur du maximum de vraisemblance de Tipping et Bishop, 1999). Pour ce qui est de ϕ, nous construisons une grille puis optimisons un critère heuristique graphique basé sur l'allure de la courbe de vraisemblance marginale attendue. L'idée de base est qu'une courbe "idéale" devrait avoir deux phases : une phase de croissance rapide (l'ajout des dimensions de signal) et une de décroissance lente (les dimensions de bruit). Cette différence de vitesse de croissance/décroissance assure que l'algorithme préfèrera ajouter une dimension de bruit plutôt que de tuer de l'information. Plus précisément, notre critère assure que le ϕ^* choisi sera à l'origine d'une courbe de vraisemblance marginale comprenant deux phases bien distinctes (avant et après le maximum), telles que la pente de la première phase est plus grande que celle de la deuxième phase. Cela assure que notre courbe ne préfèrera pas la sous-estimation plutôt que la surestimation de la dimension intrinsèque des données.

3 Simulations

Nous reprenons le schéma de simulation isotrope de Bouveyron *et al.* (2011) et calculons le pourcentage de bons choix de d pour cinquante répétitions. Nous comparons notre méthode exacte (EBMS) avec l'approximation de Laplace de la vraisemblance marginale de Minka (2000), ainsi qu'avec la méthode de maximum de vraisemblance (ML) de Bou-

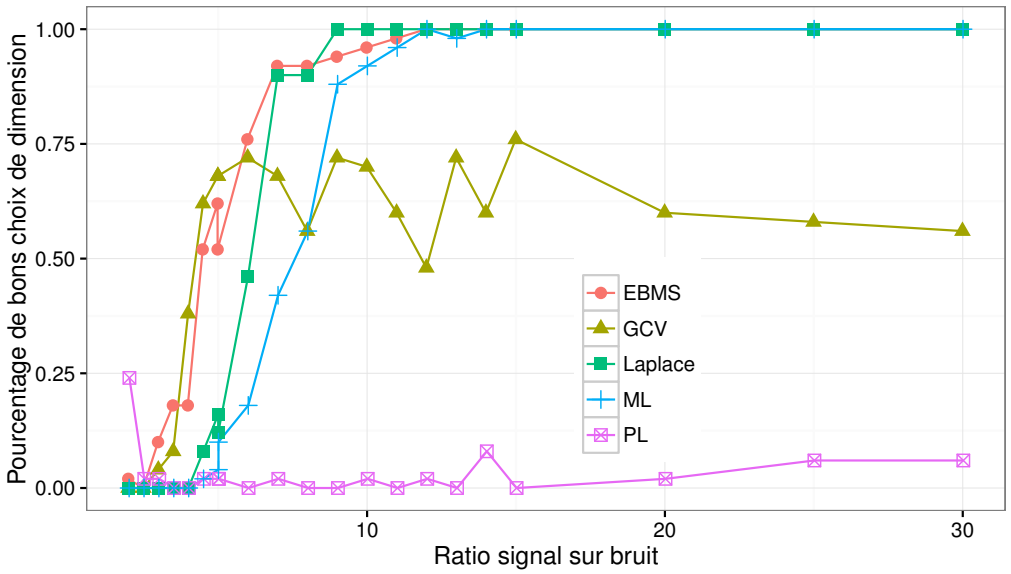

FIGURE 2 – Simulations selon le schéma de Bouveyron *et al.* (2011) avec $n = 70$ et $p = 50$.

veyron *et al.* (2011), la validation croisée généralisée (GCV) de Josse et Husson (2012) ainsi que la méthode *profile likelihood* (PL) de Zhu et Ghodsi (2006).

Dans le cadre largement non-asymptotique considéré ($n = 70, p = 50$), nous voyons que seule notre méthode est compétitive à tous niveaux de bruit.

4 Conclusion

L'ACP est essentiellement une analyse exploratoire. Le consensus général est qu'aucune méthode de sélection de dimension n'est meilleure, dans l'absolu, que les autres. Cependant, il nous paraît fondamental de continuer à explorer des pistes non-asymptotiques de décision automatique afin d'aider le praticien à se faire une idée sur la structure des données auxquelles il est confronté lorsque celles-ci sont peu nombreuses et/ou coûteuses. Notre travail, en s'éloignant des approximations asymptotiques habituellement faites en sélection de modèle bayésienne, constitue un pas dans cette direction.

Références

Ando, T. 2009, «Bayesian factor analysis with fat-tailed factors and its exact marginal likelihood», *Journal of Multivariate Analysis*, vol. 100, n° 8, p. 1717–1726.

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

Bishop, C. M. 1999, «Bayesian PCA», dans *Proceedings of the 1998 conference on Advances in neural information processing systems II*, p. 382–388.

Bouveyron, C., G. Celeux et S. Girard. 2011, «Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic pca», *Pattern Recognition Letters*, vol. 32, n° 14, p. 1706–1713.

Bouveyron, C., P. Latouche et P.-A. Mattei. 2016, «Bayesian variable selection for globally sparse probabilistic PCA», Technical report, HAL-01310409.

Cattell, R. B. 1966, «The scree test for the number of factors», *Multivariate behavioral research*, vol. 1, n° 2, p. 245–276.

Jolliffe, I. T. et J. Cadima. 2016, «Principal component analysis : a review and recent developments», *Phil. Trans. R. Soc. A*, vol. 374, n° 2065, p. 20150 202.

Josse, J. et F. Husson. 2012, «Selecting the number of components in principal component analysis using cross-validation approximations», *Computational Statistics & Data Analysis*, vol. 56, n° 6, p. 1869–1879.

Kotz, S., T. Kozubowski et K. Podgórski. 2001, *The Laplace distribution and generalizations : a revisit with applications to communications, exonomics, engineering, and finance*, 183, Springer Science & Business Media.

Kozubowski, T., K. Podgórski et I. Rychlik. 2013, «Multivariate generalized laplace distribution and related random fields», *Journal of Multivariate Analysis*, vol. 113, p. 59–72.

Mattei, P.-A. 2017, «Multiplying a Gaussian matrix by a Gaussian vector», *Statistics & Probability Letters*, vol. à paraître.

Minka, T. P. 2000, «Automatic choice of dimensionality for PCA», dans *Nips*, vol. 13, p. 598–604.

Robert, C. 2006, *Le choix bayésien : Principes et pratique*, Springer Science & Business Media.

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.

Zhu, M. et A. Ghodsi. 2006, «Automatic dimensionality selection from the scree plot via the use of profile likelihood», *Computational Statistics & Data Analysis*, vol. 51, n° 2, p. 918–930.