Modèle Bayésien à Facteurs Latents pour l'analyse de données fonctionnelles

Gabrielle Weinrott [1,*], Bénédicte Fontez [2,*], Nadine Hilgert [3,*] & Susan Holmes [4,†]

[*] *UMR 0729 MISTEA, Montpellier SupAgro-INRA,*
2 Place Viala, 34060 Montpellier cedex 2, France
[†] *Stanford Univeristy, 450 Serra Mall, Stanford, CA 94305, USA*
[1] *gabrielle.weinrott@supagro.inra.fr*
[2] *benedicte.fontez@supagro.fr*
[3] *nadine.hilgert@supagro.inra.fr*
[4] *susan@stat.stanford.edu*

Résumé. Nous proposons une approche exploratoire pour des données fonctionnelles issues de l'agronomie qui sont hétérogènes, longitudinales, et incertaines. L'ensemble de courbes est projeté dans une base d'histogrammes, ce qui permet de considérer des intervalles de temps comme variables. Un modèle Bayésien à facteurs latents est ensuite utilisé pour identifier les grandes sources de variation dans le jeu de données. L'inférence est faite à l'aide d'un algorithme type Hybrid Monte Carlo (HMC) implémenté sous STAN.
La méthode est illustrée sur un jeu de données simulées. L'approche Bayésienne permet de prendre en compte l'expertise et l'incertitude des données, ainsi que de visualiser l'incertitude des facteurs latents, et la projection des individus sur les axes de ces derniers.

Mots-clés. Données fonctionnelles, Données longitudinales, Incertitude, Inférence Bayésienne, Facteurs Latents, Visualisation.

Abstract. We suggest an exploratory approach for functional agricultural data that is longitudinal, heterogeneous, and uncertain. The set of curves are projected onto a histogram basis, which allows us to consider that the variables are intervals in time. We then use a Bayesian Latent Factor Model to identify the major sources of variation across the dataset. The inference is done using a Hybrid Monte Carlo algorithm implemented using STAN.
The method is illustrated on a simulated data set. The Bayesian approach allows us to take into account expert knowledge and data uncertainty; it also allows us to visualise the uncertainty of projection of the observations onto the axes spanned by the latent factors.

Keywords. Functional data, Longitudinal data, Uncertainty, Bayesian Inference, Latent Factors, Visualization.

1 Contexte

En sciences du vivant, les chercheurs recueillent de grandes quantités de données longitudinales et fonctionnelles à analyser et comprendre. Ces données issues de capteurs variés sont souvent entachées d'incertitude de mesure qui nécessite d'être prise en compte pour une exploration des

données plus pertinente.

L'analyse de données fonctionnelles implique généralement de réduire la dimension en projetant dans une base. Un cas particulier est la projection dans la base de Karhunen-Loève, qui est l'équivalent en données fonctionnelles d'une Analyse en Composantes Principales (ACP) (Jolliffe, 1986) en données multivariées. L'ACP fonctionnelle (Rice & Silverman, 1991) fournit des résultats intéressants car elle maximise la dispersion des observations sur un espace de dimension réduite, ce qui facilite l'objectif de visualisation, mais les graphiques sont difficiles à interpréter pour un utilisateur non-averti. Aussi, nous proposons de projeter les données fonctionnelles dans une base d'histogrammes avant d'effectuer une ACP (Deville, 1974). Cette approche converge vers les résultats d'une ACP fonctionnelle, tout en gardant des variables simples à interpréter (du type intervalles de temps).

L'ACP dans une base d'histogramme permet d'examiner les grandes sources de variabilité dans un jeu de données fonctionnelles, mais ne donne pas la possibilité d'intégrer les incertitudes de mesure. En reformulant l'ACP comme un Modèle à Facteurs Latents, il est possible de considérer une version probabiliste de la méthode (Tipping & Bishop, 1999). Dans ce travail, nous proposons d'étudier le Modèle à Facteur Latents dans un cadre Bayésien (Bishop, 1999), ce qui ouvre la possibilité d'intégrer de l'expertise et l'incertitude de mesure sous forme de lois a priori.

L'évaluation de ce modèle permet d'obtenir des intervalles de crédibilité pour l'estimation des facteurs latents, et pour les scores des observations associés à chaque facteur. De cette manière, nous pouvons visualiser les résultats de façon analogue aux sorties de l'ACP, mais avec des régions de confiances autour des projections, et un intervalle de crédibilité pour les facteurs latents.

2 Méthodes

2.1 Notation

Nous disposons d'un ensemble de courbes $X_i(t)$, avec $i \in \{1, ..., N\}$, mesurées discrètement dans le temps, ou $t \in T$ est l'ensemble des temps de mesure.

Soit $\Pi_{(l)} : X_i \to \Pi_{(l)} X_i$ le projecteur sur la base d'histogrammes, et l la taille de la fenêtre de la base.

2.2 Projection sur la base d'histogrammes

Soit $U = \{u_1, ..., u_j, ..., u_{P+1}\}$ l'ensemble des bornes des p intervalles définissants la base d'histogrammes de même longueur $l = \frac{u_{P+1} - u_1}{P}$.

Pour $t \in [u_j, u_{j+1}]$ et i une observation donnée,

$$\Pi_{(j)} X_i(t) = \frac{1}{l} \int_{u_j}^{u_{j+1}} X_i(s) ds.$$

La i^{eme} courbe projetée s'écrit alors :

$$Y_i(t) = \sum_{j=1}^{p} \Pi_{(j)} X_i(t) \mathbf{1}_{[u_j, u_{j+1}]}(t).$$

2.3 Modèle Bayésien à Facteurs Latents

2.3.1 Le Modèle

Soit la matrice $\mathbf{Y}_i$, $i \in \{1, ..., N\}$ des données projetées dans la base d'histogrammes. Le modèle s'écrit comme:

$$\mathbf{Y_i}^T \quad = \quad \mathbf{W}\beta_i + \epsilon_i \tag{1}$$

où $\mathbf{W} = (w_{j,k})_{j,k}$ est une matrice de transition $P \times Q$, β_i est un Q-vecteur avec $Q < P$ de coordonnées de $\mathbf{Y}_i$ dans l'espace de dimension réduite, et ϵ_i sont des P-vecteurs d'erreurs de mesure.

La contrainte usuelle dite triangulaire inférieure est appliquée aux facteurs latents pour forcer l'identifiabilité (Geweke & Zhou, 1996). La décomposition de la matrice $\mathbf{W}^T = \mathbf{LR}$ où $\mathbf{L}$ est une matrice orthogonale et $\mathbf{R}$ est triangulaire-supérieure permet l'estimation des éléments hors diagonale $r_{j,k}$ ou $j \in \{1, ..., P\}$, $k \in \{1, ..., Q\}$ et $j > k$. Pour récupérer une estimation de la matrice de rang inférieur, on multiplie la transposée de la matrice estimée $\mathbf{R}$ par la transposée d'une matrice de rotation. Comme dans (Rowe, 2000), l'Analyse en Composantes Principales guide notre choix de $\mathbf{L}$. Nous utilisons un modèle complet Bayésien similaire à (Ghosh & Dunson, 2009), où les scores sont des Q-vecteurs centrés et gaussiens, et les paramètres libres de la matrice triangulaire-inférieure sont normaux avec précision τ^2, et τ^2 et σ^2 suivent des lois a priori non-informatives (par exemple, une loi inverse gamma).

Le modèle complet Bayésien est :

$$\mathbf{Y_i}|\mathbf{R}^T, \beta_i, \sigma^2 \quad \overset{iid}{\sim} \quad \mathcal{N}_P\left(\mathbf{R}^T\beta_i, \sigma^2\mathbf{1}_P\right) \tag{2}$$

avec les priors suivants :

$$\begin{aligned} r_{j,k} &\overset{iid}{\sim} N\left(0, \tau^2\right) \\ \beta_i &\overset{iid}{\sim} N_Q\left(0, \mathbf{1}_Q\right) \\ \tau^2, \sigma^2 &\sim IG\left(0.001, 0.001\right) \end{aligned}$$

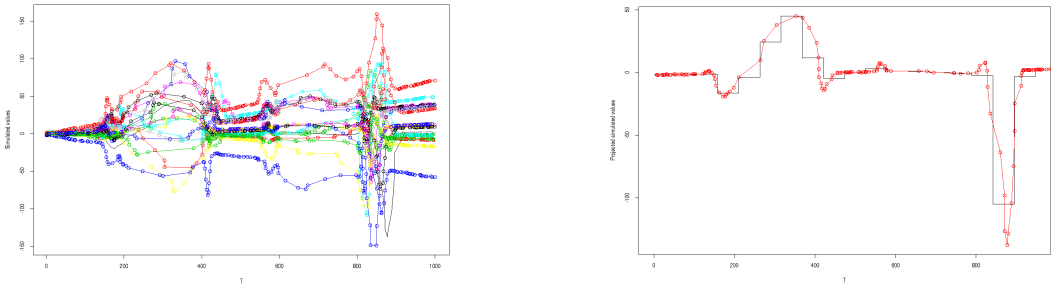
Nous avons choisi les hyperparamètres des termes de variance τ^2 et σ^2 comme le font de nombreux auteurs dans la litérature (Gelman, 2006).

2.3.2 Inference

Les paramètres à estimer sont σ^2, τ^2, les β_i pour chacune des N observations, et les éléments non-nuls de la matrice $\mathbf{R}^T$. On fait l'inférence Bayésienne avec un échantillonneur MCMC en utilisant le langage de programmation STAN (Stan, 2014) dans R (package *rstan*), et l'algorithme NUTS (Hoffman & Gelman, 2014).

3 Illustration

Nous avons simulé un ensemble de courbes $\mathbf{X}$ pour $N = 10$ observations, que nous avons ensuite projeté dans une base d'histogrammes avec $P = 14$ pour obtenir le jeu de données simulées $\mathbf{Y}$ (Figure 1).



(a) Les courbes simulées $\tilde{X}$ (b) Une courbe $\tilde{X}_i$ et sa projection $\tilde{Y}_i$

Figure 1: Exemple de données simulées

Nous avons inféré le modèle (1) avec l'algorithme NUTS dans STAN, avec 4 chaînes et 5×10^5 itérations. La première moitié des itérations est une période de chauffe. L'autre moitié permet d'estimer les variables (facteurs latents, scores, et la variance) a posteriori ainsi que la variabilité autour de l'estimation.

En représentant les scores estimés des 10 observations sur les facteurs estimés, nous avons obtenu la Figure 2. Il est possible de calculer numériquement la loi complète a posteriori, ordonner les itérations, et obtenir des quantiles à 90%. Les quantiles permettent de représenter l'incertitude par une enveloppe de crédibilité pour chaque observation.

4 Discussion

Cette approche permet de modéliser l'incertitude sur les facteurs latents d'un ensemble de courbes longitudinales. Les priors du modèle complet (2) peuvent être adaptés pour prendre en compte de l'expertise, ou d'avantage d'information sur l'incertitude des données. La Figure 2 est intéressante car elle illustre de façon directe l'impact de l'incertitude sur les scores et les facteurs latents, ce qui est novateur par rapport aux approches similaires existantes et implémentées sous R.

Nous présenterons aussi la comparaison de notre approche avec l'ACP classique (Dray & Dufour, 2007), l'ACP fonctionnelle (Ramsay, Wickham, Graves, & Hooker, 2014), ou encore avec l'ACP Bayésienne. En ce qui concerne le modèle Bayésien, il serait possible de comparer entre différentes

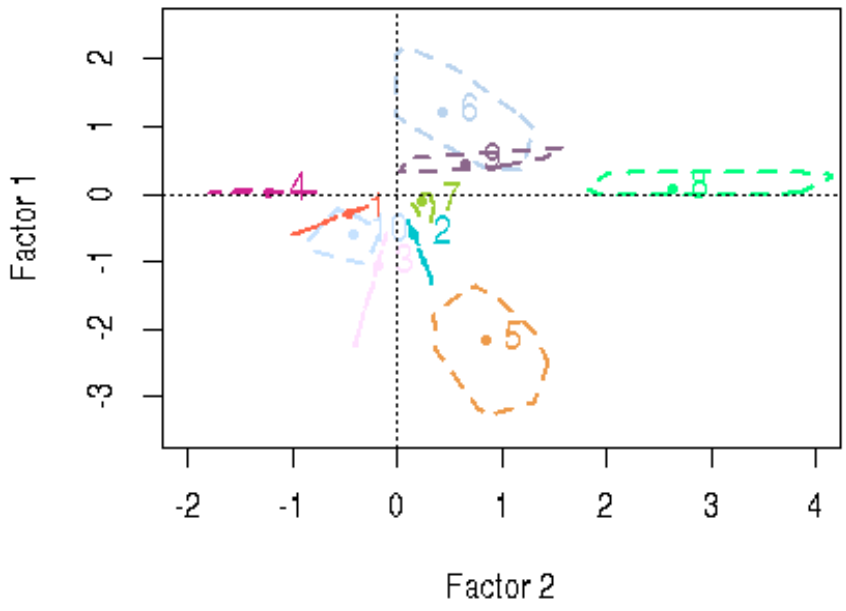


Figure 2: Les scores estimés avec leur enveloppe de crédibilité à 90 %

lois a priori non-informatives (Ghosh & Dunson, 2009), ou de comparer la performance de STAN avec d'autres algorithmes MCMC.

Remerciements

Les auteurs remercient la région Languedoc-Roussillon pour le financement partiel de ce travail par le dispositif de l'ARPE (Aide Régionale en Partenariat avec les Entreprises) sous le numéro 2015-005028.

References

Bishop, C. M. (1999). Bayesian principal component analysis. *Advances in Neural Information Processing Systems*, *11*, 382-388.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, *15*.

Dray, S., & Dufour, A. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*(4), 1-20.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515-533.

Geweke, J., & Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, *9*, 557-585.

Ghosh, J., & Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics*, *18*.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sample: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*, 1351 - 1381.

Jolliffe, I. (1986). *Principal component analysis* (Springer, Ed.).

Ramsay, J., Wickham, H., Graves, S., & Hooker, G. (2014). fda: Functional data analysis.

Rice, J., & Silverman, B. (1991). Estimating the mean and covariance structure nonparemetrically when the data are curves. *Journal of the Royal Statistical Society*, *53*, 233-243.

Rowe, D. B. (2000). A bayesian factor analysis model with generalized prior information. *Social Science Working Paper*, *1099*.

Stan, D. T. (2014). Stan modeling language users guide and reference manual, version 2.5.0 [Computer software manual]. Retrieved from `http://mc-stan.org/`

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, *61*, 611-622.