Le package *dad* d'analyse de données ternaires organisées en *folder* via les densités de probabilité

Pierre Santagostini [1], Smail Yousfi [2], Sabine Demotes-Mainard [3] & Rachid Boumaza [4]

[1] *Agrocampus-Ouest, UMR 1345 IRHS (Institut de Recherche en Horticulture et Semences), Angers F-49000, France, pierre.santagostini@agrocampus-ouest.fr*
[2] *Université Mouloud Mammeri, Tizi-Ouzou, Algérie, smail_yousfi@ymail.com*
[3] *INRA, UMR 1345 IRHS, Beaucouzé F-49071, France, sabine.demotes@angers.inra.fr*
[4] *Agrocampus-Ouest, UMR 1345 IRHS, Angers F-49000, France,, rachid.boumaza@agrocampus-ouest.fr*

Résumé. Le package **dad** de l'environnement **R** contient des fonctions pour la gestion et l'analyse de données à trois indices (occasions $\times$ individus $\times$ variables) constituées de T tableaux à n_t lignes ($t = 1, \ldots, T$) et p colonnes. Pour chaque occasion t, les lignes du $t^{ème}$ tableau correspondent à n_t observations d'un p-vecteur aléatoire X_t de densité de probabilité de carré intégrable.

Le package **dad** permet principalement de réaliser les calculs des deux méthodes suivantes :

— analyse en composantes principales fonctionnelle de densités de probabilité (`fpcad`),
— analyse discriminante fonctionnelle de densités de probabilité (`fdiscd.misclass` et `fdisc.predict`).

Les deux méthodes précédentes étant basées sur le produit scalaire de $L^2(\mathbb{R}^p)$, le package propose des fonctions calculant le produit scalaire de deux densités (`l2d`) estimées soit par la méthode du noyau gaussien soit paramétriquement dans le cas gaussien. La présentation du package est illustrée par des exemples.

Mots-clés. Données ternaires, ACP fonctionnelle de densités, analyse discriminante de densités

Abstract. The **dad** package of the **R** environment contains functions for managing and analyzing three-way data (occasions $\times$ individuals $\times$ variables) formed by T tables with n_t rows ($t = 1, ..., T$) and p columns. For each occasion t, the rows of the t^{th} table correspond to n_t observations of the random p-vector X_t whose probability density function is square integrable.

The `dad` package is mainly used to perform calculations of the following two methods :

— functional PCA of probability densities (`fpcad`),
— functional discriminant analysis of probability densities (`fdiscd.misclass` and `fdisc.predict`).

The previous two methods being based on the inner product of $L^2(\mathbb{R}^p)$, the package provides functions calculating the inner product of two densities (`l2d`) estimated either by the Gaussian kernel method or parametrically in the Gaussian case.

The presentation of the package is illustrated by examples.

Keywords. Three-way data, functional PCA of densities, discriminant analysis of densities

1 Présentation des données et objectifs

Dans le package **dad** de l'environnement **R** (2016), les données $\mathbf{X}$ (Tab. 1) auxquelles on s'intéresse sont des données à trois indices : occasions $\times$ individus $\times$ variables. Si T désigne le nombre d'occasions, pour chaque $t \in \{1, \ldots, T\}$, les lignes de $\mathbf{X}_t$ correspondent à n_t observations $\mathbf{x}_{t1}, \ldots, \mathbf{x}_{tn_t}$ d'un p-vecteur aléatoire X_t. Le terme occasion est synonyme des termes groupe, lot…

Dans l'essai de normalisation proposé par Kiers (2000), ce sont des données 3-voies (*three-way data*), à la différence près que les tailles d'échantillons n_t ne sont pas nécessairement toutes égales.

A titre d'exemples fournis dans le package **dad**, on trouve des données :

— d'archéologie (`castles.dated$stones` : pour chacun des $T = 68$ châteaux on mesure $p = 4$ caractéristiques numériques d'un lot de pierres ayant servi à le construire) ;

— de sensométrie (`roses` : pour chacun des $T = 10$ produits, 14 juges ont noté à 3 reprises, $p = 16$ de leurs caractéristiques numériques, d'où un tableau de 42 lignes et 16 colonnes par produit).

L'objectif est de décrire de façon globale les données de la table 1 en visualisant les occasions pour en apprécier les ressemblances / dissemblances. Les données de ce type peuvent être décrites au moyen de l'analyse en composantes principales (ACP) classique (la fonction `princomp` de **R**). Cependant ce type de traitement ne tient pas compte de l'organisation des données en occasions. Elles peuvent aussi être décrites au moyen de deux autres analyses distinctes :

— une portant sur les moyennes des occasions : ACP classique des T moyennes des p variables,

— l'autre portant sur les matrices de variance (ou de corrélation) des occasions : STATIS duale (Lavit et al., 1994) dont les calculs peuvent être réalisés par la fonction `DSTATIS` du package **multigroup** (Eslami et al., 2013) ou par la fonction `statis` du package **ade4**.

La méthode alternative proposée, appelée analyse en composantes principales fonctionnelles (FPCA) de densités de probabilité et rappelée en section 3, est une analyse globale qui prend en compte aussi bien les moyennes que les variances / covariances ou corrélations... Cette analyse repose sur les densités estimées (cf. Section 2) à partir des observations de la table 1.

Aux données précédentes, on ajoute une variable qualitative G à K modalités définie sur l'ensemble des T occasions. Une nouvelle occasion $T+1$ se présente. La valeur de G pour cette occasion $T+1$ est inconnue, l'objectif de la méthode appelée analyse discriminante de densités, rappelée en section 4, est de prédire cette valeur de G au vu des n_{T+1} individus correspondants pour lesquels on a observé les p variables numériques décrites dans la table 1.

TABLE 1 – Pour chaque occasion $t = 1, \ldots, T$, on observe les mêmes p variables quantitatives sur n_t individus.

Occasion		Variables $1 \; \ldots \; p$
1	$\mathbf{x}_{11}$ $\vdots$ $\mathbf{x}_{1n_1}$	$\mathbf{X}_1$
$\vdots$	$\vdots$	$\vdots$
t	$\mathbf{x}_{t1}$ $\vdots$ $\mathbf{x}_{tn_t}$	$\mathbf{X}_t$
$\vdots$	$\vdots$	$\vdots$
T	$\mathbf{x}_{T1}$ $\vdots$ $\mathbf{x}_{Tn_T}$	$\mathbf{X}_T$

2 Estimation des densités associées aux occasions et de leurs produits scalaires

A chaque occasion t ($t = 1, \ldots, T$) est associée la densité f_t, supposée de carré intégrable, qu'on estime par $\hat{f}_t$ au vu des n_t observations correspondantes. On considère les deux cas suivants :

— Si les densités de probabilité f_t sont supposées gaussiennes $N(\mathbf{m}_t, \mathbf{V}_t)$, ses paramètres sont respectivement estimés par $\widehat{\mathbf{m}}_t = n_t^{-1} \sum_i^{n_t} \mathbf{x}_{ti}$ et $\widehat{\mathbf{V}}_t = (n_t - 1)^{-1} \sum_i^{n_t} (\mathbf{x}_{ti} - \widehat{\mathbf{m}}_t)(\mathbf{x}_{ti} - \widehat{\mathbf{m}}_t)'$.

— Si les f_t sont quelconques, elles sont estimées par la méthode du noyau gaussien :

$$\hat{f}_t(\mathbf{z}) = \frac{1}{n_t |\mathbf{h}_t|^{1/2}} \frac{1}{(2\pi)^{p/2}} \sum_{i=1}^{n_t} \exp(-\frac{1}{2}(\mathbf{z} - \mathbf{x}_{ti})' \mathbf{h}_t^{-1} (\mathbf{z} - \mathbf{x}_{ti})) \tag{1}$$

où la matrice non singulière $\mathbf{h}_t$ est la fenêtre de lissage et $|\mathbf{h}_t|$ son déterminant. Cette matrice peut être soit fournie par l'utilisateur, soit directement calculée selon le critère AMISE, minimisant une approximation de l'erreur quadratique moyenne intégrée, en référence à la loi normale (Wand and Jones (1995)), soit

$$\mathbf{h}_t = h_t \widehat{\mathbf{V}}_t^{1/2} \tag{2}$$

avec :

$$h_t = \left(\frac{4}{n_t(p+2)} \right)^{\frac{1}{p+4}}. \tag{3}$$

De cette estimation des densités, on déduit une estimation des produits scalaires et distances entre densités, en s'appuyant sur la bilinéarité du produit scalaire.

3 ACP fonctionnelle de densités de probabilités

On associe à chaque occasion t une densité de probabilité f_t qu'on estime à partir des n_t observations contenues dans le tableau $\mathbf{X}_t$ (Table 1).La FPCA fournit une décomposition, optimale en un certain sens, des T densités estimées $\hat{f}_t$ sur une famille orthonormale de fonctions $\hat{u}_1, \ldots, \hat{u}_L$ ($L \leq T$) de $L^2(\mathbb{R}^p)$ (Ramsay and Silverman (1997)) :

$$\hat{f}_t = \sum_{\ell=1}^{L} \hat{a}_{\ell t} \, \hat{u}_\ell. \tag{4}$$

Les $\hat{a}_{\ell t}$ sont appelés scores principaux et les $\hat{u}_\ell$ les composantes principales. Ils sont déduits des éléments propres de la matrice $\widehat{\mathbf{W}}$ dont le terme général est le produit scalaire entre densités (Section 2) :

$$\widehat{W}_{tr} = \int_{\mathbb{R}^p} \hat{f}_t(\mathbf{z}) \hat{f}_r(\mathbf{z}) d\mathbf{z}. \tag{5}$$

Cette ACP de densités peut aussi être considérée comme une technique de positionnement multidimensionnel (MDS pour “multidimensional scaling”) ou comme une ACP fonctionnelle (FPCA). Les fondements mathématiques de cette méthode peuvent être trouvés dans plusieurs travaux (Delicado, 2011, comme MDS ; Kneip and Utikal, 2001, et Yousfi et al., 2015, comme FPCA). Une méthode d'interprétation des sorties numériques et graphiques de cette technique peut être trouvée en Boumaza et al. (2015b).

Le package **dad** permet d'effectuer tous les calculs nécessaires pour appliquer une telle méthode et en interpréter les sorties :

- la fonction **fpcad** qui réalise l'analyse en composantes principales avec comme sorties par défaut les aides à l'interprétation classiques ;
- la fonction **plot.fpcad** qui réalise les graphiques représentant les densités sur les axes factoriels ;
- la fonction **interpret.fpcad** qui retourne d'autres aides à l'interprétation se basant sur les moments des variables.

4 Analyse discriminante de densités de probabilités

Aux données précédentes, on ajoute une variable qualitative G à K modalités définie sur l'ensemble des occasions. A chaque modalité de G correspond une ou plusieurs occasions.

En notant que la valeur de G pour l'occasion $T+1$ est inconnue, l'objectif de la méthode appelée analyse discriminante de densités est de prédire cette valeur de G au vu des n_{T+1} individus correspondant à l'occasion $T+1$, sur lesquels on a observé les p variables numériques décrites dans la Table 1. Cette méthode a été introduite dans Boumaza (2004). Elle s'appuie principalement sur les distances L^2 entre chaque densité f_t, représentant l'occasion t, et chaque densité g_k, représentant la k^{ieme} modalité de G.

En notant $\{f_t,\ t \in \mathcal{T}_k\}$ les T_k densités f_t appartenant à la classe k de G, trois critères de construction de g_k sont proposés.

- Le premier consiste à considérer que tous les échantillons relatifs aux T_k occasions $t \in \mathcal{T}_k$ constituent un seul échantillon qu'on utilise pour estimer g_k.
- Le second consiste à estimer chaque f_t $(t \in \mathcal{T}_k)$ puis à calculer leur moyenne : $\hat{g}_k = \frac{1}{T_k} \sum_{t\in\mathcal{T}_k} \hat{f}_t$.
- Le troisième consiste à calculer la moyenne des $\hat{f}_t$, chacune étant pondérée par la taille de l'échantillon correspondant : $\hat{g}_k = (1/\sum_{t\in\mathcal{T}_k} n_t) \sum_{t\in\mathcal{T}_k} n_t \hat{f}_t$.

Le package **dad** permet d'effectuer tous les calculs nécessaires pour obtenir :

- le taux d'erreur de classement sur les occasions dont on connait la classe de G à laquelle appartient chacune d'elles, et ce au moyen de la fonction **fdiscd.misclass** ;
- la classe de G à laquelle appartient une occasion de classe inconnue, et ce au moyen de la fonction **fdiscd.predict**.

5 En guise de conclusion

La mise en œuvre des méthodes statistiques rappelées ci-dessus nécessite quelques fonctions de mise en forme des données ternaires qui s'appuient principalement sur les deux classes d'objets `folder` et `folderh` dont on peut trouver la description dans l'aide du package **dad**.

Bibliographie

[1] Boumaza, R. (2004), Discriminant analysis with independently repeated multivariate measurements : an L^2 approach, Computational Statistics and Data Analysis, 47, 4, 2004, 823–843.
[2] Boumaza, R., Santagostini, P., Yousfi, S. and Demotes-Mainard, S. (2015a). dad : Three-Way Data Analysis Through Densities. R package version 1.0.2., `http://CRAN.R-project.org/package=dad`.
[3] Boumaza, R., Yousfi, S. and Demotes-Mainard, S. (2015b), Interpreting the principal component analysis of multivariate density functions, Communications in Statistics - Theory and Methods, 44, 16, 3321–3339.
[4] Delicado P (2011), Dimensionality reduction when data are density functions, Computational Statistics and Data Analysis, 55, 401–420.
[5] Eslami, A., Qannari, E.M., Kohler, A. and Bougeard, S. (2013), General overview of methods of analysis of multi-group datasets, Revue des Nouvelles Technologies de l'Information, 25, 108–123.
[6] Kiers (2000), Towards a standardized notation and terminology in multiway analysis, Journal of Chemometrics, 14, 105–122.
[7] Kneip, A. and Utikal, K. (2001) Inference for density families using functional principal component analysis, Journal of the American Statistical Association, 96, 519–542.
[8] Lavit C, Escoufier Y, Sabatier R and Traissac P (1994), The ACT (STATIS method), Computational Statistics & Data Analysis, 18, 97–119.
[9] R Core Team (2016). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https ://www.R-project.org/.
[10] Ramsay, J. and Silverman, B. (1997), Functional data analysis, Springer, New York.
[11] Wand, M. and Jones, M. (1995), Kernel Smoothing, Chapman an Hall, London.
[12] Yousfi, S., Boumaza, R., Aissani, D. and Adjabi, S. (2015), Optimal bandwith matrices in functional principal component analysis of density functions, Journal of Statistical Computation and Simulation, 85, 11, 2315-2330.