1ères Rencontres R

Le 2 et 3 juillet 2012

Recueil des résumés

Inria INVENTEURS DU MONDE NUMÉRIQUE

UNIVERSITÉ DE BORDEAUX

Institut de Mathématiques de Bordeaux

Table des matières

Pourquoi R devient incontournable en enseignement, recherche et développement ?

E. Matzner-Løber[a]

[a]Lab. Math. Appl. Agrocampus Ouest et Univ. Rennes 2
Univ. Rennes 2
Av. Gaston Berger, 35043 Rennes Cedex
eml@uhb.fr

Mots clefs : Statistique, R.

Dans cette conférence, nous aborderons le logiciel R sous trois aspects : son utilisation en enseignement, en recherche et dans le monde de l'entreprise. A chacune de ces trois utilisations, il est nécessaire d'évaluer les demandes spécifiques du domaine et les réponses qu'apporte R.
Rappelons brièvement que R est multi-plateforme et multi-OS. Il est entièrement gratuit, très complet et offre à la fois des commandes mais aussi des menus déroulants. Il est donc raisonnable de penser que ce logiciel fera partie des logiciels de statistique les plus enseignés.
Sa facilité de programmation et sa forte utilisation dans le monde de la recherche en font dès aujourd'hui un langage omniprésent. On peut donc s'interroger sur les futures évolutions de R vis à vis de la recherche en statistiques, sur la création de librairies de fonctions comme outil de diffusion de la recherche.
Dans la troisième partie, nous comparerons R avec ses différents concurrents et analyserons les points qui gouvernent les choix de logiciels en entreprise (prix, interface graphique, intégration dans les base de données, etc.). Bien évidemment le logiciel est loin d'être parfait mais il comporte dès à présent des avantages qui lui valent d'être adopté par un nombre croissant d'entreprises.

Références

[1] Cornillon P-A, Guyader, A., Husson, F., Jégou, N., Josse, J., Kloareg, M., Matzner-Løber, E. et Rouvière, L. (2012) **Statistiques avec R**, troisième édition, PUR, France.
[2] Cornillon P-A, Guyader, A., Husson, F., Jégou, N., Josse, J., Kloareg, M., Matzner-Løber, E. et Rouvière, L. (2012) **R for Statistics**, Chapman, USA.
[3] Lafaye de Micheaux, P. Drouilhet, R. , Liquet, B., (2010) **Le logiciel R**, Springer, France.

Optimisation de protocoles dans les modèles non linéaires à effets mixtes avec PFIM : application aux études pharmacocinétiques chez l'enfant

C. Dumont et F. Mentré

Univ Paris Diderot, Sorbonne Paris Cité, INSERM, UMR 738
16 rue Henri Huchard
75018 Paris
cyrielle.dumont@inserm.fr

Mots clefs : Matrice d'information de Fisher, Modèles non linéaires à effets mixtes, Optimisation de protocoles, Pédiatrie, PFIM, Pharmacocinétique.

Contexte : Dans le cadre du Plan d'Investigation Pédiatrique [1] en vigueur depuis 2007, les études pharmacocinétiques (PK) de médicaments en développement doivent être réalisées chez l'enfant et il est recommandé de les analyser via des modèles non linéaires à effets mixtes (MNLEM) [2,3]. Le choix du protocole PK, consistant à trouver un compromis entre le nombre de sujets, le nombre de prélèvements par sujet et le choix des temps, influe sur les résultats de l'étude. Par conséquent, des approches basées sur l'évaluation de la matrice d'information de Fisher (M_F) [4] ont été proposées. Nous les avons implémentées dans la fonction PFIM [5,6] en R, fonction dédiée à l'évaluation et l'optimisation de protocoles dans les MNLEM avec une librairie spécifique de modèles PK/pharmacodynamiques. Il est possible dans PFIM d'avoir recours à des modèles incluant une variabilité inter-occasion et de prendre en compte des covariables discrètes sur les paramètres. La puissance attendue du test de Wald pour le test de comparaison ou d'équivalence et le nombre de sujets nécessaire pour une puissance donnée peuvent être calculés.

Objectifs : Dans ce travail, PFIM a été utilisé pour planifier des études PK chez l'enfant d'une molécule en développement. Les objectifs ont été de i) proposer une extension de M_F dans PFIM en considérant une covariance entre les effets aléatoires et étudier l'impact de sa valeur sur les erreurs standards (SE) et sur la quantité d'information ; ii) optimiser les temps de prélèvements du protocole PK ; iii) prendre en compte les données sous la limite de quantification (LOQ) ; iv) développer avec PFIM des protocoles adaptatifs [7] en deux étapes [8].

Méthode : Les concentrations de la molécule parent et de son métabolite, que l'on sait actif d'après l'étude chez l'adulte, ont été simulées chez l'enfant à l'aide d'un modèle physiologique développé avec le logiciel SIMCYP. L'extension de M_F pour les MNLEM, incluant la covariance entre les effets aléatoires [9], a été implémentée dans une version de travail de PFIM. Nous avons prédit les SE des effets fixes et des paramètres de variance du modèle PK en supposant différentes valeurs de covariances et évalué l'information totale via le déterminant de M_F. Le protocole PK, pour une prochaine étude chez 82 enfants, a été optimisé, à l'aide de l'algorithme du Federov-Wynn dans PFIM, en prenant en compte plusieurs contraintes cliniques. La LOQ n'a pas été prise en compte pour l'optimisation du protocole. Pour l'évaluation finale du protocole proposé, une approche a été

développée, consistant à calculer la proportion de données simulées sous la LOQ à chaque temps du protocole optimal. Ce protocole prenant en compte la LOQ a été comparé à celui ne tenant pas compte de la LOQ. Un protocole adaptatif en deux étapes a ensuite été développé. Cela a nécessité le développement de M_F dans PFIM dans le cas de l'optimisation en deux étapes. L'impact d'un protocole adaptatif en deux étapes est évalué, notamment quand les vrais paramètres sont différents des paramètres simulés initiaux. Une étude de simulation est en cours pour évaluer l'impact de la taille de chacune des deux cohortes sur la précision d'estimation des paramètres.

Résultats : Dans le contexte de l'évaluation de protocoles, nous avons montré que la valeur de la covariance entre les effets aléatoires du modèle PK n'affectent ni les valeurs des SE des paramètres d'effets fixes, ni celles des paramètres de variance. Cependant, la quantité d'information augmente lorsque la covariance augmente. De plus, les résultats ont montré que la taille de la covariance influe sur le protocole optimal. PFIM a permis d'éviter la mise en place de protocoles peu informatifs et de souligner l'importance d'un temps tardif. En ce qui concerne l'optimisation de protocole adaptatif, le protocole en une étape, obtenu à partir des paramètres initiaux, montre une perte d'efficacité quand les vrais paramètres sont différents des paramètres initiaux. Le protocole en deux étapes permet de compenser en partie cette perte d'information. De plus, la taille respective de chaque cohorte influence le gain d'efficacité du protocole en deux étapes.

Conclusion : PFIM dans R est un bon outil pour évaluer et optimiser des protocoles pour des analyses par MNLEM. Le nouveau calcul de M_F combinée nous permet de mener des optimisations en deux étapes. Ces protocoles sont relativement faciles à mettre en oeuvre et sont une bonne alternative pour mener des études PK chez l'enfant pour lesquelles on dispose de peu d'information.

Références:

[1]www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003066.pdf

[2] Mentré F, Dubruc C, Thénot J.P. Population pharmacokinetic analysis and optimization of the experimental design for Mizolastine solution in children. *J Pharmacokinet Pharmacodyn*, 2001; 28(3): 299-319.

[3] Tod M, Jullien V, Pons G. Facilitation of drug evaluation in children by population methods and modelling. *Clin Pharmacokinet*, 2008; 47(4): 231-243.

[4] Mentré F, Mallet A, Baccar D. Optimal design in random-effects regression models. *Biometrika*, 1997; 84(2): 429-442.

[5] Bazzoli C, Retout S, Mentré F. Design evaluation and optimisation in multiple response nonlinear mixed effect models: PFIM 3.0. *Comput Methods Programs Biomed*, 2010; 98(1): 55-65.

[6] www.pfim.biostat.fr

[7] Foo L K, Duffull S. Adaptive optimal design for bridging studies with an application to population pharmacokinetic studies. *Pharm Res*, 2012; in press.

[8] Federov V, Wu Y, Zhang R. Optimal dose-finding designs with correlated continuous and discrete responses. *Stat Med*, 2010; 31: 217-234.

[9] Ogungbenro K, Graham G, Gueorguieva I, Aarons L. Incorporating correlation in interindividual variability for the optimal design of multiresponse pharmacokinetic experiments. *J Biopharm Stat*, 2008; 18(2): 342-358.

Variables latentes dans les modèles linaires généralisés

D. Thiam[a] et **G. Nuel**[a,b]

[a]Labo de Maths Appliquées (MAP5, CNRS 8145)
Université Paris Descartes
djenaba.thiam@gmail.com

[b] Institut des Maths et Interactions (INSMI)
CNRS Paris
gregory.nuel@parisdescartes.fr

Mots clefs : Variables Latentes, modèles linaires généralisés, algorithme EM.

En sciences sociale comme en biologie, de nombreux phénomènes d'intérêt ne sont pas observés directement et son modélisés par des variables latentes [1]. Un exemple est celui de données non observées en raison d'un seuil de détection au niveau de l'appareil de mesure [2]. Une autre illustration est celle des données hétérogènes qui peuvent être traitées en introduisant une classe latente [3].

Parmi les packages disponibles sous R qui permettent la gestion des variables latentes dans les modèles linéaires généralisés (GLMs) on peut citer `flexmix` [4] dans le cadre de modèles de régression et `lcmm` [5] dans le cadre de modèles à effets mixtes. Si ces packages sont puissants et indéniablement utiles, ils ont le défaut notable de limiter les possibilités de modélisation aux cas implémentés par les développeurs. Dans le cas de `flexmix` par exemple, on ne peut considérer qu'un simple modèle de mélange (avec variables concomitante pour la classe). Que faire si cette classe latente intervient de manière hiérarchique? Comment gérer des paramètres partagés à travers différentes classes? Que faire si un GLM particulier n'est pas implémenté (ex: regression multinomiale)? Et comment peut-on introduire des variable latentes continues (autres que des effets mixtes)?

L'objectif de ce travail est de répondre à ces question en présentant une approche simple et généraliste qui permet de gérer sous R tout type de variables latentes dans les modèles linéaires généralisés sans recourir pour cela à une implémentation spécifique. Notre approche consiste à utiliser un classique algorithme Expectation-Maximization (EM) [6] sans avoir à entrer au cœur des méthodes d'estimations. La clef est une utilisation astucieuse de l'option `weights` des procédures d'estimation classiques (ex: `lm`, `glm`, `lmer` et `glmer`), ces poids étant mis à jour impérativement à l'extérieur de la procédure d'estimation.

Pour illustrer cette approche, considérons le modèle suivant:

$$\mathtt{y} \sim \mathtt{x} + \mathtt{z}$$

où le `y` est une variable de réponse ($\in \mathbb{R}^n$), `x` est une covariable ($\in \mathbb{R}^n$), et `z` une variable latente binaire ($\in \{0,1\}^n$). Si `z` était connu, un simple $\mathtt{fit} = \mathtt{lm}(\mathtt{y} \sim \mathtt{x} + \mathtt{z})$ permettrait d'ajuster ce modèle. La valeur de `z` étant manquante, on se tourne vers l'algorithme EM. Supposons qu'à une itération donnée de l'algorithme on dispose d'un paramètre θ (contient les paramètres du modèle linéaire ainsi que la proportion *a priori* ρ de $\mathtt{z}[\mathtt{i}] = 1$), il nous suffit alors de remplacer

le paramètre courant par:

$$M(\theta) = \arg\max_{\theta'} \underbrace{\sum_{\mathtt{z}} \mathbb{P}(\mathtt{z}|\mathtt{x},\mathtt{z};\theta) \log \mathbb{P}(\mathtt{y}|\mathtt{x},\mathtt{z};\theta')}_{Q(\theta'|\theta)}$$

Or cette étape est en fait équivalente à l'ajustement du modèle:

$$\begin{pmatrix}\mathtt{y}\\ \mathtt{y}\end{pmatrix} \sim \begin{pmatrix}\mathtt{x}\\ \mathtt{x}\end{pmatrix} + \begin{pmatrix}\mathtt{z}=1\\ \mathtt{z}=0\end{pmatrix} \quad \text{avec} \quad \mathtt{weights} = \begin{pmatrix}\mathtt{w}\\ 1-\mathtt{w}\end{pmatrix}$$

où $\mathtt{w} = \mathbb{P}(\mathtt{z}=1|\mathtt{x},\mathtt{z};\theta)$. On peut ainsi facilement mettre à jour les paramètres de la régression à l'aide de la commande: $\mathtt{fit} = \mathtt{lm}(\mathtt{c}(\mathtt{y},\mathtt{y}) \sim \mathtt{c}(\mathtt{x},\mathtt{x}) + \mathtt{c}(\mathtt{z}=1,\mathtt{z}=0), \mathtt{weights} = \mathtt{c}(\mathtt{w}, 1-\mathtt{w}))$. Le paramètre ρ de la variable latente peut quant à lui facilement être mis à jour directement à partir de `w`: $\rho = \mathtt{mean}(\mathtt{w})$.

Dans le cas d'une variable latente discrète, l'approche proposée est totalement équivalente à un algorithme EM classique (y compris en termes de complexité). Pour les variable latentes continues, notre approche se ramène à une approximation (on se contente de répliquer `z` pour un nombre donné de valeurs qui sont spécifiques à chaque individus et à chaque itération). La méthode est évidemment généralisable aux modèles linéaires généralisés plus complexes, seul le calcul et la mise à jours des poids restant à la charge de l'utilisateur.

Avec la méthodologie proposée, nous montrons qu'une exploitation astucieuse de l'option `weights` d'une procédure de R permet d'introduire de manière très souple des variables latentes dans cette procédure (ici l'ajustement de modèles linéaires généralisés). Au delà de cet exemple particulier, l'approche que nous suggérons devrait inciter le développeur de toute procédure statistique sous R à se poser la question de la prise en compte d'observations pondérées par des poids avec la perspective d'une future exploitation de cette fonctionnalités dans le cadre de l'algorithme EM et de ses variantes.

Références

[1] Kenneth A. Bollen (2002). Latent variables in psychology and social sciences. *Annual Review of Psychology* ,**53**, 605-634

[2] Goodman L.A(1974). The analysis of systems of qualitative variables
when some of the variables are unobservable *JAmerican Journal of Sociology* , **79**, 1179-1259

[3] Bert F. Green (1951). A general solution for the latent class model of latent structure analysis *PSYCHOMETRIKA* , **16**, 151-166

[4] Friedrich Leisch (2003). FlexMix: A general framework for finite mixture models and latent class regression in R. *Report* , **86**

[5] Cecile Proust-Lima, Benoit Liquet (2009). lcmm: an R package for estimation of latent class mixed models and joint latent class models, *R cran*.

[6] A. P. Dempster; N. M. Laird; D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm *Journal of the Royal Statistical Society* , **39**, 1-38

New mixture models and algorithms in the mixtools package

Didier Chauveau

MAPMO - Fédération Denis Poisson
Université d'Orléans and CNRS UMR 7349
BP 6759, 45067 Orléans cedex 2
didier.chauveau@univ-orleans.fr

Mots clefs : Finite mixture, nonparametric mixtures, EM algorithms

The mixtools package for the R statistical software [7] has evolved from 2006 up to the current CRAN version. Benaglia et al. [2] give a comprehensive account of mixtools capabilities as in 2009. This package provide various tools for analyzing a variety of finite mixture models, from traditional methods such as EM algorithms for uni- and multivariate Gaussian mixtures, up to more specific and recent models such as, e.g., multinomial mixtures, mixtures of regression or multivariate non-parametric mixtures.
Since then, new and different models connected to mixtures have been investigated by several authors, some of those involved in mixtools' development. For most of these model analysis, new or specific computational techniques have been progressively implemented in the development version of the package, taking advantage of its environnement. This talk, that involves joint works with the co-authors cited below, presents some of these models and illustrate mixtools' new capabilities that have been added since the publication of Benaglia et al. [2]. These new models share in common the description of the distribution of the observations by a finite mixture density

$$g(x|\boldsymbol{\theta}) = \sum_{j=1}^{m} \lambda_j f_j(x), \quad \boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{f}), \quad x \in \mathbb{R}^r,$$

where $\boldsymbol{\theta}$ is the model parameter, consisting in the *component densities* f_j's and component weights λ_j's that are positive and sum to unity. Precise specification of $\boldsymbol{f}$ depends on the model assumptions, e.g., for univariate normal mixtures f_j is the density of $\mathcal{N}(\mu_j, \sigma_j^2)$ and $\boldsymbol{f} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, the m-vectors of component means and variances.

Gaussian mixtures with constrained parameters: Motivated by mixture models issued from psychometrics, Chauveau and Hunter [5] consider the problem of linear constraints on the parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ for finite mixtures of normal components. Surprisingly, we show that even for simple linear constraints on $\boldsymbol{\mu}$ such as $\boldsymbol{\mu} = M\boldsymbol{\beta}+C$ for some unknown p-vector $\boldsymbol{\beta}$ with $p \leq m$, and known matrix M and vector $\boldsymbol{C}$, the Maximum Likelihood Estimation problem succumbs to an ECM (with Conditional-M steps) generalization of the EM algorithm. With certain types of variance constraints, a further generalization of EM known as MM (Majorization-Minorization) algorithm have also been added in mixtools.

Nonparametric MM for smoothed likelihood maximization: Benaglia et al. [1] originally designed an empirical "nonparametric EM" (npEM) algorithm for fitting multivariate nonparametric mixtures with completely unspecified component densities except for a conditional independence assumption $f_j(x) = \prod_{k=1}^{r} f_{jk}(x_k)$ which means that the (scalar) coordinates of the r-dimensional observation x are independent conditional on the component from which x

is drawn. Despite its superiority over competing methods shown by numerical evidence, this npEM algorithm which is in the spirit of an EM in its formulation lacks any sort of theoretical justification. Following this work, Levine et al. [6] have proposed and implemented a new MM algorithm which does provide an ascent property (just as a genuine EM does) with respect to a smoothed loglikelihood, at the cost of a higher computing load. Both versions are know available in mixtools.

Reliability mixture models on randomly censored data: Mixtures are also suitable to modelize lifetime data, but these data are often censored. Randomly censored data from mixture models have been considered in Bordes and Chauveau [3], both for parametric or semiparametric mixtures. They propose several algorithms, from genuine parametric EM for specific families, to parametric and semiparametric Stochastic EM (St-EM). These stochastic versions, that include an additional step for simulating the missing part of the data, provide workable estimation methods since completion of the data for the component indicators allows application of nonparametric estimates for survival data such as the Kaplan-Meier estimate. Most of these algorithms are already implemented in the development version of mixtools.

Semiparametric EM with one component known: In multiple testing and False Discovery Rate estimation, semiparametric mixtures with one component known can be used (Bordes et al. [4]). In Saby et al. [8], some new EM-like algorithms of this kind have been implemented in mixtools and tested on simulated and actual data.

References

[1] Benaglia, T., Chauveau, D., and Hunter, D. R. (2009a). An EM-like algorithm for semi-and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526.

[2] Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009b). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

[3] Bordes, L. and Chauveau, D. (2010). Some algorithms to fit some reliability mixture models under censoring. In Lechevallier, Y. and Saporta, G., editors, *Proceedings of the 19th International Conference on Computational Statistics, Paris France*. Springer.

[4] Bordes, L., Delmas, C., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Statistics*, 33:733–752.

[5] Chauveau, D. and Hunter, D. R. (2011). ECM and MM algorithm for mixtures with constrained parameters. Technical Report hal-00625285, version 1, HAL.

[6] Levine, M., Hunter, D. R., and Chauveau, D. (2011). Smoothed likelihood for multivariate mixtures. *Biometrika*, 98(2):403–416.

[7] R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[8] Saby, N., Orton, T. G., Chauveau, D., Lemercier, B., Walter, C., Schvartz, C., and Arrouays, D. (2011). Application of a mixture model approach to large-scale simultaneous hypothesis testing in soil monitoring. In *Pedometrics*.

PairedData 0.9 : Un package R en S4 pour analyser les données numériques appariées

S. Champely

Sciences et Techniques des Activités Physiques et Sportives
Centre de Recherche et d'Innovation sur le Sport
27-29, Boulevard du 11 Novembre 1918, 69622 Villerubanne cedex.
champely@univ-lyon1.fr

Mots clefs : Données appariées, Graphiques, Robustesse, Sport, S4.

Le dispositif apparié est l'un des plus utilisés en sciences du sport afin d'augmenter la puissance de comparaisons du type : avant *vs.* après entraînement, main droite *vs.* main gauche, temps réels *vs.* temps imaginés (*i.e.* un facteur intra à deux niveaux). Il est également classique de réaliser ces comparaisons sur plusieurs groupes (*i.e.* un facteur inter, le plus souvent à deux niveaux également : groupe traité *vs.* témoin).
L'analyse statistique employée est immanquablement un ou une combinaison de tests de Student appariés voire une analyse de variance à mesures répétées. Cependant, cette analyse élémentaire cache parfois diverses difficultés [8] : points extrêmes, différence de dispersion, multimodalité, hétéroscadasticité. Le problème principal est de repérer ces situations et, pour ce faire, la visualisation est un outil privilégié [9]. S'il existe dans la littérature plusieurs techniques graphiques dévolues aux données appariées ([10], [6], [3], [7]), elles sont rarement présentes dans les logiciels statistiques. Le package PairedData 0.5, en employant les outils du package ggplot2 [11], réunit ces propositions (cf. Figure 1) et les étend en autorisant la prise en compte de groupes.
Une fois les structures repérées graphiquement, le calcul de statistiques descriptives en permet une discussion plus précise. La tendance dans les publications scientifiques dans certaines disciplines (Médecine, Psychologie) des sciences du sport est de seconder les résumés habituels (moyenne, écart-type) par des tailles d'effet standardisées. Or, ces statistiques étant plus ou moins sensibles à la normalité des données, des versions robustes sont préférables. Les propositions d'Algina *et al.* [1] sont intégrées dans le package PairedData 0.5. Au delà de ces descriptions, les tests d'hypothèses permettant de comparer la centralité, mais aussi la dispersion, accompagnés par les intervalles de confiance associés, sont également peu disponibles. PairedData réunit les versions classiques (Student, Pitman, Morgan) et des alternatives robustes moins connues ([13], [12], [5], [2]).
L'un des objectifs du package est de servir d'outil pédagogique. Aussi, plusieurs jeux de données typiques des sciences du sport (biomécanique, psychologie, neurosciences) sont proposés, ainsi qu'un exemple illustrant les pièges du test de Student apparié (voir également Figure 1). Une version (RcmdrPlugin.PairedData), pour l'instant confidentielle, intégrée au package Rcmdr [4] permet aux utilisateurs moins avertis d'utiliser ces outils d'analyse de données appariées à l'aide d'une interface graphique en menus déroulants.
Enfin, une nouvelle version (0.9) en S4 du package PairedData est en construction. Basée sur la création d'un objet "paired", de type "dataframe", mais contraint à deux colonnes et possédant des "class" identiques (et même des "levels" identiques dans le cas de données catégorielles), cette version permet d'employer les génériques classiques (`show`, `summary`, `plot`) et de construire des méthodes adaptées à la fonction `t.test`, ainsi qu'à `var.test`.

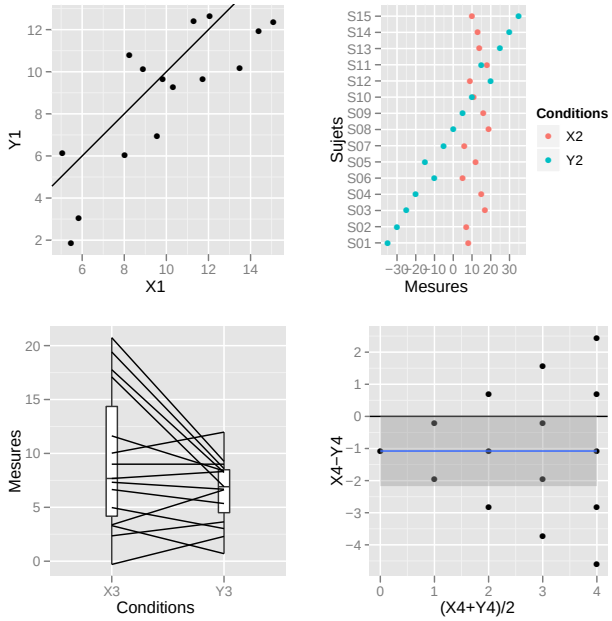


Figure 1: Divers graphiques adaptés au cas de données numériques appariées. Les données reflètent successivement diverses situations problématiques du test de Student apparié

Références

[1] Algina, J., Keselman, H.J., Penfield, R.D. (2005). Effect Sizes and their Intervals: the Two-Level Repeated Measures Case. *Educational and Psychological Measurement*, **65**, 241−258.

[2] Bonett, D.G., Seier, E. (2003). Statistical Inference for a Ratio of Dispersions using Paired Samples. *Journal of Educational and Behavioral Statistics*, **28**, 21−30.

[3] Cox, N. (2004). Speaking Stata: Graphing Agreemeent and Disagreement. *The Stata Journal*, **4**, 329−349.

[4] Fox, J. (2005). The R Commander: A Basic Statistics Graphical User Interface to R. *Journal of Statistical Software*, **14** (9), 1−42.

[5] Grambsch, P.M. (1994). Simple Robust Tests for Scale Differences in Paired Data. *Biometrika*, **81**, 359−372.

[6] McNeil, D.R. (1992). On Graphing Paired Data. *The American Statistician*, **46**, 307−310.

[7] Meek, D.M. (2007). Two Macros for Producing Graphs to Assess Agreement Between Two Variables. In Proceedings of Midwest SAS Users Group Annual Meeting.

[8] Preece, D.A. (1982). t is for Trouble (and Textbooks): a Critique of some Examples of the Paired-Samples t-Test. *The Statistician*, **31**, 169−195.

[9] Pruzek, R.M., Helmreich, J.E. (2009). Enhancing Dependent Sample Analyses with Graphics. *Journal of Statistics Education*, **17**.

[10] Rosenbaum, P.R. (1989). Exploratory Plot for Paired Data. *The American Statistician*, **43**, 108−110.

[11] Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

[12] Wilcox, R.R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*. San Diego: Academic Press.

[13] Yuen, K.K. (1974). The Two-Sample Trimmed t for Unequal Population Variances. *Biometrika*, **61**, 165−170.

Une interface graphique pour analyser des données distantes sous R

R. Coudret[a] and **G. Durrieu**[b] and **J. Saracco**[a]

[a]Équipe CQFD et Institut de Mathématiques de Bordeaux UMR CNRS 5251
INRIA et Université de Bordeaux
351 cours de la Libération, 33405 Talence
{rcoudret, Jerome.Saracco}@math.u-bordeaux1.fr

[b]Laboratoire de Mathématiques de Bretagne Atlantique UMR CNRS 6205
Université de Bretagne Sud
Campus de Tohannic, 56017 Vannes
gilles.durrieu@univ-ubs.fr

Mots clefs : Analyse de données, Base de données, Interface graphique.

Utiliser ou faire utiliser des méthodes statistiques récentes pour analyser un jeu de données peut représenter une tâche ardue. Lorsqu'il s'agit de s'intéresser à des données dont la diffusion doit être contrôlée ou qui représentent un volume important, la question de leur accès et de leur stockage peut également se poser. Dans ce cadre, grâce à l'utilisation combinée des *packages* `RMySQL` et `RGtk2`, qui permettent de relier R à MySQL et à Gtk+, nous proposons une méthode pour concevoir une interface graphique pour des algorithmes d'analyse statistique lorsque les données ne sont pas situées sur la machine de l'utilisateur. Dans une première partie, nous décrirons le système de gestion de base de données (SGBD) MySQL ainsi que le *package* `RMySQL` qui lui est associé. Nous nous intéresserons ensuite à la librairie GTK+, relative à la création d'interface graphique et au *package* `RGtk2` qui permet d'y accéder sous R. Nous terminerons par un exemple d'analyse de données qui tire profit de ces deux logiciels.

1 Accéder aux données

MySQL est un SGBD utilisable gratuitement hors d'un contexte commercial. Il possède de nombreux concurrents dont PostgreSQL et Ingres, sous licence libre. Il permet de stocker des données sur un serveur tout en les rendant disponibles via Internet. L'accès à un serveur MySQL est sécurisé de telle sorte que l'administrateur peut définir, pour chaque utilisateur, les éléments de la base de données sur lesquels il a la permission de travailler. Pour se connecter au serveur MySQL il est nécessaire d'avoir installé un client sur son ordinateur. La principale manière d'interagir avec MySQL réside en l'utilisation de scripts au format SQL (Structured Query Language). L'exécution de ces scripts, aussi appelées requêtes, peut se faire à l'aide d'une interface textuelle, disponible après exécution dans un terminal de la commande `mysql`. Le *package* `RMySQL` offre l'opportunité d'envoyer des scripts SQL au serveur MySQL, et d'en récupérer les résultats, avec R. Ceci est réalisé avec la fonction `dbSendQuery()`, sur laquelle nous nous focaliserons. Nous nous en servirons d'une part pour montrer comment il est possible de récupérer des données depuis le serveur MySQL afin de les traiter, et d'autre part pour expliquer comment faire pour enregistrer les résultats de ces analyses sur le serveur MySQL. Cette dernière fonctionnalité est avantageuse dans le cadre de travaux en équipe, par exemple.

2 Des méthodes d'analyse dans une interface graphique

GTK+ est un ensemble de librairies sous licence libre permettant la création d'interfaces graphiques. L'environnement de bureau GNOME et le logiciel de traitement d'images GIMP sont des exemples de logiciels basés sur GTK+. Les objets graphiques disponibles avec ce dernier sont très variés et comprennent entre autres des boutons, des barres de progressions, des zones d'affichage et des zones de texte. Ces objets peuvent être disposés dans une fenêtre selon la volonté du créateur de l'interface grâce à un logiciel approprié, comme par exemple Glade. Le fichier XML généré par ce logiciel peut alors être lu depuis R grâce au *package* `RGtk2` et à la fonction `gtkBuilder()`. Le comportement des objets graphiques doit alors être géré par des programmes en R grâce à la fonction `gSignalConnect()`. `RGtk2` permet également d'afficher des graphiques R dans des zones d'affichage de l'interface graphique en utilisant le *package* `cairoDevice` et la fonction `asCairoDevice()`.

3 Application à des mesures biologiques

MySQL et GTK+ sont tous deux utilisés dans l'étude de mesures, au cours du temps, de distances entre les deux parties de coquilles d'huîtres. Ces animaux sont analysés dans la baie d'Arcachon en France et sur plusieurs autres sites comme Santander en Espagne, Locmariaquer en France ou encore Tromsø en Norvège, par le laboratoire EPOC (Environnements et Paléoenvironnements Océaniques et Continentaux) UMR CNRS 5805. Les données sont acquises à une fréquence de 0.625 Hz pour chaque animal. Des liens entre la qualité de l'eau et ces signaux ont été mis en évidence (voir par exemple Tran et al. [4]). Afin de déterminer si une huître est en bonne santé, nous estimons la densité de probabilité f des distances entre les deux parties de sa coquille.
L'estimation est réalisée grâce à un estimateur à noyau $\hat{f}_{K,h}$ (voir Parzen [1]), défini quel que soit $t \in \mathbb{R}$ par :

$$\hat{f}_{K,h}(t) := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right).$$

De tels estimateurs ont déjà été étudiés dans ce contexte (voir par exemple Sow et al. [3]). Ceux-ci requièrent le choix d'un noyau K et d'une fenêtre de lissage h. Pour K, nous prenons le noyau gaussien, défini par $K(t) := \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. Pour h, nous choisissons la fenêtre de lissage critique de Silverman [2] donnée par $h_{crit} := \min_{N(\hat{f}_{K,h})=N(f)}(h)$, où la fonction N associe son nombre de modes à une densité de probabilité. Ainsi, l'utilisation de h_{crit} nécessite une hypothèse concernant $N(f)$. Nous supposons donc que $N(f) = 2$. Sur des exemples d'huîtres mourantes ou saines, nous détaillerons les caractéristiques couramment observées de $\hat{f}_{K,h_{crit}}$.

References

[1] Parzen, E. (1962). On estimation of probability density function and mode, *The Annals of Mathematical Statistics*, **91**, 115–132.
[2] Silverman, B. W. (1986). Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society. Series B (Methodological)*, **43**(1), 97–99.
[3] Sow, M., Durrieu, G., Briollais, L. (2011). Water quality assessment by means of HFNI valvometry and high-frequency data modeling. *Environmental Monitoring and Assessment*, **182**, 155–170.
[4] Tran, D., Fournier, E., Durrieu, G., Massabuau, J.-C. (2003), Copper detection in the Asiatic clam Corbicula fluminea: Optimum valve closure response. *Aquatic Toxicology*, **65**, 317–327.

Analyse de $(K+1)$ tableaux avec le logiciel `ade4`. Application en épidémiologie.

S. Bougeard[a] and **S. Dray**[b]

[a]Département d'épidémiologie
Anses (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail)
Zoopôle, 22440 Ploufragan, France
stephanie.bougeard@anses.fr

[b]Laboratoire de biométrie et biologie évolutive
CNRS - Université Lyon 1
UMR CNRS 5558, 43 bd du 11 novembre 1918, 69622 Villeurbanne, France
stephane.dray@univ-lyon1.fr

Mots clefs : Statistique, Biologie, Régression multi-tableaux, `Ade4`.

Dans de nombreux domaines, sont recueillies des données présentant une structure en plusieurs tableaux dont il convient de tenir compte à la fois pour le traitement statistique mais aussi pour l'interprétation. Nous traitons ici du cas où celles-ci sont organisées en $(K+1)$ tableaux, à savoir un tableau Y comprenant plusieurs variables à expliquer et K tableaux $(X_1, \ldots, X_K)$ comprenant chacun plusieurs variables explicatives, l'ensemble de ces variables étant mesuré sur les mêmes observations. Les variables sont supposées quantitatives, mais des variables qualitatives peuvent être intégrées après codage disjonctif. Dans la bibliographie, les principaux domaines dans lesquels ces données sont décrites sont le suivi de processus industriels, la chimiométrie, l'analyse sensorielle, les études de marché, l'écologie et l'épidémiologie (*e.g.* Figure 1).

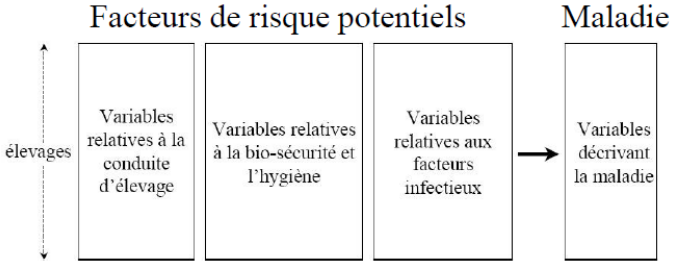


Figure 1: Exemple de données d'épidémiologie vétérinaire organisées en $(K+1)$ tableaux.

Les données multi-tableaux étant complexes, les objectifs poursuivis peuvent être multiples. Dans les domaines décrits précédemment, l'objectif majeur est de réaliser une description ainsi qu'une prédiction du tableau Y à partir des K tableaux explicatifs $(X_1, \ldots, X_K)$. Le traitement statistique de ce type de données pose actuellement problème car aucune méthode adaptée n'est actuellement implémentée dans des logiciels, qu'ils soient libres ou commerciaux. Deux types de solutions peu satisfaisantes pour l'utilisateur restent possibles : (i) simplifier la structure de ces données et appliquer une méthode développée pour deux tableaux Y et X, *e.g.* la régression PLS (package `pls` de `R`) ou l'analyse des redondances (*e.g.*, fonction `pcaiv` du package `ade4`), (ii) ou au contraire utiliser des méthodes développées pour des données plus complexes, *e.g.* l'approche PLS (package `plspm` de `R`) avec l'inconvénient d'un algorithme complexe dont la convergence n'est pas démontrée. Le logiciel d'analyse de données `ade4` [3] propose un ensemble de méthodes

d'analyse de données pour le traitement d'un seul, de deux mais aussi de K tableaux [2].

Afin de proposer aux utilisateurs des méthodes adaptées au traitement de $(K+1)$ tableaux, deux d'entre elles ont été récemment développées et implémentées dans le logiciel `ade4` : la régression PLS multibloc [4, 5] et l'analyse en composantes principales sur variables instrumentales multibloc, aussi appelée analyse des redondances multibloc [1]. La régression PLS multibloc (fonction `mbpls`) a été choisie pour sa popularité et sa stabilité pour le cas de variables explicatives nombreuses et corrélées, mais il est démontré que pour le cas d'un seul tableau Y ses principaux résultats sont ceux de la régression PLS classique. L'analyse en composantes principales sur variables instrumentales multibloc (fonction `mbpcaiv`) a été choisie pour sa bonne adaptation aux données structurées en $(K + 1)$ tableaux ayant une visée prédictive, mais peut présenter des limites en cas de quasi-colinéarité marquée au sein des tableaux explicatifs. Pour pouvoir utiliser ces deux fonctions, il convient de définir le tableau Y comme un objet de la classe `dudi` (classe d'objet `ade4` pour les données organisées en un seul tableau) et le tableau X comme un objet `ktab` (classe d'objet `ade4` pour les données structurées en K tableaux). En complément et dans l'objectif de valoriser au mieux les nombreux résultats issus des méthodes multi-tableaux, des outils d'aide à l'interprétation pour la description mais aussi pour l'explication sont spécifiquement développés. (i) Du point de vue descriptif, la fonction `summary` fournit pour chaque dimension, l'inertie et la variance expliquée de chaque tableau par les variables latentes. Pour entrer dans le détail des liens entre variables et entre tableaux en lien avec les observations, des représentations factorielles graphiques sont proposées par la fonction `plot`. (ii) Du point de vue explicatif, la fonction `testdim` permet à l'utilisateur de choisir la dimension optimale du modèle par validation croisée. Une fois cette dimension définie, il est possible de calculer et de représenter les intervalles de confiance des principaux indices d'aide à l'interprétation, *i.e.* coefficients de régression, importance globale des variables explicatives et importance globale des tableaux explicatifs.

Une application est proposée en épidémiologie vétérinaire. Les données proviennent d'une enquête analytique menée sur un échantillon de 351 lots de poulets. L'objectif est de comprendre les facteurs de risques globaux des pertes (Y) décrites par quatre variables, *i.e.* la mortalité durant la première semaine, la mortalité durant le reste de la période d'élevage, la mortalité pendant le ramassage et le transport, le taux de saisie à l'abattoir. Les variables explicatives sont organisées en 4 tableaux, *i.e.*, X_1 relatif à la structure de l'élevage, X_2 aux caractéristiques du lot la première semaine, X_3 aux caractéristiques du lot durant le reste de l'élevage et X_4 au ramassage, transport et abattage. La finalité pour l'épidémiologiste est d'avoir une vision globale des actions à mener dans les différentes phases de production afin de réduire les pertes.

Références

[1] Bougeard, S., Qannari, E.M., Rose, N. (2011). Multiblock Redundancy Analysis: interpretation tools and application in epidemiology. *Journal of Chemometrics*, **25**(9), 467-475

[2] Dray, S., Dufour, A.B., Chessel., D. (2007). The `ade4` package - II: Two-table and K-table methods. *R News*, **7**(2), 47-52

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

[4] Wangen, L.E., Kowalski, B.R. (1988). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, **3**, 3-20

[5] Wold, S. (1984). Three PLS algorithms according to SW. *Symposium MULDAST (Multivariate analysis in science and technology)*, Umea University, Sweden, 26-30

Package de sélection de modèle CAPUSHE

J-P. Baudry[a], **V. Brault**[b], **C. Maugis-Rabusseau**[c] and **B. Michel**[d]

[a]Laboratoire de Statistique Théorique et Appliquée
Université Paris 6
4 place Jussieu, 75252 Paris cedex 05
jean-patrick.baudry@upmc.fr

[b]Laboratoire de Mathématiques UMR 8628
Université Paris-Sud 11
F-91405 Orsay cedex
vincent.brault@math.u-psud.fr

[c]Institut de Mathématiques de Toulouse
INSA de Toulouse, Université de Toulouse
118 route de Narbonne F-31062 Toulouse Cedex 9
cathy.maugis@insa-toulouse.fr

[d]Laboratoire de Statistique Théorique et Appliquée
Université Paris 6
4 place Jussieu, 75252 Paris cedex 05
bertrand.michel@upmc.fr

Mots clefs : Estimation de pente guidée par les données – Saut de dimension – Sélection de modèle – Pénalisation – Heuristique de pente

La sélection de modèle est un paradigme général incluant de nombreux problèmes de statistiques. L'une des approches les plus populaires est la minimisation d'un critère pénalisé. Birgé et Massart [2] ont proposé une méthode de calibration où les pénalités sont connues à un facteur multiplicatif près : l'heuristique de pente. Des travaux théoriques valident cette méthode heuristique dans certaines situations et plusieurs articles montrent un comportement prometteur dans des cas pratiques.
Deux méthodes sont actuellement utilisées pour la calibration de cette pénalité : le saut de dimension (figure 2) et l'estimation de pente guidée par les données (figure 1). Baudry, Maugis et Michel [6] ont proposé un package matlab implémentant ces deux méthodes avec des interfaces graphiques. Nous présentons ici le package CAPUSHE pour le logiciel `R` proposant leurs implémentations avec la possibilité de représenter graphiquement les résultats pour les valider. Durant cet exposé, nous ferons un cours rappel de la théorie puis nous expliquerons l'utilisation de nos fonctions à l'aide de l'exemple proposé dans le package. Nous détaillerons également les possibilités que possède l'utilisateur pour évaluer et améliorer la qualité des résultats.
Enfin, nous serons très attentifs aux questions ou suggestions d'améliorations pour le package.

Références

[1] Birgé, L. et Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203-268.
[2] Birgé, L. et Massart, P. (2006). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33-73.

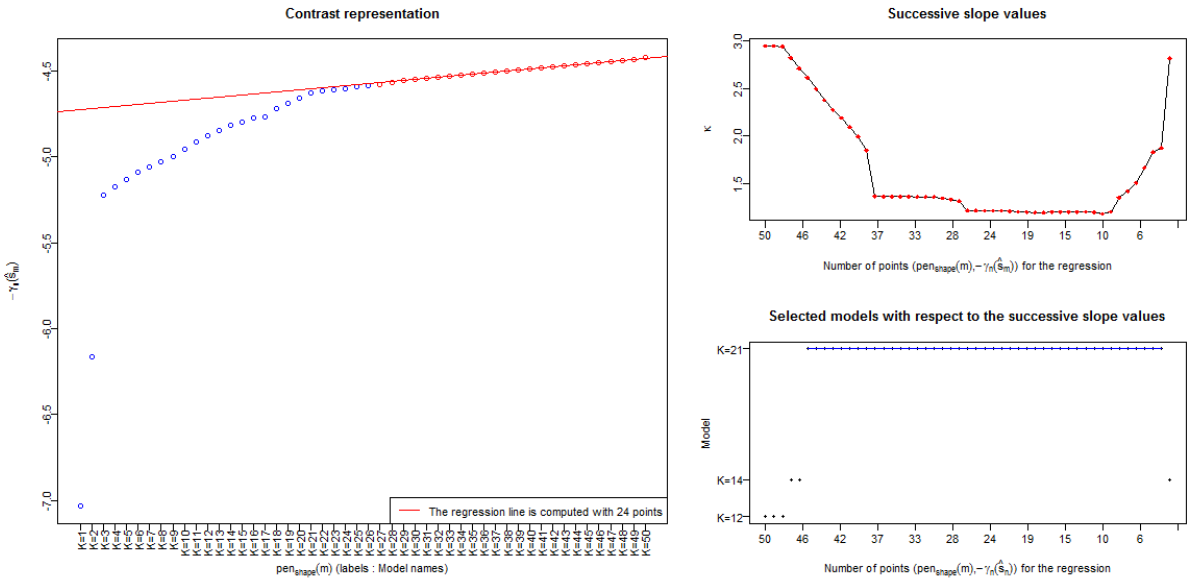


Figure 1: *Représentation graphique proposée par le package* CAPUSHE *pour vérifier la qualité des résultats obtenus par la méthode d'estimation de pente guidée par les données.*

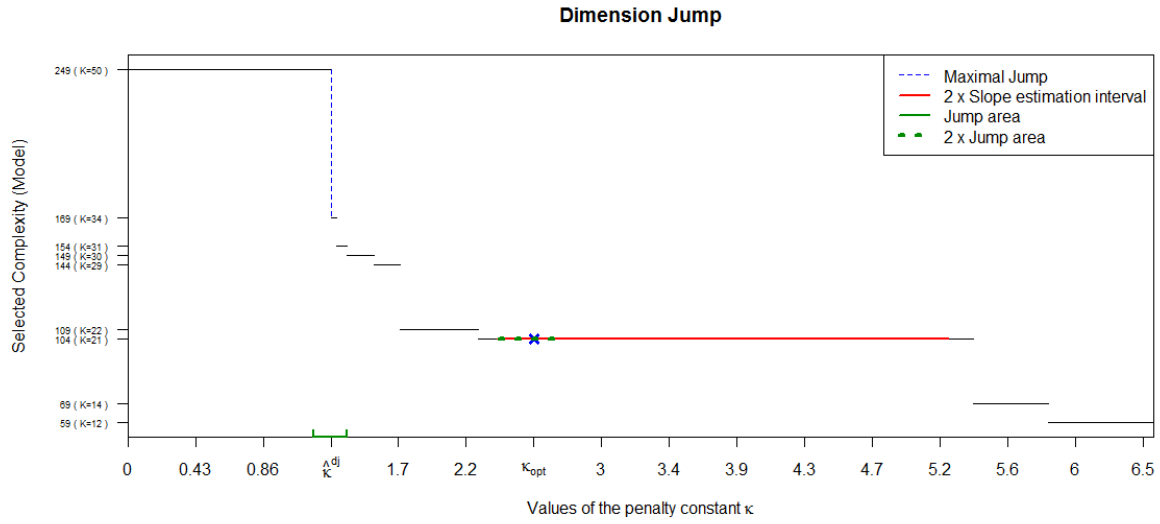


Figure 2: *Représentation graphique proposée par le package* CAPUSHE *pour vérifier la qualité des résultats obtenus par la méthode du saut de dimension.*

[3] Lebarbier, E. (2005). Detexting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85(4):717-736.
[4] Massart, P. (2007). *Concentration Inequalities and Model Selection. École d'été de Probabilités de SAint-Flour 2003. Lecture Notes in Mathematics*. Springer.
[5] Maugis, C. et Michel, B. (2009). A non-asymptotic penalized criterion for gaussian mixture model selection. *ESAIM*: P & S URL http://hal.inria.fr/docs/00/28/50/31/PDF/RR-6549.pdf
[6] Baudry, J.-P., Maugis, C. and Michel, B. (2011) Slope Heuristics: Overview and Implementation. *Statistics and Computing*, Vol 22(2), 455-470.

lcmm: un package *R* pour l'estimation des modèles mixtes à classes latentes et des modèles conjoints à classes latentes pour données répétées Gaussiennes, ordinales ou curvilinéaires et données de survie

C. Proust-Lima[a], **A. Diakité**[b] et **B. Liquet**[c]

[a,b,c]INSERM U897
ISPED, Univ. Bordeaux
F-33000 Bordeaux
[a]cecile.proust-lima@inserm.fr
[a]amadou.diakite@isped.u-bordeaux2.fr
[c]benoit.liquet@isped.u-bordeaux2.fr

Mots clefs : classification, données censurées, données hétérogènes, données longitudinales, modèles à effets aléatoires, modèles conjoints, modèles de mélange, processus latent.

Le modèle linéaire mixte [1], couramment utilisé pour décrire le changement au cours du temps d'une variable quantitative dans les études longitudinales, est fondé sur plusieurs hypothèses :

(i) normalité des erreurs et des effets aléatoires

(ii) linéarité des relations avec le marqueur longitudinal

(iii) continuité du marqueur longitudinal

(iv) homogénéité de la population d'étude

(v) données manquantes aléatoires

Le package *R* **lcmm** permet l'estimation de plusieurs extensions du modèle linéaire mixte:

- le modèle mixte à classes latentes [2-4] qui relache l'hypothèse (iv) pour analyser les changements au cours du temps d'un marqueur longitudinal en population hétérogène;
- le modèle conjoint à classes latentes [5] qui relache les hypothèses (iv) et (v) pour analyser conjointement un marqueur longitudinal et un délai d'événement censuré à droite et potentiellement tronqué à gauche;
- le modèle mixte à processus latent [6,7] qui relache les hypothèses (i),(ii) et (iii) (ainsi que l'hypothèse (iv) si besoin) pour analyser les changements au cours du temps de marqueurs quantitatifs potentiellement curvilinéaires, et de marqueurs ordinaux.

Les modèles mixtes à classes latentes consistent à explorer les profils latents de trajectoires qui peuvent exister dans une population hétérogène. Ces modèles combinent la théorie des modèles mixtes pour tenir compte de la corrélation individuelle entre les mesures répétées du marqueur et les modèles à classes latentes pour discriminer des groupes homogènes de sujets. Malgré leur intérêt en pratique, leur implémentation dans les logiciels gratuits est encore limitée. Au sein du package **lcmm**, la fonction `hlme` estime des modèles linéaires mixtes à classes latentes pour des données "gaussiennes" (i.e. sous les hypothèses (i),(ii) et (iii)), et la fonction `lcmm` estime des modèles mixtes à classes latentes pour des données continues curvilinéaires ou ordinales.

Les modèles pour données curvilinéaires et ordinales sont des modèles mixtes dits "à processus latent" qui font intervenir des fonctions de lien paramétrées pour lier les observations curvilinéaires ou ordinales à leur processus latent continu sous-jacent.

Il existe principalement deux types de modèles conjoints pour l'analyse jointe de données répétées et de données de survie : les modèles à effets aléatoires partagés dans lesquels des fonctions de la trajectoire du marqueur sont incluses dans le modèle de survie, et les modèles à classes latentes qui font l'hypothèse qu'une structure en classes latentes capture toute la corrélation entre les données répétées du marqueur et le risque d'événement. Nous avons implémenté ces derniers dans la fonction `Jointlcmm` du package **lcmm**. Notons que les modèles à effets aléatoires partagés peuvent aussi être estimés sous *R* avec le package **JM** [8].

La méthode d'estimation implémentée dans les fonctions `hlme`, `lcmm` et `Jointlcmm` est le maximum de vraisemblance par un algorithme de Marquardt [9] modifié avec des critères d'arrêt stricts (sur les dérivées premières et secondes) [3,6]. N'importe quelle forme de trajectoire peut être modélisée et des variables explicatives peuvent être incluses dans toutes les parties des modèles (avec ou sans effets spécifiques aux classes). Plusieurs fonctions de risque de base sont implémentées dans `Jointlcmm` (M-splines, Weibull, exponentiel par morceaux) et plusieurs fonctions paramétrées de lien sont implémentées dans la fonction `lcmm` pour tenir compte de la curvilinéarité (I-splines, FdR Beta, linéaire, modèle à seuils). Enfin, des fonctions sont disponibles pour évaluer l'adéquation des modèles, leur qualité prédictive, et les classifications a posteriori [5].

Références

[1] Laird N & Ware J (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-74.

[2] Verbeke G & Lesaffre E (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217-21.

[3] Muthén B & Shedden K (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55**, 463-9.

[4] Proust C & Jacqmin-Gadda H (2005). Estimation of linear mixed models with a mixture of distribution for the random-effects. *Computer Methods and Programs in Biomedicine*,**78**,165-73.

[5] Proust-Lima C, Séne M, Taylor JMG & Jacqmin-Gadda H (2012). Joint latent class models for longitudinal and time-to-event data: a review, *Statistical Methods in Medical Research*, in press.

[6] Proust C, Jacqmin-Gadda H, Taylor JMG, Ganiayre J & Commenges D (2006). A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics*,**62**, 1014-24.

[7] Proust-Lima C, Dartigues J-F & Jacqmin-Gadda H (2011). Misuse of the linear mixed model when evaluating risk factors of cognitive decline. *American Journal of Epidemiology*, **174**(9),1077-88.

[8] Rizopoulos D (2010). JM: An R package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software* **35**(9), 1-33.

[9] Marquardt D (1963). An algorithm for leastsquares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, **11**, 431-41.

saemix, an R version of the SAEM algorithm for parameter estimation in nonlinear mixed effect models

A. Lavenu[a], **E. Comets**[b] and **M. Lavielle**[c]

[a]University Rennes-I & INSERM CIC 0203
Rennes,France
audrey.lavenu@univ-rennes1.fr

[b]INSERM, UMR 738 & Univ Paris Diderot, Sorbonne Paris Cité
Paris, France
emmanuelle.comets@inserm.fr

[c]INRIA
Saclay, France
Marc.Lavielle@inria.fr

Mots clefs : Nonlinear mixed effect models, parameter estimation, SAEM algorithm, R, R package, pharmacokinetics, pharmacodynamics, longitudinal data.

Introduction: The use of modelling and simulation in clinical drug development is now well established. Regardless of whether a single outcome is considered at the end of the study, clinical trials often collect longitudinal data, with each subject providing several measurements throughout the study. Longitudinal data is a staple in particular of pharmacokinetic (PK) and pharmacodynamic (PD) studies, which are a required part of a new drug application file. Nonlinear mixed effect models can help to characterise and to understand many complex nonlinear biological processes, such as biomarkers or surrogate endpoints, and are crucial in describing and quantifying the mechanisms of drug action and the different sources of variation, e.g., the interindividual variability. Over the past decade, new and powerful estimation algorithms have been proposed to estimate the parameters of these models. The Stochastic Approximation Expectation Maximization (SAEM) algorithm has proven very efficient, quickly converging to the maximum likelihood estimators [1] and performing better than linearisation-based algorithms [2]. It has been implemented in the Monolix software [3] which has enjoyed increasingly widespread use over the last few years, more recently in the Statistics toolbox of Matlab (nlmefitsa.m), and is also available in NONMEM version 7 [4]. The objective of the present package was to implement SAEM in the R software [5].

Methods: Detailed and complete presentations of the nonlinear mixed effects model can be found in several reference textbooks, for instance [6]. We consider the following general nonlinear mixed effects model for continuous outputs:

$$y_{ij} = f(x_{ij}, \psi_i) + g(x_{ij}, \psi_i, \xi)\varepsilon_{ij} \quad , \; 1 \leq i \leq N \quad , \quad 1 \leq j \leq n_i \tag{1}$$

where y_{ij} is the jth observation of subject i, N is the number of subjects, n_i is the number of observations of subject i, x_{ij} are known regression variables, and ψ_i is the vector of individual parameters. The SAEM algorithm is used to obtain maximum likelihood estimates of the parameters of nonlinear mixed effects models without any linearisation of the model. The log-likelihood for nonlinear mixed effect models is analytically intractable since it requires integration over the unknown individual parameters. The SAEM algorithm uses an EM algorithm

[7], where the unknown individual parameters are treated as missing data, and replaces the usual E-step with a stochastic approximation step [8]. The missing parameters are simulated at each iteration via a MCMC procedure, which can be used after the algorithm has converged to obtain the conditional modes, the conditional means and the conditional standard deviations of the individual parameters.

Results: The library uses the S4 class system of R to provide a user-friendly input and output system, with methods like `summary` or `plot` for fitted objects. The package provides summaries of the results, individual parameter estimates, standard errors (obtained using a linearised computation of the Fisher information matrix) Wald tests for fixed effects, and a number of diagnostic plots, including VPC plots and npde [9]. The log-likelihood can be computed by three methods: a linearisation of the model, an importance sampling procedure, or a Gaussian quadrature. The diagnostic graphs can be tailored to the user's individual preferences by setting a number of options, and are easily exported to a file.

We illustrate the use of the library with the well known PK dataset of theophylline. These data includes the concentration versus time data collected in 12 subjects given a single oral dose of theophylline, and for whom 11 blood samples were collected over a period of 24 h. We modelled this data using a one-compartment model with first-order absorption, parameterised as k_a, V, CL. The IIV was modelled using an exponential model with diagonal variance-covariance matrix, while the residual variability was modelled with a combined error model. Many diagnostic plots are available to evaluate convergence or model adequacy, such as individual plots, using a `plot` function through which user-specific options can be set.

Conclusion: The saemix package provides the SAEM algorithm for R users. The current version handles models in analytical form, with continuous or binary covariates.

Références

[1] Delyon B., Lavielle M., Moulines E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* **27**, 94–128

[2] Girard P., Mentré F. (2005). A comparison of estimation methods in nonlinear mixed effects models using a blind analysis (oral presentation). *Meeting of the Population Approach Group in Europe (PAGE)*, Pamplona

[3] Lavielle M. (2010). *MONOLIX (MOdèles NOn LInéaires á effets miXtes) User Guide.* MONOLIX group, Orsay, France. URL: http://software.monolix.org/

[4] Beal S., Sheiner L.B., Boeckmann A., Bauer R.J. (2009). *NONMEM User's Guides. (1989-2009)*, Icon Development Solutions, Ellicott City, MD, USA

[5] R Development Core Team (2006). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria

[6] Davidian M., Giltinan D (1995). *Nonlinear models for repeated measurement data.* Chapman & Hall, London

[7] Dempster A. P., Laird N. M., Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39**, 1–38

[8] Kuhn E., Lavielle M (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics and Data Analysis* **49**, 1020–38

[9] Brendel K., Comets E., Laffont C., Laveille C., Mentré M.. Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide (2006). *Pharmaceutical Research*, **23**, 2036–49

Représentation des caractéristiques du vent estimée par une méthode à noyau

J. Chau[a] and **N. Hengartner**[b] and **N.H. Khuc**[a,c] and **E. Matzner-Løber**[c]

[a]Greenwich Statistics
209 Rue de l'Université - 75007 Paris
jeff.chau@greenwich-statistics.com

[b]Los Alamos National Laboratory
Los Alamos, NM 87545
nickh@lanl.gov

[c]Laboratoire de Mathématiques
Agrocampus ouest
65 rue de Saint-Brieuc, 35042 Rennes Cedex
nh.khuc@greenwich-statistics.com

Mots clefs : Distribution vitesse et direction du vent, estimation par noyau, variable circulaire.

Introduction

L'étude du vent est essentiel dans le choix de l'emplacement d'un parc éolien, l'optimisation de son orientation et l'évaluation du productible. Trois indicateurs sont utilisés : la rose des vents, la distribution de Weibull et la rose des énergies. Ce papier propose une alternative aux méthodes actuelles en introduisant l'estimation de la densité par une méthode à noyau et propose une amélioration des outils graphiques pour faciliter la lecture des résultats.

Estimation bivariée dont direction du vent comme variable angulaire

Dans le domaine éolien, certains auteurs ont proposés l'estimateur bivariée non paramétrique classique :

$$\hat{f}_{h_1,h_2}(x_1,x_2) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} \mathbb{K}\left(\frac{x_1 - X_{1,i}}{h_1}\right)\mathbb{K}\left(\frac{x_2 - X_{2,i}}{h_2}\right) \tag{1}$$

où en général $\mathbb{K}$ est un noyau gaussien [1]. Or la direction du vent est une variable angulaire où il y a correspondance entre les degrés 0 et 360 (0 étant le Nord). Pour l'estimation de la densité, les noyaux usuels donnent un poids élevé aux observations proches et attribuent un poids très faible aux observations éloignés. Or nous souhaitons ici que les observations à 360^o-près soient aussi pris en compte. Pour ce faire, nous proposons d'utiliser un noyau qui résulte d'un mélange de noyaux gaussiennes tronquées, noté MGT. Considérons la famille des noyaux gaussiens et définissons le noyau MGT tel que :

$$\mathbb{K}_{(m,\nu)}(\cdot) = \frac{1}{\sqrt{2\pi\nu}}\exp\left(-\frac{1}{2}\frac{(\cdot - m)^2}{\nu}\right) \qquad MGT = \frac{1}{A}\sum_{j=1}^{2} q_j \mathbb{K}_{m_j,s_j}(\cdot) \tag{2}$$

Avec A une constante telle que $\int MGT(u)du = 1$.

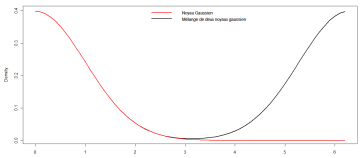


Figure 1: Noyau MGT par rapport au noyau gaussien

L'estimateur bivarié de la vitesse et de la direction du vent s'écrit donc comme suit :

$$\hat{f}_{h_1,h_2}(x_1,x_2) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} \mathbb{K}\left(\frac{x_1 - X_{1,i}}{h_1}\right) MGT\left(\frac{x_2 - X_{2,i}}{h_2}\right) \tag{3}$$

Nous avons développé l'algorithme suivant qui permet une représentation graphique de l'estimation bivariée $\hat{f}_{h_1,h_2}(x_1,x_2)$ dans l'espace de coordonées polaire (ρ,θ).

Entrée : données de vitesse et direction (ρ,θ).
Transformation des données vitesse et direction en cartésien (X,Y).
Construction d'une grille de x *et de* y *tels qu'ils couvrent le domaine de définition de* (X,Y).
Pour chaque couple (x_i,y_j),

- *Transformer le couple* (x_i,y_j) *en coordonnées polaire* (ρ_k,θ_l).
- *Appliquer l'estimateur à noyau multiplicatif bivarié composé d'un noyau gaussien et d'un noyau MGT au couple* (ρ_k,θ_l), *évalué sur les données d'entrée* (ρ,θ). *Les fenêtres sont choisies par validation croisée.*
- *Stockage de l'estimation dans la case correspondant au couple* (x_i,y_j).

Choix des lignes de contour à dessiner

Les lignes de niveau sont tracées de manière aléatoire en fonction du nombre de lignes à dessiner (fixé par défaut dans les logiciels). Nous proposons de leur donner un sens en dessinant les quantiles de la distribution, ce qui permettra un repérage des structures particulière de la répartition des vents.

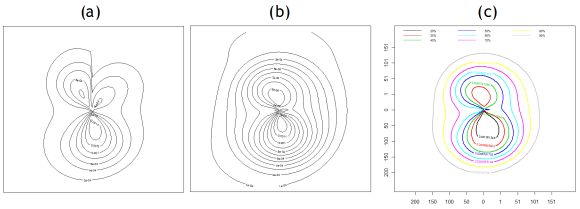


Figure 2: Problème d'estimation en 0 degré. Contour de la densité estimée (a) par deux noyaux gaussien. (b) Avec un noyau gaussien et un noyau MGT. (c) Représentation des contours en fonction des déciles. Par exemple, la ligne bleu représente le 5^e décile. Ainsi 50% des relevés de vent se situent à l'intérieur de cette ligne.

Le package dans R permettra à partir des données de vitesse et de direction du vent d'obtenir l'estimateur à noyau bivarié, sa représentation dans le repère polaire, la visualisation dynamique 3D et le traçage du graphique des contours en fonction des quantiles choisis par l'utilisateur.

Références

[1] Zhang, J., *et al.* (2011). Multivariate and multimodal wind distribution model based on kernel density estimation. *ASME 2011 5th International Conference on Energy Sustainability.*

Analyse d'images et régression non-paramétrique

P.A. Cornillon[a],**N. Hengartner**[b], **B. Thieurmel**[c] and **B. Wohlberg**[b]

[a]Université de Rennes 2
pac@univ-rennes2.fr

[b]Laboratoire national de Los Alamos
nickh@lanl.gov et brendt@lanl.gov

[c]Greenwich Statistics
benoit.thieurmel@greenwich-statistics.com

Mots clefs : Statistique, Régression non-paramétrique, Traitement d'images.

La régression non-paramétrique est un outil standard pour analyser la relation fonctionnelle entre une variable à expliquer et un ensemble de covariables. Malheureusement, la mise en œuvre de lisseurs entièrement non-paramétriques est limitée par le fléau de la dimension, ce qui explique pourquoi jusqu'à présent, cet outil a eu un succès limité en analyse d'images.

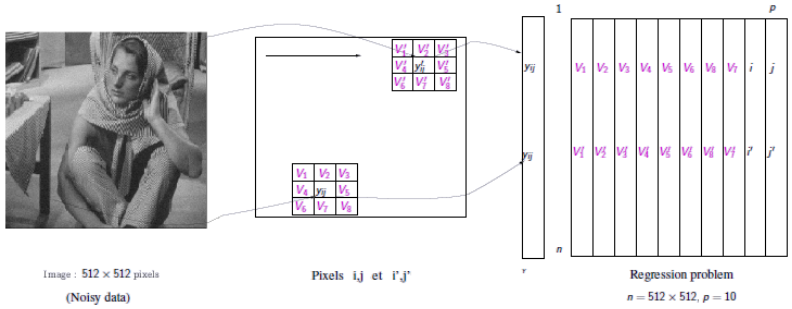
Des avancées récentes en lissage non-paramétrique [1] ont montré qu'une simple correction itérative du biais permet au lisseur de "s'adapter" à la vraie fonction de régression. Cette méthode peut donc partiellement atténuer le fléau de la dimension. Pratiquement, quand la vraie régression est lisse, il est possible d'utiliser cette méthode avec un nombre important de variables explicatives (de l'ordre de 20 à 50). On veut donc évaluer f $(\mathbb{R}^p \mapsto \mathbb{R})$ avec le modèle : $Y_i = f(X_i) + \varepsilon_i, \quad i \in \{1, \ldots, n\}$.

1. Sur-lissage des données : $\hat{Y} = SY$ donc estimateur biaisé
2. Evaluation du biais $f(X_i) - \mathbb{E}(\hat{Y}_i) = (I - S)f(X_i)$ et estimation par $(I - S)SY$
3. Correction de l'estimateur précédent : $\hat{Y}^{(2)} = \hat{Y} - (I - S)SY = (I - (I - S)^2)Y$

Cette procédure peut être itérée et on obtient alors

$$\hat{Y}^{(k)} = (I - (I - S)^k)Y$$

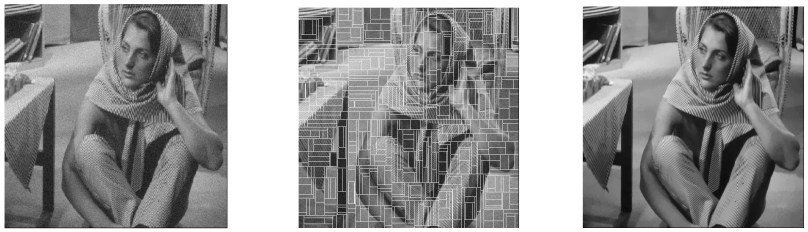
Il est donc possible d'explorer le problème du débruitage et de reconstruction d'images comme un problème de régression non-paramétrique.



Afin de pouvoir utiliser la méthode de réduction itérative du biais, nous devons préalablement partitionner l'image en patchs. On souhaite définir des régions homogènes tout en laissant une certaine liberté dans la forme. Pour se faire, nous nous basons sur l'algorithme CART [2] et les arbres de régression en expliquant l'intensité des pixels par les coordonnées (i, j) de l'image. Le modèle de débruitage est : f, $\mathbb{R}^{10} \mapsto \mathbb{R}$, $Y_i = f(x_i) + \varepsilon_i \;\; i \in \{1, \ldots, n\}$ pour chaque patch de taille n avec :

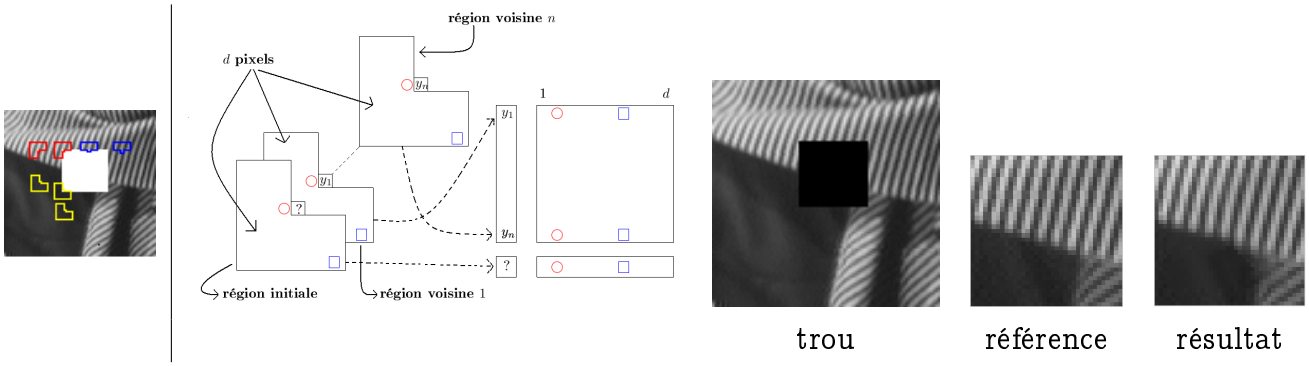
- Y_i la valeur du pixel
- x_i le vecteur des 10 variables explicatives (8 voisins + 2 coordonnées)
- f estimée par $\hat{f} = S_{\hat{k}} Y$ avec la méthode **IBR**

Afin de pallier aux effets de bord relatif à la méthode, on effectue plusieurs fois le partitionnement ainsi que l'application de la méthode IBR, en procédant entre chaque itération à une rotation d'angle α des axes I et J des coordonnées des pixels. La prédiction finale d'un pixel est la moyenne des différentes prédictions obtenues.



(a) image bruitée (b) partitionnement (c) résultat

Pour la problèmatique de reconstruction d'image, on traite les pixels un par un en formant la base de données de régression en recherchant les régions voisines à celle possédant le pixel manquant.



Nous présenterons dans cet exposé des résultats prometteurs sur un certain nombre d'images utilisées classiquement et comparerons les méthodes en utilisant le PSNR ou le SSIM.

Références

[1] P.-A. Cornillon, N. Hengartner, and E. Matzner-Lober. Recursive bias estimation and l2 boosting. Technical report, ArXiv :0801.4629, 2008.

[2] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. Classification And Regression Trees. CHAPMAN & HALL/CRC

Construction et randomisation de plans factoriels réguliers avec le package R PLANOR

H. Monod[a], **A. Bouvier**[a] and **A. Kobilinsky**[a]

[a]Unité de Mathématiques et Informatique Appliquées
INRA, UR0341 MIA-Jouy
Domaine de Vilvert, 78326 Jouy en Josas
herve.monod@jouy.inra.fr

Mots clefs : Statistique, Plan d'expériences, Plan factoriel régulier.

Les plans factoriels fractionnaires sont des plans d'expériences utiles pour de nombreux domaines d'application [1]. Nous nous intéressons ici aux fractions régulières, obtenues par des méthodes algébriques de construction et plus précisément par des relations de définition entre facteurs [3,4]. Malgré leur intérêt, les méthodes automatiques pour les construire sont encore peu présentes parmi les méthodes disponibles sous R. C'est dans le but de combler ce manque qu'a été développée le package R PLANOR, directement inspiré d'un logiciel développé dans les années 90 et 2000 par André Kobilinsky [2]. L'algorithme implémenté dans PLANOR permet de construire des plans factoriels fractionnaires réguliers, sans contrainte limitante sur le nombre de niveaux des facteurs. Il offre d'autres originalités d'une grande importance en pratique, par exemple la possibilité de prendre en compte plusieurs niveaux de variabilité ainsi que des hiérarchies entre facteurs induites par des contraintes expérimentales. Des fonctions permettent de construire et de randomiser en quelques étapes une grande diversité de plans factoriels fractionnaires orthogonaux. L'environnement R permet de prolonger facilement la construction du plan par son analyse statistique et par des représentations graphiques.
Au cours de cette communication, nous rappellerons le principe des fractions régulières, nous évoquerons les algorithmes de PLANOR et leur implémentation sous R, puis nous montrerons des exemples d'application. Au passage, nous montrerons que les généralisations implémentées dans PLANOR permettent d'unifier dans un même cadre un grand nombre de plans factoriels orthogonaux.
Remarque: PLANOR doit être déposé prochainement sur le CRAN. Le package est actuellement accessible et téléchargeable à l'adresse:
http://w3.jouy.inra.fr/unites/miaj/public/logiciels/planor/

Références

[1] Kobilinsky, A. (1997). Les plans factoriels. *In:* Plans d'expériences: applications à l'entreprise (Droesbeke, J.J., Fine, J. et Saporta, G. éds.), pp. 69-209 (Chapitre 3). Technip, Paris.
[2] Kobilinsky, A. (2005). *PLANOR : program for the automatic generation of regular experimental designs. Version 2.2 for Windows.* Technical Report. MIA Unit, INRA Jouy en Josas.
[3] Kobilinsky, A. et Monod, H. (1991). Experimental design generated by group morphisms: an introduction. *Scandinavian Journal of Statistics*, **18**, 119-134.
[4] Pistone, G. et Rogantin, M.-P. (2008). Indicator function and complex coding for mixed fractional factorial designs. *Journal of Statistical Planning and Inference*, **138**, 787-802.

Simulation and competing risks

Jan Beyersmann, Freiburg, Germany
jan@fdm.uni-freiburg.de

The analysis of survival or time-to-event data is one of the most common applications of advanced statistical techniques in medical research and beyond. Data are typically incomplete as a consequence of limited observation periods. Therefore, survival analysis is based on hazards. Hazard-based techniques also allow for analysing competing risks, i.e., time-to-first-event and type-of-first-event. A common example from cancer research is relapse-free survival, which is the time until relapse or death, whatever comes first. A competing risks analysis distinguishes between these two event types, but the presence of more than one event-specific hazard poses a challenge both in theory and practice. We discuss the use of simulation, preferably in R, to understand and analyse competing risks data, interpret the analyses and plan studies with competing risks outcomes.

Reference

Beyersmann, J., Allignol, A., and Schumacher, M. (2012) *Competing Risks and Multistate Models with R.* Springer, New York

`FrailtyPack`: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood or Parametrical Estimation

V. Rondeau[a,b] and Y. Mazroui[a] and A. Mauguen[a] and A. Diakite[a] and JR. Gonzalez[c]

[a]INSERM U897
ISPED
146 Rue Léo Saignat 33076 Bordeaux Cedex
Virginie.Rondeau@isped.u-bordeaux2.fr
Yassin.Mazroui@isped.u-bordeaux2.fr
Audrey.Mauguen@isped.u-bordeaux2.fr
Amadou.Diakite@isped.u-bordeaux2.fr

[b]Université Bordeaux Segalen, ISPED
146 Rue Léo Saignat 33076 Bordeaux Cedex

[c]Centre for Research in Environmental Epidemiology (CREAL)
Biomedical Park Research of Barcelona (PRBB)
Avda. Dr Aiguader 88, Barcelona 08003, Spain
jrgonzalez@creal.cat

Mots clefs : Statistic, Frailty models, clustered data, recurrent events

Frailty models are extensions of the Cox proportional hazards model which is the most popular model in survival analysis. In many clinical applications, the study population needs to be considered as a heterogeneous sample or as a cluster of homogeneous groups of individuals such as families or geographical areas. Sometimes, due to lack of knowledge or for economical reasons, some covariates related to the event of interest are not measured. The frailty approach is a statistical modelling method which aims to account for the heterogeneity caused by unmeasured covariates. It does so by adding random effects which act multiplicatively on the hazard function.

`FrailtyPack` is an R package[1] which allows to fit four types of frailty models, for left-truncated and right-censored data, adapted to most survival analysis issues. The aim of this talk is to present the new version of the R package `FrailtyPack`, which is available from the Comprehensive R Archive Network at http://CRAN.R-project.org/, and the various new models proposed. It depends on the R `survival` package.[2] The initial version of this package[3] was proposed for a simple shared frailty model, and was developed for more general frailty models.[4] The shared frailty model[5] can be used, when observations are supposed to be clustered into groups. The nested frailty model[6] is most appropriate, when there are two levels of hierarchical clustering. However, several relapses (recurrent events) are likely to increase the risk of death, thus the terminal event is considered as an informative censoring. Using a joint frailty model, it is possible to fit jointly the two hazard functions associated with recurrent and terminal events,[7] when these events are supposed to be correlated. The additive frailty model[8] is more adapted to study both heterogeneity across trial and treatment-by-trial heterogeneity (for instance meta-analysis or multicentric datasets study). We show how a simple multi-state frailty model can be used to study semi-competing risks while fully taking into account the clustering (in ICU) of the

data and the longitudinal aspects of the data, including left truncation and right censoring.[9] We included recently parametric hazard functions and prediction methods. Depending on the models, stratification and time-dependent covariates are allowed or not.
The frailty models discussed in recent literature present several drawbacks. Their convergence is too slow, they do not provide standard errors for the variance estimate of the random effects and they can not estimate smooth hazard function. `FrailtyPack` use a non-parametric penalized likelihood estimation, and the smooth estimation of the baseline hazard functions is provided by using an approximation by splines.
`FrailtyPack` was first written in Fortran 77 and was implemented for the statistical software R. We will present the models that `FrailtyPack` can fit and the estimation method, then we will describe all the functions and the arguments of `FrailtyPack`. Finally epidemiological illustrations will be provided using `FrailtyPack` functions. `FrailtyPack` is improved regularly in order to add new developments around frailty models.

References

[1] R Development Core Team. *R: A Language and Environment for Statistical Computing*.

[2] T. Therneau. *survival: A Package for Survival Analysis in S*, 2012. R package version 2.36-12.

[3] V. Rondeau and J.R. Gonzalez. frailtypack: A computer program for the analysis of correlated failure time data using penalized likelihood estimation. *Computer Methods and Programs in Biomedicine*, 80(2):154–164, 2005.

[4] V. Rondeau, Y. Mazroui, A. Mauguen, A. Diakite, and JR Gonzalez. FRAILTYPACK*: An* R *package for General frailty models using a semi-parametrical penalized likelihood estimation or a parametrical estimation*, 2012. R package version 2.2-22.

[5] V. Rondeau, D. Commenges, and P. Joly. Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Analysis*, 9(2):139–153, 2003.

[6] V. Rondeau, L. Filleul, and P. Joly. Nested frailty models using maximum penalized likelihood estimation. *Statistics in Medicine*, 25(23):4036–4052, 2006.

[7] V. Rondeau, S. Mathoulin-Pelissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. *Biostatistics*, 8(4):708–721, 2007.

[8] V. Rondeau, S. Michiels, B. Liquet, and J.P. Pignon. Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach. *Statistics in Medicine*, 27(11):1894–1910, 2008.

[9] B. Liquet, J.F. Timsit, and V. Rondeau. Investigating hospital heterogeneity with a multistate frailty model: application to nosocomial pneumonia disease in intensive care units. *BMC medical reserach methodology*, In press, 2012.

Analyse de données de survie en présence de censure par intervalles : le package SmoothHazard

P. Joly[a,b], **C. Touraine**[a,b], **A. Diakite**[a,b] and **T. A. Gerds**[c]

[a]Univ. Bordeaux,
Isped, Centre INSERM U897-Epidemiologie-Biostatistique,
F-33000 Bordeaux, France
pierre.joly@isped.u-bordeaux2.fr

[b]INSERM,
Isped, Centre INSERM U897-Epidemiologie-Biostatistique,
F-33000 Bordeaux, France

[c]Department of Biostatistics,
University of Copenhagen,
Denmark.

Mots clefs : Analyse des données de survie, modèle Illness-death, Censure par intervalles, vraisemblance pénalisée, modèles paramétriques.

L'analyse des données de survie est l'étude du délai de la survenue d'un évènement précis pour un ou plusieurs groupes d'individus. Cet évènement est souvent associé à un changement d'état ; il peut tout aussi bien être la mort d'un individu pour une cause déterminée, que l'apparition chez cet individu d'une certaine maladie. Les modèles multi-états permettent d'étudier l'évolution complexe de sujets qui peuvent connaître plusieurs événements. Les modèles de survie sont des cas particuliers des modèles multi-états, ne comportant que 2 états. Une des caractéristiques des données de survie est l'existence d'observations incomplètes, la censure et la troncature font partie des processus générant ce type de données. Des packages R existent pour traiter des données de survie dans des cas "classiques", c'est à dire avec censure à droite et troncature à gauche [1], [2]. Ces programmes permettent, en particulier, l'estimation des paramètres de régression. En effet, les individus ou les groupes d'individus sont susceptibles de différer pour un ou plusieurs facteurs. Ces facteurs, représentés par des variables explicatives, peuvent expliquer une différence importante de la durée de survie des individus et pour modéliser leur influence on utilise des modèles de régression. En épidémiologie, le modèle de régression le plus utilisé en analyse des données de survie est le modèle à risques proportionnels (souvent dénommé modèle de Cox [3]). De même, pour les modèles multi-états c'est souvent des modèles à intensités de transition proportionnelles qui sont utilisés.

Or dans les enquêtes épidémiologiques, les données sont souvent recueillies en temps discret. En fonction de l'évènement étudié, cela peut générer des données censurées par intervalles. Pour le cas des modèles de survie, une observation est censurée par intervalles si au lieu d'observer avec exactitude le temps de l'événement, la seule information disponible est qu'il se situe entre deux dates connues. Pour prendre en compte la censure par intervalles, les approches les plus simples sont les approches paramétriques (comme la fonction exponentielle ou la fonction de Weibull) ou par vraisemblance pénalisée. L'approche de la vraisemblance pénalisée considérée ([4], [5]) consiste à chercher les coefficients d'une base de fonctions M-splines cubiques (qui sont une variante des B-splines) pour obtenir une approximation de l'estimateur non paramétrique de la fonction d'intérêt définie comme le maximum de la vraisemblance pénalisée. Pour le choix

des paramètres de lissage nous utilisons une approximation du critère de validation croisée.

Dans le package SmoothHazard, que nous sommes en train de finaliser, il y a actuellement deux classes de modèles, le modèle de survie et le modèle Ilness-death. Dans chacune de ces classes nous proposons soit une approche paramétrique avec un modèle de Weibull soit une approche par vraisemblance pénalisée. Les approches proposées pour ces différents modèles permettent de prendre en compte des données dans un contexte très général de données censurées (en particulier par intervalles) et tronquées. De plus ces approches permettent d'obtenir des estimateurs des intensités de transition qui permettent d'estimer des incidences, des taux de décès, des probabilités de transition et des espérances de vie. Ces intensités peuvent être estimées conjointement avec des paramètres de régression pour étudier l'effet de variables explicatives. Le modèle Illness-death proposé est un modèle à 3 états utilisé, en particulier, pour traiter des problèmes de risques compétitifs. L'état 0 représente l'état "sain", l'état 1, l'état "malade" et l'état 2 l'état "décédé". On considère que la transition de l'état 0 à 1 est possiblement censurée par intervalles ou à droite et que les transitions vers l'état 2 sont observées ou censurées à droite. Quand les sujets sont observés de manière non continue, le temps de la transition entre les états 0 et 1 n'est pas connu exactement mais en plus le nombre de transition peut être inconnu. Les quatre modèles différents proposés actuellement sont programmés en Fortran 90 en ce qui concerne la recherche des paramètres qui maximisent la vraisemblance. Le package permet de choisir le modèle désiré, de spécifier s'il y a troncature à gauche, censure à droite et censure par intervalle. Pour les modèles Illness-death une seule transition peut-être censurée par intervalle. Il est possible de mettre des variables explicatives, potentiellement différentes d'une transition à une autre. Les sorties comprennent les paramètres de régression avec les résultats classiques : risques relatifs, intervalles de confiance, p-values. A la demande de l'utilisateur peuvent aussi être fournie les intensités de transition avec leurs bandes de confiance, des probabilités de transition et des espérances de vie. Ces dernières quantités peuvent être calculées pour des temps et des valeurs des variables explicatives proposées par l'utilisateur. Leur calcul est programmé en R. Les perspectives pour des versions ultérieures du package sont de proposer des classes de modèles et des approches paramétriques supplémentaires.

Références

[1] Beyersmann, J., Schumacher, M. and Allignol, A. (2011). Competing Risks and Multistate Models with R. *Use R Series*, Springer

[2] Putter, H., Fiocco, M., and Geskus, R. (2007). Tutorial in biostatistics: competing risks and multi-state models.*Statistics in Medicine*, **26** (11):2277-2432.

[3] Cox, D.R. (1972). Regression models and life tables (with discussion), *Journal Royal Statistical Society B*, **34**, 187-220.

[4] Joly, P., Commenges, D., Helmer, C. and Letenneur, L. (2002). A penalized likelihood approach for an illness–death model with interval-censored data: application to age-specific incidence of dementia, *Biostatistics*, **3** (3), 433-443.

[5] Joly, P., Letenneur, L., Alioum, A. and Commenges D. (1999). PHMPL: a computer program for hazard estimation using a penalized likelihood method with interval-censored and left truncated data, *Computer Methods and Programs in Biomedicine*, **60**, 225-231.

Planification d'essais randomisés séquentiels ayant comme critère de jugement un délai de survie à l'aide de la fonction plansurvct.func

J. Gal[a] , **A. Kramar**[b],**T. Filleron**[c]

[a]Département Recherche Clinique Innovation et Statistiques
Centre Antoine Lacassagne
33, Avenue de Valombrose 06189 NICE
jocelyn.gal@nice.unicancer.fr

[b]Unité de Méthodologie et Biostatistique
Centre Oscar Lambret
3, rue Frédéric Combemale , BP 307, 59020 Lille
a-kramar@o-lambret.fr

[c]Unité de Biostatistique
Institut Claudius Régaud
20-24 rue du Pont Saint Pierre, 31052 Toulouse
Filleron.Thomas@claudiusregaud.fr

Mots clefs : Fonction R, Nombre de sujets nécessaires, Délai de survie, Essais randomisés séquentiels.

L'estimation du nombre de sujets nécessaires dans un essai randomisé est un défi particulièrement important pour le statisticien, en particulier dans le cas où le critère de jugement est un délai de survie. En effet, la puissance de l'étude ne dépend pas du nombre de patients inclus mais du nombre d'événements observés. Il est donc primordial de déterminer le nombre de patients à inclure dans l'essai afin d'observer le nombre d'événements nécessaires. Le calcul du nombre de sujets nécessaires s'effectue donc en deux étapes :

1. Calcul du nombre d'événements à observer.
2. Calcul du nombre de patients à inclure.

Un des défauts récurrents des essais ayant comme critère de jugement la survie, est l'inclusion de trop peu de patients pour obtenir la puissance nécessaire. Il est donc impératif pour déterminer la taille de l'échantillon de prendre en considération plusieurs paramètres : la durée des inclusions, la durée de suivi ainsi que le taux de perdus de vues. Plusieurs logiciels commerciaux [1][2], permettent d'estimer la taille de l'échantillon en considérant ces différents facteurs. Certains d'entre eux permettent également de planifier des essais séquentiels [1]. Certains packages [3], comme le package gsDesign [4], ont été développés sous R afin d'utiliser certaines de ces méthodes. La fonction plansurvct.func a été implémentée afin de compléter et de simplifier l'utilisation du package gsDesign [5].

Cette fonction permet de planifier des essais de supériorité et de non-infériorité ou d'équivalence dans le cas d'analyse des données de survie en faisant varier différents paramètres :
– Type de l'essai(Supériorité, Non infériorité, Equivalence)
– Hypothèses (Taux de survie, Hazard Ratio,....)
– Proportion de patients randomisés dans le bras expérimental

– Risque de 1^{iere} Espèce
– Puissance
– Durée des inclusions
– Durée du suivi
– Délai souhaité pour l'analyse
– Taux de perdu de vue dans chaque bras de traitement
– Planification d'analyse Intermédiaire (Efficacité et/ou futilité)
    - Méthode de Pocock [6]
    - Méthode de O'Brien Fleming [7]

L'expression de la fonction sous R est la suivante :

$$plansurvct.func(design, Survhyp, pe, \alpha, \beta, duraccrual, durstudy, look, fup, dropout)$$

Les résultats obtenus en retour sont les suivants :
– Nombre d'événements à observer
– Nombre de patients à inclure
– Date d'analyse théorique sous les hypothèses nulle et alternative
– Frontière de rejet
– Si des analyses intermédiaires sont prévues
– Frontière de Rejet
– Date d'analyse sous H0 ou H1

L'utilisation de cette fonction R sera présentée à l'aide de différents exemples inspirés d'essais thérapeutiques en cancérologie. Les résultats obtenus seront comparés avec ceux de East.

Références

[1] East 5.3 A software for the design and interim monitoring of group-sequential clinical trials, Cytel Software Corporation, Cambridge, 2009.
[2] Statistical and Power Analysis Software PASS, in, NCCS,http ://www.ncss.com/pass.html.
[3] Weiliang Qiu,powerSurvEpi : Power and sample size calculation for survival analysis of epidemiological studies,Version 0.05,2009,http ://cran.r-project.org/web/packages/powerSurvEpi/index.html
[4] K. Anderson, gsDesign : Group Sequential Design-R Package Version 2.4-01, 2011,http ://CRAN.R-project.org/package=gsDesign.
[5] T. Filleron, J. Gal, A. Kramar,*Designing group sequential randomized clinical trials with time to event end points using a R function*,Computer Methods and Programs in Biomedecine 2012(In press)
[6] S.J. Pocock, Interim analyses for randomized clinical trials : the group sequential approach, Biometrics 38 (1982) 153-162.
[7] P.C. O'Brien, T.R. Fleming, A multiple testing procedure for clinical trials, Biometrics 35 (1979) 549-556.

HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data

L. Bergé[a] and **C. Bouveyron**[b]and **S. Girard**[c]

[a]Laboratoire GREthA
Université Bordeaux IV
laurent.berge@u-bordeaux4.fr

[b]Laboratoire SAMM, EA 4543
Université Paris 1 Panthéon-Sorbonne
charles.bouveyron@univ-paris1.fr

[b]Equipe Mistis
INRIA Rhône-Alpes & LJK
stephane.girard@inrialpes.fr

Mots clefs : Model-based classification and clustering, high-dimensional data, subspaces.

This paper presents the **R** package *HDclassif* which is devoted to the clustering and the discriminant analysis of high-dimensional data. The classification methods proposed in the package result from a new parametrization of the Gaussian mixture model which combines the idea of dimension reduction and model constraints on the covariance matrices. The supervised classification method using this parametrization is called high dimensional discriminant analysis (HDDA). In a similar manner, the associated clustering method is called high dimensional data clustering (HDDC) and uses the expectation-maximization algorithm for inference. In order to correctly fit the data, both methods estimate the specific subspace and the intrinsic dimension of the groups. Due to the constraints on the covariance matrices, the number of parameters to estimate is significantly lower than other model-based methods and this allows the methods to be stable and efficient in high dimensions. Two introductory examples illustrated with **R** codes allow the user to discover the *hdda* and *hddc* functions. Experiments on simulated and real datasets also compare HDDC and HDDA with existing classification methods on high-dimensional datasets. *HDclassif* is a free software and distributed under the general public license, as part of the **R** software project.

The **R** package *HDclassif* (currently in version 1.2) implements these two classification methods for the clustering and the discriminant analysis of high-dimensional data. The package is available from the CRAN at *http://CRAN.R-projects.org/package=HDclassif*.

Références

[1] L. Bergé, C. Bouveyron and S. Girard, *HDclassif : an R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data*, Journal of Statistical Software, vol. 42 (6), pp. 1-29, 2012.

HiDimDA: An R package for Supervised Classification of High-Dimensional Data

A. Pedro Duarte Silva

Faculdade de Economia e Gestão & CEGE
Catholic University of Portugal / Porto
Rua Diogo Botelho, 1327, 4169-005 Porto, Portugal
psilva@porto.ucp.pt

Keywords : Supervised Classification, Discriminant Analysis, High Dimensionality, Feature Selection.

Classical methods of supervised classification often assume the existence of a training data set with more observations than variables. However, nowadays many classification applications work with data bases where the total number of original features is larger, and often much larger, than the number of available data units. For instance, in microarray applications several thousand genes are usually collected on a few dozen individuals with known clinical conditions, in order to derive classification rules capable of supporting the diagnostic of future patients (see, e.g., Dudoit, Fridlyard and Speed (2002)). A similar pattern occurs in image recognition problems, where the information contained in hundreds of pixels is trained on a much smaller set of images belonging to well defined classes (Thomaz and Gillies (2005)).
Furthermore, in most high-dimensional classification problems the majority of the original features do not contribute to distinguish the underlying classes, and can have a large negative impact if forced into the resulting classification rules (Fan and Fan 2008). Nevertheless, the number of useful features is often still comparable to, or even larger than, the number of available training sample observations.
Therefore, effective classification methodologies for these applications require scalable methodologies of feature selection, and classification rules that can use sample information in a way that is not severely limited by the number of data units in the training sample. The most common strategy to deal with the latter problem is to adopt rules (e.g., Domingos and Pazzani (1997) Tibshirani et. al. (2003)) that treat all features independently, and ignore all sample information about their dependence structure. Recent proposals (e.g., Thomaz and Gillies (2005), Fisher and Sun (2011), Duarte Silva (2011)) try to surpass this limitation by relying on estimators of covariance matrices with good statistical properties when the number of used features is close to, or larger than, the training sample size.
In this presentation, I will describe the *HiDimDA* (High Dimensional Discriminant Analysis) R package, available on CRAN, that implements several routines and utilities for supervised k-group classification in high-dimensional settings. *HiDimDA* includes routines for the construction of classification rules with the above mentioned properties, methods for predicting new observations of unknown origin, as well as cross-validation and feature selection utilities. The selection routines of *HiDimDA* implement modern proposals for feature selection in high-dimensional classification problems (see e.g. Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), Efron, (2004), Donoho and Jin (2008), Fan and Fan (2008)), which often rely on ideas originated form the related theory of large-scale hypothesis testing.
HiDimDA can be used to construct, apply and assess k-group ($k \geq 2$) classification rules for problems with several thousand variables, dealing effectively with the problems of high dimensionality, and including rules that do not ignore the dependence structure of the data.

References

[1] Dudoit S., Fridlyard, J. and Speed TP. (2002). Comparison of discrimination methods for the classification of tumours using gene expression data. Journal of the American Statistical Association, **97**, 77-87.
[2] Thomaz, C.E. and Gillies, D.F. (2005). A maximum uncertainty lda-based approach for limited sample size problems with application to face recognition. In: 18th Brazilian Symposium on Computer Graphics and Image Processing. SIBGRAPI, 89-96.
[3] Fan J. and Fan, Y. (2008). High Dimensional Classification using Features Annealed Independence Rules. Annals of Statistics, **38**, 2605-2637.
[4] Domingos, P. and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, **29**, 103-130.
[5] Tibshirani, R., Hastie, B., Narismhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science, **18** (1), 104-117.
[6] Fisher, T.J. and Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Computational Statistics and Data Analysis, **55**, (5), 1909-1918.
[7] Duarte Silva, A.P. (2011). Two-group classification with high-dimensional correlated data: A factor model approach. *Computational Statistics and Data Analysis*, **55** (11), 2975-2990.
[8] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B* **57** (1), 289-300.
[9] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29** (4), 1165-1188.
[10] Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99** (465), 96-104.
[11] Donoho, D. and Jin, J. (2008). Higher criticism thresholding. Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences*, **105**, 14790-14795.

Rmixmod: A MIXture MODelling R package

R. Lebret[a,1] and **S. Iovleff**[a,2] and **F. Langrognet**[b]

[a]Laboratoire de mathématiques Paul Painlevé
U.M.R. 8524 - CNRS - Université Lille 1 - INRIA Lille Nord-Europe - MODAL Team
Cité Scientifique - 59655 Villeneuve d'Ascq Cedex - FRANCE
[1] remi.lebret@math.univ-lille1.fr
[2] serge.iovleff@math.univ-lille1.fr

[b]Laboratoire de mathématiques de Besançon
U.M.R. 6623 - CNRS - Université de Franche-Comté
16 route de Gray - 25030 Besançon - FRANCE
florent.langrognet@univ-fcomte.fr

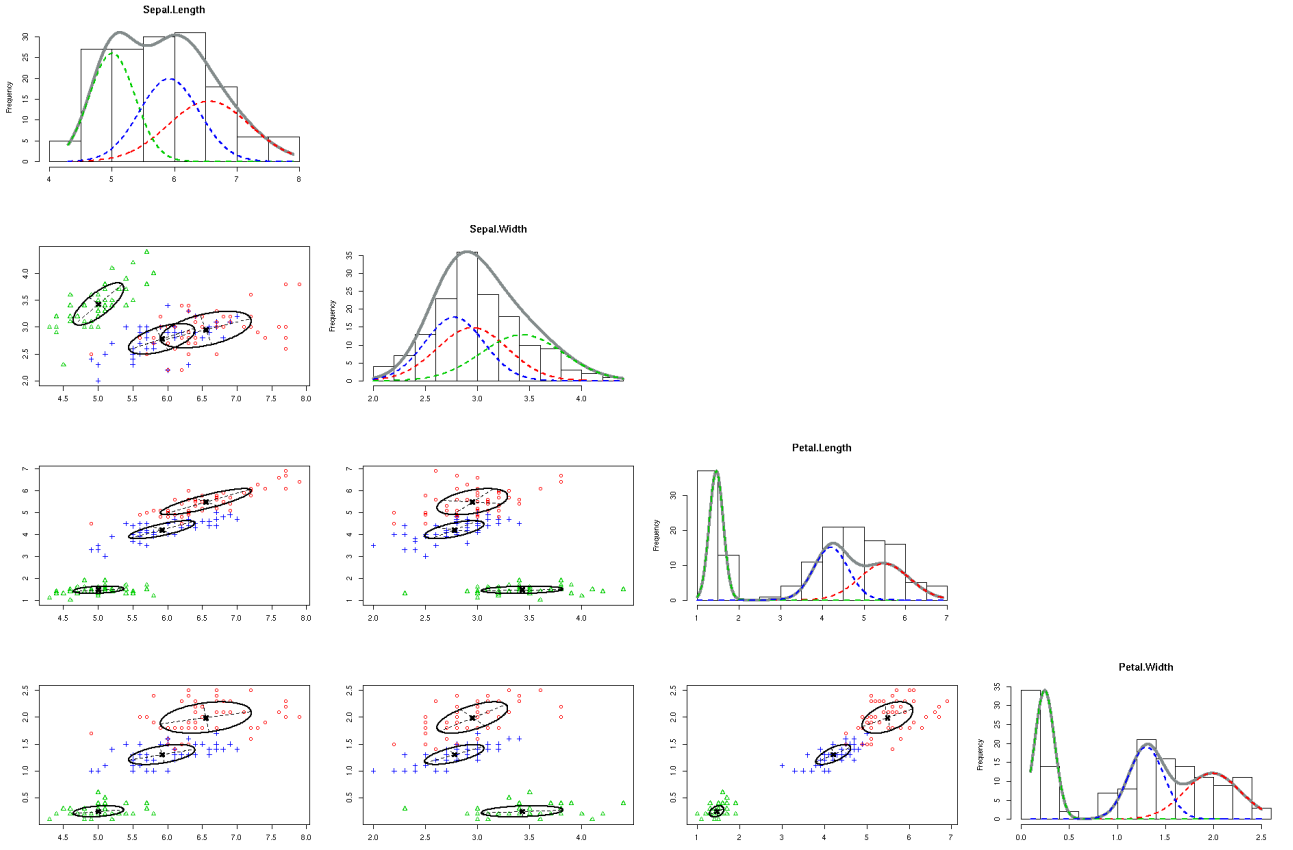
Keywords: model-based clustering, discriminant analysis, visualization, C++, R

Abstract: Mixmod [1] is a well-established software for fitting a mixture model of multivariate Gaussian or multinomial components to a given data set with either a clustering, a density estimation or a discriminant analysis point of view. It is written in C++ and its core library has been interfaced with Scilab and Matlab. It lacked an interface with R. The Rmixmod package provides a bridge between the C++ core library of Mixmod and the R statistical computing environment. Both cluster analysis and discriminant analysis can be now performed using Rmixmod. Many options are available to speficy the models and the strategy to run. Rmixmod is dealing with 28 multivariate Gaussian mixture models for quantitative data and 10 multivariate multinomial mixture models for qualitative data. Estimation of the mixture parameters is performed via the EM, the SEM or the CEM algorithms. These three algorithms can be chained and initialized in several different ways which leads to obtain original fitting strategies. Different model selection criteria are proposed according to the modelling purpose. User-friendly outputs and graphs allow for a good visualisation of the results. Rmixmod is available on CRAN.

An example of clustering in a quantitative case: The outputs and graphs of Rmixmod are illustrated on the well-known iris flower data set. `iris` is a data frame with 150 cases (rows) and 5 variables (columns) named `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, and `Species`. The first fourth variables are quantitative and the `Species` variable is qualitative with 3 modalities. Hence, it is natural to fit a three component Gaussian mixture to this data set to retrieve the true partition. That can be done with the function `mixmodCluster()`:

```
# load Rmixmod package into R environment
R> library(Rmixmod)

# run a cluster analysis on the four quantitative variables of iris with three
# clusters, all the Gaussian models, the BIC and ICL model selection criteria
R> xem <- mixmodCluster(iris[1:4], 3, models=mixmodGaussianModel(), criterion=c("BIC","ICL"))

# show a summary of the best model containing the estimated parameters, the likelihood
# and the criteria values (here the output has been truncated)
R> summary(xem)
*************************************************************
* Number of samples     =  150
* Problem dimension     =  4
```

```
*************************************************************
* Number of cluster =  3
*         Criterion =  BIC(553.4052) ICL(557.6575)
*        Model Type =  Gaussian_p_Lk_Dk_A_Dk
*        Parameters =  list by cluster
*        Cluster  1 :
              Proportion =  0.3333
                   Means =  6.5516 2.9510 5.4909 1.9904
               Variances = |     0.4282     0.1078     0.3310     0.0630 |
                           |     0.1078     0.1155     0.0879     0.0606 |
                           |     0.3310     0.0879     0.3585     0.0831 |
                           |     0.0630     0.0606     0.0831     0.0847 |
*        Cluster  2 :
[ ... ]
*        Cluster  3 :
[ ... ]
*    Log-likelihood =  -186.5112
*************************************************************

# show the partition returned by the mixmodCluster() function
R> xem["partition"]
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 3 1 3
 [75] 3 3 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1
[149] 1 1

# the plot() function has been redefined to get on the same graph:
# - a 1D representation with densities and data
# - a 2D representation with isodensities, data points and partition
R> plot(xem)
```



Reference

[1] Biernacki C., Celeux G., Govaert G., Langrognet F., (2006). Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. *Computational Statistics and Data Analysis*, vol. 51/2, pp. 587-600.

Imputation de données manquantes pour des données mixtes via les méthodes factorielles grâce à missMDA

Vincent Audigier, Francois Husson, Julie Josse

Département de mathématiques appliquées, Agrocampus, Rennes, France
*Contact author : julie.josse@agrocampus-ouest.fr

Keywords: Imputation simple, Données mixtes, Données manquantes, Méthodes factorielles, ACP, ACM

Cette présentation a pour objet une nouvelle méthode d'imputation simple de données mixtes. L'objectif est alors de compléter des tableaux comprenant à la fois des variables quantitatives et qualitatives. L'imputation repose ici sur l'utilisation de méthodes factorielles.

Toutes les méthodes d'analyse factorielle peuvent s'écrire comme une ACP (Analyse en Composantes Principales) ou une décomposition en valeurs singulières d'un tableau de données particulier. L'ACP est donc au cœur de ces méthodes. L'approche classique pour gérer les données manquantes en ACP consiste à minimiser la fonction de coût (l'erreur de reconstitution) sur tous les éléments présents. Ceci peut être effectué à travers un algorithme d'ACP itérative (aussi appelé expectation maximisation PCA, EM-PCA) décrit dans Kiers (1997). Celui-ci consiste à attribuer une valeur initiale aux données manquantes, effectuer l'analyse (ACP) sur le jeu rendu complet, compléter les données manquantes via la formule de reconstitution pour un nombre d'axes fixé, et recommencer ces deux étapes jusqu'à convergence. Les paramètres (axes et composantes) ainsi que les données manquantes sont de cette manière simultanément estimés. Par conséquent cet algorithme peut être vu comme une méthode d'imputation simple. Il souffre cependant d'un problème de surajustement. En conséquence une version régularisée de cet algorithme doit être utilisée (Josse et al., 2009; Ilin and Raiko, 2010) afin de répondre à ce problème. De même un algorithme d'ACM régularisée permet de gérer les données manquante en ACM. Il consiste à effectuer une ACP régularisée sur une matrice judicieusement pondérée (Josse et al., 2012).

L'AFDM (Analyse Factorielle des Données Mixtes) généralise l'ACP et l'ACM, elle permet de traiter à la fois des données quantitatives propres à l'ACP et des variables qualitatives propres à l'ACM. La force de l'AFDM réside donc dans la prise en compte des relations entre individus, au même titre que toutes les autres méthodes factorielles, mais aussi, et c'est là son unicité, dans les relations entre les variables quantitatives et qualitatives équilibrées, renforçant ainsi la qualité d'imputation que l'on aurait eu en utilisant séparément une imputation par ACP et une par ACM. L'équilibre entre les différents types de variables est important au risque d'altérer l'imputation.

Le package **missMDA** (Husson and Josse, 2010) permet de gérer les données manquantes dans les méthodes d'analyse factorielle. Il s'agit d'abord d'imputer les données manquantes à l'aide des fonctions du package, puis d'effectuer l'analyse à l'aide d'un logiciel adapté comme **FactoMineR** (Lê et al., 2008; Husson et al., 2011). Ainsi, **missMDA** permet d'envisager tout type d'analyse et ceci en dépit de l'absence de données.

Bien que le problème de données manquantes sur des données mixtes soit courant, peu de méthodes d'imputation sont disponibles. Une des plus récentes (2011) et offrant de bons résultats est basée sur les forêts aléatoires et donc sur des prédicteurs par arbres (Stekhoven and Buhlmann, 2011). Les comparaisons avec cette méthode d'imputation offrent des résultats comparables et encourageants autant sur des jeux réels que simulés.

Références

Husson, F. and J. Josse (2010). *missMDA : Handling missing values with/in multivariate data analysis (principal component methods)*. R package version 1.2.

Husson, F., J. Josse, S. Le, and J. Mazet (2011). *FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.16.

Ilin, A. and T. Raiko (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research 11*, 1957–2000.

Josse, J., M. Chavent, B. Liquet, and F. Husson (2012). Handling missing values with regularized iterative multiple correspondence analysis. *Journal of classification 29*, 91–116.

Josse, J., J. Pagès, and F. Husson (2009). Gestion des données manquantes en analyse en composantes principales. *Journal de la Société Française de Statistique 150*, 28–51.

Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika 62*, 251–266.

Lê, S., J. Josse, and F. Husson (2008, 3). Factominer : An r package for multivariate analysis. *Journal of Statistical Software 25*(1), 1–18.

Stekhoven, D. and P. Buhlmann (2011). Missforest - nonparametric missing value imputation for mixed-type data. *Bioinformatics 28*, 113–118.

Package 'marqLevAlg' - Algorithme de Levenberg-Marquardt en R : Une alternative à 'optimx' pour des problèmes de minimisation

M. Prague[a], **A. Diakite**[a] and **D. Commenges**[a]

[a]Épidémiologie-Biostatistique
Univ. Bordeaux, ISPED
INSERM, Centre INSERM U897
F-33000 Bordeaux, France
melanie.prague@isped.u-bordeaux2.fr

Mots clefs : Optimisation, Minimisation, Algorithm de Marcquardt-Levenberg, RDM (Distance Relative au Minimum), Package 'optimx'.

La méthode d'optimisation de Levenberg-Marquardt est particulièrement robuste et efficace. Elle est devenue un algorithme de référence pour la minimisation de fonctions. Cependant, aucune implémentation en R n'existait. Nous proposons le package 'marqLevAlg' implémentant l'algorithme de Levenberg-Marquardt sans contrainte [1]. Nous présenterons les spécificités de son utilisation ainsi qu'une comparaison avec les algorithmes existants (Nelder-mead, BFGS ...).

L'algorithme itératif de Levenberg-Marquardt permet de trouver un minimum local (éventuellement global) d'une fonction continue dérivable deux fois. Lorsque l'estimation à l'itération courante est loin du minimum, la matrice hessienne est souvent non inversible, l'algorithme de Levenberg-Marquardt permet de gonfler la diagonale pour proposer malgré tout une direction. À proximité du minimum, il correspond à l'algorithme de Newton-Raphson. L'implémentation du package 'marqLevAlg' ne comporte aucune variante particulière quant à l'algorithme lui-même.

Concernant les critères de convergence, deux critères secondaires (stabilisation des estimations et de la valeur de la fonction) et un critère principal (la Distance Relative au Minimum - RDM) sont implémentés. Leurs valeurs seuils sont modifiables par l'utilisateur. Le RDM [3] est un critère original correspondant à la norme des gradients dans la métrique des paramètres à estimer divisée par le nombre de paramètres à estimer pour s'adapter à la dimension du problème. Il est aussi interprétable comme le ratio entre l'erreur numérique et l'erreur statistique commise. Ainsi, sa valeur doit être aussi proche de zéro que possible et en tout cas inférieure à 1. Nous présenterons ce nouveau critère en insistant sur sa signification et ses propriétés d'invariance.

Nous comparerons les résultats obtenus avec 'marqLevAlg' avec ceux du package 'optimx' [2] qui fait référence dans le domaine. Nous utiliserons plusieurs exemples dont ceux disponibles dans le manuel d' 'optimx'. Avec des temps de calculs comparables, les estimations s'avèrent parfois meilleures en particulier pour des points de départ loin du minimum ou des surfaces non strictement convexes ou éloignées d'une forme quadratique. Les critères de comparaison principaux seront la valeur des estimations et de la fonction au point de convergence ainsi que la stabilité et la reproductibilité de ce minimum.

Références

[1] Marquardt, D.W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, **11**(2), 431-441

[2] Nash, J.C. and Varadhan, R. (2011). Unifying Optimization Algorithms to Aid Software System Users: optimx for R. *Journal of Statistical Software*, **49**(3), 1-14

[3] Commenges, D., Jacqmin-Gadda, H., Proust, C. and Guedj, J. (2006). A newton-like algorithm for likelihood maximization: The robust-variance scoring algorithm. *Arxiv preprint math/0610402*

Méthodologie pour le traitement des données écologiques de type inventaire avec EcoMineR

G. Bessigneul[a], **F.Collin**[b], **M. Gauthier**[c] **M. Gérard**[d] and **S. Lê**[e]

[abcde] Laboratoire de Mathématiques Appliquées
Agrocampus Ouest
65, rue de Saint Brieuc, CS 84215,
35042 Rennes Cedex, France
[a]guillaume.bessigneul@agrocampus-ouest.fr
[b]francois.collin@agrocampus-ouest.fr
[c]marion.gauthier@agrocampus-ouest.fr
[d]marianne.gerard@agrocampus-ouest.fr
[e]sebastien.le@agrocampus-ouest.fr

Mots clefs : Statistique, Données écologiques, Cartographie, Visualisation, Analyse factorielle, EcoMineR.

L'écologie vise à étudier les associations entre êtres vivants ainsi que les associations environnement/êtres vivants. Pour cela, les écologues réalisent très largement des inventaires faunistiques ou floristiques qui dénombrent les espèces répertoriées dans différents sites. Ces jeux de données sont donc fréquemment rencontrés en écologie. La description des sites par leur composition spécifique (en terme de présence/absence, abondance, biomasse, etc.) ou bien de manière complémentaire, la description des espèces par les sites qu'elles occupent, constitue l'information principale. À cela peuvent s'ajouter les coordonnées des sites qui permettent de réaliser le lien à l'organisation spatiale. Enfin, ces données peuvent être complétées par de l'information supplémentaire caractérisant les sites d'une part, et les espèces d'autre part.

Le but de notre démarche a été de proposer, à partir des questions écologiques récurrentes, une méthodologie claire et accessible à tous pour analyser ce type de données. De plus, nous avons en parallèle développé un package de fonctions R utiles à leur analyse statistique : EcoMineR.

Les questions abordées par la méthodologie proposée et EcoMineR sont les suivantes :
- Comment sont distribuées les espèces ?
- Existe-t-il, sur l'aire géographique considérée, des profils de composition spécifique des sites ou bien des profils de répartition d'espèces ? Et peut-on rapprocher ces profils des caractéristiques des espèces ou des sites ?
- Comment intégrer de l'information spatiale à ces relevés ?
- Comment coupler différents points de vue portés (moments différents, techniques différentes de mesure de l'information ou de prélèvement...) sur une même zone géographique ?

Pour répondre à ces questions, nous utilisons dans un premier temps des indicateurs numériques classiques de positions et de dispersions qui permettent de rendre compte des caractéristiques principales du relevé écologique.

Puis, nous proposons une utilisation des méthodes exploratoires multivariées telles que l'analyse des correspondances permettant de rendre compte des structures de liaison qu'on peut retrouver

entre espèces, entres sites ou entre sites et espèces. Mais ce ne sont pas les seuls outils utilisés. Ceux-ci ne peuvent pas informer d'une réalité géographique.

En effet, décrire les caractéristiques d'un inventaire uniquement par des coordonnées ne permet pas de prendre en compte les caractéristiques du milieu et tous les éléments de rupture existant (un rocher, une zone forestière...). Il est donc intéressant de proposer des outils de visualisation basés sur des cartes. Nos outils, réalisés à partir du package RgoogleMaps, permettent de projeter de l'information ponctuelle sur une carte(Figure 1). Cette information peut provenir directement de l'inventaire, mais elle peut également être issue d'analyses exploratoires. On s'intéresse ainsi à projeter des coordonnées factorielles sur une carte pour mettre en évidence les liaisons entre dimensions de variabilité et structure géographique (Figure 2).

L'ensemble de cette réflexion a donné lieu à la rédaction d'un document pédagogique disponible en ligne et la conception d'un package pour soutenir la méthodologie proposée par notre travail.

Références

1. Borcard, D., Gillet, F. & Legendre, P. (2011). *Numerical Ecology with R*. Springer.
2. Hengl, T. (2011). *A Practical Guide to Geostatistical Mapping*. University of Amsterdam
3. http ://cran.r-project.org/web/packages/RgoogleMaps/index.html
4. http ://cran.r-project.org/web/packages/gstat/index.html
5. http ://cran.r-project.org/web/packages/fields/index.html
6. http ://cran.r-project.org/web/packages/FactoMineR/index.html
7. http ://factominer.free.fr/
8. http ://pbil.univ-lyon1.fr/R/enseignement.html

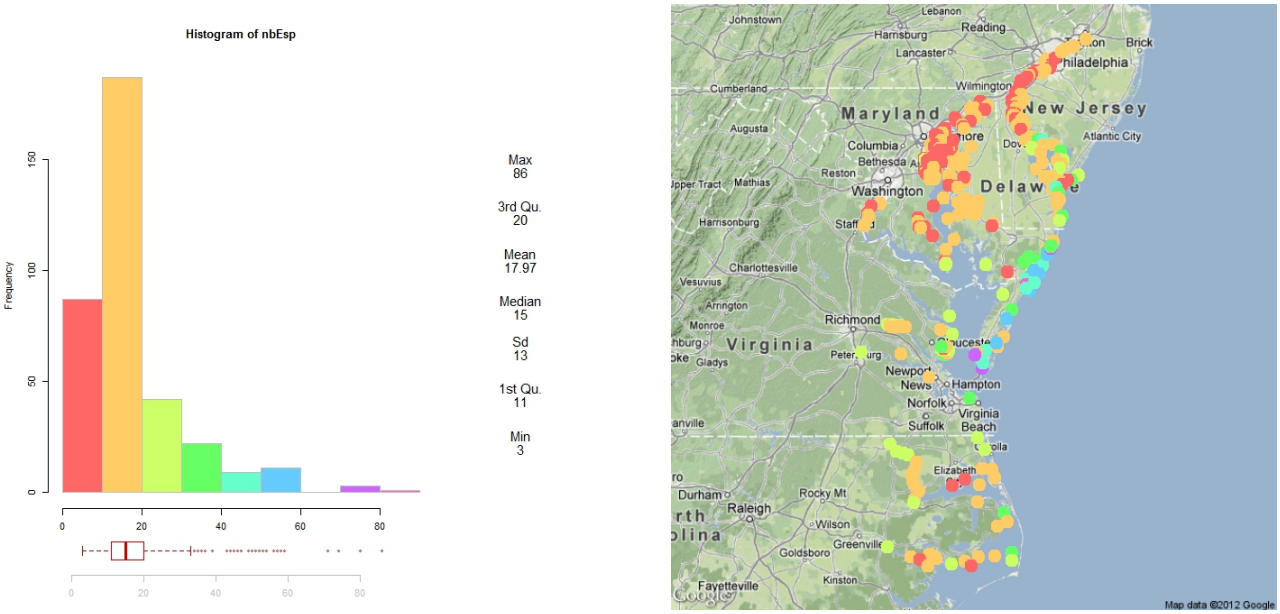


Figure 1: *Histogramme du nombre d'espèce et projection sur une carte*

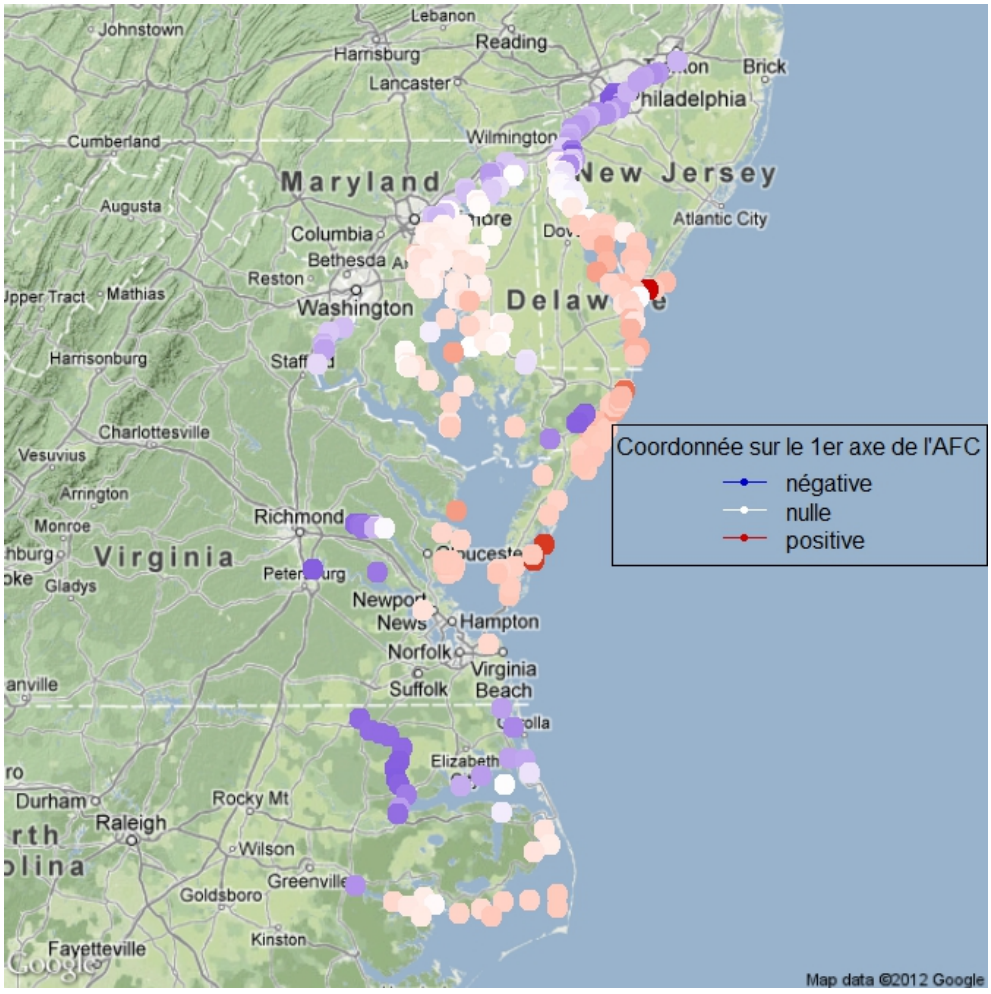


Figure 2: *Résultats d'une AFC sur l'abondance des espèces projetés sur la carte*

`SVGMapping`: an R package to map *omic* data sets onto pathways templates

R Champeimont[a,b], **C Leplat**[a], **F Chauvat**[a] and **JC Aude**[a]

[a]Integrative Biology and Molecular Genetics Department
CEA, Institute of Biology and Technology Saclay
F-91191 Gif sur Yvette, France
christophe.leplat@cea.fr
franck.chauvat@cea.fr
jean-christophe.aude@cea.fr

[b]Génomique des Microorganismes
UMR 7238 CNRS-UPMC
15 Rue de l'Ecole de Médecine, 75006 Paris, France
raphael.champeimont@upmc.fr

Mots clefs : Vizualisation, Biology, Pathways, Microarrays, High throughput assays.

High-throughput *omic* technologies are now commonly used in large-scale experimental biology. The main characteristic of these *omic* approaches is that they usually produce large amounts of data. Results obtained through these analyses are mostly interpreted or assessed in terms of given hypotheses. In most cases, huge amount of results need to be transformed (*eg* using classification methods), integrated with other biological knowledge (*eg* pathways), and explored using mainstream or dedicated visualisation tools. Then, they can be meaningfully interpreted by biologists. Visualisation is crucial for an optimal understanding of the results emerging from the concerted analysis of shared material between experimental and computational researchers.

Directed visualisation methods [1] use *prior* knowledge in their process. In biology this knowledge is often depicted by networks. For example, Momin & *al.* [2] designed a method that combines a visualisation method and a prediction process to map transcriptomic data with predicted metabolite pools into pathways. Here we report `SVGMapping` [3], an R package to map *omic* experimental data onto custom-made templates which can be used to depict metabolic pathways, cellular structures or biological processes. `SVGMapping` allows the modification of color, opacity or shape of given graphical elements. It can be applied several times on the same template to combine various *omic* data types (*eg* protein and metabolite concentrations). This package has been designed to integrate the wealth of data generated by various strains (*eg* mutants *vs* wild-type), growth conditions (*eg* before *vs* after stress) or kinetic experiments.

Templates: In the `SVGMapping` framework, a template is an SVG file where shapes are specifically labeled. Labels are specified as *attributes* assigned to any kind of shape. These labels are used to pinpoint all attribute modifications (*ie* colors or opacity) to apply on the template. We have selected the SVG format for its versatility as a web-based and portable standard that can be rendered in many other graphics formats (*eg* PNG or PDF). Furthermore, javascript code can be embedded into SVG files to provide an interactive experience to the user when viewed within a compatible browser.

Omic **data mapping:** experimental data are provided as a numeric matrix M with as many columns as conditions. Each row is labeled with a unique identifier (*eg* gene, protein or metabo-

lite). This identifier will be used to track the template shapes to modify.
For single condition experiments, M values can be uniformly bound to a set of colors or a color gradient to modify the filling or stroke colors of shapes (see figure 1). For experiments with multiple conditions, one can use pie charts or colored stripes. In both cases, the filling color of each slice/stripe is set according to the M value of the related shape identifiers (*ie* rows) and conditions (*ie* columns).
Besides this common usage, one can use M values to alter the *opacity* or the *stroke width* of shapes. Another specific use is to simulate the filling of Erlenmeyers. In this case, M values should be given as a proportion of the complete filling. This mode is of particular interest to simulate the relative concentrations of metabolites. Finally, since SVG is a web centric format, we have implemented a mechanism to add tooltips and hyperlinks (as URLs) to each shape.

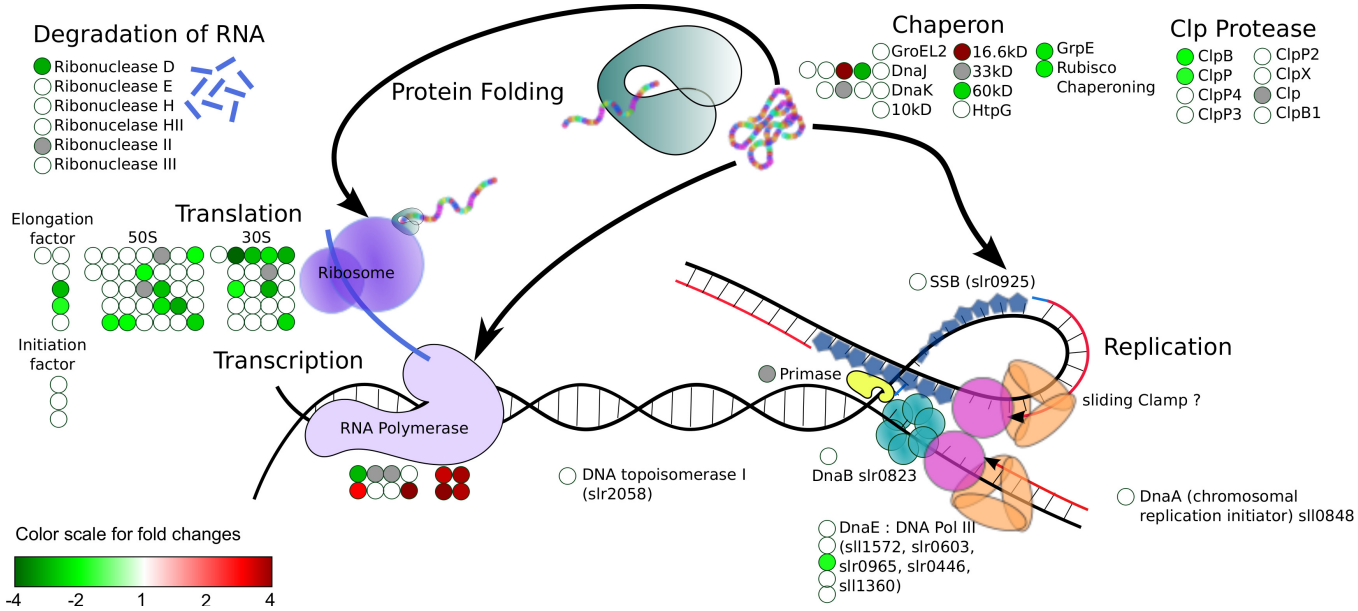


Figure 1: This figure represents DNA metabolic processes in *Synechocystis sp. PCC 6803* and illustrates a possible use of the `SVGMapping` package. Red (*resp.* green) circles depict induced (*resp.* repressed) genes involved in each biological mechanism corresponding to a microarray experiment in which cells where exposed to 3mM H_2O_2 for 30 minutes [4]. Matches between expression fold-changes and color levels can be obtained using the scale on the lower left of the figure. Circles filled in grey are related to probes that were not hybridized on the microarrays. Notice that many genes involved in protein translation were repressed while genes of RNA polymerase transcription were induced.

References

[2] Keim, D. A. & *al.* (2006).Challenges in Visual Data Analysis. *Tenth International Conference on Information Visualisation (IV'06).*

[2] Momin, A. & *al.* (2011).A method for visualization of "omic" datasets for sphingolipid metabolism to predict potentially interesting differences. *The Journal of Lipid Research*, **52**(6), 1073-83

[3] http://svgmapping.r-forge.r-project.org/

[4] Houot, L. & *al.* (2007).Cadmium triggers an integrated reprogramming of the metabolism of Synechocystis PCC6803, under the control of the Slr1738 regulator. *BMC Genomics*, **8**, 350

Multiple Factor Analysis for Contingency Tables in FactoMineR Package

Belchin Adriyanov-Kostov [a], **Mónica Bécue-Bertaut** [b], **François Husson** [c], **Daría Hernández** [d]

[a] Primary Health Care Center Les Corts, CAPSE
Mejia Lequerica, s / n, 08028 Barcelona (Spain)
badriyan@clinic.ub.es

[b] Universitat Politècnica de Catalunya
Jordi Girona 1-3, 08034 Barcelona (Spain)
monica.becue@upc.edu

[c] Agrocampus Rennes
65 rue de Saint-Brieuc, 35042 Rennes (France)
husson@agrocampus-ouest.fr

[d] Centro Mexicano de Estudios Económicos y Sociales
Napoleón 54, Col. Moderna 3500, México D.F. (México)
dari_hdez@yahoo.com.mx

Keywords: Multiple Contingency Tables, Multiple Factor Analysis, Multiple Factor Analysis for Contingency Tables, Scientometrics, FactoMineR

FactoMineR package [1] offers the most commonly used principal component methods: principal component analysis (PCA), correspondence analysis (CA), multiple correspondence analysis (MCA) and multiple factor analysis (MFA) [2]. MFA function has been recently extended to consider contingency/frequency tables as proposed by Bécue-Bertaut and Pagès (multiple factor analysis for contingency tables, MFACT) [3-4]. MFACT is used in different domains such as sensometrics, ecology and text mining.

Multiple factor analysis [2] deals with a multiple table, composed of sets of either quantitative or categorical variables balancing the influence of the different sets on the first principal axis by dividing the weight of their variables/columns by the first eigenvalue of the separate analysis of this set (PCA or MCA depending on the type of the variables). Thus, the highest axial inertia of each group is standardized to 1. MFA offers the usual results in any PCA (global representation of rows and columns) and also tools for comparing the different sets such as the superimposed representation of the rows as induced separately by every set of columns (partial rows). Initially multiple factor analysis for contingency tables [3] was proposed to simultaneously analyze several frequency/contingency tables. Afterward, it has been extended to multiple tables with a mixture of quantitative, categorical and frequency sets [4].

This method is presented through its application to a scientometric study in medicine. 457 abstracts relative to randomized clinical trials on *Systemic Lupus Erythematosus* (SLE) from 1994 to 2011 were downloaded from PubMED, the most important scientific data base in medical research. The abstracts×words matrix was constructed, keeping only the words repeated at least 10 times. The publication year was also considered. Thus, the data base juxtaposes a quantitative set (publication year) and a frequency set (abstracts×words a 457×1046 matrix). The aim was to study the evolution in the research concerning this rare disease, through the vocabulary changes. The superimposed representation provides a graph where the abstracts are seen globally (one global

point) and partially (two partial points) from the two different points of view that are the vocabulary of the abstract and its publication year. This representation is considered to look for those abstracts evidencing an important difference between both points of view to detect either pioneering works or works returning to topics treated in the past.

Figure 1 displays the first principal global map provided by MFACT. The interpretation rules are those of CA for the words and those of PCA for the quantitative columns, here reduced to the publication year. The different years are projected as supplementary categorical columns. Year is highly correlated (0.94) with the first dimension, opposing words related to symptoms and drugs (at the left) to etiology and genetics (at the right). The part of the evolution of the vocabulary linked to chronology is reflected on this axis. The early research in SLE was dedicated to detect its symptoms and to test the effectiveness of drugs already known and previously used for other diseases. The most recent research works are concerned by etiology, in particular through genetics. The second dimension opposes topics present in the research at a same moment such as symptoms and drugs at the beginning and genetics and innovative drugs such as “Belimumab” in the recent years.
Figure 2 offers the superimposed representation of the global and partial points on the first principal plane. The partial points with the highest differences on the first dimension (indicating a chronological gap) are underlined. Two recent works (noted by 4 and 5) are related to former topics (drugs and symptoms). Three works (noted by 1, 2 and 3) can be considered as pioneering works, as they are dedicated to a genetic approach at a date as early as 1996.

Figure 1. Global representation

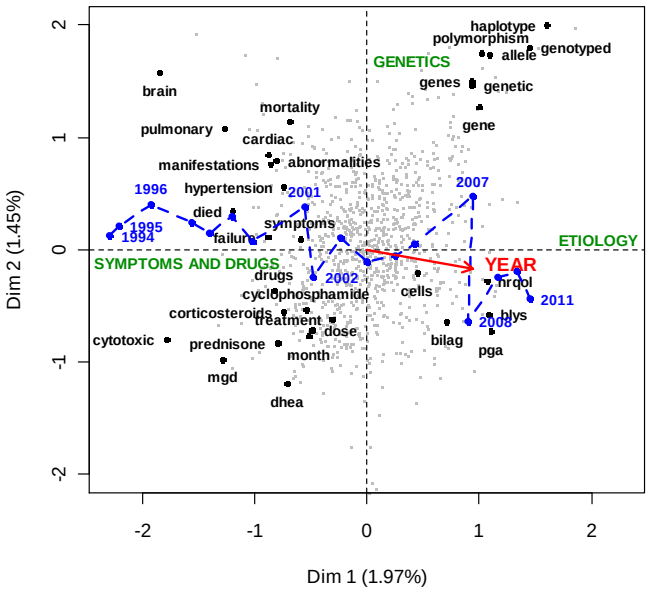
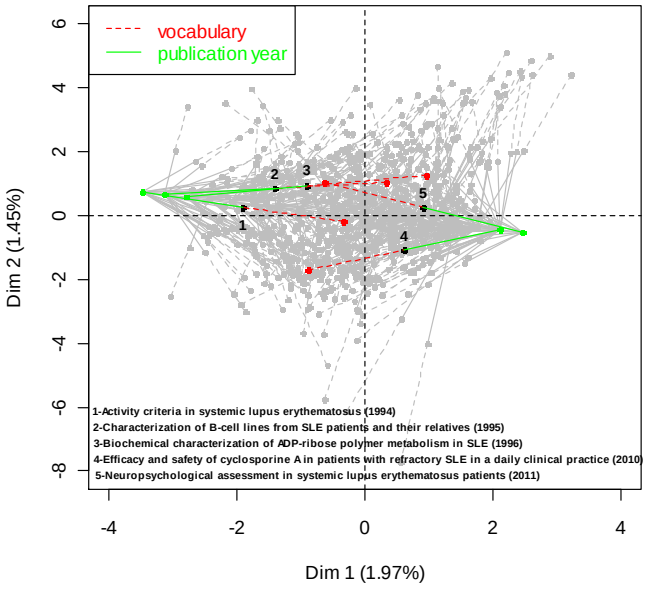


Figura 2. Superimposed representation



Références

[1] Lê, S., Josse, J., Husson, F. (2008). Factominer: An r package for multivariate analysis. *Journal of Statistical Software*, **25**(1), 1–18

[2] Escofier, B., Pagès, J. (1990). *Analyses factorielles simples et multiples: objectifs, méthodes, interprétation.* Dunod, Paris

[3] Bécue-Bertaut, M., Pagès, J. (2004). A principal axes method for comparing multiple contingency tables: MFACT. *Computational Statistics and Data Analysis*, **45**, 481–503

[4] Bécue-Bertaut, M., Pagès, J. (2008). Multiple factor análisis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics and Data Analysis,* **52**, 3255–3268

Visualisation de données multivariées : réimplémentation des fonctionnalités graphiques de la librairie `ade4`

A. Julien-Laferriere[a] and **S. Dray**[b]

[a]Laboratoire de biométrie et biologie évolutive (UMR CNRS 5558)
CNRS - Université Lyon 1
43 bd du 11 novembre 1918, 69622 Villeurbanne, France
alice.julien-laferriere@univ-lyon1.fr

[b]Laboratoire de biométrie et biologie évolutive (UMR CNRS 5558)
CNRS - Université Lyon 1
43 bd du 11 novembre 1918, 69622 Villeurbanne, France
stephane.dray@univ-lyon1.fr

Mots clefs : Analyse multivariée, Graphique, Visualisation.

Dans un grand nombre de disciplines (e.g., écologie, génétique, santé), les récents développements technologiques facilitent la collecte et la gestion de données et conduisent à l'élaboration de bases de données massives dont la structure est de plus en plus complexe (multivariée, hiérarchisée, structurée dans l'espace et/ou le temps, etc.). L'analyse et la représentation de ces données nécessitent des méthodes adaptées prenant en compte leurs caractéristiques intrinsèques. Dans ce contexte, les méthodes d'analyse multivariée fournissent un ensemble d'outils permettant de résumer l'information contenue dans de grands tableaux en identifiant les relations entre variables, et les similarités entre individus. Les résultats sont alors présentés sous la forme de graphiques, pour un nombre réduit de dimensions, permettant une exploration des principales structures identifiées dans les données.

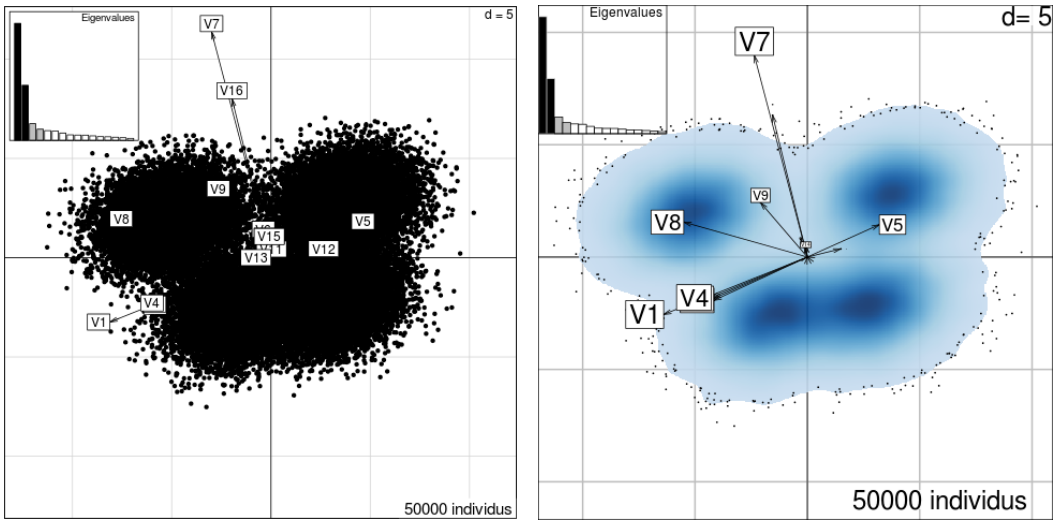
Depuis 2002, le package R `ade4` [1], développé au laboratoire de Biométrie et Biologie Évolutive, fournit un ensemble de méthodes permettant l'analyse d'un seul, de deux mais aussi de K tableaux. A ce jour, une quarantaine de méthodes différentes ont été implémentées, dont près de la moitié ont été développées par les auteurs du package. Une quarantaine de fonctions graphiques sont également disponibles afin de représenter les résultats issus de ces analyses. Près de dix ans après la première distribution d'`ade4` sur les serveurs du CRAN, nous sommes en train de mettre en place de nouvelles modalités de représentation graphique permettant une utilisation plus souple et plus intuitive du logiciel. L'objectif est d'améliorer la visualisation des données et/ou des résultats d'analyses en s'appuyant sur les nouvelles fonctionnalités offertes par R.

Cette nouvelle implémentation est réalisée avec une programmation orientée objet (S4) en s'appuyant sur une hiérarchisation des différentes représentations graphiques disponibles. Les graphiques sont alors stockées sous la forme d'objets et il est ainsi possible de les créer sans les visualiser ou de les manipuler *a posteriori*. Ces objets peuvent être combinés (juxtaposition, superposition) afin d'observer, dans une même fenêtre graphique, différents niveaux d'information.

Ces nouvelles fonctionnalités graphiques s'appuient sur l'utilisation du package `lattice` [2] qui permet d'obtenir une grande souplesse dans la production de graphiques conditionnels pour

l'exploration de données multi-dimensionnelles.

En tirant profit des fonctionnalités implémentées dans `lattice`, nous avons mis en place deux grandes classes d'objets spécifiquement associées à la représentation graphique de données sous `ade4`. La première nous permet de définir une série de graphiques élémentaires utilisés en analyse multivariée alors que la seconde classe permet de gérer une collection de graphiques obtenus par superposition ou juxtaposition. De plus, de nombreux paramètres sont disponibles afin de personnaliser facilement les graphiques et mettre ainsi en relief les principales structures associées à un jeu de données particulier (partition en groupe, structure spatiale, etc.).



(a) (b)

FIGURE 1 – Exemple et comparaison de graphiques obtenus dans `ade4` : projection des variables et individus sur les deux premiers axes d'une analyse en composantes principales (ACP). (a) Ancienne implémentation, (b) Nouvelle implémentation. En haut à gauche sont représentés les valeurs propres de l'analyse.

Sur la figure 1 sont représentés une partie des résultats d'une analyse en composantes principales sur des données fictives comportant seize variables et cinquante mille individus. La figure obtenue avec l'ancienne implémentation (Fig. 1a) ne permet pas d'observer clairement la distribution des individus. Sur la figure 1b, un des nouveaux graphiques disponibles permet de représenter les individus non par des points mais par une nappe de densité qui permet de mieux rendre compte de la distribution de ceux-ci. Enfin, ici, le graphique a été personnalisé pour augmenter la taille du titre mais aussi changer la taille des étiquettes des variables qui est proportionnelle à leurs contributions sur les deux axes représentés.

Ce travail constitue donc une étape importante pour le package `ade4` en améliorant sensiblement les fonctionnalités actuelles et en offrant un cadre général et flexible facilitant l'implémentation de futurs outils graphiques.

Références

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

[2] D. Sarkar. Lattice : multivariate data visualization with R *Springer Verlag*, 2008

Locally-Weighted Partial Least Squares Regression for infrared spectra analysis

A. Thébault and D. Juery

UMR MISTEA
INRA Montpellier
2 place Pierre Viala, 34060 Montpellier Cedex 2
Contact author : aurelie.thebault@supagro.inra.fr

Keywords : Partial least square regression, Neighbour selection, Local calibration

Recent developments in infrared spectroscopy offer advantages in terms of speed and simplicity for routine analysis of chemical or biological data. Many chemometric tools have been developped to predict chemical and biological properties of samples from their infrared spectrum. Among them, the Partial Least Squares regression (PLS) is widely used to get information from large spectral databases. However, while PLS is efficient for predicting biological or chemical parameters from homogeneous spectral databases, it performs poorly on heteregeneous databases, leading to the use of local PLS. In local PLS, the chemical or biological parameters of an unknown sample are predicted from a subset of selected, spectrally similar, samples from a largest heterogeneous spectral database.

Several R-packages such as **pls** (Mevik et al., 2012), **plspm** (Sanchez and Trinchera, 2012), **plsRglm** (Bertrand et al., 2011), **ChemoSpec** (Hanson, 2012) and **soil.spec** (Terhoeven-Urselmans, 2012) already deal with spectral data analyses but none of them propose local calibration. Based on the LOCAL calibration algorithm (Shenk et al., 1997) and the Locally Weighted Regression (LWR ; Naes et al., 1990), we developed a powerful routine associating local calibration and weighting of the calibration samples for prediction of chemical or biological parameters from large and heterogeneous spectral databases.

The Locally-Weighted PLS routine allows optimization of spectral pre-processing within a large choice of available pre-treatments. The selection of the local calibration samples can be done either through the use of a metrics (Mahalanobis distance or correlation coefficient) or through Dirichlet Process (Neal, 2000). An optimization function of the number of local calibration samples is also available. The locally-weighted PLS routine achieves a trade-off between Locally Weighted Regression (Naes et al., 1990) and LOCAL algorithm (Shenk et al., 1997). Weights can be assigned to the selected calibration samples following the Locally Weighted Regression theory. Since the best number k of PLS components is unknown *a priori*, we adopt a model averaging strategy adapted from the LOCAL algorithm: the final estimator is based on the aggregation of many PLS estimators with an increasing number of components.

The steps of the Locally-Weighted PLS will be briefly presented. In a second time, the routine will be applied on a real dataset to present main functions and options.

References

[1] Mevik, B.H., Wehrens, R., Liland, K.H. (2012). pls – Partial least squares and principal component regression. R package version 2.3-0.
http://www.cran.r-object.org/package=pls/.
[2] Sanchez, G., Trinchera, L. (2012). pls pm– Partial least squares data analysis methods. R package version 0.2-2.
http://www.cran.r-object.org/package=plspm/.

[3] Bertrand, F., Meyer, N., Maumy-Bertrand, M. (2011). plsRglm – Partial least squares regression for generalized linear models. R package version 0.7-6.
http://www.cran.r-object.org/package=plsRglm/.
[4] Hanson, B.A. (2012). ChemoSpec – Exploratory chemometrics for spectroscopy. R package version 1.50-2.
http://www.cran.r-object.org/package=ChemoSpec/.
[5] Terhoeven-Urselmans, T. (2012). soil.spec – Spectral data exploration and regression functions. R package version 1.4.
http://www.cran.r-object.org/package=soil.spec/.
[6] Shenk, S., Berzaghi, P., Westerhaus, M.O. (1997). Investigation of a LOCAL calibration procedure for near infrared instruments. J. Near Infrared Spectrosc. 5, 223-232.
[7] Naes, T., Isaksson, T, Kowalski, B. (1990). Locally weighted regression and scatter correction for near-infrared reflectance data. Analytical Chemistry 62(7), 664-673.
[8] Neal, R.M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. J. Computational and Graphical Statistics 9(2), 249-265.

Application de modèles non paramétriques sous R pour l'analyse et le suivi de la qualité de l'eau

M. Sow[a,b], G. Durrieu[c], D. Tran[a,b], P. Ciret[a,b] et J.C. Massabuau[a,b]

[a]Univ. de Bordeaux. EPOC, UMR 5805, F-33120 Arcachon, France
[b]CNRS, EPOC, UMR 5805, F-3312 Arcachon, France
{m.sow, d.tran, p.ciret, jc.massabuau}@epoc.u-bordeaux1.fr

[c] Laboratoire de Mathématiques de Bretagne Atlantique UMR CNRS 6205
Université de Bretagne Sud
Campus de Tohannic, 56017 Vannes
gilles.durrieu@univ-ubs.fr

Mots clefs : Biologie, Environnement, Régression non paramétrique, Estimateur à noyau, Valvometrie HFNI, R.

A l'heure où nos sociétés sont pleinement conscientes que la protection de l'environnement est un enjeu de société majeur, l'étude des milieux aquatiques côtiers apparaît comme une question privilégiée. Il s'agit de préserver des zones riches et sensibles particulièrement à risque. Dans ce contexte, des règlementations et de nombreux contrôles de la qualité de l'eau sont mis en place. De nouvelles techniques et de nouveaux outils doivent être imaginés et développés. Parmi ces techniques de contrôles, les bioindicateurs sont de plus en plus utilisés et sont très efficaces par leurs capacités à révéler la présence de traces (concentrations très faibles) de contaminants comme l'accumulation dans des tissus animaux ou végétaux spécifiques ou des modifications sur des structures populationnelles. Nous travaillons ici sur le développement d'un outil d'analyse haute fréquence (10 Hz en continue pendant au moins un an) du comportement d'huîtres, qui permet d'aborder leur éthologie et l'expression de leurs rythmes biologiques, mais qui est aussi un formidable moyen de surveillance de la qualité de l'eau. Le cas des huîtres est particulièrement intéressant car se sont des animaux sédentaires qui peuvent être témoins de pollutions locales (bioindicateurs) et on les trouve partout dans le monde, des tropiques aux pôles (possibilité d'application très large).

L'objet de ce travail est de modéliser et analyser l'important volumes de données acquis à haute fréquence en implémentant des modèles de régression non paramétriques et des codes de calculs sous R ([1]-[4]) pour l'analyse des données collectées sur différents sites (bassin d'Arcachon, Nouvelle Calédonie, Espagne, Bretagne, Norvège, Russie, …) dans le but de mettre en place un outil de surveillance en ligne de la qualité de l'eau basé sur l'analyse en continue du comportement de bivalves. Plus précisément, à partir d'un échantillon composé de 54000 couples indépendants de variables aléatoires $(T_1, Y_1),\ldots,(T_n, Y_n)$ qui sont le temps en heure et l'écartement valvaire en mm, nous considérons le modèle de régression non paramétrique donné, pour $i = 1,\ldots,54000$, par

$$Y_i=m(T_i)+\varepsilon_i.$$

Dans ce modèle intervient une fonction m inconnue à estimer qui exprime la valeur moyenne de l'écartement valvaire de nos bivalves en fonction du temps T et un terme aléatoire d'erreur ε de loi inconnue et indépendant de T. Nous proposons plusieurs estimateurs non paramétriques de la fonction m.

Le but des méthodes statistiques et des codes de calculs développés (R et scripts Bash) a d'abord été de mettre en évidence et d'extraire des rythmes biologiques. Elles ont été ensuite utilisées pour les présenter sous forme graphique simple afin de permettre, par la caractérisation de perturbations de ces rythmes, de détecter *in fine* une éventuelle pollution du milieu. Dans un premier temps, nous représentons graphiquement les périodes d'ouvertures et fermetures des bivalves sur l'ensemble des données disponibles depuis mars 2006 au niveau de la jetée d'EYRAC. Ces enregistrements permettent de mettre en évidence des rythmes biologiques liés aux rythmes des marées chez l'huître ([4]). Une de nos hypothèses de travail est que la modification de ces rythmes biologiques pourrait alors être symptomatique d'un problème au niveau de la qualité de l'eau. La figure 1 nous montre que les activités de fermetures sont corrélées à la marée et plus précisément nous pouvons observer que les huîtres ferment leurs valves à l'étale de basse mer.

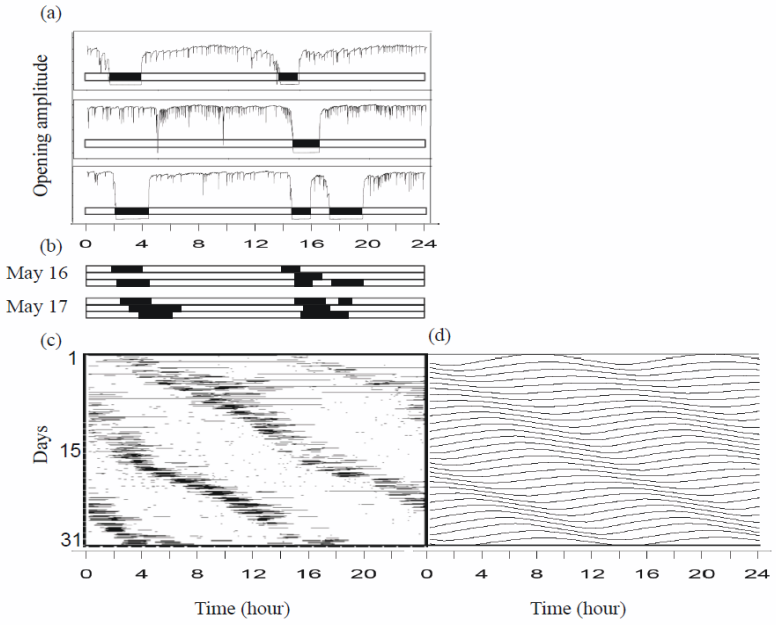


Figure 1. Principe du rythme biologique lié à la marée :
a) activité de fermetures / ouvertures pour 3 différents huîtres (trait noir fermeture et trait blanc ouverture) ;
b) superposition de fermetures / ouvertures de 3 huîtres différentes (16 et 17 Mai) ;
c) représentation de l'activité de fermeture/ouverture sur 31 jours (pour chaque jour, il y'a 16 lignes représentant les périodes de fermetures/ouvertures pour 16 huîtres) ;
d) Représentation de l'évolution de la hauteur d'eau par heure.

Actuellement et grâce aux résultats de nos travaux, l'acquisition, le transfert, et le traitement des données fonctionnent de manière automatique pour les différents sites où un système est positionné dans le monde. Les enregistrements et les résultats du traitement statistique sont accessibles sur le site web "L'œil du mollusque" (http://molluscan-eye.epoc.u-bordeaux1.fr/).

Références

[1] Coudret R., Durrieu G., Saracco J. (2012). Estimateurs a noyau bimodaux d'une densité bimodale et comparaison avec d'autres estimateurs non paramétriques, *Proc de la société Française de Statistique*, sous presse.

[2] Durrieu G., Nguyen T.M.N., Sow M. (2009). Comparaison d'estimateurs de régression non paramétriques : application en valvometrie, *Proc. de la société Française de Statistique*, http://hal.inria.fr/docs/00/38/67/16/PDF/p147.pdf.

[3] Schmitt F.G., De Rosa M., Durrieu G., Sow M., Ciret P., Tran D., Massabuau J.C. (2011). Statistical study of bivalve high frequency microclosing behavior: scaling properties and shot noise analysis, *International Journal of Bifurcation and Chaos*, 21, 3565-3576.

[4] Sow M. Durrieu G., Briollais L., Ciret P., Massabuau J.C. (2011). Water quality assessment by means of HFNI valvometry and high-frequency data modeling, *Environmental Monitoring and Assessment*, 182, 155-170.

OKM : une librairie R pour la classification recouvrante

G. Cleuziou et **L. Rousseau**

Laboratoire d'Informatique Fondamentale d'Orléans
Rue Léonard de Vinci - B.P. 6759
F-45067 ORLEANS Cedex 2
guillaume.cleuziou@univ-orleans.fr rousseauleo1@gmail.com

Mots clefs : Classification automatique, classification recouvrante, réallocation dynamique.

1 Introduction

La classification automatique ou clustering consiste à organiser un ensemble d'individus $X = \{x_1, ..., x_n\}$ en classes (ou clusters) de telle sorte que des individus qui se ressemblent soient regroupés au sein d'un même cluster et des individus dissemblables appartiennent à des clusters différents. De nombreuses stratégies de classification ont été envisagées ces soixante dernières années, chacune présentant son lot d'avantages ou d'inconvénients selon la nature ou la quantité de données à traiter, leur dimensionalité, ou la forme des résultats (dendrogrammes, partitions strictes ou floues, concepts, etc.). La présente contribution s'intéresse aux méthodes de partitionnement dites par réallocation dynamique, dont l'algorithme des k-moyennes en est le plus éminent exemple. Les travaux que nous avons menés ont consisté à généraliser le modèle sous-jacent à l'approche k-moyennes afin de générer des classes recouvrantes, c'est-à-dire autorisant chaque individu à appartenir à plusieurs classes. Le modèle OKM (*Overlapping k-means*) [1] est brièvement exposé, s'en suit la présentation d'une première librairie R associée à ce modèle.

2 L'approche OKM

La méthode des k-moyennes est guidée par un critère objectif (moindres carrés) que l'on peut interpréter comme une quantification de l'erreur commise en résumant une classe d'individus à un unique représentant. La minimisation de cette erreur passe ainsi par la recherche des k meilleurs représentants de classes. Nous avons étendu ce principe au cas où chaque individu peut appartenir à plusieurs classes et donc être représenté par plusieurs représentants de classes. Le critère objectif (moindre carré généralisé) sous-jacent à la méthode OKM est donné par

$$J(\Pi, C) = \sum_{i=1}^{n} \|x_i - \phi_{\Pi,C}(x_i)\|^2 \quad avec \quad \phi_{\Pi,C}(x_i) = \frac{\sum_{j=1}^{k} \mathbb{1}_{\{x_i \in \pi_j\}}.c_j}{\sum_{j=1}^{k} \mathbb{1}_{\{x_i \in \pi_j\}}}$$

Dans cette formalisation, Π représente l'ensemble des clusters $\{\pi_j\}_{j=1}^k$, C l'ensemble des représentants $\{c_j\}_{j=1}^k$ et $\phi_{\Pi,C}(x_i)$ une combinaison des représentants des classes de x_i ; la combinaison utilisée dans OKM correspond au centre de gravité de ces représentants. Finalement, l'erreur associée à une classification recouvrante Π est quantifiée par la somme des distances (euclidiennes) entre chaque individu et la combinaison de ses représentants dans la classification.
La minimisation du critère $J()$ est assurée par une approche itérative classique en deux étapes : (1) affectation (ici multiple) de chaque individu aux classes puis (2) mise à jour des représentants des clusters ; chacune des deux étapes assure la décroissance du critère objectif.

3 La librairie R : OKM

Nous avons développé une première librairie[1] R intégrant non seulement l'algorithme OKM mais également une version pondérée (WOKM : *Weighted*-OKM) [2] permettant d'aboutir à des clusters ellipsoïdaux et limitant l'importance des recouvrements entre classes.
Deux fonctions sont proposées (`okm()` et `wokm()`) et correspondent à des approches généralisantes de k-moyennes ; nous avons donc veillé à conserver la forme de la fonction `kmeans()` présente dans la librairie R *stats* installée par défaut :

- `okm(X, centers, iter.max = 10, nstart = 1, visu = FALSE)`
- `wokm(X, centers, iter.max = 10, nstart = 1, B = 2, visu = FALSE)`

Les arguments utilisés sont : `X` : un ensemble d'individus décrits dans une matrice ou un dataframe (e.g. de taille $n \times p$) ; `centers` : un nombre de clusters (entier) ou un ensemble de centres initiaux (matrice ou dataframe) ; `iter.max` : le nombre maximum d'itérations autorisées dans l'algorithme (fixé à 10 par défaut) ; `nstart` : le nombre d'exécutions souhaitées (initialisations différentes) dans le cas où `centers` est un nombre (une seule exécution par défaut) ; `visu` : valeur logique permettant de faire apparaître les détails sur la convergence du critère objectif au cours de l'exécution (FALSE par défaut) ; `B` : paramètre (> 1) permettant de contrôler l'importance de la pondération dans l'approche WOKM.

Chacune des deux fonctions retournera un objet constitué de 4 composantes : `Clusters` : une matrice binaire d'appartenances ($n \times k$) ; `Representatives` : une matrice ($k \times p$) où chaque ligne décrits un représentant de classe ; `Withiness` : la valeur du critère objectif à la dernière itération $J()$; `Overlaps` : le nombre moyen de classes d'appartenance sur l'ensemble des individus.

4 Conclusion

Les premiers développements liés à l'approche OKM étant très récents, de nombreuses variantes sont actuellement à l'étude (nouveaux modèles de pondération, adaptation à la norme L_1, paramétrage des recouvrements, variantes à noyaux, etc.) chacune d'elle permettant de répondre à des besoins applicatifs réels. Une seconde version de la librairie *okm*, plus complète, est actuellement en préparation et sera prochainement déposée sur le site du CRAN.

Références

[1] G. Cleuziou. An extended version of the k-means method for overlapping clustering. In 19th International Conferenceon Pattern Recognition (ICPR'2008), pages 1 4, 2008.

[2] G. Cleuziou. Two variants of the OKM for Overlapping Clustering. Springer, 2010.

[1] Disponible sur : www.univ-orleans.fr/lifo/Members/cleuziou/

New tools for studying psychotherapies

Tiba Delespierre*, Jean-Michel Thurin, Monique Thurin, Bruno Falissard

Unité INSERM 669, 97 bd de Port Royal, 75679 Paris, France

Background

The Psychotherapy Practice Based Research' Network (PPBRN, Thurin and al., 2007) is organized in 3 sub-groups, Alzheimer, Borderline and Autism, the last one being the most active. The methodology of the network deals with process-outcome intensive case studies, fitted in the innovative field of Mixed Methods, which combines quantitative and qualitative approaches.

Right now, 41 autism patients have been included and more will follow, one hundred being the ultimate objective. In order to take full advantage of the repeated measures of mixed data, a Descriptive Analysis Plan (DAP) was specially designed for the network. The DAP uses **R** ad hoc functions which combine scores and short sentences. These functions visualize through time the whole therapy process dynamic. The DAP uses also Principal Component Analysis (PCA) and hierarchical clustering to create relevant clinical subgroups.

Aim/Purpose

The Descriptive Analysis Plan (DAP), coupled with a functions library, written with **R** uses 21 steps and three levels of analysis:
- an intensive case study level designed with **R** ad hoc functions;
- group and subgroups levels using **R** functions from the cluster and FactoMineR packages;
- a case by case comparison level, both patients being in the same subgroup or in different subgroups designed with **R** ad hoc functions.

Today all 41 patients have been studied as intensive cases and united in a solid psychotherapies data base. As more patients are added, these already existing cases compose an efficient reflection mean to realize the two aggregating levels.

Method

The study of each case starts with the extensive notes of the psychotherapist during the first three talks, and then in two sessions at 2, 6 and 12 months, completed by a quantitative evaluation of changes with validated instruments: Behavioral Summarized Scale (BSE, Barthelemy and al., 1997), Autism Psychodynamic Evaluation of Changes (APEC, Haag and al., 2010), Child Psychotherapy Process Q-sort (CPQ, Schneider and Jones, 2007).

The complete versatility and modularity of **R** gives the programmer all the tools needed to explore and analyze the patient's progress through time. Repeated measures of the BSE, APEC and CPQ scores, as well as details about their textual translation provide the programmer objective elements of description of the psychotherapeutic process.

The patients' scores are followed at 2, 6 and 12 months and related to the main characteristics of the psychotherapy. The first steps of the DAP present CPQ items subsets data frames and are purely descriptive. The second steps use hierarchical clustering and try explaining more precisely the patients' trajectories with 3 or 4 measures through time. Finally these trajectories are classified into 9 classes for each process variable and result score and compared to one another.

Results/Discussion

The last steps of the DAP will follow different patients subgroups through time as a whole and compare subgroups between one another. Already, with the group of 41 patients used to design the DAP, some patients showed significant progress with the BSE and APEC instruments, some a lot more than others.

Defining good clinical criteria through pertinent clustering, selecting similar cases and comparing patients in each clinical subgroup will give some answers to the PPBRN therapists on how and when the therapy works best.

Test de la vraisemblance entre deux motifs de points

F. Bonneu[a], **M. Chavent**[b], **A. Gegout-Petit**[b], **L. Guerin-Dubrana**[c,d] et **A. Labenne**[c,d]

[a] Université d'Avignon et des Pays de Vaucluse.
Laboratoire de Mathématiques d'Avignon (EA 2151), F-84018 Avignon, France.

[b]Univ. Bordeaux, IMB, UMR 5251, INRIA, F-33400 Talence, France.

[c]INRA, ISVV, UMR 1065 SAVE, F-33140 Villenave d'Ornon, France.

[d]Univ. Bordeaux, ISVV, UMR 1065 SAVE, Bordeaux Sciences Agro,
F-33140 Villenave d'Ornon, France.
amaury.labenne@bordeaux.inra.fr

Mots clefs : Statistique spatiale, biologie végétale, test non paramètrique, test de Monte-Carlo

L'esca-BDA est la principale maladie du bois de la vigne. Environ 11% du vignoble français est improductif suite à une recrudescence de cette pathologie depuis le début des années 2000. Le seul produit de lutte homologué contre l'esca-BDA, l'arsénite de sodium, a en effet été interdit à cette période. Cette maladie serait provoquée par un complexe de micro-organismes (essentiellement des champignons) dont les rôles respectifs restent à ce jour mal connus. Dans le cadre du projet CASDAR sur les maladies du bois, l'objectif de la partie épidémiologie est de caractériser la structure spatiale et spatio-temporelle de l'esca-BDA à l'échelle de la parcelle. Une des caractéristiques de l'esca est que l'apparition de symptômes n'entraine pas forcément la mort du cep, en effet, un cep peut exprimer des symptômes d'esca une année et redevenir sain l'année suivante. Inversement, un cep sain peut mourir sans jamais avoir exprimé de symptômes par le passé.

L'étude spatiale réalisée sur 13 parcelles en Aquitaine et en Bourgogne entre 2004 et 2010 a permis de révéler différentes structures d'agrégation de la maladie. Une des principales questions est de savoir si la structure spatiale des ceps une année $n+1$ est semblable à la structure spatiale des ceps symptomatiques l'année n. Pour mettre en place le test, on commence par estimer la probabilité pour chaque cep de la parcelle d'exprimer des symptômes l'année n. Ainsi, on construit une variable aléatoire de Bernoulli associée à l'événement pour chaque cep. On obtient donc une carte de probabilités représentant l'ensemble de la parcelle. Ensuite, on simule n_{sim} motifs de points grâce aux probabilités obtenues. Pour chaque motif i simulé, on calcule la statistique V_i égale à la vraisemblance des variables de Bernoulli simulées. Ces simulations nous permettent de proposer un test de Monte-Carlo au risque de première espèce α de cette statistique V. On calcule la même statistique (V_2) sur le motif de points observé de l'année $n+1$. On présentera des résultats de simulations et les conclusions des tests sur les données des 13 parcelles. La mise en œuvre de ce test a été réalisée grâce au logiciel R et au package *spatstat*.

Références

[1] Diggle, P.J. (2003). Statistical Analysis of spatial point patterns. Arnold.
[2] A.Baddeley and R.Turner. Spatstat:an R package for analyzing spatial point patterns. *Journal of Statistical Software* **12**: 6 (2005) 1-42.

The R Journal

M. Plummer[a]

[a]Editor in Chief
The R Journal
martyn.plummer@gmail.com

Mots clefs : libre accès, évaluation par les pairs

The R Journal is a publication of the R Foundation for Statistical Computing [1]. Following on from its predecessor *R News* it is the journal of record for developments in the R project and news from the R community. The R Journal also publishes peer-reviewed articles on topics of interest to users or developers of R. The R Journal is open access and all articles are licensed under a Creative Commons license [2].

The R Journal intends to reach a wide audience and have a thorough review process. Papers are expected to be reasonably short, clearly written, not too technical, and of course focused on R.

Références

[1] The R Journal `http://journal.r-project.org`
[2] Creative Commons Attribution 3.0 Unported license `http://creativecommons.org/licenses/by/3.0/`

A study of daily mobility patterns highlighting R workflow fluidity

H. Commenges

Université Paris Diderot-Paris 7
UMR 8504 Géogaphie-cités
13, rue du Four, 75006 Paris
hcommenges@parisgeo.cnrs.fr

Keywords: Geography, Daily mobility, Spatial analysis, Sequential analysis

Introduction: Daily mobility and transport demand have always been described and forecast through the concept of trip, defined since early transportation studies [1] as the movement linking two activities. Over the last 15 years a growing number of studies have tried to change the perspective in order to better grasp daily mobility. Among these relatively new approaches one deserve a particular attention: the so called "activity based" approach. This approach, closely linked to Swedish time-geography, aims to comprehend daily mobility not as a juxtaposition of trips but as a chain of interdependent trips and activities [2].

Time-geographic and activity-based studies require the processing of a high variety of objects (activities, individuals, spatial units) and the computing of a high variety of measurements: spatial, temporal and ordinary (i.e. non spatial, non temporal). That is the reason why, until a few years ago, geographers needed to manipulate several software packages, one for each kind of information:

- Geographic Information System (GIS) such as ArcGIS or QGIS,
- Traditional statistical analysis software such as SAS or SPSS,
- Graph manipulation software such as Tulip or Gephi,
- Sequential analysis software such as Optimize or CHESA.

With R software, combined with the growing number of specialized packages, it is now possible to manipulate all kinds of geographical and non geographical information into the same platform, and to produce nice graphical outputs without any post-production. **The poster aims to highlight this workflow fluidity** by describing all the stages necessary to build a description and classification of mobility patterns.

Data and methods: The study focuses on daily mobility of Île-de-France (IdF) region's inhabitants. Two kinds of information sources are used: first, the IdF region municipalities base map, released by the regional urbanism and planning institute (IAU-IdF, Institut d'Aménagement et d'Urbanisme d'Île-de-France). It is a polygon shape file representing the 1300 municipalities of the region. Second, the 2002 global travel survey (EGT, Enquête Globale Transport), which is a classical daily mobility survey that exists since the 1960s. It is a questionnaire face-to-face survey producing an information on the trips realized by a sample of 22000 IdF inhabitants. It collects all the trips realized the day before the interview.

Data are processed in two main ways to achieve a description and classification of mobility patterns: on the one hand, the municipalities base map is used to create an adjacency matrix

(with spdep package), then a topological distance matrix (with igraph package). This processing avoid absolute zonal classification (e.g. Paris, 1st crown, 2nd crown) in order to propose a relative classification peculiar to each individual: home, adjacent, near and far municipalities. On the other hand the table of trips produced by the EGT survey is processed in order to create sequences of trips and activities realized during the day by sampled individuals. The created sequence object is then manipulated with TraMineR package and used as a basis for description and classification of individuals' mobility patterns [3].

Finally, the whole exercise pretends to be an illustration of the possibilities offered by R as an integrated platform that fluidifies the workflow. It makes possible the processing of all kinds of objects and it produces high quality graphic and cartographic representations, which is crucial for geographers.

References

[1] Weiner, E. (1997). *Urban transportation planning in the United States: an historical overview*, U.S. Department of Transportation, 5th Edition.

[2] Stopher, P. (1992). Use of an activity-based diary to collect household travel data, *Transportation*, 19(2), 159-176

[3] Gabadinho, A., Ritschard, G., Mueller,N.S., Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4), 1-37.

POSTER : Caractérisation d'événements à partir de signaux relatifs au comportement d'un élément combustible en situation accidentelle

L. Pantera et **V. Lefrançois**

Département d'Etudes des Réacteurs
Laboratoire de Préparation et Réalisation des Essais
Commissariat à l'Energie Atomique
CEA Cadarache, 13108 Saint-Paul-Lez-Durance
laurent.pantera@cea.fr

Mots clefs : nucléaire, combustible, traitement du signal, classification, diagnostic

CABRI est un réacteur expérimental destiné aux études de sûreté en soutien au parc électronucléaire. Le travail présenté entre dans le cadre du programme d'essais "CABRI International Program (CIP)". Celui-ci a pour objectif d'étudier le comportement des éléments de combustible d'un Réacteur à Eau sous Pression ou REP (PWR Pressurized Water Reactor en anglais) à haut taux de combustion lorsqu'ils sont soumis à un accident d'insertion de réactivité correspondant à l'éjection d'une barre de contrôle. Chaque essai porte sur l'étude de comportement d'un seul élément combustible appelé également crayon. Pour fixer les idées, celui-ci est constitué d'un ensemble de pastilles d'oxyde d'uranium empilées sur une hauteur d'un mètre environ et placé dans une gaine en alliage de zirconium de 9.5 mm de diamètre et de 0.57 mm d'épaisseur. Les essais consistent à soumettre l'élément combustible à des puissances pouvant atteindre les 20 GW. Ces montées en puissance sont réalisées sur de très courtes durées (10 à 100 ms). Après chaque essai, les expérimentateurs sont amenés dans le cadre d'un rapport préliminaire à se prononcer sur l'absence ou pas de rupture de la gaine de l'élément combustible testé. Ce diagnostic est effectué à partir des signaux obtenus en ligne, notamment à partir de deux microphones (fréquence d'acquisition égale à 1 MHz) placés en amont et en aval du dispositif d'essai. Le poster présenté souhaite montrer comment au sein d'une équipe d'expérimentateurs très mobilisés autour de l'élaboration d'expériences, il a été possible grâce à la souplesse du langage R de réorganiser les mesures de nos anciens essais pour pouvoir les analyser et les utiliser dans nos futures prévisions. Nous avons dans un premier temps récupéré tous les signaux suivis en ligne pour chaque portion d'une seconde d'essai [1] et effectué une découpe de ces secondes expérimentales en événements en se basant sur la variation de la variance [2] des signaux des microphones. Chaque événement a été stocké dans une base de données relationnelle (PostgreSQL) [3] puis nous avons extrait de ces signaux des indicateurs sur lesquels nous avons pu effectuer des analyses en composantes principales suivies de classifications [4] en utilisant la distance euclidienne classique calculée sur les premiers facteurs retenus. L'agencement de ces méthodes a pu ensuite être intégré dans une interface développée en JAVA [5], langage utilisé en interne pour effectuer le suivi en temps réel des essais. Les signaux analysés sont non stationnaires, la caractérisation des événements par discrétisation du spectre de Fourier ne suffit pas pour obtenir une classification qui mettrait en évidence les événements de rupture : des événements de non rupture restent mal classés. Afin d'affiner nos classements, nous cherchons maintenant à définir des indices de classification qui nous permettraient de caractériser la répartition des fréquences dans le temps de chaque événements [6].

Références

[1] Zeileis, A., Grothendieck, G., Ryan, J., Andrews, F. Package zoo, http://zoo.R-Forge.R-project.org/, consulté le 24 avril 2012

[2] Scrucca, L.(2004). qcc: An R package for quality control charting and statistical process control, R news 4:11-17

[3] Urbanek, S. Package RJDBC, http://www.rforge.net/src/contrib/Documentation/RJDBC.pdf, consulté le 24 avril 2012

[4] Chessel, D., Dufour, A.-B. and Thioulouse J. (2004). The ade4 package - I : One-table methods, R news 4:5-10

[5] Urbanek, S. Package rjava, http://www.rforge.net/rJava/, consulté le 24 avril 2012

[6] Shumway, R.H., (2003) Time-frequency clustering and discriminant analysis, Statistics & Probability Letters, 63(3), 307-314

La prise en compte de l'environnement par les agriculteurs : une analyse avec le package "ClustOfVar"

V. Kuentz-Simonet[a], **S. Lyser**[a],**M. Chavent**[b], **J. Saracco**[b], **J. Candau**[a], and **P. Deuffic**[a]

[a] Irstea, UR ADBX,
50 avenue de Verdun, 33612 Cestas Cedex, France
{vanessa.kuentz-simonet,sandrine.lyser,jacqueline.candau,philippe.deuffic}@irstea.fr

[b] INRIA Bordeaux Sud Ouest,
Institut de Mathématiques de Bordeaux,
Université de Bordeaux,
351 cours de la libération, 33405 Talence Cedex, France
{marie.chavent, jerome.saracco}@math.u-bordeaux1.fr

Mots clefs : classification de variables, variables synthétiques, perception environnementale, typologie.

En statistique exploratoire multidimensionnelle, la classification des observations est couramment utilisée pour établir des profils-types. Une stratégie classique consiste à réaliser une analyse factorielle des données puis à appliquer une méthode de classification sur les scores des individus mesurés sur les composantes principales obtenues. Cependant certains auteurs (De Soete et Carroll, 1994 ou Vichi et Kiers, 2001) ont souligné les effets néfastes de cette procédure dite « tandem analysis ». Dans le cas particulier de données quantitatives, ils montrent que l'ACP identifie parfois des composantes qui contribuent peu à la détection d'une structure dans les observations ou qui au contraire masquent l'information taxinomique. En effet, on peut concevoir que des informations relatives à la structure des observations puissent être masquées par la création de ces composantes orthogonales qui visent à reconstruire au mieux la variance contenue dans le nuage de points initial. L'approche par classification de variables proposée ici (Chavent et al. 2011) est une alternative à la première étape d'analyse factorielle pour la typologie d'observations. Cette méthode permet de réduire la dimension du tableau en supprimant l'information redondante. En réorganisant les variables en classes homogènes, elle conduit à construire simultanément des variables synthétiques sans imposer de contraintes d'orthogonalité. De plus, dans chaque classe de variables, les coordonnées des modalités des variables qualitatives la composant nous permettent de visualiser ces variables synthétiques comme une sorte de gradient. Il est alors possible d'interpréter et de labelliser ces variables synthétiques aisément. Ainsi la compréhension de la typologie des observations est simplifiée.

La méthodologie proposée est illustrée à partir d'une enquête réalisée en 2005 par des sociologues d'Irstea auprès d'agriculteurs français. L'objectif de cette étude était de "cerner la façon dont les agriculteurs conçoivent la protection de l'environnement en relation avec leur activité" (Candau et al., 2005). L'approche proposée permet de mettre en évidence une structure dans les réponses des agriculteurs, en résumant l'information par 9 variables synthétiques (relatives au cadre de vie, à la perception du métier, aux problèmes de l'environnement, etc.). Dans un second temps, cette analyse est complétée par une typologie en 7 groupes (intéressés par le changement, attentifs à la protection de l'environnement, adeptes de la déprise agricole, etc.). Cette démarche nous permet d'appréhender la perception de l'environnement par cette catégorie socioprofessionnelle et ainsi de mesurer l'importance accordée aux problématiques environnementales parmi les autres préoccupations actuelles.

Références

[1] Candau, J., Deuffic, P., Ginelli, L., Lewis, N., Lyser, S., 2005, La prise en compte de l'environnement par les agriculteurs. Résultats d'enquête. CNASEA, 83 p.

[2] Chavent M., Kuentz V., Liquet B., Saracco J., 2011, Clustering of variables via the PCAMIX method, International Classification Conference, St Andrews, Ecosse.

[3] De Soete G., Carroll J.D., 1994, K-means clustering in a low-dimensional Euclidean Space, in *New Approaches in Classification and Data Analysis*, Diday E., et al. (Eds.).

[4] Vichi M., Kiers H.A.L., 2001, Factorial k-means analysis for two-way data, *Computational Statistics and Data Analysis*, vol 37, n° 1, p 49-64.

R dans un Environnement Pédagogique Virtuel (EPV) : démarche pédagogique et retour d'expérience dans une école d'ingénieurs en agriculture

M. CANNAVACCIUOLO, A. FADIL, et P. HUYNH

Département de Sciences Fondamentales et Méthodes
Groupe ESA - PRES L'UNAM
55, Rue Rabelais – BP 30748 49007 Angers Cedex 01
m.cannavacciuolo@groupe-esa.com; a.fadil@groupe-esa.com; p.huynh@groupe-esa.com

Mots clefs : Virtualisation, Environnement pédagogique, Logiciel libre, Statistiques, Informatique.

Contexte (les savoirs) : L'Ecole Supérieure d'Agriculture d'Angers forme des ingénieurs dans les secteurs aussi variés que l'agronomie, l'environnement, la production animale, l'agroalimentaire et la gestion d'entreprise dont certains secteurs sont ouverts à l'international.
Dans le cadre d'une analyse prospective, une enquête auprès de nos entreprises partenaires nous a permis d'identifier leurs besoins professionnels et les compétences requises de nos ingénieurs d'aujourd'hui et de demain. Ainsi, nos ingénieurs doivent être dans leur entreprise force de proposition et maître d'ouvrage dans des projets informatiques (système d'information, base de données, progiciel de gestion intégrée …). Ces ingénieurs doivent aussi mettre en œuvre des méthodes informatiques de statistique décisionnelle et du data mining. Conjointement à ces nouveaux besoins des entreprises, l'ingénieur ESA doit toujours savoir mener des expérimentations (échantillonnage, plans d'expériences, statistiques inférentielles).

Constat (les savoir-faire) : La démarche pédagogique adoptée par les enseignants du département Sciences Fondamentales et Méthodes (SFM) de l'ESA apporte une harmonisation des outils et une rigueur dans l'application des méthodes. Elle cherche à améliorer l'autonomie, et positionner les apprenants dans une posture plus professionnalisante.

Matériel et méthodes : Depuis les années 2000, le taux d'équipement de nos étudiants en ordinateur portable ne cesse d'augmenter pour atteindre aujourd'hui 100% en cycle Master.
L'hétérogénéité du matériel informatique et des logiciels qu'ils utilisent nous oblige à reconsidérer nos pratiques d'enseignement. L'utilisation des ordinateurs portables des étudiants dans les enseignements nécessite un environnement informatique pédagogique homogène, peu couteux et respectant les contraintes légales de copyright. Cet environnement doit être facile d'installation et d'utilisation. Enfin, il doit être compatible avec les solutions de la DSI et ne pas trop modifier les pratiques des enseignants.
Pour répondre à ces besoins, le département SFM a préinstallé R avec la suite bureautique LibreOffice (http://www.libreoffice.org) pour le formatage / saisie de données, les logiciels pour la conception de bases de données et des progiciels comme vTigerCRM / OpenERP dans un environnement virtuel (EPV). Cet EPV fonctionne sous Lubuntu (variante d'Ubuntu utilisant un environnement de bureau allégé LXDE; https://wiki.ubuntu.com/Lubuntu). Il est ensuite déployé par une simple copie sur les portables des étudiants ayant une installation gratuite de VirtualBox (licence GNU GPL ; http://www.virtualbox.org). Les applications scientifiques sélectionnées dans l'EPV devaient satisfaire les trois phases de la réalisation d'un projet d'étudiant : conception, développement et traitement (Tableau 1).

Discipline/phase	**Conception**	**Développement**	**Traitement**
Modélisation	Modelio		
BD	MySQL Workbench		
Gestion intégrée		OpenERP	
Statistiques ***data mining***			
RO			

Tableau 1 : Recouvrement des outils par discipline et phase d'un projet étudiant

Retour d'expérience « utilisation de R dans l'EPV » : L'EPV est déployé depuis 3 ans dans le tronc commun du cycle Master (une centaine d'étudiants par semestre). En exemple, il est utilisé en Travaux Dirigés et dans des études de cas en analyse de données multidimensionnelles (packages ade4 [1] et FactomineR [2]) et de recherche opérationnelle (package lpSolve [3] et fonctions personnalisées). Après une première phase d'appropriation de l'EPV, les étudiants développent ensuite différentes stratégies combinant l'utilisation du tableur (acquisition et formatage des données), l'utilisation de R en CLI (adaptation de scripts disponibles à partir d'une bibliothèque fournie par les enseignants) et l'utilisation de R en GUI (Rcommander [4]). Ce constat permet de compléter l'utilisation combinatoire par l'apprenant du tableur, de R en CLI et de R en GUI abordée par Cornillon et al. [5]. Ces stratégies dépendent des objectifs pédagogiques et de l'aisance de l'apprenant à manipuler les lignes de commandes et/ou les interfaces graphiques. Ce constat nous a amené à installer Rstudio (http://rstudio.org/), un environnement de développement intégré pour R, facilitant la saisie, l'exécution de scripts et la visualisation des résultats.

Discussion : L'utilisation de l'EPV sur les ordinateurs portables apporte une souplesse et une mobilité dans la réalisation des cours aux enseignants, dans l'apprentissage aux étudiants et dans une large mesure une maîtrise des coûts d'investissement. Les étudiants étrangers peuvent aussi personnaliser l'environnement de bureau à leur langue maternelle. R permet d'aborder la plupart des méthodes statistiques nécessaires à la formation de nos ingénieurs. Toutefois, les méthodes d'enquête par questionnaire dans l'EPV nécessitent des outils spécifiques de création et de paramétrage de questionnaires. Pour répondre à ces besoins, notre département développe RQuest une application GUI en Python et en Qt (version alpha-test) qui interface R avec SQLlite. Cette expérience est aussi appliquée dans une thèse au Laboratoire d'Ecologie Végétale du Groupe ESA.

Références

[1] Chessel, D. and Dufour, A.B. and Thioulouse, J. (2004). The ade4 package-I- One-table methods. R News. 4: 5-10.

[2] Husson, F. Josse, J. Le, S. and Mazet, J. (2011). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. R package version 1.16. http://CRAN.R-project.org/package=FactoMineR

[3] Berkelaar, M. and others (2011). lpSolve: Interface to Lp_solve v. 5.5 to solve linear/integer programs. R package version 5.6.6. http://CRAN.R-project.org/package=lpSolve

[4] Fox, J. <jfox@mcmaster.ca>, et al. (2011). Rcmdr: R Commander. R package version 1.6-4. http://CRAN.R-project.org/package=Rcmdr

[5] Cornillon, P.-A. and Matzner-Lober, E. (2009). Atouts et faiblesses de R en enseignement, recherche et industrie. 41èmes Journées de Statistique, SFdS, Bordeaux.

The npde library for R to compute normalised prediction distribution errors

Emmanuelle Comets[a,b], **Thi Huyen Tram Nguyen**[a,b] and **France Mentré**[a,b]

[a]INSERM, UMR 738
Paris, France
emmanuelle.comets@inserm.fr

[b]Univ Paris Diderot
Sorbonne Paris Cité
Paris, France

Mots clefs : Non-linear mixed effect models; model evaluation; normalised prediction distribution errors

Objectives: Over the last few years, several new approaches including VPC (Visual Predictive Check) [1], prediction discrepancies (pd) [2] and normalised prediction distribution errors (npde) [3] have been proposed to evaluate nonlinear mixed effect models. npde are now included in the output of NONMEM [4] and Monolix [5], and we created a R library to facilitate the computation of pd and npde using simulations under the model [6]. We propose a new version of this library with methods to handle data below the limit of quantification (BQL) [7] and new diagnostic graphs [8].

Methods: BQL data occurr in many PK/PD applications, particularly in HIV/HCV trials where multi-therapies are now so efficient that viral loads become undetectable after a short treatment period. These data are generally omitted from diagnostic graphs, introducing biases. Here, we propose to impute the pd for a BQL observation by sampling in $U(0,p_{BQL})$ where p_{BQL} is the model-predicted probability of being BQL. To compute the npde, censored observations are first imputed from the imputed pd, using the predictive distribution function obtained by simulations, then npde are computed for the completed dataset [3].

New graphical diagnostics include a graph of the empirical cumulative distribution function of pd and npde. Prediction intervals, obtained using simulations under the model, can be added to each graph to assess how the distribution of observed data and metrics compare to the expected distribution under the model. Tests can be performed to compare the distribution of the npde relative to the expected standard normal distribution. In addition, graphs and tests to help selecting covariate models have been added [9].

These extensions were implemented in a new version of the npde library. The new library uses S4 classes from R to provide an easier user-interface to the many new graphs, while remaining mostly compatible with the previous version. Exceptions are that computing the pd in addition to the npde is now a default option. Several new options are also available in the computations.

Results: We illustrate the new library on data simulated using the design of the COPHAR3-ANRS 134 trial. In the trial, viral loads were measured for 6 months in 34 naive HIV-infected patients after initiation of a tri-therapy, and up to 50% of data were BQL. Ignoring BQL data results in biased and uninformative diagnostic plots, which are much improved when pd are imputed. Adding prediction intervals is very useful to highlight departures from the model.

Conclusion: Version 2 of the npde library implements a new method to handle BQL data, as well as new graphs, including VPC and prediction intervals for distributions.

References

[1] Holford N (2005). The Visual Predictive Check: superiority to standard diagnostic (Rorschach) plots. *14th meeting of the Population Approach Group in Europe, Pamplona, Spain*, (Abstr 738).

[2] Mentré F and S Escolano S (2006). Prediction discrepancies for the evaluation of nonlinear mixed-effects models. *J Pharmacokinet Biopharm*, **33**, 345-67.

[3] Brendel K, Comets E, Laffont C, Laveille C, Mentré F (2006). Metrics for external model evaluation with an application to the population pharmacokinetics of gliclazide. *Pharm Res*, **23**, 2036-49.

[4] Beal S, Sheiner LB, Boeckmann A, Bauer RJ (2009). NONMEM User's Guides. (1989-2009), Icon Development Solutions, Ellicott City, MD, USA.

[5] Lavielle M (2010). MONOLIX (MOdèles NOn LInéaires à effets miXtes) User Guide. MONOLIX group, Orsay, France. URL: http://software.monolix.org/

[6] Comets E, Brendel K, Mentré F (2008). Computing normalised prediction distribution errors to evaluate nonlinear mixed-effect models: The npde add-on package for R. *Comput Meth Prog Biomed*; **90**, 154-66.

[7] Nguyen THT, Comets E, Mentré F (2011). Prediction discrepancies (pd) for evaluation of models with data under limit of quantification. *20th meeting of the Population Approach Group in Europe, Athens, Greece*, (Abstr 2182).

[8] Comets E, Brendel K, Mentré F (2010). Model evaluation in nonlinear mixed effect models, with applications to pharmacokinetics. *J-SFdS*: **1**, 106-28.

[9] Brendel K, Comets E, Laffont C, Mentré F (2010). Evaluation of different tests based on observations for external model evaluation of population analyses. *J Pharmacokinet Pharmacodyn*; **37**,49-65.

Utilisation du logiciel R pour l'identification de nouvelles cibles et régulateurs du protéasome

C. Pellentz[a], C. Saveanu[b], A. Jacquier[b] and A. Peyroche[a]

[a] CEA Saclay
IBITECS, SBIGEM, LMARGe
F-91191, France
celine.pellentz@cea.fr
anne.peyroche@cea.fr

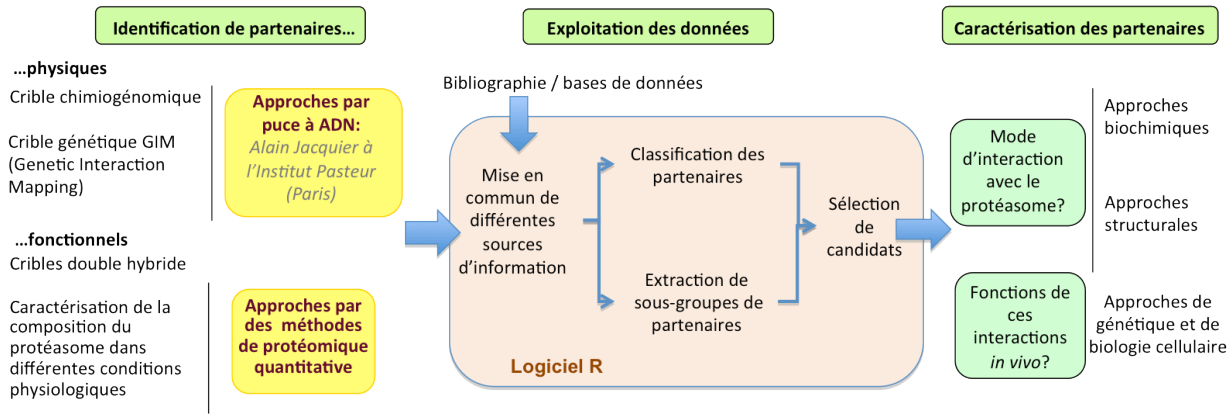
[b] Institut Pasteur
Génétique des Interactions Macromoléculaires
25-28, Rue du Docteur Roux
75724 Paris Cedex 15 France
cosmin.saveanu@pasteur.fr
alain.jacquier@pasteur.fr

Mots clefs : Biologie, Analyse de cribles et de données, Hiérarchisation, Clusterisation.

Le protéasome est une protéase multimérique composée de différents sous-complexes obtenus par l'assemblage ordonné de dizaines de polypeptides. Il est présent chez toutes les cellules eucaryotes où il constitue l'unité catalytique du système Ubiquitine-protéasome (UPS). Son activité de dégradation des protéines, à la fois cytoplasmique et nucléaire est essentielle à de nombreux processus cellulaires [1]. Le protéasome est une machinerie plastique et dynamique qui est elle-même soumise à régulation [2]. Des dysfonctionnements de l'UPS participent à la pathogenèse de maladies telles que les cancers ou les maladies neurodégénératives. Il semble donc important d'identifier i) les processus cellulaires dans lesquels le protéasome est impliqué, ii) les facteurs influençant son activité.

Dans ce but, nous développons un projet pour identifier et caractériser les partenaires physiques et fonctionnels du protéasome par une approche multi-technique chez la levure *S. cerevisiae*. Nous déterminons ensuite s'ils sont fonctionnellement conservés chez les Mammifères.

Ce projet peut être divisé en trois étapes :



1/ L'identification de partenaires

La première étape consiste à identifier des partenaires physiques et fonctionnels du protéasome par la réalisation de cribles à grande échelle utilisant le modèle cellulaire *Saccharomyces cerevisiae*.

2/L'exploitation des données grâce au logiciel R

L'utilisation du logiciel R permet d'exploiter les résultats des cribles et de croiser ces données avec d'autres données d'interactions issues de résultats du laboratoire ou de ressources bibliographiques.

Les cribles utilisés pour identifier les partenaires fonctionnels du protéasome reposent sur la technologie des puces à ADN. La quantité des données quantitatives qui en sont issues doit être gérée et exploitée pour contrôler la reproductibilité des cribles et distinguer les répondants du bruit de fond. Les résultats de ses cribles peuvent ensuite être hiérarchisés pour essayer d'identifier des groupes répondants de façon similaire ou opposé.

Le deuxième aspect de l'utilisation du logiciel R dans ce projet est le regroupement de données expérimentales d'origines variées mais ayant pour point commun l'étude des partenaires du protéasome. L'intérêt de cette approche est double : i) chaque technique expérimentale présente des biais et bruits de fond différents, comparer des résultats de techniques très différentes permet de s'affranchir du bruit de fond de chacune et d'être plus sensible ; ii) les techniques à grande échelle, qui sont de plus en plus utilisées, génèrent une très grande quantité de données qu'il est impossible de mémoriser et comparer aux données existantes antérieures sans outil informatique, compiler ces données grâce au logiciel R permet de les pérenniser et de pouvoir donner du poids à des données qui ont peu de poids isolées mais qui deviennent significatives quand elles sont retrouvées dans d'autres jeux de données.

Enfin, la classification des interactants potentiels identifiés suivant différents critères permet de faire ressortir des candidats potentiels non identifiés auparavant. Classer les interactants selon le critère du nombre de techniques différentes l'ayant mis en évidence permet par exemple de s'affranchir du bruit de fond de chaque technique. On peut ainsi identifier de nouveaux acteurs potentiels du fonctionnement du protéasome et/ou mettre en évidence de nouvelles connexions entre le système UPS et d'autres voies cellulaires.

La banque d'interactants ainsi générée peut ensuite être interrogée pour identifier des interactants potentiels du protéasome dans son ensemble ou des interactants spécifiques d'une sous-unité donnée (parmi la trentaine de sous-unités au total) du protéasome.

3/Caractérisation de candidats potentiels
Enfin, la dernière étape consiste en la caractérisation de ces candidats potentiels, en combinant des approches biochimiques et génétiques classiques.

Cette démarche originale pourrait permettre d'identifier des partenaires du protéasome non encore identifiés grâce à la puissance de la comparaison de données issues de techniques très différentes qui permet de distinguer le bruit lié à chaque technique des interactants probablement significatifs.

Références
[1] Bedford L., Paine S., Sheppard P.W., Mayer R.J., Roelofs J. (2010) Assembly, structure and function of the 26S proteasome. Trends in Cell Biology 20:391-401
[2] Glickman M.H. and Raveh D., (2005) Proteasome plasticity. FEBS Lett, 579(15): p3214-23

Le logiciel **R** en neuro-imagerie fonctionnelle

Pierre Lafaye de Micheaux
Département de mathématiques et Statistique
Université de Montréal

L'imagerie cérébrale fonctionnelle est une discipline qui désigne l'ensemble des techniques issues de l'imagerie médicale: citons l'imagerie par résonance magnétique fonctionnelle (IRMf), la tomographie par émission de positrons (TEP), l'électroencéphalographie (EEG), la magnétoencéphalographie (MEG), l'imagerie du tenseur de diffusion, etc. Ces techniques visent à observer le cerveau d'un individu pendant qu'il exécute une tâche cognitive. Durant cette opération, on mesure certains signaux produits par l'activité cérébrale. Suivant les techniques et les outils mathématiques employés, on cherche à retrouver, avec plus ou moins de précision, quelle région du cerveau était particulièrement active et à quel moment de la tâche cognitive. On cherche aussi à déceler des réseaux de connectivité fonctionnelle. Il faut savoir que l'activité cérébrale résulte du fonctionnemment biologique de cellules nerveuses appelées les neurones. Ces cellules spécialisées, réparties dans ce que l'on appelle le cortex cérébal, sont capables d'émettre des signaux électriques qui peuvent être mesurés, soit directement, soit indirectement via leur activité métabolique. Ce sont ces signaux, et leur localisation dans le cerveau, qui traduisent l'activité cérébrale. On cherche ainsi à caractériser la dynamique temporelle de l'activité cérébrale, au moyen d'outils statistiques et logiciels. Dans cette présentation, nous commencerons par donner quelques notions succinctes sur le fonctionnement cérébral, ce qui nous conduira à présenter le type de données qu'il est possible d'analyser. Nous montrerons ensuite comment le logiciel **R** peut être utilisé, via certains *packages* spécialisés, pour visualiser ou effectuer certaines analyses statistiques sur ces données, comme la régression linéaire ou l'analyse en composantes indépendantes. Nous mettrons plus particulièrement l'accent sur l'IRMf, technique relativememt récente, pour laquelle des données réelles assez volumineuses seront utilisées.

Analysing eye movement data using Point Process models

Simon Barthelmé[a], Hans Trukenbrod[b], Ralf Engbert[b], Felix Wichmann[c]

[a]**Computer Science**
Technical University and Bernstein Center for Computational Neuroscience
Franklinstr. 28-29, 10487 Berlin
simon.barthelme@bccn-berlin.de

[b]**Psychology**
University of Postdam
Karl-Liebknecht-Str. 24/25, 14476 Potsdam OT Golm, Germany
{ralf,hans}.{engbert,trukenbrod}@uni-potsdam.de

[c]**Computer Science**
University of Tübingen
Institute for Computer Sciences, Eberhard Karls University, 72076 Tübingen Germany
felix.wichmann@uni-tuebingen.de

Keywords: point processes, eye movements, spatial statistics

The measure and analysis of eye movements is crucial to neuroscience and psychology [10]. Eye movements are extremely useful from a methodological point of view - among other things, it is relatively easy to train animals to respond using eye movements, and the neurophysiological pathways involved are relatively well understood [6]. Eye movements are also tremendously interesting as an object of study in their own right, because they are the most immediate means we have to explore our visual environment.

In humans, the eyes do not move constantly but rather alternate between periods of relative stability, called fixations, and periods of movement. Very often the analysis is not concerned with movements but rather with fixations, and most especially where fixations occur. For example, in so-called "free-viewing" experiments, subjects view natural images, with no specific instructions - they are free to look wherever they like. Where they choose to fixate is far from random: subjects focus on similar locations, and exactly why they do that is an old debate in neuroscience and psychology [11, 8].

Some authors have argued that eye movements are controlled by a cortical *saliency map* [5], which represents interesting locations in the visual field, and that "interestingness" is computed very

early in the visual cortex using local image information. Following these ideas, models have been developed that seek to predict fixations from local image features [4]. Exactly what is being predicted and how is a source of some confusion in the literature, and many different methods have been proposed, with no unifying framework so far [12].
We argue that the right framework is to be found in the tools of spatial statistics [2]. A set of fixations is in essence spatial data - a set of points in space. For such data, appropriate statistical models are known as point processes: a point process is a probability distribution over finite subsets of a spatial domain. There is an extensive literature on how point process models can be used to analyse point patterns, reviewed for example in [3] and [7].
While the literature on point processes focuses mostly on studying the outcome of one point process, fixation data is better thought of as arising from many related point processes - for example, one process per image, or one process per subject, etc. The interesting questions often have to do with how fixation patterns vary (across subjects, across images) and whether they are common factors.
We have developed a R package called **mpp**, for "multiple point processes", which aims to facilitate Bayesian inference for such problems. It builds on the **spatstat** package [1] and uses **INLA** for approximate inference [9]. We will show how **mpp** can be used to explore some simple hierarchical point process models applied to fixation locations.

Acknowledgements
This work has benefited from funding from the BMBF (Foerderkennzeichen 01GQ1001B).

References

[1] **Adrian Baddeley and Rolf Turner. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005. ISSN 1548-7660.**

[2] **Peter J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Hodder Education Publishers, 2 edition.**

[3] **Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan. *Statistical Analysis and Modelling of Spatial Point Patterns (Statistics in Practice)*. Wiley-Interscience, 1 edition, March 2008.**

[4] **L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3):194–203, March 2001.**

[5] **C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985.**

[6] **Richard J. Krauzlis. The control of voluntary eye movements: New perspectives. *The Neuroscientist*, 11(2):124–137, April 2005.**

[7] **Jesper Møller and Rasmus P. Waagepetersen. Modern statistics for spatial point processes*. *Scandinavian Journal of Statistics*, 34(4):643–684, December 2007.**

[8] Antje Nuthmann and John M. Henderson. Object-based attentional selection in scene viewing. *Journal of vision*, 10(8), 2010.

[9] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

[10] Alexander C. Schütz, Doris I. Braun, and Karl R. Gegenfurtner. Eye movements and perception: A selective review. *Journal of vision*, 11(5), 2011.

[11] Benjamin W. Tatler, Mary M. Hayhoe, Michael F. Land, and Dana H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5), May 2011.

[12] Niklas Wilming, Torsten Betz, Tim C. Kietzmann, and Peter König. Measures and limits of models of fixation selection. *PLoS ONE*, 6(9):e24038+, September 2011.

A common signal detection model describes threshold and supra-threshold performance

K. Knoblauch[a]

[a]INSERM U846, Stem Cell and Brain Research Institute
Department of Integrative Neurosciences
18 avenue du Doyen Lépine, 69500, Bron
ken.knoblauch@inserm.fr

Mots clefs : Statistics, Psychophysics, GLM, Signal Detection Theory

Psychophysical experiments are typically based on analyzing observer choices as a function of some stimulus dimension in order to make inferences about the underlying sensory and/or decision processes that account for the observer's choices. Modern psychophysical theory derives from Signal Detection Theory (SDT) [1, 2] in which the observer's performance depends on a noise contaminated decision variable that in association with a criterion that determines the rates of both successful classifications and errors. When the decision rule can be characterized as a linear predictor, the framework can be formalized as a Generalized Linear Model (GLM) with a binomial family, facilitating estimating of model parameters by maximum likelihood. The Gaussian, equal-variance model is the most commonly employed which leads naturally to the use of a probit link. The largest body of psychophysical work is based on discrimination of small (threshold) stimulus differences, yielding measures of perceptual strength that do not easily extrapolate to predict performance for large (supra-threshold) differences or for appearance, either.

Some recent techniques do permit an extension of this approach to the supra-threshold domain. For example, Maximum Likelihood Difference Scaling (MLDS) is a psychophysical method and fitting procedure that permits scaling of large stimulus differences based on paired-comparisons of stimulus intervals [3, 4]. In the task, observers are presented with quadruples (or triads) of stimuli distributed along a physical scale. For quads, the observers judge between which pair the difference is greatest; for triads, the judgment is more similar to a bisection task. We review the decision model and the maximum likelihood fitting procedure. Because the decision rule is linear, the fitting procedure can be simplified by reformulating it as a GLM. The **MLDS** package [4] available on CRAN provides tools for fitting and evaluating data sets that arise from these experiments. The resulting scales have interval properties (equal differences along the scale are perceptually equal).

Previous studies have shown that MLDS scales are qualitatively consistent with discrimination performance [5] and that differences along MLDS scales are inversely proportional to reaction times [6]. The decision rule in MLDS is based on an equal-variance, Gaussian signal detection model. We show that when properly parameterized, difference scales also predict the traditional SDT measure of discrimination, d'. We demonstrate this for an experiment in which we used MLDS to quantify the watercolor effect [7], a long-range perceptual filling-in phenomenon. The results imply that a common signal detection model suffices to account for both discrimination performance and appearance. Since SDT provides a common metric for relating threshold behavior to neural response mechanisms, the results have important implications for relating perceptual to neural responses.

Références

[1] Green, D.M., Swets, J.A.(1966/1974) *Signal Detection Theory and Psychophysics*. Robert E. Krieger Publishing Company, Huntington.

[2] Macmillan, N.A., Creelman, C.D. (2005) *Detection Theory: A User's Guide*, 2nd edition, Lawrence Erlbaum Associates, New York.

[3] Maloney, L.T., Yang, J.N. (2003) Maximum Likelihood difference scaling. *Journal of Vision* **3(8)**, 573–585). `http://www.journalofvision.org/3/8/5`.

[4] Knoblauch, K., Maloney, L.T. (2008) MLDS: Maximum likelihood difference scaling in R. *Journal of Statistical Software* **25**, 1–26 `http://www.jstatsoft.org/v25/i02`.

[5] Rhodes, G., Maloney, L.T., Turner, J., Ewing, L. (2007) Adaptive face coding and discrimination around the average face. *ÊVision Research* **47**, 974–989.

[6] Brown, A. M., Lindsey, D. T., Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of Vision*, **11(12)**, 1–21, `http://www.journalofvision.org/content/11/12/2`.

[7] Devinck, F., Knoblauch, K. (2012). A common signal detection model accounts for both perception and discrimination of the watercolor effect. *Journal of Vision*, **12(3)**, 1–14, `http://www.journalofvision.org/content/12/3/19`,

Discovering the relevant variables in a large clinical database by back-fitting fixed effects in a mixed linear model: Study of a long-term electrophysiological survey of cochlear implanted patients

Rafael Laboissière[a], **Michel Mazzuca**[a], **Hung Thaï-Van**[a], & **Lionel Collet**

[a]Groupe Parole, Audiologie, Communication et Santé
Centre de Recherche en Neurosciences de Lyon
⟨rafael.laboissiere@inserm.fr⟩

Mots clefs : auditory evoked potentials, maturation, large data set, mixed-models

Clinical surveys are routine procedures that are typically run in Hospital services and intended to evaluate patients after surgery or therapeutic treatment. They differ from clinical trials, where specific drugs or new therapies are investigates according to a planned protocol and with a predetermined cohort. They also differ from basic research experiments, which are done in even narrowly controlled setups in laboratories. Clinical surveys, on the other hand, can include a huge amount of patients, can last for a very extended amount of time (sometimes even without a fixed time to be terminated) and can include a large amount of independent factors and dependent measured variables.

In this paper, we present the analysis of a large data set collected on cochlear implanted patients at the Audiology Service of the Édouard Herriot Hospital in Lyon. The analysis on a reduced fraction of this data set has already been published elsewhere (Thai-Van et al., 2007). A cochlear implant (CI) is a surgically implanted electronic device that restores the sense of audition to profoundly deaf patients. It acts as a transducer between the sound captured by an external microphone, placed closely to the patient's ear, and the neural cells at the extremity of the auditory nerve. The transducer device is an linear array of electrodes that is introduced in the scala tympani of the cochlea. After surgery, the patients do systematic visits to the hospital service, during which electrophysiological tests are carried out in order to verify the transmission of the auditory information from the implant to the brain. In these tests, specific electrodes are activated and the evoked potentials of the brainstem auditory relays are measured using scalp electrodes (Guiraud et al., 2007).

The first obstacle when trying to analyze such a data set comes from the fact that the information are scattered through medical records and are not always in a format that a statistician would expect them to be. A preliminary work of data formatting and quality control must then be done. Already at this point of the study, the R software presents itself as a powerful, if not essential, tool. The data set can be organized according to several independent factors, namely: the sex of the patient (male of female), the age at implantation (in days), the duration of privation before implantation (in days), the side of the implanted ear (left or right), the number of the activated electrode (in our case, #5 and #20, respectively more basal or more apical). Also, the patients are followed longitudinally and the main independent variable in the study is the duration of CI use at the date of visit (in days). The dependent measures are the latency of the waves III (response from the superior olive) and V (response from the inferior colliculus). The main goal of the study is to evaluate the neuronal maturation of the auditory pathways with the duration of CI use and how this maturation depends on the considered independent factors.

The data set included 232 patients (112 females and 120 males), some of them implanted bilaterally. The visits dates range from one month after the surgery until 14 years and a total

of over 13,000 independent measures are present in the data set. The data is well suited for a linear mixed-model analysis (Bates et al., 2011), in which the intercept latency for each patient is considered as a random factor. When the analysis is performed with an omnibus model including all the factors and their interactions, we end up with 43 factors and the analysis become almost unfeasible. In order to retain the factors that really matter in explaining the data, we back-fitted the fixed effects using the LMERConvenienceFunctions package of R (Newman et al., 2011; Tremblay and Tucker, 2011). This technique estimates the F statistics of each factor for the fitted model and the algorithm proceeds from the highest-order interaction terms towards the individual factors. The factor with the smallest value of F is considered for removal. A statistical test is then made between the models with and without this factor by using the log-likelihood ratio test. If the resulting p-value is below a given threshold α, then the factor is kept, otherwise it is discarded. This is done progressively until there are no factors left. Finally, an ANOVA table of the reduced model is computed using conservative estimations of the degrees of freedom of the denominator.
Using this technique we could find evidence for the following effects: (1) a difference in the response of electrodes #5 and #20, related to the anatomical position in the cochlear modiolus; (2) a latency difference between males and females, that correlates with anthropometric data; (3) a fast maturational rate followed by a standing plateau and a later increase of the latencies along the duration of CI use; (4) a lack of later increase in the latency time with CI duration use when the interval III–V is considered; and (5) a difference in the behavior of right *vs.* left implanted ears that interacts with the age at implantation. This later effect, the most interesting one found in thiss study, was confirmed in a subset of the data, where patients were selected to form matched groups in age and ear side.

References

Bates D, Maechler M, and Bolker B (2011) *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-41 (URL: `http://CRAN.R-project.org/package=lme4`).

Guiraud J, Besle J, Arnold L, Boyle P, Giard MH, Bertrand O, Norena A, Truy E, and Collet L (2007) Evidence of a tonotopic organization of the auditory cortex in cochlear implant users. *J Neurosci* 27(29): 7838–7846.

Newman AJ, Tremblay A, Nichols ES, Neville HJ, and Ullman MT (2011) The influence of language proficiency on lexical semantic processing in native and late learners of English. *J Cogn Neurosci* PMID #21981676.

Thai-Van H, Cozma S, Boutitie F, Disant F, Truy E, and Collet L (2007) The pattern of auditory brainstem response wave V maturation in cochlear-implanted children. *Clin Neurophysiol* 118(3): 676–89.

Tremblay A and Tucker BV (2011) The effects of n-gram probabilistic measures on the processing and production of four-word sequences. *The Mental Lexicon* (accepted for publication).

Analyse non paramétrique de séquences de potentiels d'action. Construction de modèles et de tests de qualité d'ajustement.

Christophe Pouzat

MAP5 - Mathématiques Appliquées à Paris 5
Université Paris-Descartes
45, rue des Saints-Pères, 75006, Paris
christophe.pouzat@parisdescartes.fr

Mots clefs : Neurosciences, fonctions splines, vraisemblance pénalisée.

Les neurosciences contemporaines utilisent de plus en plus d'enregistrements extra-cellulaires multiples effectués avec des matrices d'électrodes. Ces enregistrements, une fois pré-traitées par une étape de tri des potentiels d'action, fournissent au neurophysiologiste et au statisticien de longues séquences de potentiels d'actions venant de plusieurs neurones identifiés. Notre communication sera consacrée à une méthode d'analyse pour ce type de données.

Les processus ponctuels sont reconnus depuis plus de 40 ans comme une formalisation pertinente des données [1]. Suivant les travaux pionniers de David Brillinger [2,3] nous modélisons directement l'*intensité conditionnelle* (ou l'*intensité stochastique*) du processus ponctuel et nous employons une discrétisation du temps qui ramène le problème à une régression binomiale. Cette discrétisation est également appelée « approximation probabiliste » par Berman et Turner [4]. Les lacunes de nos connaissances sur la biophysique des neurones nous amènent à adopter une approche non-paramétrique ; c'est-à-dire que nous développons concrètement notre prédicteur linéaire sur une base de fonctions splines, comme proposé par Kass et Ventura [5]. Mais nous nous distinguons de ces derniers en employant une vraisemblance pénalisée, c'est-à-dire de « vraies » splines de lissage [6,7]. Notre approche est mise en œuvre dans le paquet `STAR` (*Spike Train Analysis with R*), disponible sur `CRAN` et « construit sur » le paquet `gss` (*general smoothing spline*) de Chong Gu [7].

`STAR` permet, une fois une estimation non-paramétrique de l'intensité conditionnelle obtenue, de tester la qualité de l'ajustement du modèle aux données avec les tests proposés par Y. Ogata [8]. Nous proposons également un nouveau test basé sur l'identification de la différence entre le processus de comptage observé et l'intensité conditionnelle intégrée avec un mouvement brownien standard (après une transformation du temps adéquate). `STAR` permet également de simuler des processus ponctuels – suivant une intensité conditionnelle estimée – avec la méthode de l'« éclaircissage » (*thinning*) [9].

Plusieurs exemples d'applications, sur des données provenant de différents laboratoires, seront présentés.

Références

[1] D. H. Perkel, G. L. Gerstein, G. P. Moore (1968). Neuronal spike trains and stochastic point processes. I the single spike train. *Biophysical Journal*, **7**, 391-418.
[2] D. R. Brillinger (1988). Maximum likelihood analysis of spike trains of interacting nerve cells.*Biological Cybernetics*, **59**(3), 189-200.
[3] D. R. Brillinger (1992). Nerve Cell Spike Train Data Analysis : A Progression of Technique.

Journal of the American Statistical Association, **87**(418), 260-271.
[4] M. Berman, T. R. Turner (1992). Approximating Point Process Likelihoods with GLIM. *Applied Statistics*, **41**, 31-38.
[5] R. E. Kass, V. Ventura (2001). A spike-train probability model. *Neural Computation*, **13**, 1713-1720.
[6] G. Wahba (1990). Spline Models for Observational Data. *SIAM*.
[7] C. Gu (2002). Smoothing Spline Anova Models. *Springer*.
[8] Y. Ogata (1988). Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical Association*, **83**, 9-27.
[9] Y. Ogata (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, **IT-29**, 23-31.

DiscreteTS : two hidden-Markov models for time series of count data

J. Alerini[a], **M. Olteanu**[b] and **J. Ridgway**[b]

[a] PIREH (Pôle Informatique de Recherche et d'Enseignement en Histoire)
Université Paris 1
1 Rue Victor Cousin, 75005 Paris, France
julien.alerini@univ-paris1.fr

[b] SAMM (Statistique, Analyse et Modélisation Multidisciplinaire), EA 4543
Université Paris 1
90 Rue de Tolbiac, 75013 Paris, France
madalina.olteanu@univ-paris1.fr, James.Ridgway@ensae-paristech.fr

Mots clefs : Integer-valued time series, hidden Markov models, autoregressive regime-switching models.

Time series of count data are encountered often in Humanities and Social Sciences. Modeling this kind of data is a challenging topic for the statistician : autoregressive structure, over-dispersion in zero, existence of several unobserved regimes controlling the process.

One common approach used for modeling integer-valued time series are the hidden Markov models. However, the available R packages such as HiddenMarkov [1] or HMM [2] are implemented for usual distributions only. Moreover, none of this packages performs estimation for autoregressive Markov-switching models.

Two new models were recently introduced in [3] and [4] :

1. ZIP-HMM (Hidden Markov models with zero-inflated Poisson distributions) were proposed in order to take into account the over-dispersion in zero. This model is a usual hidden Markov model, except that the Poisson distribution of the observed process conditionally to the hidden state was replaced by a mixture of a Poisson and a Dirac distribution.

2. INAR(p)-HMM (Integer-valued autoregressive models with Markov-switching regimes) were introduced as a parallel to the autoregressive hidden-Markov models existing already in the continuous case [5]. The observed process is supposed to behave as an integer-valued autoregressive INAR(p) [6], whose parameters are controlled by the states of a hidden Markov chain.

For both models, the estimation procedure is achieved through the EM algorithm. These models were implemented in a R-package called DiscreteTS. The package provides the possibility of either simulating these models, or of estimating them starting from a given time-series. A toy example on medieval historical data is also provided.

References

[1] http://cran.r-project.org/web/packages/HiddenMarkov/index.html
[2] http://cran.r-project.org/web/packages/HMM/index.html

[3] Olteanu M., Ridgway J. (2012). Hidden Markov models for time series of counts with excess zeros. *Proceedings of ESANN 2012*, 133-138
[4] Ridgway J. (2011). Hidden Markov models for time series of count data. *Rapport de stage*
[5] Hamilton J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357-384.
[6] Al-Osh M.A. and Alzaid A.A. (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability*, **vol.27(2)**, 314-324

BiiPS : un logiciel pour l'inférence bayésienne dans les modèles graphiques utilisant des méthodes de Monte Carlo séquentielles

A. Todeschini[a], **F. Caron**[b], **P. Legrand**[c] and **P. Del Moral**[d]

[a]Inria Bordeaux - Sud-Ouest, équipe ALEA
adrien.todeschini@gmail.com

[b]Inria Bordeaux - Sud-Ouest, équipe ALEA
IMB, Institut de Mathématiques de Bordeaux, UMR CNRS 5251
Université de Bordeaux
francois.caron@inria.fr

[c]Université Bordeaux Segalen
Inria Bordeaux - Sud-Ouest, équipe ALEA
IMB, Institut de Mathématiques de Bordeaux, UMR CNRS 5251
pierrick.legrand@u-bordeaux2.fr

[d]Inria Bordeaux - Sud-Ouest, équipe ALEA
IMB, Institut de Mathématiques de Bordeaux, UMR CNRS 5251
Université de Bordeaux
pierre.del-moral@inria.fr

Mots clefs : Méthodes de Monte Carlo séquentielles, filtrage particulaire, systèmes de particules en interaction, modèles graphiques, langage BUGS, estimation de paramètres, poursuite de cibles, filtrage de signaux, volatilité stochastique, modèles biologiques.

L'un des principaux facteurs du succès des méthodes de Monte Carlo par chaînes de Markov (MCMC) en inférence bayésienne est qu'elles peuvent être mises en œuvre avec peu d'effort dans une grande variété de cas. De nombreux logiciels ont été développés, comme BUGS et JAGS qui ont contribué à populariser les méthodes bayésiennes. Ces logiciels permettent aux utilisateurs de définir leur modèle statistique dans un langage appelé langage BUGS, puis exécutent des algorithmes MCMC en boîte noire. Une nouvelle génération d'algorithmes, basés sur des systèmes de particules en interaction, a fait son apparition ces vingt dernières années. Bien que ces méthodes, dites "particulaires" ou "Monte Carlo séquentielles", soient devenues une classe très populaire de méthodes numériques, il n'existe pas de tel logiciel "boîte noire" pour cette classe de méthodes. Le logiciel BiiPS, acronyme pour **Bayesian Inference with Interacting Particle Systems**, vise à combler ce manque. A partir d'un modèle graphique défini en langage BUGS, il met en œuvre automatiquement des algorithmes particulaires et fournit des résumés statistiques des distributions *a posteriori*. Dans cette présentation, nous mettrons en évidence quelques-unes des fonctionnalités du logiciel BiiPS, son interface R, et des applications du logiciel au suivi de cible, à l'estimation de la volatilité stochastique en finance et à la calibration de modèles proie-prédateur.

Références

[1] N. De Freitas, A. Doucet, Arnaud and N. Gordon (2001). Sequential Monte Carlo Methods in Practice. Springer.
[2] P. Del Moral (2004). Feynman-Kac formulae: genealogical and interacting particle systems

with applications. Springer. Series: Statistics for Engineering and Information Science.
[3] P. Del Moral and A. Doucet (2012). Sequential Monte Carlo & Genetic particle models. Theory and Practice. Chapman & Hall. Series: Mathematics and Statistics.
[4] A. Doucet and A.M. Johansen (2010). A tutorial on particle filtering and smoothing: fifteen years later. In D. Crisan and B. Rozovsky editors, Oxford Handbook of Nonlinear Filtering. Oxford University Press.
[5] C. Andrieu, A. Doucet, and R. Holenstein (2010). Particle Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269-342.

Rotation orthogonale en ACP de données mixtes. Le package PCAmixdata et une application en sociologie culturelle.

M. Chavent[a], **V. Kuentz-Simonet**[b], **Z. Lakatos**[c] and **J. Saracco**[a]

[a] INRIA Bordeaux Sud Ouest,
Institut de Mathématiques de Bordeaux,
Université de Bordeaux,
351 cours de la libération, 33405 Talence Cedex, France
{marie.chavent, jerome.saracco}@math.u-bordeaux1.fr

[b]Irstea, UR ADBX,
50 avenue de Verdun, 33612 Cestas Cedex, France
vanessa.kuentz-simonet@irstea.fr

[c] BME-GTK
Université Polytechnique et Economique de Budapest,
Faculté des Sciences Economiques et Sociales,
Stoczek utca 2, 1111 Budapest, Hongrie
sultan.lakatos@gmail.com

Mots clefs : Données mixtes, mélange de variables quantitatives et qualitatives, analyse en composantes principales, analyse des correspondances multiples, rotation, sociologie culturelle.

La méthode PCAMIX est une méthode d'analyse de données qui permet en particulier de déterminer des composantes principales pour des données mixtes, c'est à dire pour un mélange de variables quantitatives et de variables qualitatives. L'approche PCAMIX inclut ainsi comme cas particulier l'analyse en composantes principales (ACP) et l'analyse des correspondances multiples (ACM). Kiers (1991) a présenté une manière de faire de la rotation orthogonale dans le cadre de la méthode PCAMIX.
Dans cette communication, nous donnons une nouvelle présentation de la méthode PCAMIX dans laquelle les composantes principales et les "squared loadings" sont obtenus via une décomposition aux valeurs singulières. Nous proposons alors une procédure efficace pour faire de la rotation de type Varimax en PCAMIX. Nous donnons en particulier une solution directe de l'angle optimal de rotation. Un package appelé "PCAmixdata" comprenant la méthode PCAMIX ainsi que la rotation en PCAMIX a été implémenté dans R. Ce package disponible auprès des auteurs et sera bientôt sur le site du CRAN. Notons que ce package permet aussi de gérer des données manquantes.

Nous illustrerons avec des données simulées le bon comportement numérique de l'algorithme proposé dans ce package et le gain en terme de temps de calculs par rapport à l'approche de Kiers (1991).
Nous présenterons aussi une application sur des données réelles permettant d'illustrer l'intérêt de faire de la rotation en ACM (cas particulier de PCAMIX lorsque toutes les variables sont qualitatives). Cette application est issue d'une collaboration avec un doctorant hongrois en sociologie, Zoltan Lakatos, dont les travaux portent sur l'évolution des valeurs culturelles dans les sociétés. Plus précisément, il s'intéresse à une critique empirique de la thèse sociologique du post- matérialisme du politologue américain Ronald Inglehart. Les données sur lesquelles il travaille sont issues d'une enquête globale (World Values Survey) sur les valeurs culturelles, initiée et dirigée par Ronald Inglehart (enquêtes individuelles menées au niveau national dans une centaine de pays, par vagues successives depuis 1981). Une ACM au niveau des répondants est appliquée pour mettre en évidence les limitations des thèses d'Inglehart sur le contenu et l'évolution des valeurs culturelles. Grâce à la rotation, il devient possible d'identifier deux dimensions distinctes, à savoir "religieux vs. laïque" et "autoritaire vs. libertaire" que les méthodes d'Inglehart (ACP appliquées aux moyennes nationales) traitent comme appartenant à une seule et même dimension sous l'étiquette "valeurs traditionnelles vs. modernes". Les résultats obtenus par l'ACM avec rotation constituent un pas important vers le dépassement des typologies

réductionnistes en sociologie des valeurs présentant un schéma unidirectionnel et linéaire de l'évolution socioculturelle.

Références

[1] Chavent, M., Kuentz-Simonet, V. Saracco, J. (2012). Orthogonal rotation in PCAMIX. To appear in *Advances in Data Analysis and Classification*.

[2] Kiers, H.A.L. (1991) Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56:197-212.

[3] LAKATOS, Zoltan (2012) The Cultural Values-Economic Growth Nexus: A Critical Reassessment. Doctoral thesis: Faculty of Social Sciences, Eotvos Lorand University of Sciences (ELTE TaTK), Budapest.

MAINT.Data: Parametric Modelling and Analyzing Interval Data in R

A. Pedro Duarte Silva[a] and **P. Brito**[b]

[a]Faculdade de Economia e Gestão & CEGE
Catholic University of Portugal / Porto
Rua Diogo Botelho, 1327, 4169-005 Porto, Portugal
psilva@porto.ucp.pt

[b]Faculdade de Economia & LIAAD-INESC TEC
Universidade do Porto
Rua Dr. Roberto Frias, 4200-464 Porto, Portugal
mpbrito@fep.up.pt

Keywords : Symbolic data, Interval data, Parametric modelling of interval data, Statistical tests for interval data, Skew-Normal distribution.

In multivariate data analysis, data is usually represented in a $n \times p$ data-array where n "individuals" take exactly one value for each of p descriptive variables. Symbolic Data Analysis (see, e.g. Diday and Noirhomme-Fraiture (2008), Noirhomme-Fraiture and Brito (2011)) enlarged the classical framework, proposing a model where variability associated to each single observation is directly taken into account. New variable types - interval, categorical multi-valued and modal variables - have been introduced, which may take multiple, possibly weighted, values for each variable. We focus on the analysis of interval data, i.e., where elements are described by variables whose values are intervals of $\mathbb{R}$.
Parametric inference methodologies based on probabilistic models for interval variables are developed in Brito and Duarte Silva (2011). Under this approach, each interval is represented by its midpoint and log-range, for which Normal and Skew-Normal (Azzalini and Dalla Valle (1996)) distributions are assumed. The main advantage of the Normal model lies in that it allows for a straightforward application of classical inference methods, and permits a direct modelling of the variables' covariance structure. If the intervals' midpoints are looked at as "location indicators" of the variables' values, assuming that they follow a joint Normal distribution corresponds to the usual Gaussian assumption for classical data; the log transformation of the ranges allows overcoming the difficulties created by their limited domain. It then follows that the marginal distributions of the midpoints are Normal and those of the ranges are Log-Normal. In a second step, we also consider the Skew-Normal distribution, which alleviates some of the known limitations of the multivariate Normal, by introducing skewness parameters. The intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix, which are represented by five different possible configurations. In the most general formulation we allow for non-zero correlations among all midpoints and log-ranges; other cases of interest are:

- The interval variables are independent, but for each variable, the midpoint may be correlated with its range;
- Midpoints (respectively, ranges) of different variables may be correlated, but no correlation between midpoints and ranges is allowed;
- Midpoints (respectively, ranges) of different variables may be correlated, the midpoint of

each variable may be correlated with its range, but no correlation between midpoints and ranges of different variables is allowed.

We present the package MAINT.DATA, available on CRAN, which implements the proposed methodologies in R using S4 classes and methods. Its basic class, *IData*, represents $n \times p$ data sets of interval variables, and has interval data specializations of traditional R methods such as *print*, *summary* and indexing and assignment operators.
Maximum Likelihood Estimation and Multivariate Analysis of Variance for interval data are performed by two other *IData* methods, named *mle* and *MANOVA*. These methods create objects of class *IdtSngDE* (Single Distribution Estimates), *IdtHomMxE* (Homocedastic Mixture Estimates) or *IdtHetMxE* (Heterocedastic Mixture Estimates), representing the results of the analysis performed. The *IdtSngDE*, *IdtHomMxE* and *IdtHetMxE* classes have further methods for inspecting results, including the computation of standard errors and tests for the models and configurations assumed. Furthermore, the latter two classes have respectively a *lda* (class *IdtHomMxE*) and *qda* (class *IdtHetMxE*) method, implementing Linear and Quadratic Discriminant Analysis of interval data.
Planned extensions of MAINT.DATA include the implementation of a *lm* (Linear Models) method for the *IData* class, and the inclusion of other established estimation methodologies, such as the Generalized Method of Moments (Hansen (1982)), as feasible alternatives to maximum likelihood estimation for problems where the optimization of the interval data likelihood is too demanding in time and computer resources.

References

[1] Azzalini, A., Dalla Valle, A. (1996). The multivariate Skew-Normal distribution, *Biometrika*, **83**, (4), 715–726.
[2] Brito, P., Duarte Silva, A.P. (2011). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, **39**, (1), 3–20.
[3] Noirhomme-Fraiture, M., Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining* , **4**, (2), 157–170.
[4] Diday, E., Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester.
[5] Hansen, L.P.(1982). Large sample properties of Generalized Method of Moments estimators. *Econometrica*, **50**, 1029-1054.

Unravelling 'omics' data with the R package `mixOmics`

K-A. Lê Cao[a], **I. González**[b] and **S. Déjean**[b]

[a]Queensland Facility for Advanced Bioinformatics
University of Queensland
4072 St Lucia, QLD, Australia
k.lecao@uq.edu.au

[b]Institut de Mathématiques
Université de Toulouse et CNRS
UMR 5219, F-31062 Toulouse, France
ignacio.gonzalez@math.univ-toulouse.fr
sebastien.dejean@math.univ-toulouse.fr

Mots clefs : multivariate statistics, data integration, high-throughput biological data.

Recent advances in high throughput 'omics' technologies enable quantitative measurements of expression or abundance of biological molecules of a whole biological system. The transcriptome, proteome and metabolome are dynamic entities, with the presence, abundance and function of each transcript, protein and metabolite being critically dependent on its temporal and spatial location.

With `mixOmics`, we are currently establishing a global analytical framework to extract relevant information from high throughput 'omics' platforms such as genomics, proteomics and metabolomics. Specifically, the statistical methodologies developed and implemented in the R package focus on the so-called multivariate projection-based approaches, which can handle such large data sets, deal with multicollinearity and missing values. These methodologies enable dimension reduction by projecting these large data sets into a smaller subspace, to capture the largest sources of variation in the biological studies. These techniques enable exploration, visualisation of the data and lead to biological insights.

Principal Component Analysis (PCA) is a commonly used dimension reduction technique to highlight expression patterns that might be due to biological variation, or systematic platform bias in a single data set. Recently, we have proposed another variant based on Independent Component Analysis (IPCA) [1]. By applying Lasso penalisation [2] on the PCA or IPCA components, we further reduce the dimension of the data by selecting the relevant information (the measured biological entities) related to the biological study. Both approaches are unsupervised, i.e. the focus is to identify the genes, proteins or metabolites with similar information without taking into account experimental knowledge on class labels of the samples. A supervised approach was also developed based on Partial Least Square Discriminant Analysis (PLS-DA) [3] to select discriminative biological entities across several groups of samples.

Whilst single omics analyses are commonly performed to detect between-groups difference from either static or dynamic experiments, the integration or combination of multi-layer information is required to fully unravel the complexities of a biological system. Data integration relies on the currently accepted biological assumption that each functional level is related to each other.

Therefore, considering all the biological entities (transcripts, proteins, metabolites) as part of a whole biological system is crucial to unravel the complexity of living organisms.

To that purpose, we have developed integrative approaches, such as regularized Canonical Correlation Analysis (rCCA) [4], sparse PLS [5,6] to highlight or understand the relationship between two types of biological entities. We have demonstrated on several biological studies that this integrative analyses of large scale omics datasets could generate new knowledge not accessible by the analysis of a single data type alone.

All methodologies are implemented in `mixOmics` along with S3 methods for an easy use of the package and an easy interpretation via graphical tools [6]. Our website gives more information about the methodologies and how to use the package (http://www.math.univ-toulouse.fr/~biostat/mixOmics/). For the non R specialist, a web application was also developed and made available to the research community (http://mixomics.qfab.org).
In this presentation, I will cover the recent developments of `mixOmics`, illustrate the use of the methodologies to various biological studies and demonstrate the usefulness of the graphical tools to give biological meaning to the obtained results.

Références

[1] Yao, F., Coquery J. and Lê Cao K.-A. (2012). Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinformatics*, 13:24.
[2] Tibshirani, R. (2007). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1): 267-288.
[3] Lê Cao K.-A., Boitard, S. and Besse, P. (2011). Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12:253.
[4] González I., Déjean S., Martin P.G.P., Gonçalves O., Besse P. and Baccini A. (2009) Highlighting Relationships Between Heteregeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis. *Journal of Biological Systems* **17**(2), pp 173-199.
[5] Lê Cao K.-A., Rossouw, D., Robert-Granié C., Besse, P. (2008). A Sparse PLS for Variable Selection when Integrating Omics data. *Statistical Applications in Genetics and Molecular Biology*: Vol. 7 : Iss. 1, Article 35.
[6] Lê Cao K.-A., Martin P.G.P, Robert-Granié C., Besse, P. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* 10: 34.
[7] Lê Cao K.-A., González I. and Déjean S (2009). integrOmics: an R package to unravel relationships between two omics data sets. *Bioinformatics* **25**(21):2855-2856. *Note: the package has since been renamed mixOmics.*

Les cartes auto-organisatrices de Kohonen appliquée à l'étude des communautés de micro-algues des cours d'eau

M. Bottin[a], **J.L. Giraudel**[b], **J. Guéguen**[a], **S. Boutry**[a], **I. Lavoie**[c], **N. Antunes**[d] and **J. Tison-Rosebery**[a]

[a]Unité REBX
Irstea
50 avenue de Verdun, 33612 Cestas cedex, France
Marius.Bottin@irstea.fr

[b]EPOC
Université de Bordeaux - UMR CNRS 5805
Rue Doyen Joseph LAJUGIE 24019 PERIGUEUX Cedex, FRANCE

[c]Centre Eau Terre Environnement
Institut National de la Recherche Scientifique
490 rue de la Couronne,Québec, Québec,Canada

[d] UMR PACEA/ALEA.
Université de Bordeaux - UMR CNRS 5199 / INRIA
Université Bordeaux 1 - 351 cours de la libération 33405 Talence Cedex

Mots clefs : Écologie, Diatomées, Communautés, Cartes auto-organisatrices de Kohonen, Bioindication.

Les diatomées sont des algues microscopiques libres ou fixées sur les galets ou végétaux des rivières. Les différentes espèces pouvant composer ces communautés présentent des préférences environnementales différentes, ce qui permet une description des conditions écologiques à partir de l'observation des espèces présentes. Elles sont donc utilisées en routine pour l'évaluation de l'état écologique des cours d'eau depuis une vingtaine d'années [1, 2]. Afin de mettre en œvre la Directive Cadre Européenne sur l'Eau, Irstea a participé au développement de nouvelles méthodes d'évaluation, grâce à l'analyse de grandes bases de données d'échelle nationale. Du fait de leur grande diversité et de la complexité des phénomènes agissant sur les communautés de diatomées, l'analyse de leur répartition requiert l'utilisation de méthodes innovantes telles que la logique floue ou les réseaux de neurones. En effet, les méthodes statistiques classiques s'avèrent souvent incapables de décrire une part acceptable de la variation de ces communautés.

Une cartes auto-organisatrice de Kohonen (Self-Organizing Kohonen Maps : SOM) est un réseau de neurones à apprentissage non supervisé. Cette technique permet à la fois une classification objective des communautés et l'étude simultanée de gradients biologiques et environnementaux. Un de ses avantages principaux est la prise en compte de gradients non-linéaires complexes, qui explique en partie sa grande efficacité pour décrire les communautés biologiques [3]. Aussi, cette technique a été utilisée pour décrire, entre autres, les variations spatiales [4] ou l'impact des pesticides [5] sur les communautés de diatomées. En revanche, comme pour les autres réseaux de neurones, de nombreux paramètres doivent être réglés, de façon adaptée aux différents types de jeux de données (nombre de neurones, mesures de dissimilarité, nombres d'étapes, fonctions de voisinages ...).

Des packages existent sous R pour la réalisation des cartes de Kohonen, mais ils sont à notre connaissance assez peu utilisés pour l'ordination de communautés biologiques. Des solutions existantes sous d'autres logiciels (en particulier MATLAB) paraissent mieux adaptées à ce type de données. Afin de pouvoir adapter plus particulièrement cette technique à nos besoins, nous avons développé un ensemble de fonctions sous R permettant la réalisation et la visualisation de SOM. L'utilisation conjointe de R et du langage C nous a permis d'atteindre une efficacité de calcul mieux adaptée à la grande taille de nos tableaux de données, tout en gardant une relative flexibilité de mise en œuvre. Les possibilités graphiques de ce logiciel nous ont de plus permis de développer des outils de visualisation et d'analyses de SOM, eux aussi mieux adaptés à nos besoins.
Les différentes fonctions ont été compilées dans un nouveau package, contenant aussi d'autres techniques (comme les « fuzzy pattern trees »), que nous espérons soumettre aux CRAN d'ici la fin de l'année 2012.

Nous discuterons plus particulièrement de l'algorithme utilisé dans ce package et des possibilités qu'il offre, notamment dans le domaine de l'écologie. Aussi, nous agrémenterons cet exposé d'exemples dans lesquels ces fonctions sont déjà utilisées pour la typologie des communautés :
– des cours d'eau français de métropole.
– des cours d'eau de la Guadeloupe et de la Martinique pour l'application de la Directive Cadre Européenne sur l'Eau.
– de cours d'eau canadiens dans le cadre de la réalisation d'une nouvelle version de l'« Indice Diatomées de l'Est du Canada ».

Références

[1] J. Prygiel, M. Coste, and J. Bukowska. Review of the major diatom-based techniques for the quality assessment of rivers. state of the art in europe. In *Use of algae for monitoring rivers III*, pages 224–238. Agence de I'Eau Artois-Picardie Press Douai Cedex, 1999.

[2] M. Coste, S. Boutry, J. Tison-Rosebery, and F. Delmas. Improvements of the biological diatom index (BDI) : Description and efficiency of the new version (BDI-2006)mo. *Ecological Indicators*, 9(4) :621–650, 2009.

[3] J. L. Giraudel and S. Lek. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling*, 146(1-3) :329–339, 2001.

[4] J. Tison, Y. S. Park, M. Coste, J. G. Wasson, L. Ector, F. Rimet, and F. Delmas. Typology of diatom communities and the influence of hydro-ecoregions : A study on the french hydrosystem scale. *Water Research*, 39(14) :3177–3188, 2005.

[5] S. Morin, M. Bottin, N. Mazzella, F. Macary, F. Delmas, P. Winterton, and M. Coste. Linking diatom community structure to pesticide input as evaluated through a spatial contamination potential (phytopixal) : A case study in the neste river system (south-west france). *Aquatic Toxicology*, 94(1) :28–39, 2009.

Comparison of network inference packages and methods for multiple network inference

Nathalie Villa-Vialaneix[a], **Nicolas A. Edwards**[b], **Laurence Liaubet**[b]
Nathalie Viguerie[c], **Magali SanCristobal**[b]

[a]Laboratoire SAMM - Université Paris 1 (Panthéon-Sorbonne)
90 rue de Tolbiac, 75013 Paris - France
nathalie.villa@univ-paris1.fr

[b]INRA, UMR444 - Laboratoire de Génétique Cellulaire
F-31326 Castanet Tolosan cedex, France
nicolas.ae@free.fr, {laurence.liaubet,magali.san-cristobal}@toulouse.inra.fr

[c]Inserm UMR1048, Obesity Research Laboratory
I2MC, Institute of Metabolic and Cardiovascular Diseases
CHU Rangueil, Toulouse
nathalie.viguerie@inserm.fr

Keywords: network inference, transcriptomic data, gene co-expression network, Gaussian graphical model, multiple graphical structure

Integrative and systems biology is a very promising tool for deciphering the biological and genetic mechanisms underlying complex traits. In particular, gene networks are used to model interactions between genes of interest. They can be defined in various ways, but a standard approach is to infer a co-expression network from genes expression measured by means of sequencing techniques (for example, microarrays). Among methods used to perform the inference, **Gaussian graphical models** (GGM) are based on the assumption that the gene expressions are distributed as Gaussian variables, and Σ is their covariance matrix. Non-zero partial correlations between two genes are modeled by network edges, and are directly obtained from the inverse of Σ. But it turns out that estimating the inverse of Σ leads to an ill posed problem, since this kind of data leads to a number of observations (typically less than one hundred) that is usually much smaller than the number of variables (the number of genes/nodes in the network can range from a few hundred to several thousands). To overcome this difficulty, the seminal papers [8, 9] were the basis for the R package **GeneNet**, in which the partial correlation is estimated either by means of a bootstrap approach (not available in the package anymore) or of a shrinkage approach. More recently, the ability to handle genomic longitudinal data was also added as described in [7]. Then, [6] and later [3] introduced sparse approaches, both implemented in the R package **glasso** (graphical LASSO). Similarly, [4] describes the methods implemented in the package **parcor** that provides several regularization frameworks (PLS, ridge, LASSO...) to infer networks by means of Gaussian graphical models. Finally, [2, 1] describe several extensions of the Gaussian graphical model implemented in the package **simone** such as latent variable models and time-course transcriptomic data.
In systems biology, an interesting issue is to link gene functioning to an external factor. Thus, transcriptomic data are often collected in different experimental conditions. One must then understand which genes are correlated *independently* from the condition and which ones are correlated *depending* on the condition, under the plausible biological assumption that a common functioning should exist regardless of said condition. A simple naive approach would be to infer a different network from each sample, and then to compare them. Alternative approaches

are described in [2, 1] and implemented in **simone**: the log-likelihood can be penalized by a modified group-LASSO penalty or the empirical covariance matrix can be modified by adding a component depending on all samples. The purpose of this communication is to present a full comparative case study of this problem on two real data sets.
The first dataset has been collected during the DiOGenes project[1]: a few hundreds human obese individuals were submitted to a 8 weeks low calorie diet. The expressions of pre-selected genes as well as physiological variables (age, weight, waist size...) were collected *before* and *after* the diet (see [5] for further information). The underlying issue is to understand how the diet has affected the correlations between all these variables. The second data set has been collected during the Delisus project[2]: the expression of several thousands genes were collected from 84 pigs (in both *Landrace* and *Large White* breeds). The underlying issue is to understand how the breed affects the correlations between a set of selected genes which were found to be differentially expressed for the breed.
The comparison is lead by using independent inference from the packages **GeneNet**, **glasso** and **simone** or by using the different joint models included in **simone** or even by proposing new joint approaches based on the aforementionned packages. Networks are inferred from the previously described real datasets or from simulated datasets that mimic the real ones. The proximity between networks inferred from different methods or from different conditions is assessed by means of common edge counts, or, when available, by the accuracy of the inferred network when compared to the true one. A biological discussion about the relevance of the inferred networks will also be provided.

References

[1] J. Chiquet, Y. Grandvalet, and C. Ambroise. Inferring multiple graphical structures. *Statistics and Computing*, 21(4):537–553, 2011.

[2] J. Chiquet, A. Smith, G. Grasseau, C. Matias, and C. Ambroise. SIMoNe: Statistical Inference for MOdular NEtworks. *Bioinformatics*, 25(3):417–418, 2009.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[4] N. Kraemer, J. Schaefer, and A.L. Boulesteix. Regularized estimation of large-scale gene regulatory networks using Gaussian Graphical models. *BMC Bioinformatics*, 10:384, 2009.

[5] T.M. Larsen, S.M. Dalskov, M. van Baak, S.A. Jebb, A. Papadaki, A.F.H. Pfeiffer, J.A. Martinez, T. Handjieva-Darlenska, M. Kunešovà, M. Pihlsgård, S. Stender, C. Holst, W.H.M. Saris, and A. Astrup. Diets with high or low protein content and glycemic index for weight-loss maintenance. *New England Journal of Medicine*, 363:2102–2113, 2010.

[6] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistic*, 34(3):1436–1462, 2006.

[7] R. Opgen-Rhein and K. Strimmer. Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65, 2006.

[8] J. Schäfer and K. Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

[9] J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implication for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:1–32, 2005.

[1] supported by funding from the European Communities (DiOGenes, FP6-513946, MolPAGE, LSHG-CT-2004-512066 and ADAPT, HEALTH-F2-2008-2011 00), Fondation pour la Recherche Médicale and Région Midi-Pyrénées `http://www.diogenes-eu.org`

[2] funded by the ANR, `http://www.inra.fr/les_partenariats/programmes_anr/genomique/genanimal/appel_a_projets_2007/delisus`

Représentation, analyse et simulation de processus ponctuels spatio-temporels

E. Gabriel[a]

[a]Département de Mathématiques
Université d'Avignon et des Pays de Vaucluse
33 Rue Louis Pasteur, 84000 Avignon
edith.gabriel@univ-avignon.fr

Mots clefs : Epidémiologie, Processus ponctuels spatio-temporels, Simulation, Statistique spatiale.

Un semis de points est un ensemble de sites géographiques, irrégulièrement distribués dans une région S, en lesquels des évènements sont relevés, par exemple la localisation d'arbres dans une forêt régénérée naturellement (Diggle, 2003). Un semis de points peut être modélisé par une réalisation d'un processus stochastique spatial représenté par un ensemble de variables aléatoires : $Y(S_m)$, $S_m \in S$, où $Y(S_m)$ est le nombre d'évènements ayant lieu dans la sous-région S_m de S. La simulation de processus ponctuels spatiaux est essentiellement implémentée dans les packages R spatstat (Baddeley et Turner, 2005) et splancs (Rowlingson et Diggle, 1993).

Beaucoup de processus spatiaux ont aussi une composante temporelle qui doit être prise en compte lors de la modélisation du phénomène sous-jacent (par exemple la distribution de cas d'une maladie ou l'estimation du risque de pollution de l'air). Les processus ponctuels spatio-temporels doivent alors être privilégiés comme modèles potentiels par rapport aux processus purement spatiaux. Il existe une vaste littérature sur l'analyse de processus ponctuels dans le temps (e.g. Cox et Isham, 1980 ; Daley et Vere-Jones, 2003) et dans l'espace (e.g. Cressie, 1991 ; Diggle, 2003 ; Møller et Waagepetersen, 2003). Les méthodes d'analyse de processus ponctuels spatio-temporels sont beaucoup moins établies (voir Diggle, 2006 ; Gabriel et Diggle, 2009 ; Cressie et Wikle, 2011) bien qu'il y ait une large littérature sur l'utilisation de modèles de processus ponctuels dans des domaines spécifiques comme la sismologie (Zhuang, Ogata et Vere-Jones, 2002).

Gabriel *et al.* (2012) proposent un package R, stpp, pour la représentation graphique, l'analyse et la simulation de processus ponctuels spatio-temporels. Il s'agit de présenter ce package. Après une brève description des processus ponctuels spatio-temporels, certains modèles seront présentés : des modèles classiques comme le processus de Poisson et le processus de Cox et des modèles souvent utilisés en épidémiologie comme les processus d'infection et de contagion. L'objectif ici est de proposer des algorithmes pour leur simulation, un outil statistique basé sur les propriétés d'ordre 2 du processus pour leur analyse et des outils de représentation graphique en dimension 2 et en dimension 3.

Références

[1] Baddeley, A., Turner, R. (2005). spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, **12**(6), 1-42.
[2] Cox, D., Isham, V. (1980). *Point Processes*, Chapman and Hall, London.
[3] Cressie, N. (1991). *Statistics for spatial data*, New-York, Wiley.
[4] Cressie, N., Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data*, Hoboken, Wiley.
[5] Daley, D., Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, Second

Edition, Springer, New York.
[6] Diggle, P. (2003). *Statistical analysis of spatial point patterns*, second edition. Arnold.
[7] Diggle, P. (2006). *Spatio-temporal point processes: methods and applications*, In Semstat2004, B. Finkenstadt, L. Held and V. Isham (eds), 1-45, London: CRC Press.
[8] Gabriel, E., Diggle, P. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, **63**, 43–51.
[9] Gabriel, E., Rowlingson, B., Diggle, P. (2012). STPP: Plotting, simulating and analysing Spatio-Temporal Point Patterns. *Submitted*.
[10] Møller, J., Waagepetersen, R. (2003). *Statistical inference and simulation for spatial point processes*, Monographs on Statistics an Applied Probability 100. Chapman & Hall.
[11] Rowlingson, B., Diggle, P. (1993). Splancs: Spatial Point Pattern Analysis Code in S-Plus. *Computers and Geosciences*, **19**, 627-655.
[12] Zhuang, J., Ogata, Y., Vere-Jones, D. (2002). Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, **97**, 369–380.

Package `CPMCGLM` : Correction de la pvaleur engendré par la recherche d'un codage d'une variable explicative dans un modèle linéaire généralisé

J. Riou[a,b] and **B. Liquet**[b]

[a]Equipe Biometrie
Danone Research
RD 128, Avenue de la Vauve, 91767 Palaiseau Cedex, FRANCE
jeremie.riou@isped.u-bordeaux2.fr

[b]Equipe Biostatistique
ISPED
Université de Bordeaux - CR INSERM U897
146 rue Léo Saignat, 33076 Bordeaux Cedex, FRANCE
benoit.liquet@isped.u-bordeaux2.fr

Mots clefs : Procédure de Bonferroni, Modèles Linéaires Généralisés, Codages Multiples, Multiplicité des tests, Méthodes de rééchantillonage

Une pratique courante en modélisation consiste à transformer une variable quantitative en variable catégorielle. Cette transformation se base normalement sur des seuils scientifiquement reconnus. Mais, dans de nombreux cas, les seuils ne sont pas connus et il est nécessaire de déterminer le meilleur codage possible. Ce choix de codage, se fait en testant de nombreuses combinaisons de seuils jusqu'à obtenir la meilleure d'entre elle. Cette procédure entraîne un problème de multiplicité, nécessitant une correction de la p_{valeur} afin de ne pas surestimer l'association entre la variable codée et la variable à expliquer.
C'est ce que propose de faire le package `CPMCGLM` dans le cadre de modèles linéaires généralisés. Les méthodes de corrections utilisées dans le package sont la procédure basée sur l'inégalité de Bonferroni, et des procédures de rééchantillonage. Ces dernières basées sur la permutation et le bootstrap paramétrique sont plus précises, puisqu'elles nous permettent de prendre en compte la corrélation qui existe entre les tests réalisés [1]. Une méthode de correction asymptotiquement exacte est également utilisée dans le cadre de codage binaire, et/ou continu [2,3].
Dans le package, les transformations de Box-Cox, les transformations binaires, et les transformations catégorielles sont disponibles. L'utilisateur peut soit rentrer les transformations qu'il veut utiliser, soit utiliser une des stratégies de codage disponibles. La fonction `CPMCGLM()` nous fournit en sortie le codage retenu, ainsi que les p_{valeur} ajustées et non ajustée correspondant à ce codage.

Références

[1] Liquet, B. and Riou, J. (2012). Correction of significance level after multiple coding in the Generalized Linear Model. [Soumis].
[2] Liquet, B. and Commenges, D. (2001). Correction of the p-value after multiple coding of an explanatory variable in logistic regression. *Statistics in Medicine*, $20 : 2815 - 2826$.
[3] Liquet, B. and Commenges, D. (2005). Computation of the p-value of the minimum of score tests in the generalized linear model, application to multiple coding. *Statistics & Probability Letters*, $71 : 33 - 38$.
[4] Yu, K., Liang, F., Ciampa, J., and Chatterjee, N. (2011). Efficient p-value evaluation for

resampling based tests. *Biostatistics*, 12(3) : 582 – 593.

clogitLasso: an R package for L^1 penalized estimation of conditional logistic regression models

M. Avalos[a,b] and **H. Pouyes**[a,c]

[a]INSERM, ISPED, Centre INSERM U897–Epidemiologie–Biostatistique,
F–33000 Bordeaux, France
[b]Univ. Bordeaux, ISPED, Centre INSERM U897–Epidemiologie–Biostatistique,
F–33000 Bordeaux, France
marta.avalos@isped.u-bordeaux2.fr

[a]INSERM, ISPED, Centre INSERM U897–Epidemiologie–Biostatistique,
F–33000 Bordeaux, France
[c]Univ. de Pau,
Pau, France
helene.pouyes@isped.u-bordeaux2.fr

Keywords: lasso, penalized conditional likelihood, matching, epidemiology.

The conditional logistic regression model is the standard tool for the analysis of epidemiological studies in which one or more cases (the event of interest), are individually matched with one or more controls (not showing the event). These situations arise, for example, in matched case–control studies and self–matched case–only studies (such as the case–crossover [1], the case–time–control [2] or the case–case–time–control [3] designs).

Usually, odds ratios are estimated by maximizing the conditional log–likelihood function and variable selection is performed by conventional manual or automatic selection procedures, such as stepwise. These techniques are, however, unsatisfactory in sparse, high-dimensional settings in which penalized methods, such as the lasso (*least absolute shrinkage and selection operator*) [4], have emerged as an alternative. In particular, the lasso and related methods have recently been adapted to conditional logistic regression [5].

The R package *clogitLasso* implements, for small to moderate sized samples (less than $3,000$ observations), the algorithms discussed in [5], based on the stratified discrete-time Cox proportional hazards model and depending on the *penalized* package [6]. For large datasets, *clogitLasso* computes the highly efficient procedures proposed in [7, 8], based on an IRLS (iteratively reweighted least squares) algorithm [9] and depending on the *lassoshooting* package [10]. The most common situations that involve 1:1, 1:M and N:M matching are available.

The talk outlines the statistical methodology behind *clogitLasso* as well as its practical application by means of three real data examples arising from Epidemiology.

References

[1] Maclure, M. (1991). The case–crossover design: a method for studying transient effects on the risk of acute event. *American journal of epidemiology* **133**, 144-153.
[2] Suissa, S. (1995).The case-time-control design. *Epidemiology* **6**, 248-53.
[3] Wang, S., Linkletter, C., Maclure, M., Dore, D., Mor, V., Buka, S., Wellenius, GA. (2011).

Future cases as present controls to adjust for exposure trend bias in case-only studies. *Epidemiology* **22**, 568-74.
[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.
[5] Avalos, M., Grandvalet, Y., Duran-Adroher, N., Orriols, L., Lagarde, E. (2012). Analysis of multiple exposures in the case-crossover design via sparse conditional likelihood. *Stat Med* **15**.
[6] Goeman, J. (2010). L^1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, **52**:70–84.
[7] Avalos, M., Pouyes, H., Grandvalet, Y., Wittkop, L., Orriols, L., Letenneur, L., Lagarde, E. (2012). High–dimensional variable selection in individually matched case-control studies. Technical Report. ISPED, Univ Bordeaux Segalen. Bordeaux, France.
[8] Avalos, M., Orriols, L., Pouyes, H., Grandvalet, Y., Lagarde, E. (2012). Variable selection in the case–crossover design via Lasso with application to a registry-based study of medicinal drugs and driving. Technical Report. ISPED, Univ Bordeaux Segalen. Bordeaux, France.
[9] Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, **46**:149–192.
[10] Jörnsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T., Nordlander, B., Sander, C., Gennemark, P., Funa, K., *et al.*. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular Systems Biology*, **7**:486.

Estimation de l'indice des valeurs extrêmes en présence de covariables

Antoine Schorgen[a]

[a]Département de Mathématiques
Université de Strasbourg et CNRS
IRMA, UMR 7501
7 Rue René Descartes
67084 Strasbourg Cedex
schorgen@unistra.fr

Mots clefs : Valeurs Extrêmes, Estimateur à noyau, Covariables.

Cet exposé sera basé sur l'article [1] écrit en collaboration avec Laurent Gardes et Armelle Guillou et portera sur l'estimation de l'indice des valeurs extrêmes en présence de covariables (aussi appelé indice de queue conditionnel).

Nous nous plaçons dans le domaine de Fréchet où les queues de distribution sont lourdes : la fonction de distribution conditionnelle de Y sachant x se modélise alors de la façon suivante :

$$\overline{F}(y,x) = y^{-1/\gamma(x)} L(y,x), \tag{1}$$

où x est fixée dans un espace métrique muni d'une distance d, et $L(\cdot,x)$ est une fonction à variations lentes.

Supposons que l'on dispose d'une suite $\{(Y_i,x_i), i=1,\cdots,n\}$ de couples indépendants provenant du modèle (1), la classe d'estimateurs proposés est basée sur un estimateur à noyau des quantiles conditionnels $\widehat{q}_n(\alpha,x) := \widehat{\overline{F}}_n^{\leftarrow}(\alpha,x)$ où

$$\widehat{\overline{F}}_n(y,x) = \sum_{i=1}^{n} H\left(\frac{d(x,x_i)}{h_{1,n}}\right) K\left(\frac{y-Y_i}{h_{2,n}}\right) \Bigg/ \sum_{i=1}^{n} H\left(\frac{d(x,x_i)}{h_{1,n}}\right),$$

avec $h_{1,n}$ et $h_{2,n}$ deux suites positives non aléatoires, $H(\cdot)$ un noyau asymétrique sur $[0,1]$ et $K(\cdot)$ un noyau intégré. Cet estimateur a été introduit par Ferraty et Vieu (2006).
Pour chaque x fixé, les observations utilisées sont celles dont les covariables se trouvent dans le voisinage de x, noté $B(x,h_{1,n})$ et désignons par m_x le nombre de ces observations.

L'estimateur proposé de notre indice peut s'écrire sous la forme suivante :

$$\check{\gamma}_\theta(x) = \int_0^{k_x/m_x} \Psi_\theta(\alpha, k_x/m_x, x) \log \widehat{q}_n(\alpha,x) d\alpha,$$

avec $k_x \in (1, m_x)$ et Ψ_θ une fonction convenablement choisie. Sous des hypothèses classiques en théorie des valeurs extrêmes, nous avons établi la normalité asymptotique de notre estimateur.

Afin d'étudier son comportement à distance finie, nous avons effectué des simulations en choisissant explicitement les noyaux H et K ainsi que la fonction Ψ_θ. Toutes les simulations ont été effectuées sous R, avec plusieurs difficultés à surmonter. Tout d'abord la présence de covariables nous incite à travailler localement : obtenir des tailles de voisinages acceptables implique une

taille d'échantillon totale conséquente ($n = 5000$). De plus, la présence de multiples facteurs à optimiser par validation croisée (taille de voisinage m_x, fenêtre pour les noyaux, nombre de valeurs extrêmes k_x) rend l'algorithme coûteux en temps. Pour réduire les temps de calcul, il a été nécessaire de programmer les éléments clés à l'aide du langage C et de découper les estimations en plusieurs blocs parallèles. La présentation abordera plus particulièrement ces aspects pratiques liés à l'optimisation de la programmation des simulations.

À titre d'exemple, nous illustrons sur la Figure 1, après optimisation des différents paramètres, le comportement de notre estimateur pour une fonction $\gamma(x)$ fixée (courbe en trait plein) et une distribution conditionnelle de type Burr pour les observations.

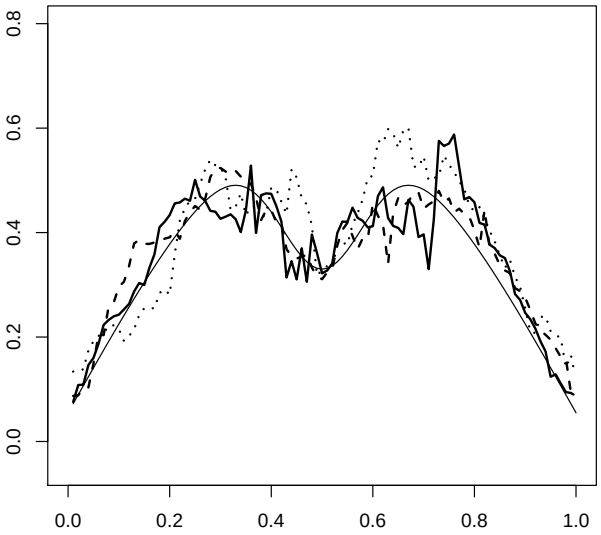


FIGURE 1 – Estimation de $\gamma(.)$: Médiane (plein), Quantiles de niveaux 10% (tirets) et 90% (pointillés) du carré de l'erreur L^2 et vraie function (—) pour 100 échantillons de taille 5000.

Références

[1] Gardes L., Guillou A., Schorgen A. (2012). Estimating the conditional tail index by integrating a kernel conditional quantile estimator, à paraître dans *Journal of Statistical Planning and Inference*.

[2] Daouia, A., Gardes, L., Girard, S., Lekina, A. (2010). Kernel estimators of extreme level curves, *Test*, **20**(2).

[3] Ferraty, F., Vieu, P. (2006). Nonparametric Functional Data Analysis : Theory and Practice, Springer Series in Statistics, Springer.

Modélisation bayésienne avec JAGS et R

M. Plummer[a]

[a]Centre International de Recherche sur le Cancer
150 Cours Albert Thomas
69372 Lyon Cedex 08
plummerm@iarc.fr

Mots clefs : BUGS, MCMC, réseau bayésien

JAGS (Just Another Gibbs Sampler) est un logiciel conçu pour l'analyse des modèles hiérarchiques bayésiens utilisant la méthode de Monte Carlo par chaînes de Markov (MCMC). Le paquetage rjags est une interface orientée objet entre R et la bibliothèque JAGS. Un objet de classe « jags.model » est construit à partir d'une description du modèle hiérarchique dans le langage BUGS sous la forme d'un réseau bayésien. Les variables aléatoires sont représentées par des nœuds ; les liens orientés définissent une factorisation de la distribution jointe de ces nœuds. Cette approche de la modélisation a été popularisée par le logiciel WinBUGS, dont JAGS est un clone multiplateforme. Elle permet la création des modèles complexes d'une façon modulaire. Une fois crée, un objet de classe « jags.model » peut générer des échantillons de la distribution a posteriori des nœuds non-observés (ou paramètres du modèle). Ces échantillons sont représentés par la classe « mcmc » du paquetage coda.

Tout comme R, les capacités de JAGS peuvent être augmentées par le chargement dynamique des modules qui mettent à disposition de l'utilisateur de nouvelles fonctions, distributions et statistiques sommaires (« monitors »). Un module peut aussi appliquer de nouvelles méthodes d'échantillonnage plus adaptées à certaines classes de modèle. Par exemple, le module « mix » permet une meilleure analyse des modèles mixtes gaussiens grâce à la méthode de Tempered Transitions [1]. Le module « glm » applique des méthodes spécifiques aux modèles linéaires généralisés [2,3].

Références

[1] Neal, R., 1996. Sampling from multimodal distributions using tempered transitions, *Statistics and Computing*, **6**, 353-355

[2] Holmes, C., Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression, *Bayesian Analysis*, **1**, 148-168

[3] Fruhwirth-Schnatter, S., Fruhwirth, R., Held, L., Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing* **19**, 479-492

Index des auteurs