

“

1ères Rencontres R

”

2-3 juil. 2012
France


Sciencesconf.org

Table des matières

Lundi 2 juillet 2012 - 09:15 - 10:00

Amphi Pitres : Conférence invitée

Pourquoi R devient incontournable en enseignement, recherche et développement, E Matzner-lober.....	1
---	---

Lundi 2 juillet 2012 - 10:05 - 11:05

Amphi Gintrac : Modèles mixtes

Optimisation de protocoles dans les modèles non linéaires à effets mixtes avec PFIM : application aux études pharmacocinétiques chez l'enfant, C Dumont [et al.]	2
Variables latentes dans les modèles linéaires généralisés, D Thiam [et al.]	4
New mixture models and algorithms in the mixtools package, D Chauveau.....	6

Amphi Pitres : Visualisation & Graphiques

PairedData 0.9 : Un package R en S4 pour analyser les données numériques appariées, S Champely	8
Une interface graphique pour analyser des données distantes sous R, R Coudret [et al.]	10
Analyse de (K+1) tableaux avec le logiciel ade4. Application en épidémiologie., S Bougeard [et al.]	12

Lundi 2 juillet 2012 - 11:30 - 12:30

Amphi Pitres : Modélisation

Capushe : package de sélection de modèle, V Brault [et al.]	14
lcm : 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 [et al.]	16
saemix, an R version of the SAEM algorithm for parameter estimation in nonlinear mixed effect models, A Lavenu [et al.]	18

Amphi Gintrac : Etude de cas

Représentation des caractéristiques du vent estimée par une méthode à noyau, N Khuc [et al.]	20
Analyse d'images et régression non-paramétrique, B Thieurmél [et al.]	22
Construction et randomisation de plans factoriels réguliers avec le package R PLANOR, H Monod.....	24

Lundi 2 juillet 2012 - 14:00 - 14:45

Amphi Pitres : Conférence invitée

Simulation and competing risks, J Beyersmann.....	25
---	----

Lundi 2 juillet 2012 - 14:50 - 15:50

Amphi Pitres : Modèles multi-états et survie

FrailtyPack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood or Parametrical Estimation, V Rondeau [et al.]	26
Analyse de données de survie en présence de censure par intervalles : le package SmoothHazard, P Joly [et al.]	28
Planification d'essais randomisés séquentiels ayant comme critère de jugement un délai de survie à l'aide de la fonction plansurvct.func, J Gal	30

Amphi Gintrac : Classification

HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data, C Bouveyron [et al.]	32
HiDimDA: An R package for Supervised Classification of High-Dimensional Data, P Duarte silva	33
Rmixmod: A MIXture MODelling R package, R Lebrete.....	35

Lundi 2 juillet 2012 - 16:20 - 17:30

Amphi Pitres : Lightning talks

Imputation de données manquantes pour des données mixtes via les méthodes factorielles grâce à missMDA, V Audigier [et al.]	37
Package 'marqLevAlg' - Algorithme de Levenberg-Marquardt en R : Une alternative à 'optimx' pour des problèmes de minimisation., M Prague [et al.]	39
Méthodologie pour le traitement des données écologiques de type inventaire avec EcoMineR, G Bessigneul [et al.]	41
SVGMMapping: an R package to map omic data sets onto pathways templates, R Champeimont [et al.]	44
Multiple Factor Analysis for Contingency Tables in FactoMineR Package, B Kostov [et al.]	46
Visualisation de données multivariées: réimplémentation des fonctionnalités graphiques de la librairie ade4, A Julien-laferriere [et al.]	48
Locally-Weighted Partial Least Squares Regression for infrared spectra analysis, A Thébault [et al.]	50
Application de modèles non paramétriques sous R pour l'analyse et le suivi de la qualité de l'eau, M Sow.....	52
Okm : une librairie R pour la classification recouvrante, G Cleuziou [et al.]	54
New tools for studying psychotherapies, T Delespierre.....	56
Test de la vraisemblance entre deux motifs de points, A Labenne [et al.]	58
The R Journal, M Plummer.....	59

Lundi 2 juillet 2012 - 20:00 - 23:00

Atrium : Posters

A study of daily mobility highlighting R workflow fluidity, H Commenges.....	60
Caractérisation d'événements à partir de signaux relatifs au comportement d'un élément combustible en situation accidentelle, L Pantera [et al.]	62
La prise en compte de l'environnement par les agriculteurs : une analyse avec le package "ClustOfVar", V Kuentz-simonet [et al.]	64
R dans un Environnement Pédagogique Virtuel (EPV) : démarche pédagogique et retour d'expérience dans une école d'ingénieurs en agriculture, A Fadil.....	65
The npde library for R to compute normalised prediction distribution errors, E Comets [et al.]	67
Utilisation du logiciel R pour l'identification de nouvelles cibles et régulateurs du protéasome, C Pellentz [et al.]	69

Mardi 3 juillet 2012 - 09:00 - 09:45

Amphi Pitres : Conférence invitée

Le logiciel R en neuro-imagerie fonctionnelle, P Lafaye de micheaux.....	71
--	----

Mardi 3 juillet 2012 - 09:50 - 11:10

Amphi Pitres : Neurosciences

Analysing eye movement data using Point Process models, S Barthelmé.....	72
A common signal detection model describes threshold and supra-threshold performance, K Knoblauch.....	75
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, R Laboissière [et al.]	77
Analyse non paramétrique de séquences de potentiels d'action. Construction de modèles et de tests de qualité d'ajustement., C Pouzat	79

Mardi 3 juillet 2012 - 09:50 - 10:30

Amphi Gintrac : Modèles de Markov cachés et modèles graphiques

DiscreteTS : two hidden-Markov models for time series of count data, J Alerini [et al.]	81
BiiPS : un logiciel pour l'inférence bayésienne dans les modèles graphiques utilisant des méthodes de Monte Carlo séquentielles, A Todeschini [et al.]	83

Mardi 3 juillet 2012 - 10:30 - 11:10

Amphi Gintrac : Analyse des données

Rotation orthogonale en ACP de données mixtes. Le package PCAmixdata et une application en sociologie culturelle., M Chavent [et al.]	85
MAINT.Data: Parametric Modelling and Analyzing Interval Data in R, A Duarte silva [et al.]	87

Mardi 3 juillet 2012 - 11:40 - 12:25

Amphi Pitres : Conférence invitée

Unravelling 'omics' data with the R package mixOmics, K Lê cao [et al.]	89
---	----

Mardi 3 juillet 2012 - 14:00 - 15:00

Amphi Pitres : Bioinformatique

Les cartes auto-organisatrices de Kohonen appliquées à l'étude des communautés de micro-algues des cours d'eau, M Bottin [et al.]	91
Comparison of network inference packages and methods for multiple network inference, N Villa-vialaneix [et al.]	93
Représentation, analyse et simulation de processus ponctuels spatio-temporels, E Gabriel.....	95

Amphi Gintrac : Biostatistique & Modélisation

Package CPMCGLM : Correction de la p-valeur engendré par la recherche d'un codage d'une variable explicative dans un modèle linéaire généralisé, J Riou [et al.]	97
clogitLasso: an R package for L1 penalized estimation of conditional logistic regression models, M Avalos [et al.]	99
Estimation de l'indice des valeurs extrêmes en présence de covariables, A Schorgen.....	101

Mardi 3 juillet 2012 - 15:05 - 15:50

Amphi Pitres : Conférence invitée

Modélisation bayésienne avec JAGS et R, M Plummer	103
---	-----

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, Fridlyand 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 from 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., Fridlyand, 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.