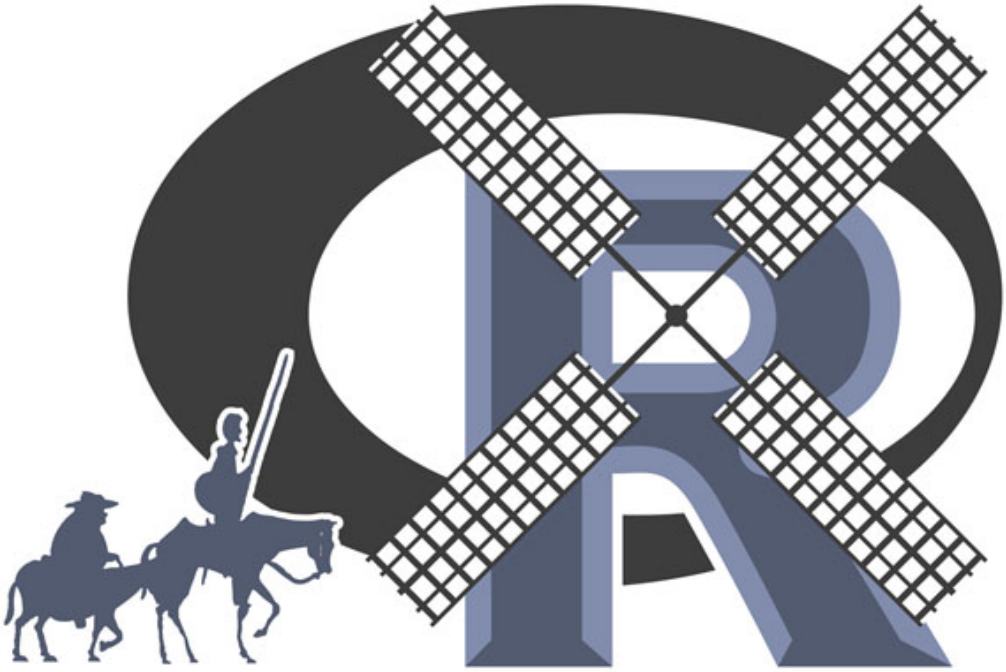
The R User Conference, useR! 2013
July 10-12 2013
University of Castilla-La Mancha, Albacete, Spain

Book of Contributed Abstracts

Compiled 2013-07-01



Contents

Wednesday 10th July 6

Bioinformatics, 10:30 **6**
- Integrating R with a Platform as a Service cloud computing platform for Bioinformatics applications. . 7
- Simulation of molecular regulatory networks with graphical models 8
- GOsummaries: an R package for showing Gene Ontology enrichment results in the context of experimental data 9
- Analysis of qPCR data in R 10

Computational Challenges in Molecular Biology I, 10:30 **11**
- The GenABEL suite for genome-wide association analyses 11
- Making enzymes with R 12
- Use of molecular markers to estimate genomic relationships and marker effects: computation strategies in R 13
- High Content Screening Analysis in R 14

Environmental statistics I, 10:30 **15**
- rClr package - low level access to .NET code from R 15
- Reproducible Research in Ecology with R: distribution of threatened mammals in Equatorial Guinea. . 16
- Using R for Mapping the Spatial Extent of Meteorological and Hydrological Drought Events 17

Statistics/Biostatistics I, 10:30 **18**
- Three-component decomposition of coal spectrum in R 18
- Method of comparison of actions of the liquidators of the accident on Chernobyl Nuclear Power Plant on the basis of fragmentation of their routes and encryption it in a form similar to the DNA 19
- Differential expression analysis of RNA-seq data at base-pair resolution in multiple biological replicates 20
- Statistical inference for Hardy-Weinberg equilibrium with missing data 21

Computational Challenges in Molecular Biology II, 12:20 **22**
- What did we learn from the IMPROVER Diagnostic Signature Challenge? 22
- Deciphering the tRNA operational code - using R 23
- Big Data and Reproducibility – Building the Bridge 24
- Topology-based Hypothesis Generation on Causal Biological Networks using igraph 25

Econometric Software, 12:20 **26**
- Hansel: A Deducer Plug-In for Econometrics 26
- Robust standard errors for panel data: a general framework 27
- Rsiopred: An R package for forecasting by exponential smoothing with model selection by a fuzzy multicriteria approach 28
- AutoSEARCH: Automated General-to-Specific Model Selection 29

Environmental statistics II, 12:20 **30**
- Driving R to the air quality industry. NanoEnvi Analyst: a tool for designing large-scale air quality plans for improvement in ambient air quality 30
- Sequential Design of Experiments for model selection: an application to the energy sector 31
- Emission inventory supported by R: dependency between calorific value and carbon content for lignite . 32

Statistics/Biostatistics II, 12:20 **33**
- Leveraging GPU libraries for efficient computation of Gaussian process models in R 33
- TriMatch: An R Package for Propensity Score Matching of Non-Binary Treatments 34
- KmL3D: K-means for Joint Trajectories 35
- Stochastic Modeling of Claim Frequency in the Ethiopian Motor Insurance Corporation: A Case Study of Hawassa Disrict 36

Database applications, 16:30 **37**
- Introducing SimpleDataManager - A simple data management workflow for R 37
- SenseLabOnline: Combining agile data base administration with strong data analysis 38
- ffbase: statistical functions for large datasets 39

Statistics/Biostatistics III, 16:30 **40**
cold: a package for Count Longitudinal Data . . . 40
kPop: An R package for the interval estimation of the mean of the selected populations. . . . 41
GLM - a case study: Antagonistic relationships between fungi and nematodes . . . 42
R Packages for Rank-based Estimates . . . 43

Time Series Analysis, 16:30 **44**
Heart Rate Variability analysis in R with RHRV . . . 44
Massively Parallel Computation of Climate Extremes Indices using R . . . 45
Segmentor3IsBack: an R package for the fast and exact segmentation of Seq-data . . . 46
hts: R tools for hierarchical time series . . . 47

Using R for Teaching I, 16:30 **48**
Teaching statistics interactively with Geogebra and R . . . 48
RKTeaching: a new R package for teaching Statistics . . . 49
genertest: a package for the developing exams in R . . . 50
Flexible generation of e-learning exams in R: Moodle quizzes, OLAT assessments, and beyond . . . 51
Teaching R in the Cloud . . . 52

Thursday 11th July 53

Machine learning I, 10:00 **53**
BayesClass: An R package for learning Bayesian network classifiers . . . 53
Constructing fuzzy rule-based systems with the R package "frbs" . . . 54
bbRVM: an R package for Ensemble Classification Approaches of Relevance Vector Machines . . . 55
Classification Using C5.0 . . . 56

Marketing/Business Analytics I, 10:00 **57**
Extending the Reach of R to the Enterprise . . . 57
Big-data, real-time R? Yes, you can. . . . 58
Large-Scale Predictive Modeling with R and Apache Hive: from Modeling to Production . . . 59
Non-Life Insurance Pricing using R . . . 60

Official statistics I, 10:00 **61**
ReGenesees: symbolic computation for calibration and variance estimation . . . 61
Big data exploration with tabplot . . . 62
rwiot: An R package for Input-Output analysis on the World Input Output Database (WIOD) . . . 63
Make Your Data Confidential with the sdcMicro and sdcMicroGUI packages . . . 64

Statistical Modelling I, 10:00 **65**
MRCV: A Package for Analyzing the Association Among Categorical Variables with Multiple Response Options . . . 65
Different tests on lmer objects (of the lme4 package): introducing the lmerTest package. . . . 66
Implementation of advanced polynomial chaos expansion in R for uncertainty quantification and sensitivity analysis . . . 67
Dhglm & frailtyHL : R package for double hierarchical genralized linear models and frailty models . . . 68

Machine learning II, 11:50 **69**
rknn: an R Package for Parallel Random KNN Classification with Variable Selection . . . 69
Patterns of Multimorbidity: Graphical Models and Statistical Learning . . . 70
ExactSampling: risk evaluation using exact resampling methods for the k Nearest Neighbor algorithm . . . 71
Classifying High-Dimensional Data with the The HiDimDA package . . . 72

Marketing/Business Analytics II, 11:50 **73**
Groupon Impact Report: Using R To Power Large-Scale Business Analytics . . . 73
Statistics with Big Data: Beyond the Hype . . . 74
Using survival analysis for marketing attribution (with a big data case study) . . . 75
Big Data Analytics - Scaling R to Enterprise Data . . . 76

Official statistics II, 11:50 **77**
Using R for exploring sampling designs at Statistics Norway . . . 77
Application of R in Crime Data Analysis . . . 78
Maps can be rubbish for visualising global data : a look at other options. . . . 79
The use of demography package for population forecasting . . . 80

Statistical Modelling II, 11:50 **81**
Shape constrained additive modelling in R . . . 81
Semiparametric bivariate probit models in R: the SemiParBIVProbit package . . . 82
"RobExtremes": Robust Extreme Value Statistics — a New Member in the RobASt-Family of R Packages 83
Generalized Bradley-Terry Modelling of Football Results . . . 84

Biostatistics: Regression Methodology, 16:30 **85**
Copula sample selection modelling using the R package SemiParSampleSel . . . 85
Robust model selection for high-dimensional data with the R package robustHD . . . 86
HGLMMM and JHGLM: Package and codes for (joint)hierarchical generalized linear models . . . 87
Fitting regression models for polytomous data in R . . . 88

Programming, 16:30 **89**
An exposé of naming conventions in R . . . 89
Statistical Machine Translation tools in R . . . 90
Reference classes: a case study with the poweRlaw package . . . 91
Combining R and Python for scientific computing . . . 92

R in companies, 16:30 **93**
Shiny: Easy web applications in R . . . 93
rapport, an R report template system . . . 94
Seamless C++ Integration with Rcpp Attributes . . . 95
The R Service Bus: New and Noteworthy . . . 96

R in the Central Banks, 16:30 **97**
Outliers in multivariate incomplete survey data . . . 97
Use of R and LaTeX for periodical statistical publications . . . 98
Solving Dynamic Macroeconomic Models with R . . . 99

Kaleidoscope I, 18:20 **100**
packdep: network abstractions of CRAN and Bioconductor . . . 100
The Beatles Genome Project: Cluster Analysis of Popular Music in R . . . 101
The secrets of inverse brogramming . . . 102

Kaleidoscope II, 18:20 **103**
Mapping Hurricane Sandy Damage in New York City . . . 103
Unlocking a national adult cardiac surgery audit registry with R . . . 104
Renjin: A new R interpreter built on the JVM . . . 105

Friday 12th July 106

GUIs/Interfaces, 10:00 **106**
Using Lazy–Evaluation to build the G.U.I. . . . 106
Survo for R - Interface for Creative Processing of Text and Numerical Data . . . 107
Using R in teaching statistics, quality improvement and intelligent decision support at Kielce University of Technology . . . 108

High performance computing, 10:00 **109**
Facilitating genetic map construction at large scales in R . . . 109
Elevating R to Supercomputers . . . 110
R in Java: Why and How? . . . 111
Rhpc: A package for High-Performance Computing . . . 112

Modelling and Optimization, 10:00 **113**
DCchoice: a package for analyzing dichotomous choice contingent valuation data . . . 113
Systems biology: modeling network dynamics in R . . . 114
Evolutionary multi-objective optimization with R . . . 115
An integrated Solver Manager: using R and Python for energy systems optimization . . . 116

Visualization/Graphics I, 10:00 **117**
Radar data acquisition, analysis and visualization using reproducible research with Sweave . . . 117
Network Visualizations of Statistical Relationships and Structural Equation Models . . . 118
tableR - An R based approach for creating table reports from surveys . . . 119
likert: An R Package for Visualizing and Analyzing Likert-Based Items . . . 120
Design of likert graphics with lattice and mosaic . . . 121

High performance computing II, 11:50 **122**
Open Source Product Creation, Bosco Team . . . 122
Practical computer experiments in R . . . 123
Symbiosis - Column Stores and R Statistics . . . 124
Memory Management in the TIBCO Enterprise Runtime for R (TERR) . . . 125

Reproducible Research, 11:50 **126**
TiddlyWikiR: an R package for dynamic report writing. . . . 126
Synthesis of Research Findings Using R . . . 127
compreGroups updated: version 2.0 . . . 128

Statistical Modelling III, 11:50 **129**
BayesVarSel. An R package for Bayesian Variable Selection. . . . 129
Bayesian learning of model parameters given matrix-valued information, using a new matrix-variate Gaussian Process. . . . 130
FluDetWeb: an interactive web-based system for the early detection of the onset of influenza epidemics 131
Looking for (and finding!) hidden additivity in complete block designs with the hiddenf package. . . . 132

Visualization/Graphics II, 11:50 **133**
A ggplot2 builder for Eclipse/StatET and Architect . . . 133
Visualizing Multivariate Contrasts . . . 134
metaplot: Flexible Specification for Forest Plots . . . 135
GaRGoyLE: A map composer using GRASS, R, GMT and Latex . . . 136

Regular Posters **137**
Asymmetric Volatility Transmission in Airline Related Companies in Stock Markets . . . 137
A R tool to teach descriptive statistics . . . 138
Using R to estimate parameters from multiple frames . . . 139
Calibration in Complex Survey using R . . . 140
R/Statistica Interface . . . 141
AMOEBA+ with R . . . 142
Software developments for non-parametric ROC regression analysis . . . 143
An R-package forWeighted Smooth . . . 144
Using R as continuous learning support in Sea Sciences degree . . . 145
Variable selection algorithm implemented in FWDselect . . . 146
Panel time series methods in R . . . 147
Teaching introductory statistics to students in economics: a comparison between R and spreadsheet . . 148
TestR: R language test driven specification . . . 149
Small area data visualization using ggplot2 library . . . 150
R as a Data Operating System for the Cloud . . . 151
TPmsm: Estimation of the Transition Probabilities in 3-State Models . . . 152
Climate Analysis Tools - An operational environment for climate products . . . 153
seq2R: Detecting DNA compositional change points . . . 154
NPRegfast: Inference methods in regression models including factor-by-curve interaction . . . 155
Pharmaceutical market analysis with R . . . 156
Standardisation on Statistics: ISO Standards and R Tools . . . 157
Quantitative Text Analysis of readers' contributions on Japanese daily newspapers . . . 158
Analysis of data from student surveys at Kielce University of Technology using R Commander and R Data Miner . . . 159

Statistical analysis with R of an effect of the air entrainment and the cement type on fresh mortar properties . . . 160
gxTools: Multiple approaches integrated in automated transcriptome analysis . . . 161
A cloud infrastructure for R reports . . . 162
On thinning spatial polygons . . . 163
Statistical analysis in R of environmental and traffic noise in Kielce . . . 164
Using R for dosimetry extremum tasks . . . 165
Data mining with Rattle . . . 166
intRegGOF: Modelling with the aid of Integrated Regression Goodness of Fit tests. . . . 167
An R script to model monthly climatic variables with GLM to be used in hydrological modelling . . . 168
Using R2wd package to automatize your reporting from R to Microsoft Word document - An application of automatic report for a survey in telecommunication . . . 169
Automation of spectroscopic data processing in routine tests of coals using R . . . 170
A Web-based Application as a Dynamical Tool for Clinical Trial Researchers . . . 171
Analysis of load capacity of pipes with CIPP liners using R Rattle package . . . 172
Efficiency analysis of companies using DEA model with R . . . 173
Introducing statistic and probability concepts with R in engineering grades . . . 174
Biomarker Discovery using Metabolite Profiling Data: Discussion of different Statistical Approaches. . . . 175
edeR: Email Data Extraction using R . . . 176
Reproducible and Standardized Statistical Analyses using R . . . 177
hwriterPlus: Extending the hwriter Package . . . 178
Application of the nearest neighbour indices in spatstat R package for Persian oak (Quercus brantii var. persica) ecological studies in Zagros woodlands, Iran . . . 179
Point process spatio-temporal product density estimation with R . . . 180
Spatio-Temporal ANOVA for replicated point patterns using R . . . 181
Estimation of parameters using several regression tools in sewage sludge by NIRS . . . 182
Recipe for the implementation of a population dynamics bayesian model for anchovy: Supercomputing using doMC , rjags and coda R packages . . . 183

Classifying High-Dimensional Data with the The HiDimDA package

A. Pedro Duarte Silva[1,2*]

1. Faculdade de Economia e Gestão / Catholic University of Portugal
2. Centro de Estudos em Economia e Gestão
*Contact author: psilva@porto.ucp.pt

Keywords: Discriminant Analysis, High Dimensionality, Variable Selection, Large Covariance Estimation.

Classical methods of supervised classification often assume the existence of a training data set with more observations than variables. However, nowadays many applications work with data bases where the total number of original features is much larger than the number of available data units. Nevertheless, in most high-dimensional classification problems the majority of the original variables do not contribute to distinguish the underlying classes, but 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 variable selection, and classification rules that work well in the few observations / many variables settings. A common strategy to achieve the later goal is to adopt rules that ignore the dependence structure of the data (*e.g,*. Fan and Fan (2008)). Recent proposals (*e.g.*, Thomaz, Kitani, and Gilies (2005), Fisher and Sun (2011), Duarte Silva (2011)) rely instead on rules based on estimators of covariance matrices with good statistical properties under such conditions.

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, as well as cross-validation and variable selection utilities.

HiDimDA can be used to construct, apply and assess k-group 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

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.

Fan, J. and Fan, Y. (2008) "High dimensional classification using Features Annealed Independence Rules." *The Annals of Statistics* **38**. 2605-2637.

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.

Thomaz, C.E; Kitani, E.C. and Gilies, D.F. (2006) "A maximum uncertainty LDA-based approach for limited sample size problems – with application to face recognition." *Journal of the Brazilian Computer Society* ***12*** *(2), 7-18.*