



The R User Conference, useR! 2011
August 16-18 2011
University of Warwick, Coventry, UK

Book of Contributed Abstracts

Compiled 2011-07-28

Contents

Tuesday 16th August	5
Kaleidoscope Ia, 09:50-10:50	5
Spectroscopic Data in R and Validation of Soft Classifiers: Classifying Cells and Tissues by Raman Spectroscopy	6
Revisiting Multi-Subject Random Effects in fMRI	7
Putting the R into Randomisation	8
Kaleidoscope Ib, 09:50-10:50	9
Bringing the power of complex data analysis methods into R	9
Using the Google Visualisation API with R	10
Experimenting with a TTY connection for R	11
Kaleidoscope Ic, 09:50-10:50	12
The R Ecosystem	12
Rc2: R collaboration in the cloud	13
RStudio: Integrated Development Environment for R	14
Portfolio Management, 11:15-12:35	15
R in the Practice of Risk Management Today	15
Stress Testing with R-Adamant	16
Handling Multiple Time Scales in Hedge Fund Index Products	17
Bioinformatics and High-throughput Data, 11:15-12:35	18
AFLP: generating objective and repeatable genetic data	18
The nz.seq package for handling genetic sequences in the Netezza Performance Server	19
Classification of Coverage Patterns	20
Finite Mixture Model Clustering of SNP Data	21
High Performance Computing, 11:15-12:35	22
GPU computing and R	22
Deploying and Benchmarking R on a Large Shared Memory System	23
The CUtil package which enables GPU computation in R	24
OBANSof: integrated software for Bayesian statistics and high performance computing with R	25
Reporting Technologies and Workflows, 11:15-12:35	26
R2wd: writing Word Documents from R	26
FRAD - Fast Results Analysis and Display	27
The Emacs Org-mode: Reproducible Research and Beyond	28
Efficient data analysis workflow in R	29
Teaching, 11:15-12:35	30
Teaching Statistics to Psychology Students using Reproducible Computing package RC and supporting Peer Review Framework	30
Teaching applied statistics to students in natural sciences using the R Commander	31
Automatic generation of exams in R	32
Modelling Systems and Networks, 16:00-17:00	33
An S4 Object structure for emulation - the approximation of complex functions	33
The structmcmc package: Structural inference of Bayesian networks using MCMC	34
Computation of generalized Nash equilibria	35
Computational Physics and Chemometrics, 16:00-17:00	36
Segmented regression in thermo-physics modeling	36
Sparse Bayesian kernel projections for classification of near-infrared spectroscopy data	37
Recovering Signals and Information From Radio Frequencies Using R (A high school student's experience)	38
Visualisation, 16:00-17:00	39
animatoR: dynamic graphics in R	39
Graphical Syntax for Structables and their Mosaic Plots	40
RMB: Visualising categorical data with Relative Multiple Barcharts	41

Official and Social Statistics, 16:00-17:00	42
ObsSensitivity: An R package for power analysis for sensitivity analyses of Observational Studies	42
Visualizing Multilevel Propensity Score Analysis	43
Dimensionality Reduction and Variable Selection, 16:00-17:00	44
ClustOfVar: an R package for the clustering of variables	44
Variable Screening and Parameter Estimation for High-Dimensional Generalized Linear Mixed Models Using L1-Penalization	45
gamboostLSS: boosting generalized additive models for location, scale and shape	46
Business Management, 16:00-17:00	47
SCperf: An inventory management package for R	47
Using R to test transaction cost measurement for supply chain relationship: A structural equation model	48
Integrating R and Excel for automatic business forecasting	49
 Wednesday 17th August	 50
Kaleidoscope IIa, 09:50-10:50	50
Using R to quantify the buildup in extent of free exploration in mice	50
Changepoint analysis with the changepoint package in R	51
Clustering patterns in streamflow to produce regionally or anthropogenically similar groups	52
Panel Discussion I, 09:50-10:50	53
R User Group Panel	53
Kaleidoscope IIb, 09:50-10:50	54
RTextTools	54
The Role of R in Lab Automation	55
Using R data functions with TIBCO Spotfire	56
Spatio-Temporal Statistics, 11:15-12:35	57
Spatio-Temporal Bayesian Modelling using R	57
Applying geospatial techniques to temporal data	58
Structured Additive Regression Models: An R Interface to BayesX	59
Molecular and Cell Biology, 11:15-12:35	60
The R package isocir for Isotonic Inference for Circular Data. Applications to Problems Encountered in Cell Biology.	60
CircNNTSR: An R Package for the Statistical Analysis of Circular Data Based on Nonnegative Trigonometric Sums	61
Summary statistics selection for ABC inference in R	62
Power and minimal sample size for multivariate analysis of microarrays	63
Mixed Effect Models, 11:15-12:35	64
Kenward-Roger modification of the F-statistic for some linear mixed models fitted with lmer	64
lqmm: Estimating Quantile Regression Models for Independent and Hierarchical Data with R	65
lcmm: an R package for estimation of latent class mixed models and joint latent class models	66
Mixed-effects Maximum Likelihood Difference Scaling	67
Programming, 11:15-12:35	68
Tricks and Traps for Young Players	68
Software design patterns in R	69
Random input testing with R	70
An Open Source Visual R Debugger in StatET	71
Data Mining Applications, 11:15-12:35	72
Predicting the offender's age	72
Leveraging Online Social Network Data and External Data Sources to Predict Personality	73
Using R to Model Click-Stream Data to Understand Users' Path To Conversion	74

Development of R, 16:00-17:00	75
Packaging R for Ubuntu: Recent Changes and Future Opportunities	75
Interpreter Internals: Unearthing Buried Treasure with CXXR	76
R's Participation in the Google Summer of Code 2011	77
Geospatial Techniques, 16:00-17:00	78
Converting a spatial network to a graph in R	78
Spatial modelling with the R-GRASS Interface	79
sos4R - Accessing SensorWeb Data from R	80
Genomics and Bioinformatics, 16:00-17:00	81
MALDIquant: Quantitative Analysis of MALDI-TOF Proteomics Data	81
QuACN: Analysis of Complex Biological Networks using R	82
Investigate clusters of co-expressed and co-located genes at a genomic scale using CoCoMap	83
Regression Modelling, 16:00-17:00	84
Beta Regression: Shaken, Stirred, Mixed, and Partitioned	84
Regression Models for Ordinal Data: Introducing R-package ordinal	85
Multiple choice models: why not the same answer? A comparison among LIMDEP, R, SAS and Stata.	86
R in the Business World, 16:00-17:00	87
Odeseus vs. Ajax: How to build an R presence in a corporate SAS environment	87
A Validation/Qualification Solution for R	88
R as a statistical tool for human factor engineering	89
Hydrology and Soil Science, 17:05-18:05	90
GWSDAT (GroundWater Spatiotemporal Data Analysis Tool)	90
IntR – Interactive GUI for R	91
Visualisation and modelling of soil data using the aqp package	92
Biostatistical Modelling, 17:05-18:05	93
survAUC: Estimators of Prediction Accuracy for Time-to-Event Data	93
Higher-order likelihood inference in meta-analysis using R	94
Gaussian copula regression using R	95
Psychometrics, 17:05-18:05	96
Multinomial Processing Tree Models in R	96
Detecting Invariance in Psychometric Models with the psychotree Package	97
Investigating multidimensional unfolding models using R2WinBUGS	98
Multivariate Data, 17:05-18:05	99
Tests for Multivariate Linear Models with the car Package	99
missMDA: a package to handle missing values in and with multivariate exploratory data analysis methods	100
MAINT.DATA: Modeling and Analysing Interval Data in R	101
Interfaces, 17:05-18:05	102
Web 2.0 for R scripts & workflows: Tiki & PluginR	102
Browser Based Applications Supported by R in Pipeline Pilot	103
A new task-based GUI for R	104
Thursday 18th August	105
Financial Models, 09:50-10:50	105
Computational aspects of continuous-time-arma (CARMA) models: The ctarma package	105
robKalman - An R package for robust Kalman filtering revisited	106
(Robust) Online Filtering in Regime Switching Models and Application to Investment Strategies for Asset Allocation	107
Ecology and Ecological Modelling, 09:50-10:50	108
Using R for the Analysis of Bird Demography on a Europe-wide Scale	108
Using OpenBUGS and lmer to study variation in plant demographic rates over several spatial and temporal scales	109
An effort to improve nonlinear modeling practice	110

Generalized Linear Models, 09:50-10:50	111
brglm: Bias reduction in generalized linear models	111
Large Scale, Massively Parallel Logistic Regression in R with the Netezza Analytics Package	112
The binomTools package: Performing model diagnostics on binomial regression models	113
Reporting Data, 09:50-10:50	114
uniPlot - A package to uniform and customize R graphics	114
sparkTable: Generating Graphical Tables for Websites and Documents with R	115
compareGroups package, updated and improved	116
Process Optimization, 09:50-10:50	117
Six Sigma Quality Using R: Tools and Training	117
Process Performance and Capability Statistics for Non-Normal Distributions in R	118
R-Package JOP: Optimization of Multiple Responses	119
Inference, 11:15-12:35	120
Density Estimation Packages in R	120
The benchden Package: Benchmark Densities for Nonparametric Density Estimation	121
Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions	122
An algorithm for the computation of the power of Monte Carlo tests with guaranteed precision	123
Population Genetics and Genetics Association Studies, 11:15-12:35	124
Simple haplotype analyses in R	124
Mixed models of large pedigrees in genetic association studies	125
Graphical tools for assessing Hardy-Weinberg equilibrium for bi-allelic genetic markers	126
Neuroscience, 11:15-12:35	127
Detecting Drug Effects in the Brain	127
Statistical Parametric Maps for Functional MRI Experiments in R: The Package fmri	128
neuRosim an R package for simulation of fMRI magnitude data with realistic noise	129
Data Management, 11:15-12:35	130
It's a Boy! An Analysis of Tens of Millions of Birth Records Using R	130
Challenges of working with a large database of routinely collected health data: Combining SQL and R	131
Demographic: Classes and Methods for Data about Populations	132
Correcting data violating linear restrictions using the deducorrect and editrules packages	133
Interactive Graphics in R, 11:15-12:35	134
iWebPlots: Introducing a new R package for the creation of interactive web-based scatter plots	134
Rocessing: Interactive Visualizations in R	135
Easy Interactive ggplots	136
RnavGraph and the tk canvas widget	137
Kaleidoscope IIIa, 14:00-15:00	138
Using R for systems understanding - a dynamic approach	138
Using multidimensional scaling with Duchon splines for reliable finite area smoothing	139
Studying galaxies in the nearby Universe, using R and ggplot2	140
Panel Discussion II, 14:00-15:00	141
Panel discussion: Challenges Bringing R into Commercial Environments	141
Kaleidoscope IIIb, 14:00-15:00	142
microbenchmark: A package to accurately benchmark R expressions	142
Vector Image Processing	143
Regular Posters	144
Late-breaking Posters	170

MAINT.DATA: Modeling and Analysing Interval Data in R

A. Pedro Duarte Silva^{1,*}, Paula Brito²

1. Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa at Porto, Porto, Portugal

2. Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto, Porto, Portugal

*Contact author: psilva@porto.ucp.pt

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

In the classical model of multivariate data analysis, data is represented in a $n \times p$ data-array where n “individuals” (usually in rows) take exactly one value for each variable (usually in columns). Symbolic Data Analysis (Diday and Noirhomme-Fraiture (2008), Noirhomme-Fraiture and Brito (2011)) provided a framework where new variable types allow to take directly into account variability and/or uncertainty associated to each single “individual”, by allowing multiple, possibly weighted, values for each variable. New variable types - interval, categorical multi-valued and modal variables - have been introduced. 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) where 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 intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix, which are represented by five different possible configurations.

In this work, we introduce the package MAINT.DATA, which implements the proposed methodologies in R. It introduces a data class for representing interval data. MAINT.DATA includes functions for modeling and analysing interval data, in particular maximum likelihood estimation and statistical tests for the different considered configurations. Methods for (M)ANOVA and Linear and Quadratic Discriminant Analysis of this data class are also provided.

References

- Azzalini, A. and Dalla Valle, A. (1996). The multivariate Skew-Normal distribution, *Biometrika* 83 (4), pp. 715–726.
- Brito, P. and Duarte Silva, A.P. (2011). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, (in press).
- Noirhomme-Fraiture, M. and Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining* (in press).
- Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester.