CRAN Task View: Chemometrics and Computational Physics

Maintainer: Katharine Mullen
Contact: katharine.mullen at stat.ucla.edu
Version: 2015-02-18

Chemometrics and computational physics are concerned with the analysis of data arising in chemistry and physics experiments, as well as the simulation of physico-chemico systems. Many of the functions in base R are useful for these ends.

The book *Chemometrics with R* by Ron Wehrens, ISBN: 978-3-642-17840-5, Springer, 2011, provides an introduction to multivariate statistics in the life sciences, as well as coverage of several specific topics from the area of chemometrics; the examples in the book are possible to reproduce using the packages ChemometricsWithR and ChemometricsWithRData.

The book *Modern Statistical Methods for Astronomy With R Applications* by Eric D. Feigelson and G. Jogesh Babu, ISBN-13: 9780521767279, Cambridge, 2012, provides an introduction to statistics for astronomers and an overview of the foremost methods being used in astrostatistical analysis, illustrated by examples in R.

The book by Kurt Varmuza and Peter Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics,* ISBN 978-1-420-05947-2, CRC Press, 2009, is associated with the package chemometrics.

A special issue of R News with a focus on R in Chemistry was published in August 2006. A special volume of Journal of Statistical Software (JSS) dedicated to oscopy and Chemometrics in R was published in January 2007.

Please let us know if we have omitted something of importance, or if a new package or function should be mentioned here.

Linear Regression Models

- Linear models can be fitted (via OLS) with `lm()` (from stats). A least squares solution for `x` in `Ax = b` can also be computed as `qr.coef(qr(A), b)`.
- The package nnls provides a means of constraining `x` to non-negative or non-positive values; the package bvls allows other bounds on `x` to be applied.
- Functions for isotonic regression are available in the package Iso, and are useful to determine the unimodal vector that is closest to a given vector `x` under least squares criteria.
- Heteroskedastic linear models can be fit using the `gls()` function of the nlme package.

Nonlinear Regression Models

- The `nls()` function (from stats) as well as the package minpack.lm allow the solution of nonlinear least squares problems.
- Correlated and/or unequal variances can be modeled using the `gnls()` function of the nlme package and by nlreg.

Curve Resolution

- The PTAk package provides functions for Principal Tensor Analysis on k modes. The package includes

also some other multiway methods: PCAn (Tucker-n) and PARAFAC/CANDECOMP.
- Multivariate curve resolution alternating least squares (MCR-ALS) is implemented in the package ALS.
- The alsace package provides MCR-ALS support for Liquid chromatography with PhotoDiode Array Detection (LC-DAD) data with many injections, with features for peak alignment and identification.
- The package drc provides functions for the analysis of one or multiple non-linear curves with focus on models for concentration-response, dose-response and time-response data.

Partial Least Squares

- The package pls implements Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR).
- The package lspls implements the least squares-partial least squares (LS-PLS) method.
- Penalized Partial Least Squares is implemented in the ppls package.
- Sparse PLS is implemented in the package spls package.
- The gpls package implements generalized partial least squares, based on the Iteratively ReWeighted Least Squares (IRWLS) method of Brian Marx.
- Package plspm contains, in addition to the usual functions for PLS regression, also functions for path modeling.

Principal Component Analysis

- Principal component analysis (PCA) is in the package stats as functions `princomp()`. Some graphical PCA representations can be found in the psy package.
- The homals package provides nonlinear PCA and, by defining sets, nonlinear canonical correlation analysis (models of the Gifi-family).
- A desired number of robust principal components can be computed with the pcaPP package. The package elasticnet is applicable to sparse PCA. The package fpca can be applied to restricted MLE for functional PCA.
- See the Multivariate task view for further packages dealing with PCA and other projection methods.

Factor Analysis

- Factor analysis (FA) is in the package stats as functions `factanal()`; see Psychometrics task view for details on extensions.

Independent Component Analysis

- Independent component analysis (ICA) can be computed using fastICA.

Clustering

- The Cluster task view provides a list of packages that can be used for clustering problems.

Variable Selection

- Stepwise variable selection for linear models, using AIC, is available in function `step()`; package leaps implements leaps-and-bounds variable selection, by default using Mallow's Cp. stepPlr provides stepwise variable selection for penalized logistic regression.
- Variable selection based on evolutionary algorithms is available in package and subselect. The latter

also provides simulated annealing and leaps-and-bounds algorithms, as well as local refinements.
- Package varSelRF provides variable selection methods for random forests. Cross-validation-based variable selection using Wilcoxon rank sum tests is available in package WilcoxCV, focused on binary classification in microarrays. Package clustvarsel implements variable selection for model-based clustering.
- The BioMark package implements two meta-methods for variable selection: stability selection (applying a primary selection method like a t-test, VIP value or PLSDA regression coefficient) to different subsets of the data, and higher criticism, which provides a data-driven choice of significance cutoffs in statistical testing.

Self-Organizing Maps

- The kohonen package implements self-organizing maps as well as some extensions for supervised pattern recognition and data fusion. The som package provides functions for self-organizing maps.

Differential Equations

- See the DifferentialEquations task view packages dealing with differential equations.

Calibration

- The investr package facilitates calibration/inverse estimation with linear and nonlinear regression models.
- The chemCal package provides functions for plotting linear calibration functions and estimating standard errors for measurements.
- The quantchem package provides functions for statistical evaluation of calibration curves by different regression techniques.
- The drm package and the nlreg package are useful for nonlinear calibration models.
- The package represent calculates the 'representativity' of two multidimensional datasets, which involves comparison of the similarity of principal component analysis loading patterns, variance-covariance matrix structures, and data set centroid locations.

Cellular Automata

- The simecol package includes functions for cellular automata modeling. One-dimensional cellular automata are also possible to model with the package CellularAutomaton.

Thermodynamics

- The CHNOSZ package provides functions for calculating the standard Gibbs energies and other thermodynamic properties, and chemical affinities of reactions between species contained in a thermodynamic database.

Interfaces to External Libraries

- The package rcdk allows the user to access functionality in the Chemistry Development Kit (CDK), a Java framework for cheminformatics. This allows the user to load molecules, evaluate fingerprints, calculate molecular descriptors and so on. In addition, the CDK API allows the user to view structures in 2D. The rcdklibs package provides the CDK libraries for use in R.
- The rpubchem package gives access to PubChem data (compounds, substance, assays).

- ChemmineR is a cheminformatics toolkit for analyzing small molecules in R. Its add-on packages include fmcsR for mismatch tolerant maximum common substructure matching, eiR for accelerated structure similarity searching; bioassayR for analyzing bioactivity data, and ChemmineOB for accessing OpenBabel functionalities from R.

Spectroscopy

- The hyperSpec packages allows analysis of hyperspectral data, i.e., spectra plus further information such as spatial information, time, concentrations, etc. Such data are frequently encountered in the analysis of Raman, IR, NIR, UV/VIS, NMR, etc., spectroscopic data sets.
- The ChemoSpec package collects user-friendly functions for plotting spectra (NMR, IR, etc) and carrying top-down exploratory data analysis, such as HCA, PCA and model-based clustering.
- The GitHub package: HyperChemoBridge interconverts ChemoSpec and hyperSpec objects
- The Peaks package implements functions for spectrum manipulation, ported from the ROOT/TSpectrum class.
- The speaq package implements the hierarchical Cluster-based Peak Alignment (CluPA) and may be used for aligning NMR spectra.
- Software for the book by Donald B. Percival and Andrew T. Walden, *Spectral Analysis for Physical Applications,* ISBN 978-0-521-43541-3, Cambridge University Press, 1993, is found in the package sapa.
- The package TIMP provides a problem solving environment for fitting separable nonlinear models in physics and chemistry applications, and has been extensively applied to time-resolved spectroscopy data.
- The package prospectr provides functions for pretreatment and sample selection of visible and near infrared diffuse reflectance spectra.
- The resemble includes functions for spectral dissimilarity analysis and memory-based learning (a.k.a. local modeling) for non-linear modeling in spectral datasets.

Mass Spectrometry

- The MSnbase defines infrastructure for mass spectrometry-based proteomics data handling, plotting, processing and quantification.
- The MALDIquant provides tools for quantitative analysis of MALDI-TOF mass spectrometry data, with support for baseline correction, peak detection and plotting of mass spectra.
- The OrgMassSpecR package is for organic/biological mass spectrometry, with a focus on graphical display, quantification using stable isotope dilution, and protein hydrogen/deuterium exchange experiments.
- The FTICRMS package provides functions for Analyzing Fourier Transform-Ion Cyclotron Resonance Mass Spectrometry Data.
- The titan provides a GUI to analyze mass spectrometric data on the relative abundance of two substances from a titration series.
- The Bioconductor packages MassSpecWavelet, PROcess, and xcms are designed for the analysis of mass spectrometry data.
- The apLCMS package is designed for the processing of LC/MS based metabolomics data.
- The xMSanalyzer package allows merging apLCMS sample processing results from multiple sets of parameter settings, among other features.
- The MSPrep package is for post-processing of metabolomic data, including summarization of replicates, filtering, imputation, and normalization.

- The metaMS package is an MS-based metabolomics data processing and compound annotation pipeline.

Functional Magnetic Resonance Imaging

- Functions for I/O, visualization and analysis of functional Magnetic Resonance Imaging (fMRI) datasets stored in the ANALYZE or NIFTI format are available in the package AnalyzeFMRI. The package fmri contains functions to analyze fMRI data using adaptive smoothing procedures.

Fluorescence Lifetime Imaging Microscopy

- Functions for visualization and analysis of Fluorescence Lifetime Imaging Microscopy (FLIM) datasets are available in the package TIMP.

Carbon Dating

- The package Bchron creates chronologies based on radiocarbon and non-radiocarbon dated depths.

X-Ray Diffractograms

- The diffractometry package provides baseline identification and peak decomposition for x-ray diffractograms.

Astronomy and astrophysics

- The astrodatR package collects 19 datasets from contemporary astronomy research, many of which are described in the aforementioned textbook 'Modern Statistical Methods for Astronomy with R Applications'.
- The astrolibR package presents an R interface to low-level utilities and codes from the Interactive Data Language (IDL) Astronomy Users Library .
- The CosmoPhotoz package performs photometric redshift estimation using generalized linear models.
- tThe CRAC collects R functions for cosmological research, with its main functions being similar to the python library, cosmolopy.
- The RobPer package calculates periodograms based on (robustly) fitting periodic functions to light curves.
- The package snapshot contains functions for reading and writing N-body snapshots from the GADGET code for cosmological N-body/SPH simulations.
- The package UPMASK performs unsupervised photometric membership assignment in stellar clusters using, e.g., photometry and spatial positions.
- The moonsun package provides functions for basic astronomical calculations.
- The solaR package provides functions to determine the movement of the sun from the earth and to determine incident solar radiation.
- The FITSio package provides utilities to read and write files in the FITS (Flexible Image Transport System) format, a standard format in astronomy.
- The ringscale package implements a method for detection of faint companions around young stars in speckle patterns of VLT/NACO cube mode images.
- The stellaR package manages and displays stellar tracks and isochrones from the Pisa low-mass database.
- The astroFns provides miscellaneous astronomy functions, utilities, and data.
- The cosmoFns contains standard expressions for distances, times, luminosities, and other quantities

useful in observational cosmology, including molecular line observations.
- The astro package provides tools for astronomy; functions provided may be grouped into 4 main areas: cosmology, FITS file manipulation, the Sersic function and general (plotting and scripting) tools.
- The celestial package includes a number of common astronomy conversion routines, particularly the HMS and degrees schemes.
- The SCEPtER package is used to estimate stellar mass and radius given observational data of effective temperature, [Fe/H], and astroseismic parameters.
- The Astrostatistics and Astroinformatics Portal Software Forum is an R-centric collection of information and discussion regarding software for statistical analysis in astronomy.

Optics and Scattering Approximations

- The planar package provides code to simulate reflection and transmission at a multilayer planar interface.
- The cda package calculates the polarizability tensor for the dipoles associated with a set of ellipsoidal nanoparticles, and solves the coupled-dipole equations by direct inversion of the interaction matrix.
- The dielectric package defines some physical constants and dielectric functions commonly used in optics and plasmonics.

Energy Modeling

- The solaR package provides functions to simulate and model systems involved in the capture and use of solar energy, including photovoltaics.

Positron Emission Tomography

- The PET package implements different analytic/direct and iterative reconstruction methods for positron emission tomography (PET) data.

Water and Soil Chemistry

- The AquaEnv package is a toolbox for aquatic chemical modelling focused on (ocean) acidification and CO2 air-water exchange.
- See the Environmetrics task view for further related packages related to water and soil chemistry.

`CRAN packages:`

- ALS (core)
- AnalyzeFMRI
- AquaEnv
- astro
- astrodatR
- astroFns
- astrolibR
- Bchron
- BioMark
- bvls
- cda
- celestial
- CellularAutomaton

- chemCal (core)
- chemometrics
- ChemometricsWithR
- ChemometricsWithRData
- ChemoSpec
- CHNOSZ
- clustvarsel
- cosmoFns
- CosmoPhotoz
- CRAC
- dielectric
- diffractometry
- drc
- drm
- elasticnet
- fastICA
- FITSio
- fmri
- fpca
- FTICRMS
- homals
- hyperSpec
- investr
- Iso (core)
- kohonen (core)
- leaps
- lspls
- MALDIquant
- minpack.lm
- moonsun
- nlme
- nlreg
- nnls (core)
- OrgMassSpecR
- pcaPP
- Peaks
- PET
- planar
- pls (core)
- plspm
- ppls
- prospectr
- psy
- PTAk (core)
- quantchem
- rcdk
- rcdklibs
- represent

- resemble
- ringscale
- RobPer
- rpubchem
- sapa
- SCEPtER
- simecol
- snapshot
- solaR
- som
- speaq
- spls
- stellaR
- stepPlr
- subselect
- TIMP
- titan
- UPMASK
- varSelRF
- WilcoxCV

Related links:

- CRAN Task View: DifferentialEquations
- CRAN Task View: Psychometrics
- CRAN Task View: Cluster
- CRAN Task View: Multivariate
- CRAN Task View: Environmetrics
- R News: R in Chemistry
- Journal of Statistical Software: Spectroscopy and Chemometrics in R
- Bioconductor Package: alsace
- Bioconductor Package: bioassayR
- Bioconductor Package: ChemmineOB
- Bioconductor Package: ChemmineR
- Bioconductor Package: fmcsR
- Bioconductor Package: gpls
- Bioconductor Package: eiR
- Bioconductor Package: MassSpecWavelet
- Bioconductor Package: metaMS
- Bioconductor Package: MSnbase
- Bioconductor Package: PROcess
- Bioconductor Package: xcms
- apLCMS
- Interactive Data Language (IDL) Astronomy Users Library
- Astrostatistics and Astroinformatics Portal Software Forum
- Chemistry Development Kit (CDK)
- MSPrep
- OpenBabel

- PubChem
- xMSanalyzer
- GitHub package: HyperChemoBridge