Chemometric analysis of spectroscopic data in R: hyperSpec

C. Beleites and V. Sergo

CENMAT, Materials and Natural Resources Dept.,
University of Trieste, Via Valerio 6/a, Trieste, I-34127, Italy

While several software solutions for the chemometric analysis of spectral data of biomedical samples exist, none could meet all our key requirements. We present a new software package, *hyperSpec*, that greatly facilitates the analysis of hyperspectral data in the statistical environment R [1]. *hyperSpec* makes R a convenient platform for the analysis of spectroscopic data sets, including spectral images and maps. It takes care of data import and export, supplies the means to plot the data (spectra, false-color maps, time series, depth profiles, calibration curves) and can be used for handling and processing spectra. The actual statistical calculations are performed by functions supplied by R: chemometric methods such as regression, classification and cluster analysis are readily available in R, as well as the means for validation, determination of confidence intervals, etc..

In *hyperSpec*, the spectra can be stored with arbitrary amounts of meta-information such as position, sample numbers, constituent concentrations, diagnoses, etc. Also, spectral maps/images need neither be rectangular nor evenly spaced, and may be combined with spectra without spatial information.

Specialized and customized methods are needed for the chemometric analysis of biomedical spectroscopic data sets (e.g. validation schemes that take into account varying numbers of spectra per patient and diagnosis, or robust statistical methods). *hyperSpec* can easily be extended or used together with other R packages.

Programming statistical data analysis methods is an error-prone task, and special care is needed to ensure computational correctness. We therefore chose a software environment that has a well-tested basis of statistical routines [2]. A standardized interface for different data types (classes) and statistical methods greatly facilitates the flexible interaction between specialized data (like *hyperSpec* data sets) and specialized data analysis methods in R. Also interaction with Matlab is easy using *R.matlab* [3]. Besides reading and writing .mat files, computations can be executed in Matlab. Also, packages for the parallelization of R code exist.

hyperSpec allows scripting of the data analysis so that computationally intensive calculations can be run as batch jobs but there are also functions for basic user interaction on the plots. We think that graphical user interfaces (GUIs) should be tailored to specific tasks to allow efficient working. GUIs can be built using *hyperSpec*'s facilities to handle spectra together with other R packages that supply GUI elements, particular data analysis methods, etc.

We present two examples of how *hyperSpec* is used: a cluster analysis of a Raman map of chondrocytes in cartilage, and a linear calibration of fluorescence emission of quinine. *hyperSpec* is hosted at http://hyperspec.r-forge.r-project.org.

References:

[1] R Development Core Team: "R, A language and environment for statistical computing", Vienna, R Foundation for Statistical Computing, ISBN 3-900051-07-0, URL http://www.R-project.org. (2009).

[2] K. B. Keeling, R. J. Pavur, *Computational Statistics & Data Analysis* 51, 3811 – 3831 (2007).

[3] H. Bengtsson: "R.matlab - Local and remote Matlab connectivity in R", Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Sweden, URL http://www.maths.lth.se/help/R/R.matlab/ (2005).