ChemoSpec: an R Package for the Chemometric Analysis of Spectroscopic Data

Bryan A. Hanson

Dept. of Chemistry & Biochemistry
hanson@depauw.edu
www.github.com/bryanhanson/ChemoSpec

DePauw University, Greencastle Indiana USA

Abstract

ChemoSpec is a collection of functions for plotting spectra (such as NMR & IR spectra) and carrying out various forms of top-down exploratory data analysis, including hierarchical cluster analysis (HCA), principal components analysis (PCA) and model-based clustering (via mclust). 53 classes are used and the data is stored in a Spectra object created during data import. Two-dimensional and several-dimensional methods are provided for visualizing score plots, including interactive plots and graphical MANOVA methods. Robust methods appropriate for this type of high-dimensional data are employed and diagnostics plots are available. These functions rely heavily on methods and functions discussed in Varmuza & Filzmoser, as well as ideas from the ggobi group. ChemoSpec is designed to facilitate comparison of samples from treatment and control groups such as typically found in ecological or medical investigations. It is designed to be user friendly and suitable for people with limited background in R, as it was written to support undergraduate research projects. ChemoSpec functions will be demonstrated using data from the author’s research on plant stress using metabolomics.

Motivation

ChemoSpec was written to support research in plant metabolomics. This project investigates the effect of stress (think climate change) on plants using chemical and ecological methods; much of the data is multivariate. In particular, we have or anticipate NMR, MS and IR measurements on plant extracts as a way of fingerprinting them. Methods were needed for exploratory data analysis on multivariate data of plants of different genotypes grown in different environments, such as low water or high salinity.

Package Features

• User friendly: many internal checks, error feedback, accommodates color-blind users.
• Consistent interface, plot formats, labeling etc.
• Collation of spectral files into one (as a Spectra object).
• Basic plotting for comparison/inspection of spectra singly or by group membership.
• Binning of data sets; removal of uninteresting regions; removal of problematic samples.
• Normalization & scaling of spectra.
• HCA of spectra or of PCA scores.
• Classical or robust PCA; plot scores by group membership or not; confidence ellipses by group; 2D and 3D static & interactive plots; diagnostics, score plot, loading plots.
• Model-based clustering (via mclust).
• You might say ChemoSpec is one big wrapper!

Demonstration Data Set

The data set used in the examples here is composed of infrared (IR) spectra of the leaf surface of the common “weed” purslane (Portulaca oleracea). A variety collected in South Carolina was grown in low and high water conditions as part of a larger study on climate change and plant stress.

The Spectra Object: An S3 class list of 9

<table>
<thead>
<tr>
<th>type</th>
<th>name</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr</td>
<td>freq</td>
<td>frequency values</td>
</tr>
<tr>
<td>chr</td>
<td>freq</td>
<td>wavenumber range (no. samples x freq values)</td>
</tr>
<tr>
<td>chr</td>
<td>names</td>
<td>sample names</td>
</tr>
<tr>
<td>chr</td>
<td>group membership</td>
<td>groups for plotting sample data</td>
</tr>
<tr>
<td>chr</td>
<td>sym</td>
<td>symbol code for plots</td>
</tr>
<tr>
<td>chr</td>
<td>sym</td>
<td>symbol to be used in lieu of symbols</td>
</tr>
<tr>
<td>chr</td>
<td>titles for frequencies &amp; intensities</td>
<td></td>
</tr>
<tr>
<td>chr</td>
<td>desc</td>
<td>description of data set</td>
</tr>
</tbody>
</table>

Graphic Manova of Cuticle IR Spectra

References


Acknowledgements

The development of ChemoSpec was supported by the DePauw University Faculty Development Committee in the form of a sabatical leave and a Fisher Fellowship. Thanks also of course to the R community for answers to many questions along the way.