News

The Newsletter of the R Project

Volume 6/3, August 2006

Editorial

by Ron Wehrens and Paul Murrell

Welcome to the third issue of R News for 2006, our second special issue of the year, this time with a focus on uses of R in Chemistry. Thanks go to our guest editor, Ron Wehrens, for doing all of the heavy lifting to get this issue together. Ron describes below the delights that await in the following pages. Happy reading!

Paul Murrell
The University of Auckland, New Zealand
paul.murrell@R-project.org

R has become the standard for statistical analysis in biology and bioinformatics, but it is also gaining popularity in other fields of natural sciences. This special issue of R News focuses on the use of R in chemistry. Although a large number of the Bioconductor packages can be said to relate to chemical subdisciplines such as biochemistry or analytical chemistry, we have deliberately focused on those applications with a less obvious bioinformatics component.

Rather than providing a comprehensive overview, the issue gives a flavour of the diversity of applications. The first two papers focus on fitting equations that are derived from chemical knowledge, in this case with nonlinear regression. Peter Watkins and Bill Venables show an example from chromatography, where the retention behaviour of carboxylic acids is modelled. Their paper features some nice examples of how to initialize the optimization. In the next paper, Johannes Ranke describes the **drfit** package for fitting dose-response curves.

The following three papers consider applications that are more analytical in nature, and feature spectral data of various kinds. First, Bjørn-Helge Mevik discusses the **pls** package, which implements PCR and several variants of PLS. He illustrates the package with an example using near-infrared data, which is appropriate, since this form of spectroscopy would not be used today but for the existance of multivariate calibration techniques. Then, Chris Fraley and Adrian Raftery describe several chemical applications of the **mclust** package for model-based clustering. The forms of spectroscopy here yield images rather than spectra; the examples focus on segmenting microarray images and dynamic magnetic resonance images. Ron Wehrens and Egon Willighagen continue with a paper describing self-organising maps for large databases of crystal structures, as implemented in the package **wccsom**. To compare the spectral-like descriptors of crystal packing, a specially devised similarity measure has to be used.

The issue concludes with a contribution by Rajarshi Guha on the connections between R and the Chemistry Development Kit (CDK), another open-source project that is rapidly gaining widespread popularity. With CDK, it is easy to generate descriptors of molecular structure, which can then be used in R for modelling and predicting properties. The paper includes a description of the **rcdk** package, where

Contents of this issue:

the key point is the connection between R and Java (which underlies CDK).

Ron Wehrens

Institute for Molecules and Materials
Analytical Chemistry
The Netherlands
R.Wehrens@science.ru.nl

Non-linear regression for optimising the separation of carboxylic acids

by Peter Watkins and Bill Venables

In analytical chemistry, models are developed to describe a relationship between a response and one or more stimulus variables. The most frequently used model is the linear one where the relationship is linear in the parameters that are to be estimated. This is generally applied to instrumental analysis where the instrument response, as part of a calibration process, is related to a series of solutions of known concentration. Estimation of the linear parameters is relatively simple and is routinely applied in instrumental analysis. Not all relationships though are linear with respect to the parameters. One example of this is the Arrhenius equation which relates the effect of temperature on reaction rates:

$$k = A \exp\left(-E_a/RT\right) \times \exp(\epsilon) \qquad (1)$$

where k is the rate coefficient, A is a constant, E_a is the activation energy, R is the universal gas constant, and T is the temperature in degrees Kelvin. As k is the measured response at temperature T, A and E_a are the parameters to be estimated. The last factor indicates that there is an error term, which we assume is multiplicative on the response. In this article we will assume that the error term is normal and homoscedastic, that is, $\epsilon \sim \mathrm{N}(0, \sigma^2)$

One way to find estimates for A and E_a is to transform the Arrhenius equation by taking logarithms of both sides. This converts the relationship from a multiplicative one to a linear one with homogeneous, additive errors. In this form linear regression may be used to estimate the coefficients in the usual way.

Note that if the original error structure is not multiplicative, however, and the appropriate model is, for example, as in the equation

$$k = A \exp\left(-E_a/RT\right) + \epsilon \qquad (2)$$

then taking logarithms of both sides does not lead to a linear relationship. While it may be useful to ignore this as a first step, the optimum estimates can only be obtained using non-linear regression techniques, that is by least squares on the original scale and not in the logarithmic scale. Starting from initial values for the unknown parameters, the estimates are iteratively refined until, it is hoped, the process converges to the maximum likelihood estimates.

This article is intended to show some of the powerful general facilities available in R for non-linear regression, illustrating the ideas with simple, yet important non-linear models typical of those in use in Chemometrics. The particular example on which we focus is one for the response behaviour of a carboxylic acid using reverse-phase high performance liquid chromatography and we use it to optimise the separation of a mixture of acids.

Non-linear regression in general is a very unstructured class of problems as the response function of the regression may literally be any function at all of the parameters and the stimulus variables. In specific applications, however, certain classes of non-linear regression models are typically of frequent occurrence. The general facilities in R allow the user to build up a knowledge base in the software itself that allows the fitting algorithm to find estimates for initial values and to find derivatives of the response function with respect to the unknown parameters automatically. This greatly simplifies the model fitting process for such classes of models and usually makes the process much more stable and reliable.

The working example

Aromatic carboxylic acids are an important class of compounds since many are pharmacologically and biologically significant. Thus it is useful to be able to separate, characterise and quantify these types of compounds. One way to do this is with chromatography, more specifically, reverse phase high performance liquid chromatography (RP-HPLC). Due to the ionic nature of these compounds, analysis by HPLC can be complicated as the hydrogen ion concentration is an important factor for separation of these compounds. Waksmundzka-Hajnos (1998) reported a widely used equation that models the separation of monoprotic carboxylic acids (i.e. containing a single hydrogen ion) depending on the hydrogen ion concentration, $[H^+]$. This is given by

$$k = \frac{(k_{-1} + k_0([H^+]/K_a))}{(1 + [H^+]/K_a)} + \epsilon \qquad (3)$$

where k is the capacity factor, k_0 and k_{-1} are the k values for the non-ionised and ionised forms of the