Journal of Statistical Software

January 2007, Volume 18, Issue 9. http://www.jstatsoft.org/

ImpuR: A Collection of Diagnostic Tools Developed in R in the Context of Peak Impurity Detection in HPLC-DAD but Potentially Useful with Other Types of Time–Intensity Matrices

Christian Ritter
Université Catholique de Louvain

Jean Gilliard
GSK Biologicals

Abstract

HPLC-DAD systems generate time intensity (absorbance) matrices called spectro-chromatograms. Under good experimental conditions, spectro-chromatograms of elution peaks of pure analytes are bilinear products of a time peak and an absorbance spectrum. Co-eluting impurities create deviations from this pure bilinear structure. Unfortunately, other imperfections, such as scan averaging, large optical windows, imperfect lamp alignment, mobile phase fluctuations, etc. also create departures from the pure bilinear structure. This makes it hard to distinguish low concentration impurities from artifacts and hampers safe detection of contaminants. There are two main ways to deal with such artifacts: removal and simulation, and **ImpuR** provides R functions to do both and to integrate both approaches.

More specifically, **ImpuR** provides a set of tools to explore time-intensity matrices with respect to their bilinear structure and departures from it. It includes exploratory graphs for bilinear matrices (bilinear residual graphs and singular value decompositions), spectral dissimilarity curves via window-evolving factor analysis with heteroscedasticity correction and the sine method, methods for removal of artifacts, and a comprehensive simulation tool to assess the impact of potential artifacts and to allow for the construction of guide curves for use with the sine method.

Keywords: HPLC-DAD, window-evolving factor analysis, spectral dissimilarity traces, spectroscopic artifacts, peak purity.

1. Introduction

Historically, the routines contained in **ImpuR** were originally written in 1994 and 1995 as

part of a research effort to assess and improve the detection limit for purity control by HPLC-DAD. The main elements are data pre-treatment , purity analysis via multiple methods, an enhanced graphical display, peak simulation `Simul.DAD`, and spectral dissimilarity analysis.

This article has two currents: A discussion of a practical problem in analytical chemistry, purity control by HPLC, and the introduction of a set of statistical package to analyze bilinear structures which was originally developed to data arising in the chemical problem but which can also be used in other contexts.

Practically, the paper starts with the point of view of "purity control". We explain the main question and we show what data arising in this context look like ideally and in practice and what our software can do with them. The discussion centers on two integrated displays, one focusing on bilinear decomposition, the other on spectral similarity.

Then, we change the point of view and talk about the implementation in the R language (R Development Core Team 2006) of a collection of tools to deal with bilinear data matrices. We introduce a suitable data class, we describe some principal exploration and analysis tools, and we explain the implementation of these tools in the context of "purity control". However, the purpose of this explanation is less aimed at "purity control" but more at giving hints on how to adapt the package to other contexts future users may be confronted with.

In the context of the special JSS issue on spectroscopy, the loose collection of tools was ported to R and re-formated for easier use. A part of these tools is generic, such as the definition of the data class and elementary functions operating on objects of this class. These tools are bundled in the source file `ImpuR.R`. Another part is specific to the treatment HPLC-DAD data and has to be adapted for use in different contexts. This part is contained in the source file `HPLCDAD.R`.

Users who wish to learn using the tools should examine the source files of this article which are contained in the sub-folder Rnw. The source files are written in Sweave format. They contain the code used to create the data sets and the graphical displays used in this paper. The paper itself can be re-generated by running the source file `Driver.R` in the top folder.

1.1. HPLC-DAD for purity control

HPLC-DAD stands for "high performance liquid chromatography with diode array detector". In this technique, the substance to analyze is dissolved in a solvent blend and injected in a solvent stream. The solvent stream, called the mobile phase, then passes through a column filled with a gel, called the stationary phase. The gel lets the solvent pass but retains the analyte. If the analyte consists of a blend of several types of molecules, the retention time of each type of molecule is characteristic for it. Eventually, the analyte also passes through the column, ideally separated in its constituents. When the analyte comes out of the other end of the column (it elutes), it can be detected as a change in UV absorbance. This is done automatically in what is called the flow cell. Traditionally, the entire UV absorbance, or the UV absorbance at a specific wavelength is registered leading to absorbance peaks and the curve of absorbance versus time, called the chromatogram, shows a succession of peaks corresponding to the different eluting constituents of the analyte.

Instead of registering overall absorbance or absorbance at a specific wavelength, modern instruments record an entire UV spectrum. For this, the UV light, commonly emitted by a deuterium lamp, passes through the flow cell and is then split via a grating and projected onto a diode array. The diode array is electronically scanned and the photon count is reg-

istered. The result is an intensity spectrum (transmittance). Comparison with a spectrum without analyte allows conversion to "absorbance" scale. This means, that an HPLC-DAD system splits an analyte into its constituents and produces for each chromatographic peak a data matrix of UV absorbance spectra over time. Under ideal conditions, each chromatographic peak corresponds to exactly one constituent. In practice, however, it can happen that two constituents elute almost at the same time, forming a composite peak (co-elution). Nevertheless, it is rare that these constituents co-elute perfectly and that they have almost identical spectra. Therefore, the data matrices corresponding to co-eluting constituents will most of the time show traces of this fact and techniques for analyzing "peak purity" in HPLC-DAD are looking for such traces.

The particular challenge in "peak purity control" is that even very small impurities should be detected. In the certification of production processes for pharmaceutical products, one aims at maximum impurity levels below one percent and ideally below 0.1 percent. Detecting such small concentrations is at the technical limit of this technique. The work, this paper is based on was undertaken about 10 years ago to better understand, quantify and extend this technical limit.

A good overview of methods for peak purity control implemented in current HPLC-DAD systems can be found by web search on the key words `Peak Purity` and `HPLC`.

2. Ideal case in HPCL-DAD

The time intensity (absorbance) matrices created by HPLC-DAD systems are called spectro-chromatograms. Under good experimental conditions, spectro-chromatograms of elution peaks of pure analytes are bilinear products of a time peak and an absorbance spectrum (Beer Lambert Law). In terms of data analysis of such matrices, this implies that they have one significant principal component, the remainder is noise.

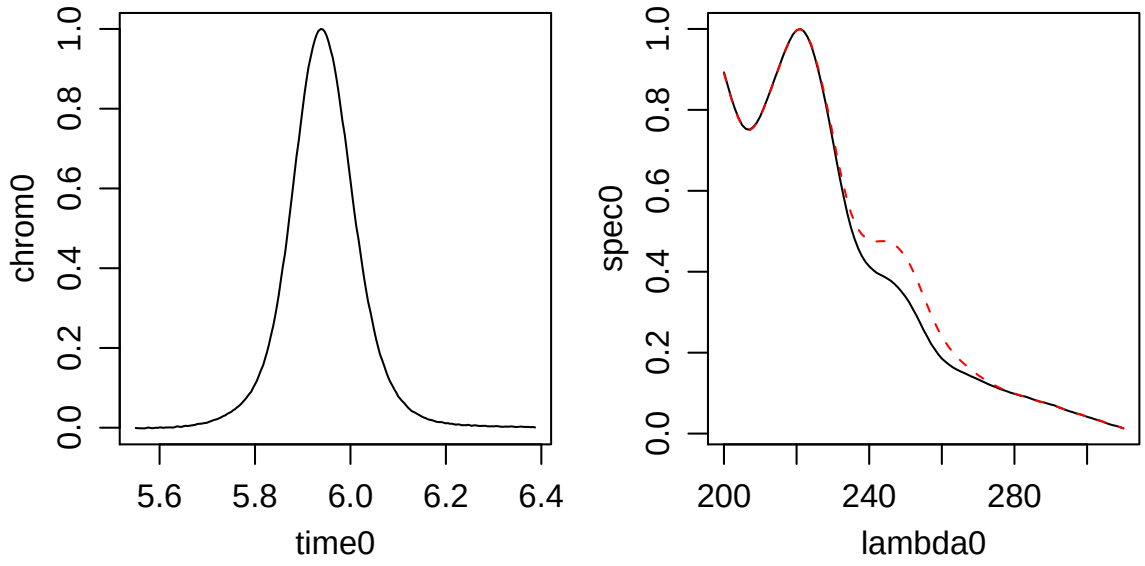

Figure 1: Alprazolam: Chromatogram and spectra of pure substance and a simulated impurity

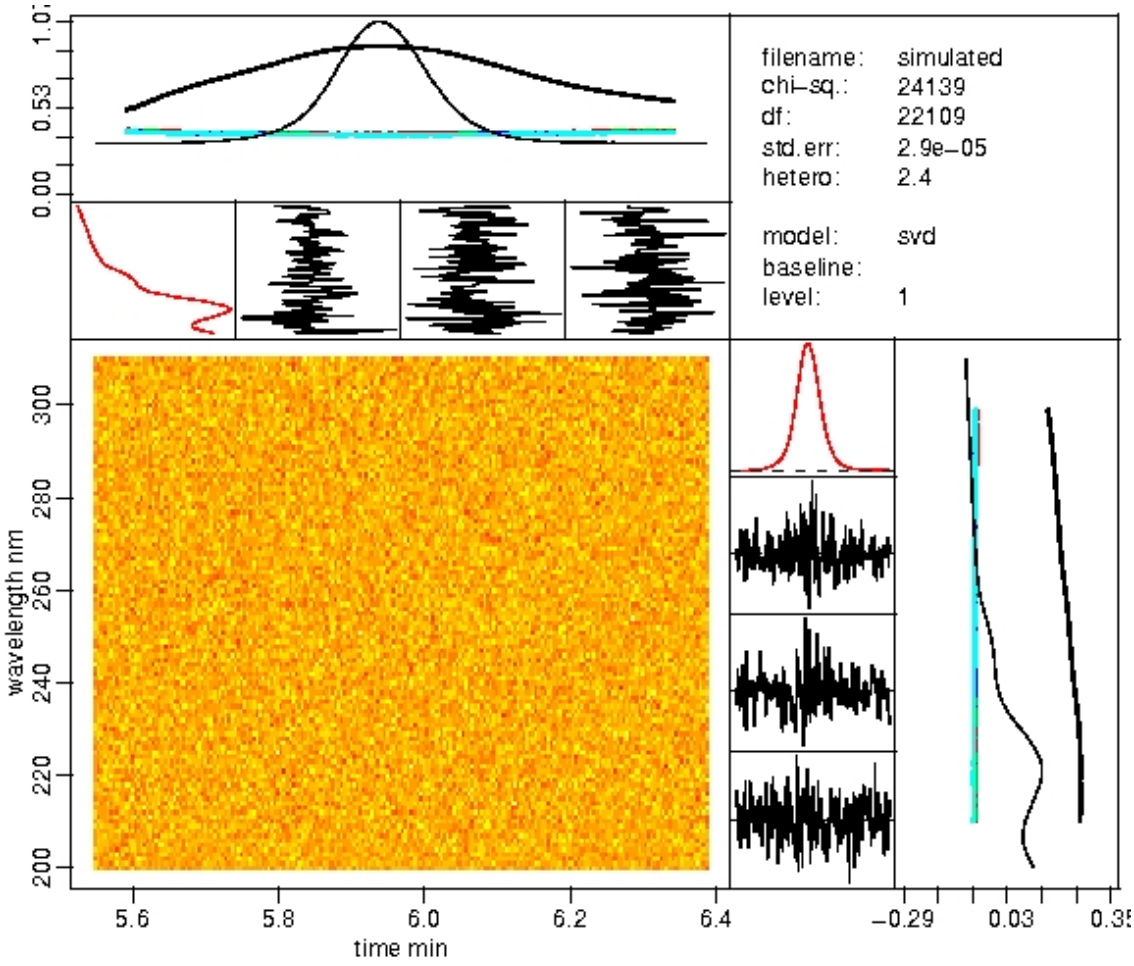

Figure 2: Ideal case without impurity

The noise comes at first from the diode array. The diode array registers photon count and these counts are Poisson distributed. Commonly, the counts are quite high and baseline noise levels (no analyte) are typically on the order of 0.00001 to 0.00005 units of absorbance (one unit of absorbance is a ten-fold reduction in intensity). Units of absorbance are denoted by UA. On the Beckman Gold system we used in 1995, the baseline noise level was about 0.00003UA. At the center of a chromatographic peak, intensity (the photon count) is reduced, leading to a higher relative error. Common practice suggests to use dilutions of the analyte which lead to 0.1 UA for the peak of interest. This implies a signal to noise ratio of close to 2000-10000. The noise is hetero-scedastic and one can estimate for a system whose baseline noise level (expressed as a standard deviation) is about 0.00003UA, the noise level increase to a bit more than 0.0001UA under a peak of 1UA. This makes for a hetero-scedasticity slope of about 3. Other noise sources exist such as flow instabilities through the column and the detector, mechanical vibration affecting the optical system and noise in the electronic amplification system. Some can be assimilated with the baseline noise, others require special attention.

As an example, Figure 1 shows a typical chromatogram, the spectrum of alprazolam, and the

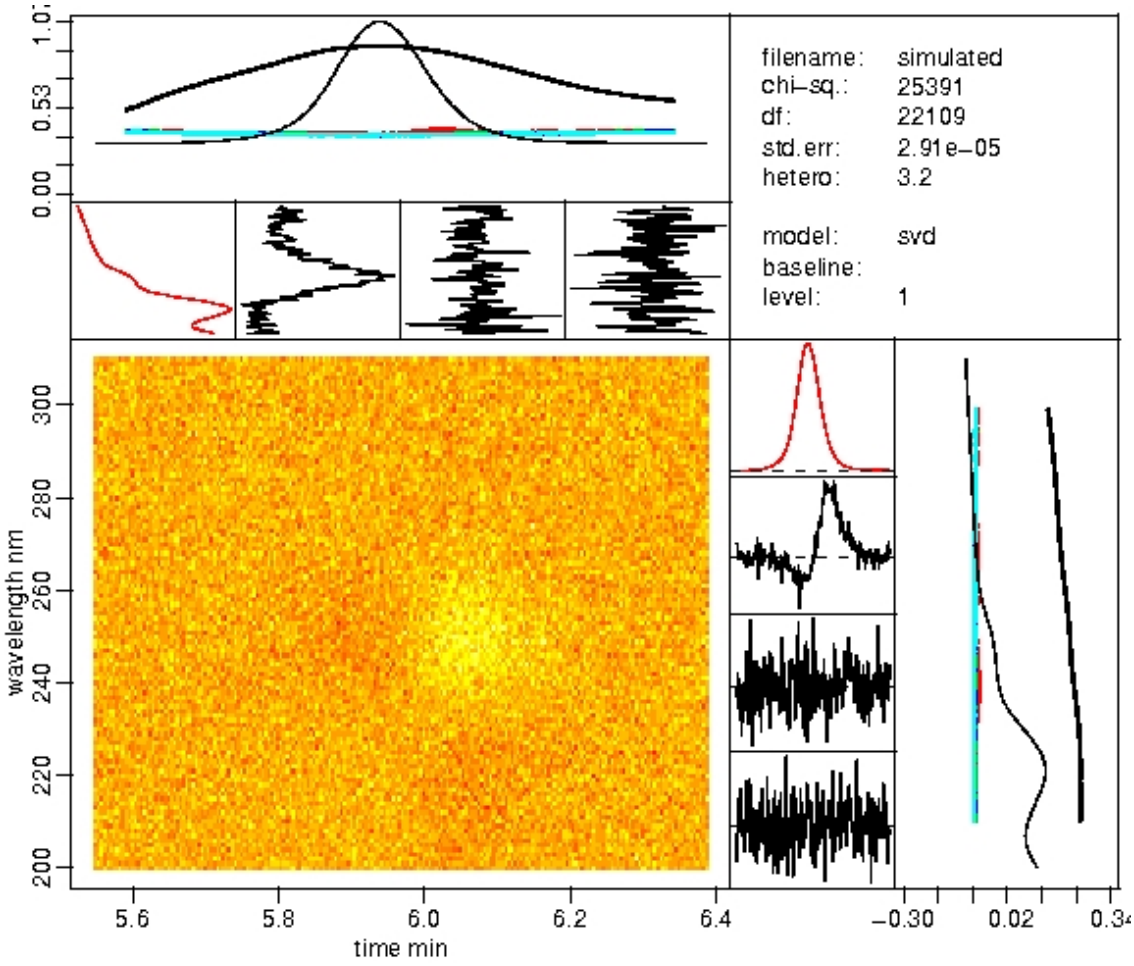

Figure 3: Ideal case with impurity

spectrum of an impurity.

Figure 2 shows an overview panel for peak purity analysis. It consists of several parts: A residual image, traces of bilinear components, traces of a "window evolving factor analysis" (Wefa), and summary information. The residual image represents the residual pattern of the bilinear model formed by the first k bilinear components indicated by "level". They are also identified as "red" traces.

The bilinear components can be obtained by different approaches. The simplest is singular value decomposition (indicated as "svd"), but other techniques such as successive "peeling" can be used as well. Window evolving factor analysis stands for sliding window principal components. For a perfect one-component bilinear matrix, Wefa yields a single important value, all others represent noise.

For the simple example of a pure simulated peak, the peak purity display for a level 1 bilinear model shows random residuals. There is one important pair of singular vectors, the other three pairs are noise. The window evolving factor analysis traces for times and wavelengths confirm that there is a single component.

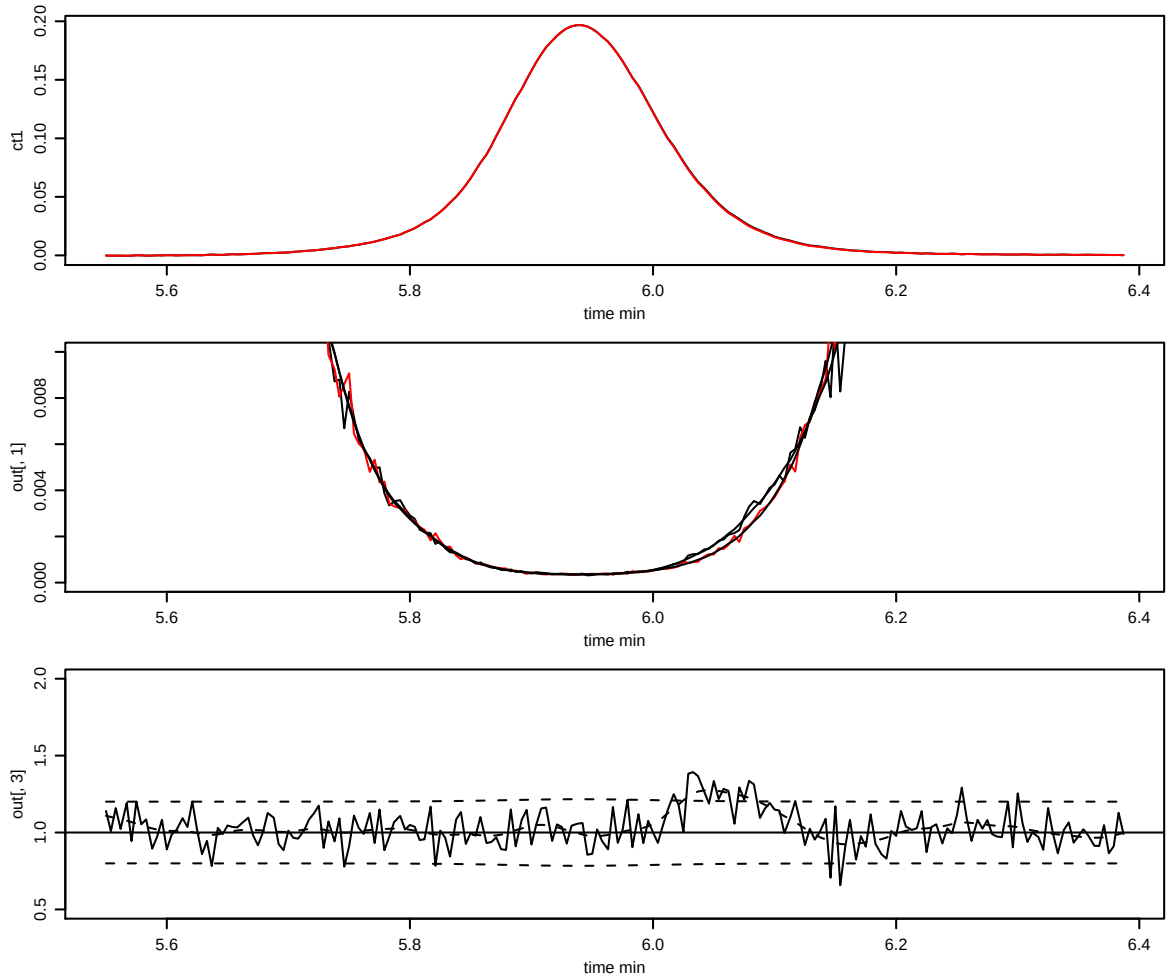

Figure 4: Spectral dissimilarity curves for a simulated pure peak of alprazolam and of a simulated impure peak containing 0.4% of an impurity separated at 0.8FWHM.

Co-eluting impurities (impurities which pass through the column in almost the same time, typically metabolites or isomers) create peaks (in time) which are imperfectly separated from the peak of the main substance. If their concentrations are high, this shows up as shoulder patterns in the chromatogram. Since these impurities are of different chemical structure, their UV spectra are also different from the UV spectrum of the main substance. This implies that the shoulder structure depends on the wavelength. By virtue of Beer's law, one can presume that a spectro-chromatogram corresponding to the combined absorption of the main substance and a co-eluting impurity consists of the sum of two single bilinear structures. In terms of data analysis, there are therefore two significant principal components plus (hetero-scedastic) noise.

Figure 3 shows the analysis of the same simulated spectro-chromatogram as above but with 0.4% of an impurity separated from the main component by 0.8FWHM (full width at half maximum). The spectrum of the impurity has been shown in Figure 1.

We see that the impurity causes characteristic patterns in all diagnostics. It is clearly visible in the residual image and in the singular vectors. It is barely visible in the Wefa traces. This

could suggest that in general the full singular vectors have higher detection power than the Wefa traces, however, the singular vectors are also more vulnerable to long-time distortions unrelated to impurities such as drifts in the baseline. Since Wefa works on a smaller time window, it is less affected by this.

The display suggests that 0.4 percent is about the lowest level of this type of impurity which can be detected at 0.8 FWHM under ideal conditions. The statistical significance of the second principal component for a (still ideal) combination of a main peak and a co-eluting impurity peak obviously depends on the "chromatographic resolution" (the time distance between the peaks measured in units of "full width at half maximum", for example, but other definitions exist taking asymmetry and varying peak width into account), the "spectroscopic dissimilarity" (the high dimensional "angle" between the main spectrum and the impurity spectrum), the concentration of the impurity, and the noise level. Perfectly co-eluting impurities (zero resolution) cannot be detected without an independently acquired target spectrum of the main component. The same holds for co-eluting impurities with very similar spectra as the main component.

Peak purity can also be assessed using spectral dissimilarity curves. One way to measure the difference between two normalized spectra is to calculate the (high-dimensional) angle between them, or equivalently, when the spectra are similar, the sine. This can be used to examine a spectro-chromatogram. For this, we extract a spectrum which would be representative if the chromatographic peak were pure. In this case, the best spectrum we can obtain is the one at the summit of the peak, that is the apex spectrum. In order to smooth out noise, it is even better to use the average over several time points around the apex, but this is a technical detail.

If the peak is "pure", all spectra should be the same (except for noise). If we know something about the noise, we can even calculate what the "theoretical" sine between the extracted representative spectrum and each spectrum of the spectro-chromatogram should be. In general, since the signal to noise ratio diminishes as a function of the chromatographic peak height, the spectral dissimilarity sine between the extracted spectrum and the spectrum for each time point will describe a bath-tub shaped curve. In the center (under the peak), this curve will be close to zero. Toward the extremes, it will fluctuate much more and approach one. An approximate value of zero corresponds to spectra which are practically identical to the reference spectrum, and an approximate value of one implies that they are completely unrelated.

When an impurity is present to the right (in time) of the apex of the main peak, the apex spectrum will still be representative of the main component, but the spectra to the right of the apex will be modified. This means that, compared to the spectral dissimilarity curve of a pure peak, the curve of the impure peak will lift off to the right of the apex. Comparison of the shape of the pure curve and an actual curve can therefore reveal the presence of an impurity.

Figure 4 shows the spectral dissimilarity curve for the simulate pure and impure peaks introduced above. We see how the black curve corresponding to the impure peak deviates from the red curve. This also translates into a peak in the ratio of the curves which can be interpreted as an impurity signal. This shows that also the spectral dissimilarity curves can be used to obtain information on peak purity.

3. Practical limits to detecting co-eluting impurities

At which concentrations can impurities be detected and what can be expected from peak purity analysis? 500mg is a typical dose for the active ingredient of a drug. If one supposes that concentrations of 1mg and less can be considered as harmless, detection of impurities down to about 0.1% are considered as sufficient. Going back to HPLC-DAD, the validity of Beer's law can be shown up to approximately 0.1 UA. For higher absorbance levels, the response of the system begins to show non-linearity and therefore departures from the ideal structure even if no impurity is present.

For an impurity with similar absorbance as the main component (as can be expected from a metabolite), this means that 0.1% of concentration will lead to a peak of 1/1000th of the size of the main peak and thus of 0.0001 UA absorbance. This is just two to three times higher than the baseline noise. If, in addition, chromatographic resolution is low (that is the impurity peak elutes almost simultaneously with the main peak) the relative signal introduced by the impurity is very small compared to the main substance and no detection is possible.

The simulation study of the previous section showed this practical limitation of HPLC-DAD for purity analysis: The difference between the spectrum of the main component and the impurity corresponded to a "Gaussian" peak of about 10% of the total height of the spectrum of the impurity which in turn corresponded to 0.4% of the main component. This implies that the "data impurity" was only about 0.04% of the main spectrum. We saw that detection under perfect conditions was still possible, but only barely.

At higher chromatographic resolution and spectral dissimilarity, detection improves but detecting signals related to impurities with similar spectra below 1% of the main substance remains difficult and require perfect observation conditions and data treatment.

On the other hand, at concentrations above 1% impurities whose spectra are clearly different from the spectrum of the main substance and which are well separated are readily detectable. The challengeing range is therefore between 0.1% and 1% at imperfect separation. The methods discussed in this paper are aimed at this application range. They were originally published inRitter, Gilliard, Cumps, and Tilquin (1995), Gilliard and Ritter (1997a), and Gilliard and Ritter (1997b). The first of these articles deals with a post treatment of window evolving factor analysis to remove the non-informative signature of heteroscedasticity from the singular value traces, the second article deals with the identification of artifacts, and the third with the realistic simulation of spectrochromatograms. In a general sense, these methods allow to visualize and characterize very small departures from otherwise perfectly bilinear data matrices.

```
R> al1 <- Read.DAD("al1.txt")
```

4. Real HPLC-DAD data

Figure 5 shows a bilinear diagnostics of an observed spectrochromatogram of alprazolam at about 0.2UA maximum absorbance. We see that the result is not what we might expect. There are clear deviations from a perfect bilinear structure although the sample was pure. How can we characterize these deviations and how could we make abstraction of them in order to spot real impurities if there were present?

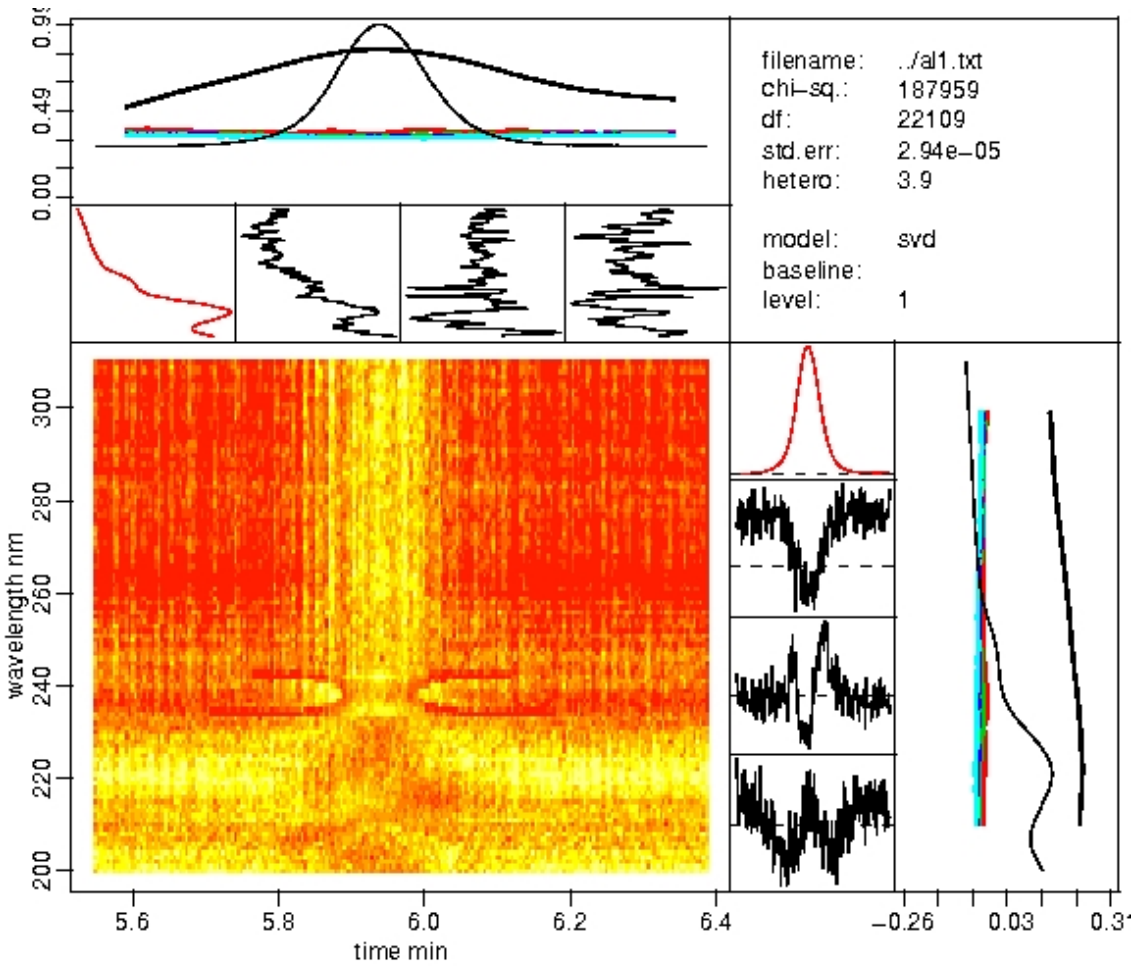

Figure 5: Observed spectro-chromatogram of a pure substance (alprazolam).

The first "feature" of the diagnostic graph is the appearance of an "X" shaped pattern in the residual graph. It only occurs under specific observation conditions and corresponds to a "defect" in the data acquisition process, probably a saturation of the diode array (bit loss). The zone in which the line pattern appears is a zone in which the deuterium lamp has high intensity. Therefore, without the analyte, "bit loss" occurs leading to a lower than "expected" intensity. When the analyte elutes, absorption reduces the intensity and no more "bit loss" occurs. That is, the absorption is under-evaluated compared. This is what we see: red stripes where red stands for residuals below zero.

The absolute intensity of the pattern is small compared to the rest, but it is clearly above the noise level. It is therefore as strong as signals which would be produced by a small impurity but it does not have the same "pattern". Now, the pattern can only be seen on the residual image, not on the summary traces. This is why the inspection of the residual pattern is important in practice to avoid false positives. A pattern like the "X" has typical repercussions on the trace diagnostics: It will create singular vectors which show patterns looking like second derivatives of the main peak. In terms of the Wefa traces, there will be small signals of an additional component at the beginning and the end of the peak.

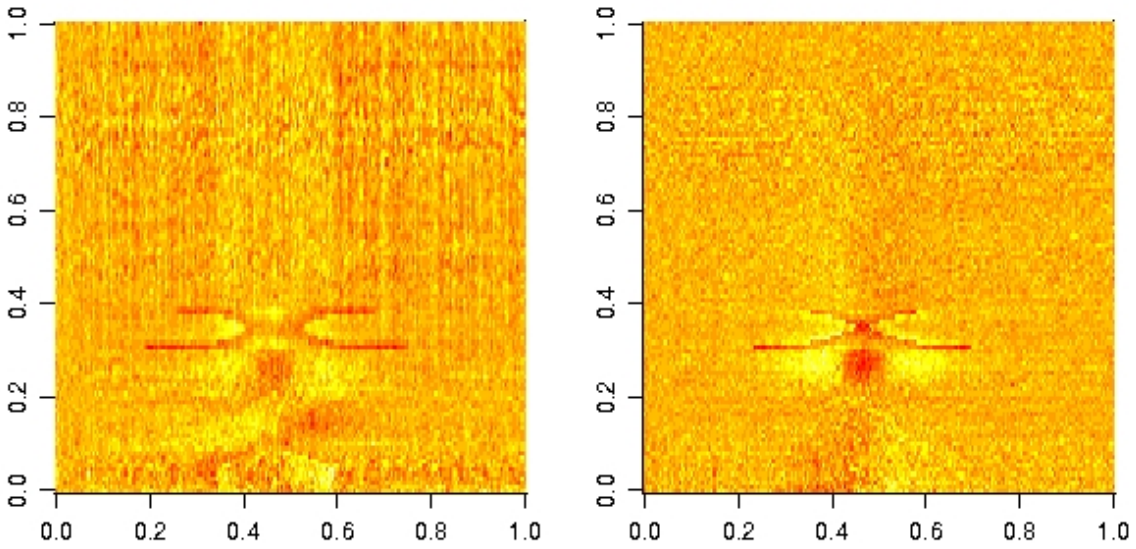

Figure 6: Observed spectrochromatogram of a pure substance (alprazolam) treated for baseline fluctuations in time and wavelength and simulated version

The second feature we see is a vertical stripe pattern. This corresponds to fluctuations on a time scale of seconds and can be associated with flow instabilities of the mobile phase. It will introduce patterns in the singular vectors which show an image of the main peak and the main spectrum disturbed by noise. The Wefa trace for time will at the same time show a higher baseline noise (the first component starts higher even before the real peak starts). Moreover, there is indication that the intensity level was not equal for all wavelength before the peak started. This corresponds to baseline fluctuations in wavelength direction.

If we look more carefully, we can still identify two further patterns: An asymmetry around the chromatographic peak (with respect to time), and succession of elevations and depressions under the peak but in the wavelength dimension. Also the sources of these patterns are known: They are related to the scan time and to the optical window. The scan time is the time it takes to "read" the diodes. The shorter the scan time, the lower the intensity and thus the higher the relative noise of one scan. Several scans can be combined to regain precision, however, the intensity of electronic noise accumulates. Therefore, the scan time is usually set to give a good compromise. A finite scan time implies that diodes corresponding to higher wavelengths are "read" at different times than diodes corresponding to lower wavelengths. If the scan time is not negligible with respect to the speed with which the peak elutes, this introduces a time-asymmetric disturbance of the bilinear matrix structure.

Finally, due to optical constraints, each diode does not only "see" the light of a single wavelength but instead an "average" over what is called the optical window. In typical HPLC-DAD instruments data resolution is 1nm but the optical window is often more than 4nm wide. What the diode sees is therefore a smoothed version of the true intensity spectrum. This is then converted to absorbance (by taking logarithms). Since the logarithm of an average is not equal to the average of logarithms, this introduces a nonlinear deviation.

Both, the effects of scan time and optical window are small. However, in practice they are

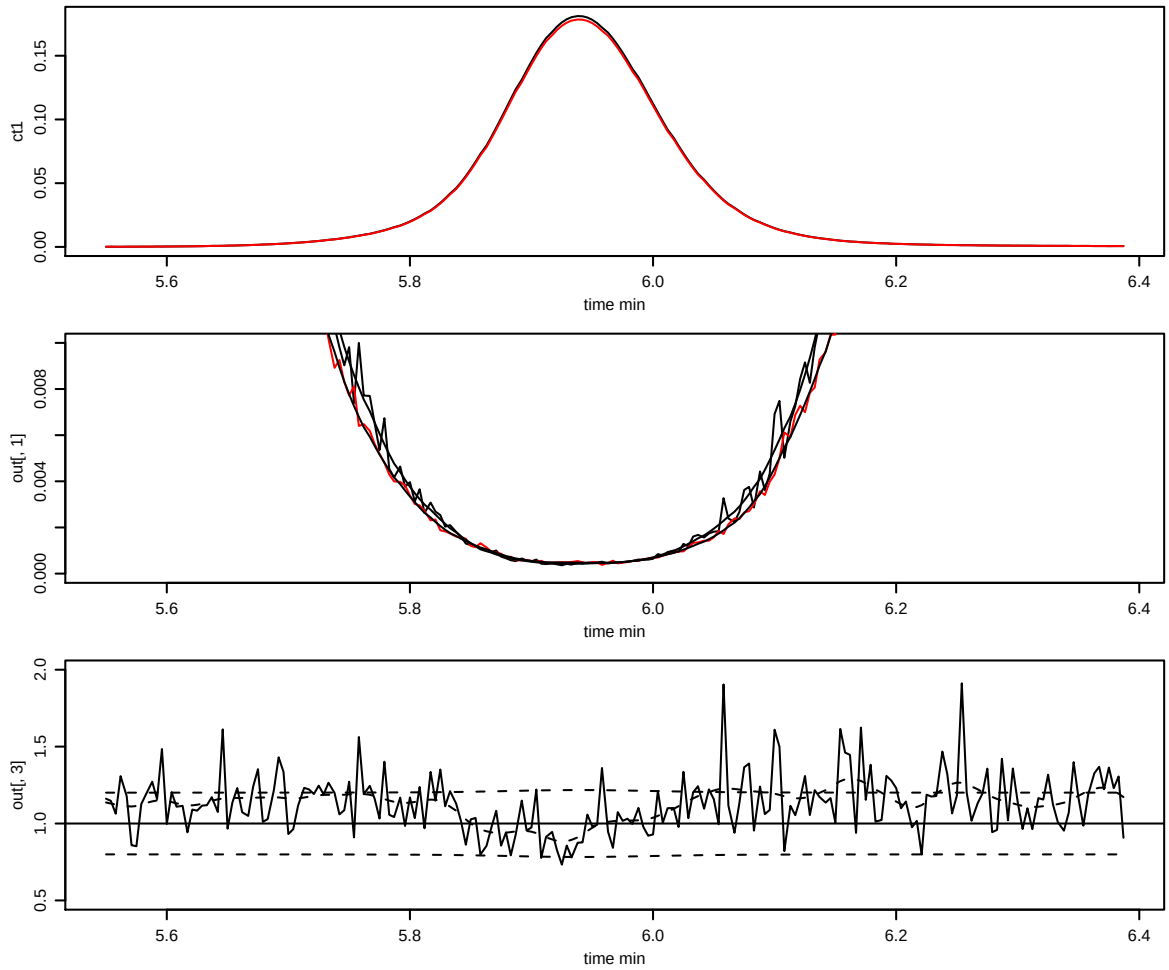

Figure 7: Spectral dissimilarity plot comparing the cleaned real data with the corresponding simulated version.

higher than the base noise level and can be in strength of similar size as small impurities. It is therefore important that they are taken into account if departures from ideal bilinear behavior are interpreted.

There are two main approaches for dealing with this: Removal and simulation. Removal means that, if the processes which cause the artifacts are known sufficiently well, algorithms can be designed which pre-process the data matrix to remove them. After that, the cleaned data matrix can be inspected. In simulation, one tries to extract the spectrum and the chromatogram from the spectrochromatogram as if it were "pure". Then one can build up an artificial data matrix according to the model one has in mind, that is, including all the instrumental artifacts which are known to act for the measurement instrument. This yields a second "synthetic" spectrochromatogram which resembles the true data in the main spectrum and chromatogram and in the main identified artifacts. The difference is that it is guaranteed to contain only one "true" component. Analysis is now based on comparing the true data matrix and its simulated counterpart. This can be done by studying the peak purity displays in parallel. It can also be done by spectral comparison.

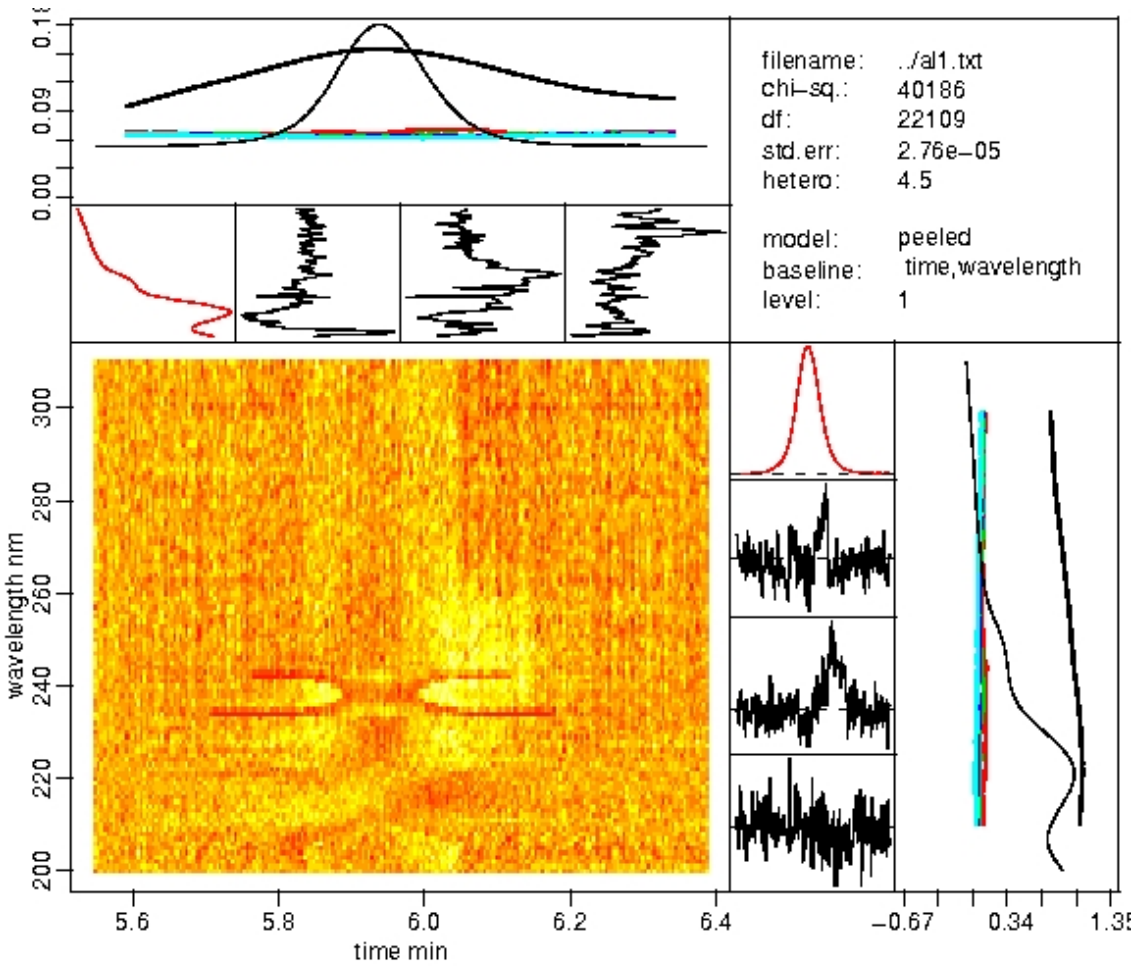

Figure 8: Real data after addition of simulated 0.5% impurity.

We found (Gilliard and Ritter (1997a), and Gilliard and Ritter (1997b)) that both approaches should be combined to achieve optimum performance. The effects of fluctuating time and wavelength baselines should be removed by pre-treatment and the effects due to scan time, optical window, and others such as the "X" should be simulated. Figure 6 shows the residual image of the cleaned true data matrix and of a simulated version making appropriate assumptions about scan time, optical window, and bit-loss. Although some differences remain, the similarity is quite striking and suggests that the patterns in the true data are more likely caused by "artifacts" than by true impurities.

The corrected true data and the simulated counterparts can now be compared using spectral dissimilarity traces. This is shown in Figure 7. We see that the spectral dissimilarity trace of the simulated data (red curve) is very similar to the trace of the true data. We conclude again that the data do not suggest the presence of an impurity.

What happens if a true impurity is present in data like ours? This is simulated in Figure 8. Here, an impurity of 0.5% with the spectrum introduced earlier was added at 1FWHM to the right of the main peak (still co-eluting with the main compound). The peak purity display shows a weak signal. The associated spectral dissimilarity display in Figure 4 reproduces

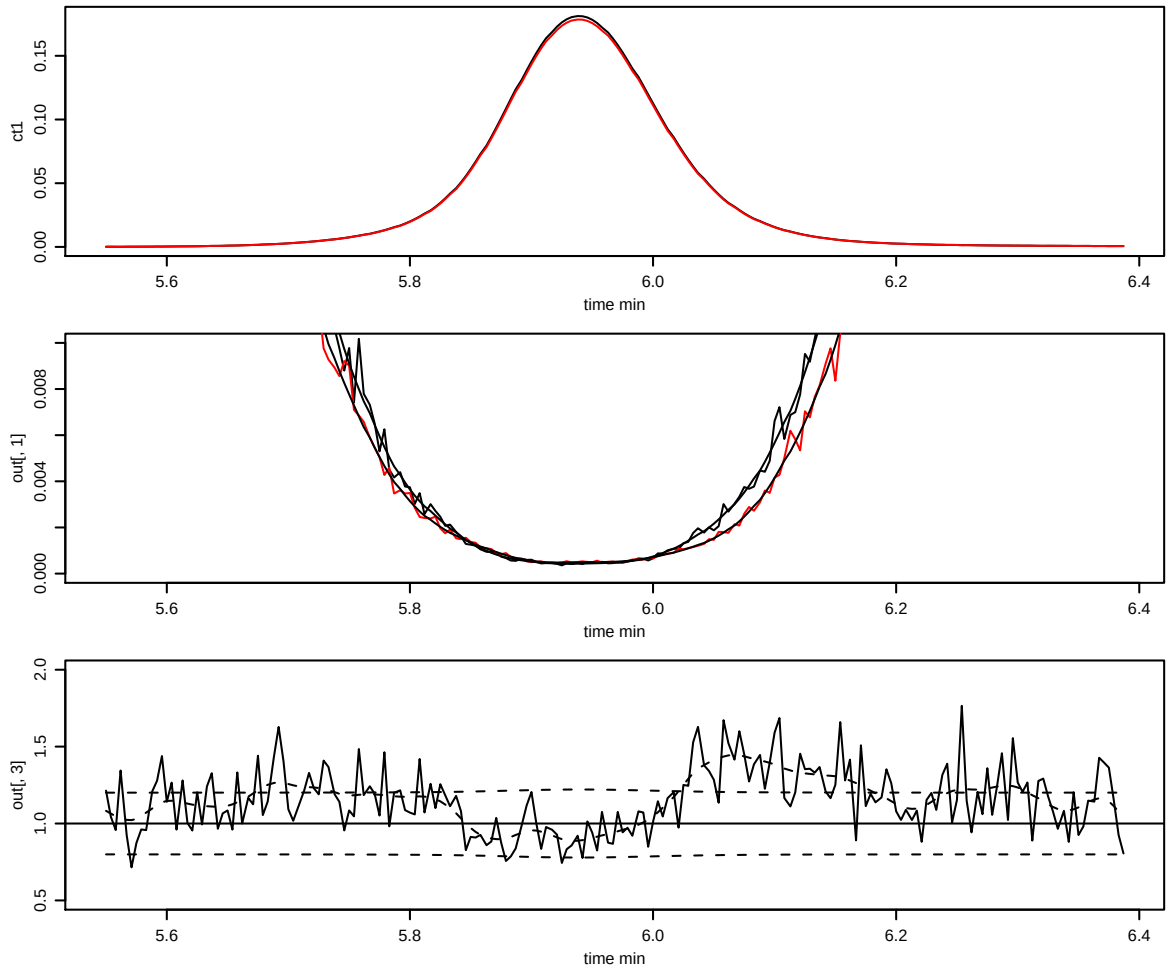

Figure 9: Spectral dissimilarity display comparing real data after addition of 0.5% simulated impurity with realistically simulated pure data matrix.

this weak signal which would be hard to identify without the guide curve obtained from the simulated pure spectrochromatogram. This shows that weak impurities are hard to detect and identify using a single technique, but that the situation improves when several techniques are used simultaneously.

5. Philosophy and structure of the package

The easiest way to learn to use the package is to examine the files in the `Rnw` subdirectory. They are the source of the paper in Sweave form. The code contained in them can be run to create the graphs shown in the paper.

Technically, the package is divided into two source files `ImpuR.R` and `HPLCDAD.R` serving two distinct purposes: `ImpuR.R` contains the tools which are general, that is which will also be useful when working with data which do not come from HPLC-DAD. The other source file

contains elements which are specific to HPLC-DAD. For simplified installation, these source files have also been converted to packages **ImpuR** and **HPLCDAD**.

The package has at its center a data class called "spectime" which can hold many types of information useful when analyzing matrices containing spectra over time.

```
"init.spectime" <-
function ()
{
    setClass("spectime",
       representation=representation(
         data = "matrix", # "data matrix (time x lambda)",
         time = "numeric", # time (numeric row names)
         wavelength = "numeric", # lambda (numeric col names)
         timeunit = "character", # time unit
         wavelengthunit = "character", # wavelength unit
         ct = "matrix", # component intensities over time
         spec = "matrix",# normalized spectra of components
         std.err = "numeric", # baseline std.error
         hetero = "numeric", # heteroscedasticity factor
         WEAtime = "matrix", # window evolving analysis for time
         WEAwavelength = "matrix", # window evolving analysis for lambda
         specsine = "numeric", # spectral dissimilarity
         filename = "character", # source data
         header = "character", # header information
         treatments = "list" # information on pre-treatment
         ),
      prototype=prototype(
         data = matrix(), # set to null matrix
         time = numeric(), # set to null vector
         wavelength = numeric(), # lambda (numeric col names)
         timeunit = "", # time unit
         wavelengthunit = "", # wavelength unit
         ct = NULL, # set to null
         spec = NULL, # set to null
         std.err = NULL, # set to null
         hetero = 0, # set to zero
         WEAtime = NULL, # set to null
         WEAwavelength = NULL, # set to null
         specsine = NULL, # spectral dissimilarity
         filename = "",   # set to empty
         header = "",     # set to empty
         treatments = NULL # set to null-list
         ))}
```

Several functions operate on this data structure. At first there are elementary routines facilitating its manipulation (such as sub-setting, adding, subtracting, printing, creating image and contour graphs). Then there are two principal functions: `PeakPurity` and `plot.spectime`.

`PeakPurity` is the principal container for bilinear analyses. Currently, two types of bilinear decomposition are foreseen: singular value decomposition and peeling. Furthermore, `PeakPurity` adds Wefa traces.

The second key function is `plot.spectime` with the arguments:

```
"plot.spectime" <-
  function(object, # a spectime object
           imageit = TRUE, # if a residual image should be drawn
           contourit=FALSE, # if contours should be added
           level=1, # the level of reconstruction
           showfit="resids" # what to show as an image
           )
  ...
```

It creates the combined peak purity display. `plot.spectime` creates a display with 6 zones: an image, displays of bilinear components in time and wavelength, displays of Wefa traces in time and wavelength, and some summary information. The image zone can be used to display the residuals of a bilinear model of k components where k is determined by the `level` argument (default for `showfit="resids"`). One can also show the bilinear model fit using a selected set of components. For example, to show the bilinear model corresponding only to the second and third component, one can give `showfit=2:3`.

In the specific case of working with spectrochromatograms in HPLC-DAD, several other functions are useful. They are grouped in the source file `HPLCDAD.R`. The tasks implemented here are primarily simulation (`Simul.DAD` and `Clone.DAD`), data pre-treatment (`Treat.DAD`), and spectral dissimilarity analysis (`Specsim` and `BasicImpurityCheck`).

The arguments of the simulation tool are as follows:

```
Simul.DAD<-
  function(
           resol, # measure of separation between main and impurity
           chrom, # chromatogram of main peak
           prin.spec, # spectrum of main peak
           times , # times
           wavelengths, # wavelengths
           timeunit="", # time units
           wavelengthunit="", # wavelength units
           conc = 0, # concentration of impurity
           UAmax = 0.2, # maximum absorbance
           boxcar.width = 1, # number of boxcar spectra in averageing
           impur.spec = prin.spec, # spectrum of impurity
           backslope=0, # slope of baseline drift
           backshift=0, # shift of baseline
           w = FWHM(chrom,times), #
           std.err = rep(0.00003,length(wavelengths)), # base noise profile
           hetero=0, # heteroscedasticity factor
           ncanal = 1, # number of channels for optical window
```

```
            scanrate=0, # simulating the effect of a finite scan time
            perturb = 0, # size of baseline fluctuation (approx noise?)
            perturbrho = 0.3, # autocorrelation of baseline fluct.
            make.cross = FALSE, # make "cross" lines
            bit.loss = 2, # size of bit loss
            deuter = lampnull(wavelengths), # making lamp response
            boxcar = FALSE)
...
```

Most parameters are self explanatory. For general use, to simulate impurity containing spectrochromatograms, the resolution, the concentration of the impurity, and its spectrum need to be known. The resolution is measured in multiples of w which, by default is the full width at half maximum. Simul.DAD has been specifically designed for the HPLC-DAD system we were working with in 1995. Users who would like to apply it in their situation, should carefully examine which of the options remain relevant in their situation and add elements which are missing. The parameters `backslope`, `backshift`, `perturb` and `perturbrho` can be used to simulate a drifting and fluctuating time baseline.

`Clone.DAD` is a simplified interface to Simul.DAD when one just wishes to create a spectrochromatogram of a pure component which matches as much as possible a given dataset. It is useful when one wishes to check whether departures from a perfect bilinear structure with a single pure component are likely artifacts or a real impurity.

Spectrochromatograms can be affected by fluctuating baselines in time and wavelength, in addition it may be necessary to limit attention to a part of the data. This can be done by using the function `Treat.DAD`. It relies on a collection of service functions for identification of the time and wavelength baselines and to deconvolve the spectra.

Potential users will find it easy to adapt Treat.DAD to specific settings by replacing the baseline subtraction routines in an adequate manner.

5.1. Specific details on some of the algorithms

Wefa

The algorithm for window evolving factor analysis works as follows: For a sliding window of $2 \cdot k + 1$ spectra where k is controlled by the `wefamargin` argument, a principal component analysis is computed and a chosen number of eigenvalues are retained and passed on to the caller. These, taken as functions of time or wavelength form the Wefa traces. Wefa traces of spectro-chromatograms with a single bilinear component and homoscedastic noise would show a single important trace (in the time direction, one trace would form a "mountain" under the chromatographic peak whereas the others would be flat lines). When impurities are present, these make that second or third traces lift off from the noise level. This is the impurity signal.

Since spectrochromatograms are heteroscedastic, also the traces (in time direction) of the "noise" components show characteristic patterns. These characteristic patterns are not interesting for peak purity control and hamper the proper reading of the traces. This disturbance can be removed by a simple procedure in which the average of higher order traces is subtracted from the lower order traces. This procedure and some related methods are described in Ritter *et al.* (1995).

Peeling

Peeling is one of the methods to extract bilinear components implemented in the software. The data matrix is taken either in row or column mode and the maximum is found. This defines either a maximum or a minimum chromatogram or spectrum. Then a slice determined by the parameters `threshold` and `margin` is selected either "before" (peel direction "left") or "after" (peel direction "right") the maximum. By default, this slice is taken from the maximum, if the parameter "threshold" is set lower, at 0.9 for example, the slice is taken with reference to the point where 90% of the maximum are reached.

From this average chromatogram or spectrum, a corresponding bilinear complement (spectrum or chromatogram) is computed by regression. If the parameter `favor.positive` is set, an ad-hoc attempt is made to force the component to be (mostly) non-negative.

Compared to singular value decomposition, "peeling" allows more control about how to select the components. It therefore permits refining a purity analysis by focusing on the beginning or the end of a chromatographic or spectroscopic peak.

Optical window effects

The effects of an optical window which is wider than the data interval can be treated with the functions convolve and deconvolve contained in the `HPLCDAD.R` file. By default, both work by first converting from absorbance to transmittance. This is done, since the actual "smoothing" due to the optical window happens in "intensity" that is in transmittance. In practice, the characteristic of the lamp is added via the parameter `lampnull`.

Deconvolution of a spectrum works by finding a first "guess", convolving it using the convolution routine with the user determined parameters of the optical window(`ncanal`) and the lamp characteristic and comparing it with the actual spectrum. An adjustment related to the difference is then applied to the guess and the iteration continues. This is an adaptation of the old iterative procedure by Burger and van Cittert (1932).

6. Conclusion and further work

As stated at the outset, **ImpuR** was written and optimized in the specific context of peak purity analysis of HPLC-DAD data. However, time-intensity matrices arise also in different contexts and bilinear analyses may also there be useful. **ImpuR** was therefore designed for easy adaptation to other settings. Users should first try to use **ImpuR** as is and make adaptations to `Read.DAD`, `baseline.time`, `baseline.lamba`, `Error.Components` as needed. Both modes of bilinear analysis "peal" and "svd" allow different modeling levels (1 to 4) which will often prove sufficient.

ImpuR is part of a wider effort to make tools available which were written for specific contexts but which may have wider applicability. The next stage in this project is to include routines for analysis of nonlinear regression situations. This includes a special adaptation of Markov Chain Monte Carlo simulations called the Griddy Gibbs sampler. We expect a release of a first version of the toolbox under the name of **KritterBox** around mid 2007.

References

Burger HC, van Cittert PH (1932). "Wahre und scheinbare Intensitätsverteilungen in Spektrallinien." *Mitteilungen aus dem Physikalischen Institut der Reichsuniversität Utrecht*, pp. 772–780.

Gilliard J, Ritter C (1997a). "Simulations of Liquid Chromatography-Diode Array Detection Data Including Instrumental Artifacts for the Evaluation of Mixture Analysis Techniques." *Journal of Chromatography A*, **758**, 1–18.

Gilliard J, Ritter C (1997b). "Use of Simulated Liquid Chromatography-Diode Array Detection Data for the Definition of a Guide Curve in Peak Purity Assessment by Spectral Comparison." *Journal of Chromatography A*, **786**, 1–11.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ritter C, Gilliard J, Cumps J, Tilquin B (1995). "Corrections for Heteroscedasticity in Window Evolving Factor Analysis." *Analytica Chimica Acta*, **318**, 125–136.

Affiliation:

Christian Ritter
Institut de Statistique
Université Catholique de Louvain
20 voie du Roman Pays
B-1348 Louvain-la-Neuve, Belgium
E-mail: ritter@stat.ucl.ac.be
URL: http://www.stat.ucl.ac.be/~ritter/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 18, Issue 9 *Submitted:* 2006-10-24
January 2007 *Accepted:* 2007-01-11