mrbin

Matthias S. Klein

Version 1.7.0

Contents

1 Introduction

This package is a collection of functions for processing and analyzing metabolomics data.
The namesake function mrbin() uses spectral binning to convert 1D or 2D Nuclear Magnetic Resonance (NMR) data into a matrix of values suitable for further data analysis and performs basic processing steps in a reproducible way. Negative values, a common issue in NMR data, are replaced by positive values. All used parameters are stored in a readable text file and can be restored from that file to enable exact reproduction of the data at a later time.
The atnv algorithm for replacing negative values in NMR data sets can be employed using atnv().
NMR plotting functions are found in mrplot().
Artificial Neural Network features can be analyzed using Feature Impact Assessment (FIA) using the function fia().

1.1 Installation

To install mrbin, please install the latest version of R first. Then install mrbin.

To install the latest stable version of mrbin from CRAN:

```
> install.packages("mrbin")
```

After installation, load the package as follows:

```
> library(mrbin)
```

2 mrbin: Magnetic Resonance Binning, Integration and Normalization

The main NMR binning functions of this package are controlled via the mrbin() function. Results returned include the final bin list and a set of used parameters.

2.1 Prerequisites and Considerations

To use this package, you will need NMR data in the Bruker file format accessible on your computer. Please make sure your data is Fourier transformed, phase corrected, baseline corrected, and correctly referenced. The data has to be stored in folders according to standard Bruker folders, that means foldername/1/pdata/1 etc. Experiment numbers and processing numbers can be freely chosen.

This package has been tested for 1D ^{1}H NOESY and 2D ^{1}H-^{13}C HSQC spectra.

Before starting mrbin, take a look at your NMR data, for example in Bruker Topspin, and decide on the following parameters (you will be able to set the values of these parameters for your data during running mrbin):

- Bin area: Area where signals are observed in your data set
- Bin width: Should match roughly the width of a singlet peak in your data set. Given in ppm.
- Bin height (only 2D): Should match roughly the height of a singlet peak in your data set. Given in ppm.
- Solvent area: Area to exclude to remove solvent artifacts
- Additional areas to be removed: Any other area containing artifacts, such as streaks surrounding strong peaks.

mrbin will also show you preview plots for these parameters during the run.

2.2 Running mrbin in Interactive Mode

You can start mrbin using the following code:

```
> mrbinResults<-mrbin()
```

This will start a series of questions that will guide you through the parameters to be used.

mrbin() returns an (invisible) object of type mrbin, containing the following variables:

- bins: A matrix containing bin data for all samples, Depending on the option you chose, the data will be cleaned up and scaled.
- parameters: A list containing all parameters used to create the bin matrix.
- metadata: A list containing metadata, if provided.
- transformations: A character vector containing information on the data transformation and scaling that has been performed, for example reference scaling, PQN, atnv, log transform, etc.
- changeLog: A data.frame containing information on documented changes that were made to the data, including time stamps.
- changeValues: A list containing control values, enabling verifying changes by checkmrbin(mrbinResults)

Please be aware that the generated bin values are calculated as the average value of all data points within each bin, multiplied by the area of the bin. This avoids the issue of different numbers of data points within a bin in different spectra, and corrects for bins that are different in their size. Therefore, the bin intensities can be viewed as pseudo-integrals of the contained signals.

Several files may be written to the chosen directory:

- A .Rdata file containing the generated mrbin data object
- A .pdf file containing quality control plots of the raw binned data
- A .pdf file containing quality control plots of the processed binned data
- A .pdf file containing preview plots of the chosen signal-to-noise ratios (SNR) of selected spectra
- A .txt file containing all parameters and potential warning messages from the mrbin run. This file can be reloaded to mrbin to recreate a dataset for reproducibility

The generated quality control plots show four panels:

- Previews of a few selected spectra, overlaid with the areas of the bins generated (in green) to make sure all peaks of interest are covered
- Boxplots of bin intensities per bin (feature) to aid in spotting data quality issues
- Boxplots of bin intensities per sample to aid in spotting data quality issues
- A PCA plot of the generated bin data to aid in spotting data quality issues

The preview plots of the signal-to-noise ratios (SNR) of selected spectra are provided to help chose an SNR level matching the used data. Please note that the visual representation of SNR levels is approximate, as noise removal is performed on the binned data, but not on the raw spectra. To achieve a satisfactory preview of SNR levels, each spectrum is scaled in a way so its peak heights are equivalent to the respective bin value. The median of these ratios is used as a scaling factor for the whole spectrum.

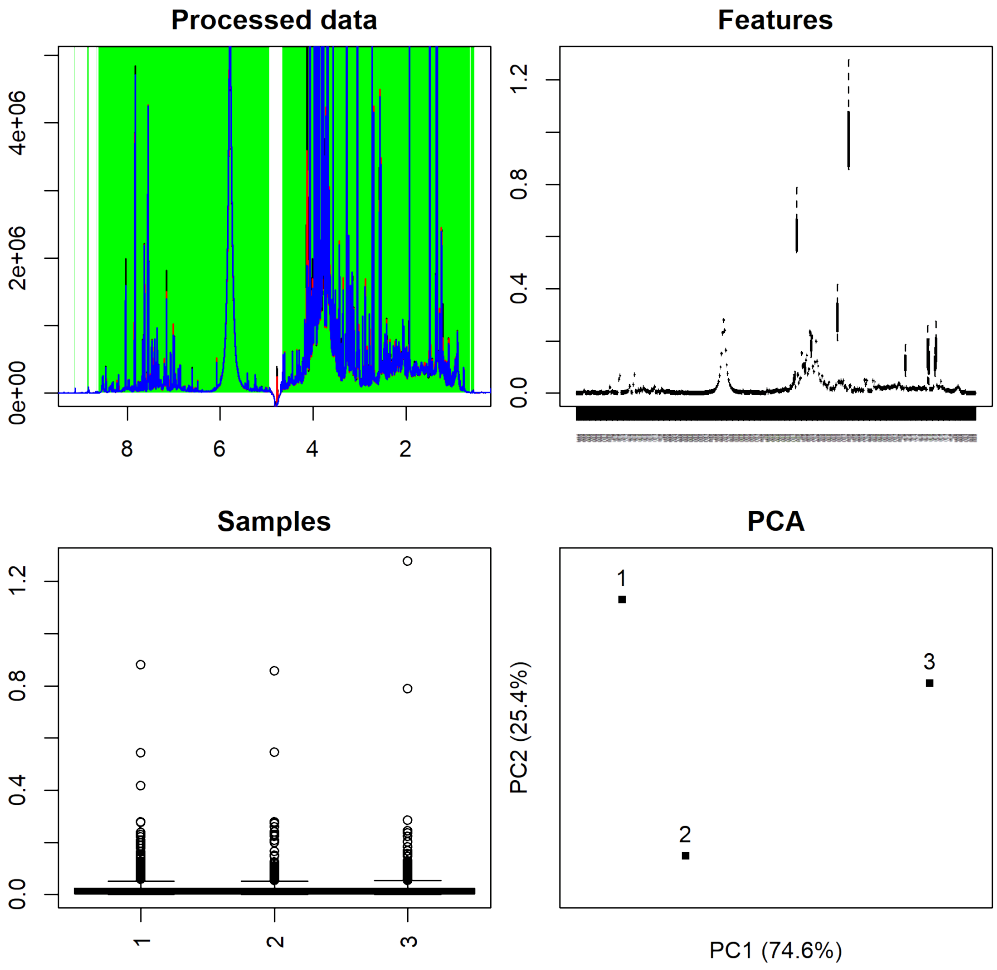
2.3 Setting Parameters at the Command Line

Parameters can also be submitted at the command line. With the default setting of silent=FALSE, the user will be guided through the user input questionnaire to view, and make adjustments to, the parameters submitted at the command line.

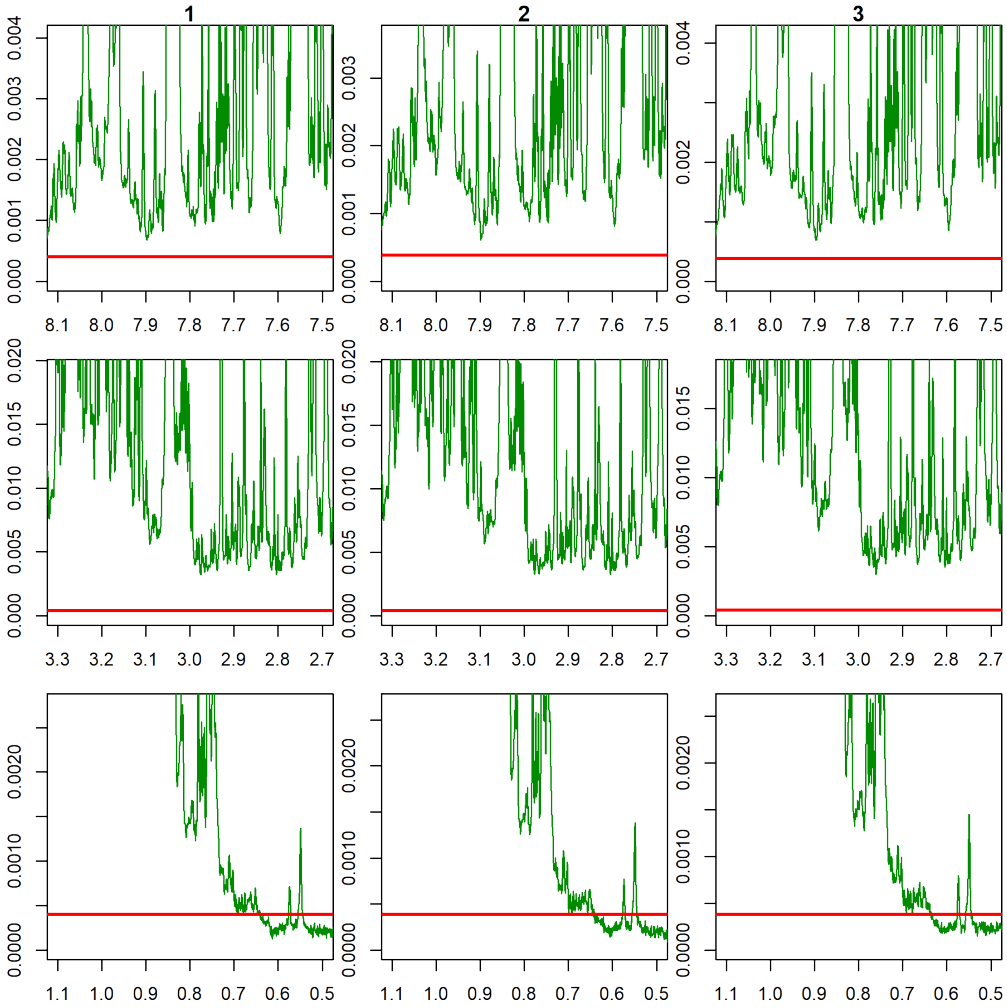
When using silent=TRUE, this will set up all parameters and run all steps without asking for further user input. This should only be used after carefully reviewing all parameters in interactive mode, as data issues might be missed.

The following example provides parameters for analyzing a 1D data set.

```
> mrbinResults<-mrbin(parameters=list(dimension="1D",
+              binwidth1D=.02,
+              referenceScaling="Yes",
+              removeSolvent="Yes",
+              solventRegion=c(4.95,4.65),
+              noiseRemoval="Yes",
+              signal_to_noise1D=35,
+              noiseThreshold=0.75,
+              PQNScaling="No",
+              fixNegatives="No",
+              logTrafo="No",
+              NMRfolders=c(system.file("extdata/1/10/pdata/10",package="mrbin"),
+                           system.file("extdata/2/10/pdata/10",package="mrbin"),
+                           system.file("extdata/3/10/pdata/10",package="mrbin"))
+      ))
```



Approximate signal-to-noise ratios (SNR) will be shown for up to three selected spectra:



In the example shown above, spectral data from the mrbin package is used. To use your own data, change the folder names to the full folder names holding the files "1r" (or "2rr" for 2D data) as follows:

```
> mrbinResults<-mrbin(parameters=list(
+    NMRfolders=c("C:/Bruker/TopSpin3.6.1/data/guest/nmr/sample_1/10/pdata/10",
+                 "C:/Bruker/TopSpin3.6.1/data/guest/nmr/sample_2/10/pdata/10",
+                 "C:/Bruker/TopSpin3.6.1/data/guest/nmr/sample_3/10/pdata/10")
+    ))
```

2.4 Processing of mrbin Objects

All processing algorithms, such as noise removal, scaling, etc., can be performed within mrbin(). However, it is also possible to run these (optional) steps later. For this, you should select "No" in mrbin() when asked about noise removal, PQN scaling, atnv, and log transform. Please note that scaling to the reference substance happens at the raw spectrum level and needs to be done during running mrbin(); it cannot be performed later. Here is a short overview of implemented commands, which should be performed in the order shown below:

Set signal-to-noise ratios (SNR) interactively:

```
> mrbinResults<-setNoiseLevels(mrbinResults)
```

To remove noise signals using the newly defined SNR levels, use removeNoise(). Noise levels are individually checked for each bin of each spectrum. Expected noise levels are calculated for each bin of each spectrum, based on the estimated noise level of the respective spectrum, corrected for the area of each bin and the number of data points within this bin (inverse square root relationship). Noise removal needs to be performed before any other scaling or transformation that changes bin values.

```
> mrbinResults<-removeNoise(mrbinResults)
```

Remove negative values by atnv scaling (see below for details):

```
> mrbinResults<-atnv(mrbinResults)
```

Probabilistic Quotient Normalization (PQN) scaling, this is similar, but better, than scaling to the total sum. Use for urine or tissue extracts:

```
> mrbinResults<-PQNScaling(mrbinResults)
```

Logarithm transform (will usually only work after atnv transform):

```
> mrbinResults<-logTrafo(mrbinResults)
```

Plot results:

```
> plotResults(mrbinResults)
```

2.5 Adding Metadata to mrbin Objects

You can add or edit metadata, including metabolite identities, for an mrbin object as follows:

```
> mrbinResults<-metadatamrbin(mrbinResults)
```

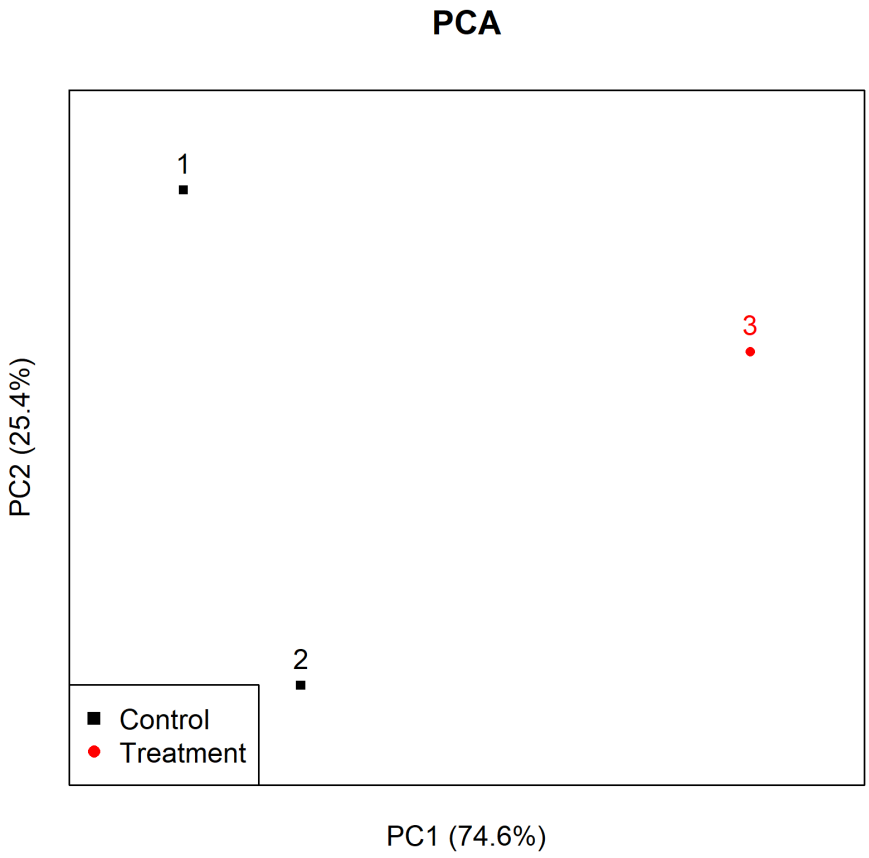
If you choose "edit" for any of the categories, an editor window may open. After editing text, close this window and select "Save changes" to save your edits. Make sure to keep the opening and closing quotation marks.

Alternatively, you can edit metadata fields in the command line:

```
> mrbinResults<-metadatamrbin(mrbinResults,metadata=list(
+   projectTitle="Test project",
+   factors=factor(c("Control","Control","Treatment"))))
```

If you choose to define treatment groups in this step, you can plot a PCA with color coded group information as follows:

```
> plotPCA(mrbinResults)
```

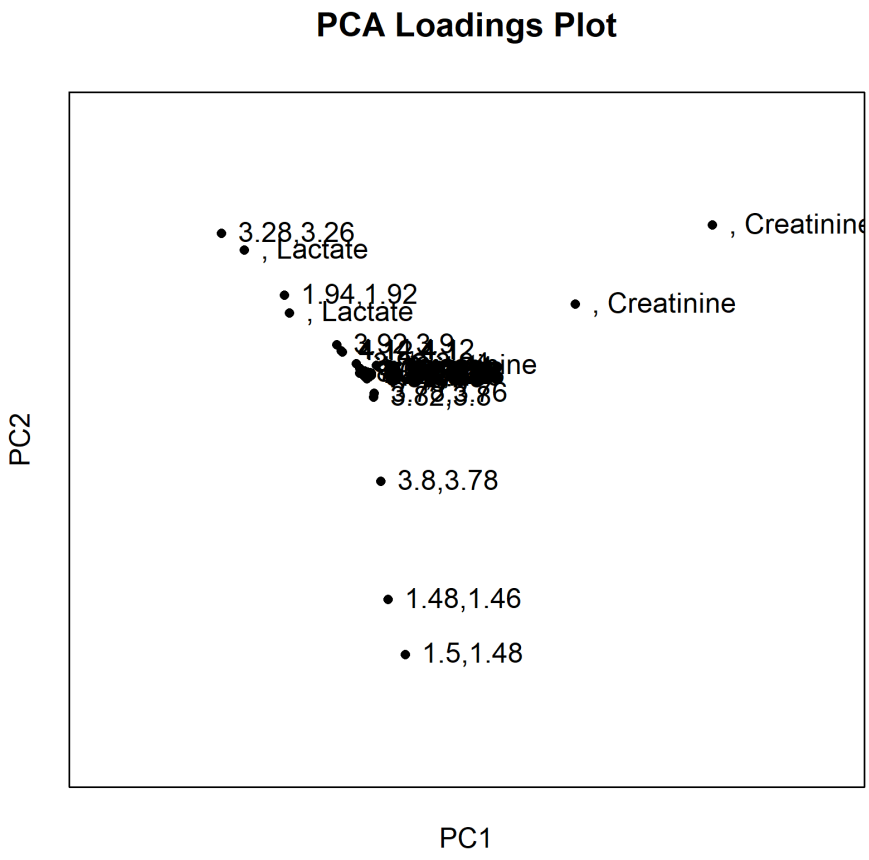


The results can be annotated with potential metabolite identities using the command metadatamrbin(), or in the command line as follows. Left, right, top, and bottom borders of the metabolite signal need to provided as a matrix, along with a vector of metabolite names. The provided data will then be used to annotate the bins of the dataset. Bins might be assigned to multiple metabolites if they cover the areas of several potential signals.

```
> mrbinResults<-editmetabolitesmrbin(mrbinResults,borders=matrix(c(
+       1.346,1.324,21,23,
+       3.052,3.043,30.5,33.5,
+       4.066,4.059,57,59.5
+     ),ncol=4,byrow=TRUE),metabolitenames=c(
+     "Lactate",
+     "Creatinine",
+     "Creatinine"
+     ))
```

The loadings plot of the PCA can be displayed as follows:

```
> plotPCA(mrbinResults,loadings=TRUE,annotate=TRUE)
```



2.6 Editing mrbin Objects

As mrbin objects use a documented change log, ideally you would use only functions from the mrbin package to change them (e.g. log transform), as these are documenting changes. If you wish to perform other edits, please do so using the command editmrbin(), as in the following example:

```
> mrbinResults<-editmrbin(
+   mrbinResults,
+   bins=mrbinResults$bins,#omit this line if no changes are made to bins
+   parameters=list(noiseThreshold=0.75),
+   metadata=list(projectTitle="Test project"),
+   comment="Changed title and noise parameters",
+   transformations="Scaling"#If you change bins, provide a short explanation
+   )
```

To see the change log, use:

```
> mrbinResults$changeLog
```

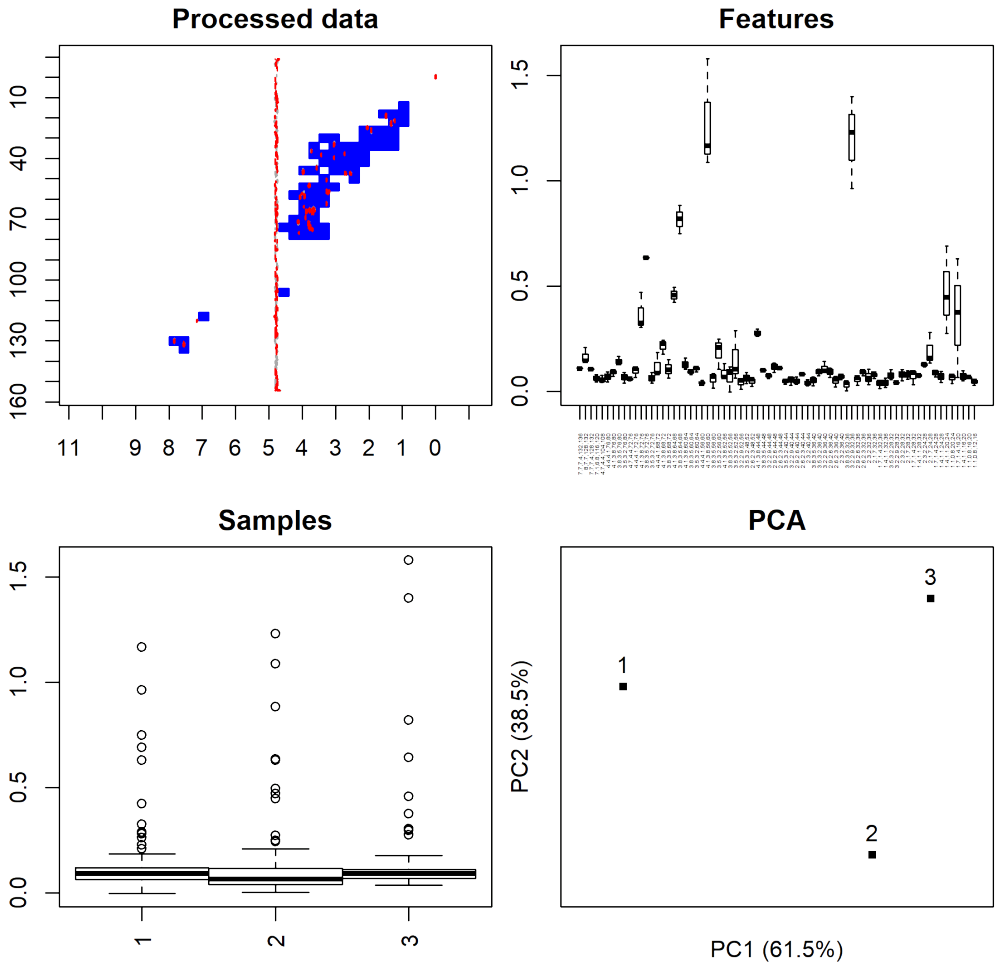
To see a summary of scalings and transformations, and check for undocumented edits to the data, use:

```
> checkmrbin(mrbinResults)
```

2.7 Example: 2D Data

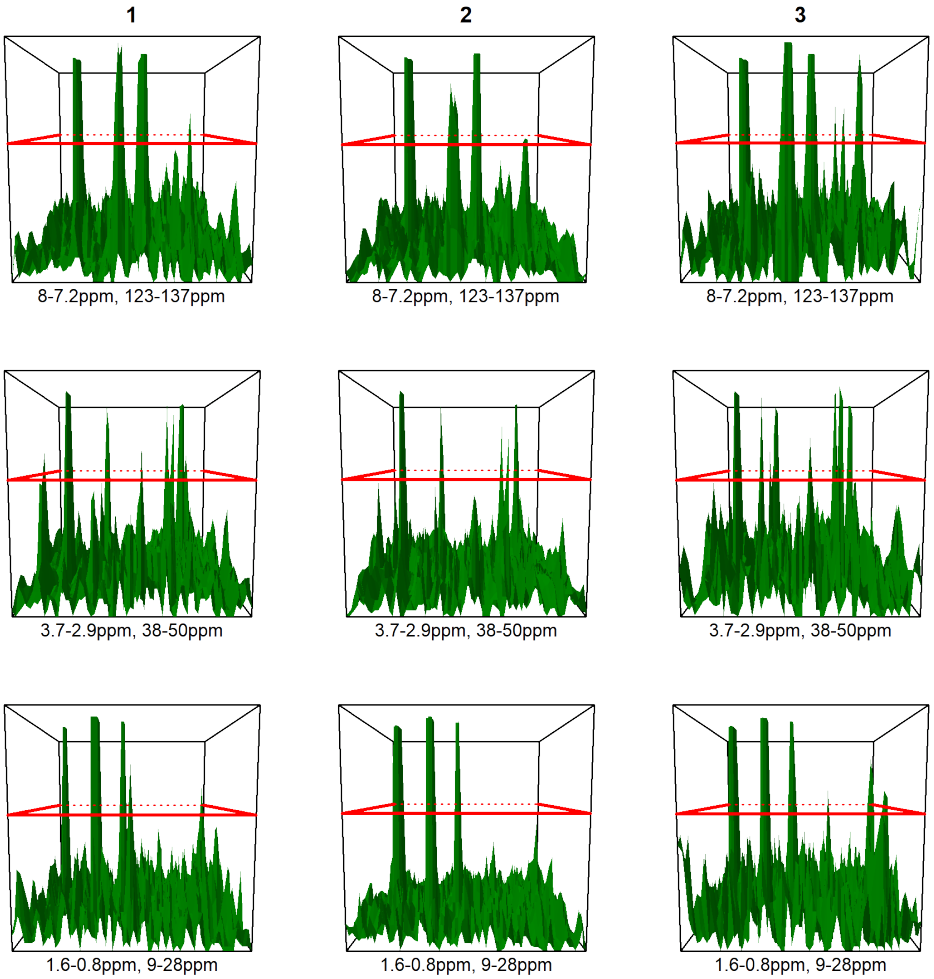
The following example provides parameters for analyzing a 2D data set.

```
> mrbinResults2D<-mrbin(setDefault=TRUE,parameters=list(dimension="2D",
+                 binwidth2D=0.3,
+                 binheight=4,
+                 removeSolvent="Yes",
+                 solventRegion=c(4.95,4.65),
+                 noiseRemoval="Yes",
+                 signal_to_noise2D=5,
+                 noiseThreshold=0.75,
+                 PQNScaling="No",
+                 fixNegatives="No",
+                 logTrafo="No",
+                 NMRfolders=c(system.file("extdata/1/12/pdata/10",package="mrbin"),
+                         system.file("extdata/2/12/pdata/10",package="mrbin"),
+                         system.file("extdata/3/12/pdata/10",package="mrbin"))
+                 ))
```



The used approximate signal-to-noise levels will be visually plotted and saved for up to three selected spectra:

```
> setNoiseLevels(mrbinResults2D,plotOnly=TRUE)
```



The binned NMR data can be found in mrbinResults$bins. This numeric matrix contains bins in columns and samples in rows, and can be directly used for further data analysis.

The binned data can also be loaded from the hard drive later (if an output file was specified):

```
> load("C:/Users/User/Documents/mrbin.Rdata")
```

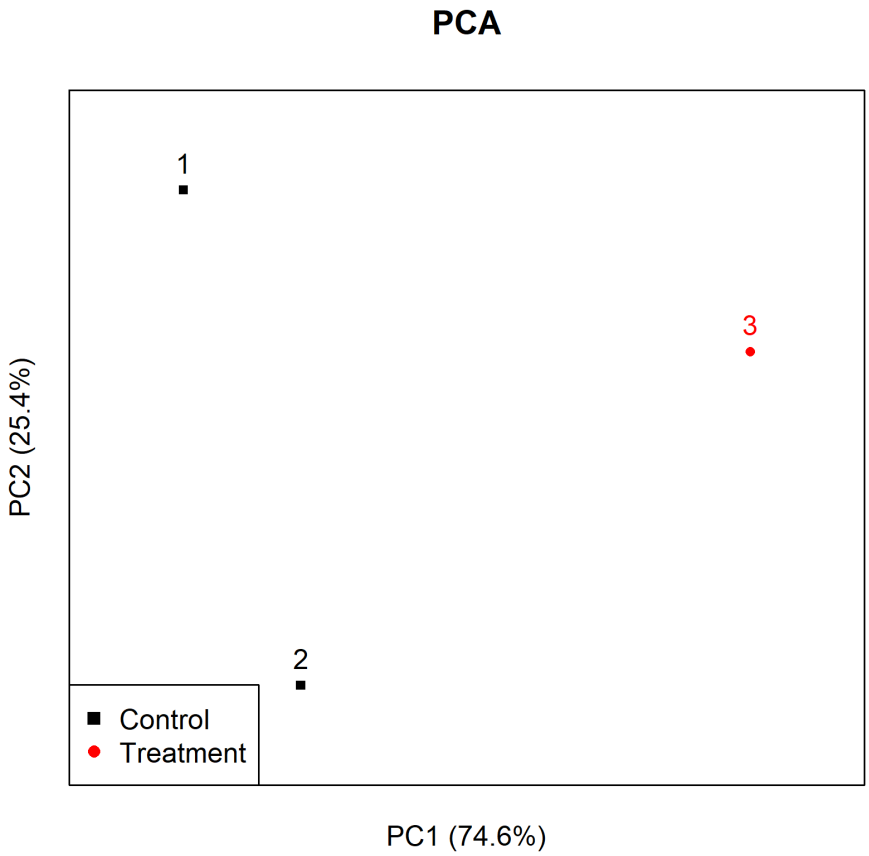
Column names include left and right borders or each bin separated by commas. For 2D data, column names include left, right, top, and bottom border of each bin.

You can edit metadata fields in the command line:

```
> mrbinResults2D<-metadatamrbin(mrbinResults,metadata=list(
+   projectTitle="Test project",
+   factors=factor(c("Control","Control","Treatment"))))
```

If you define treatment groups in this step, you can plot a PCA with color coded group information as follows:

```
> plotPCA(mrbinResults2D)
```

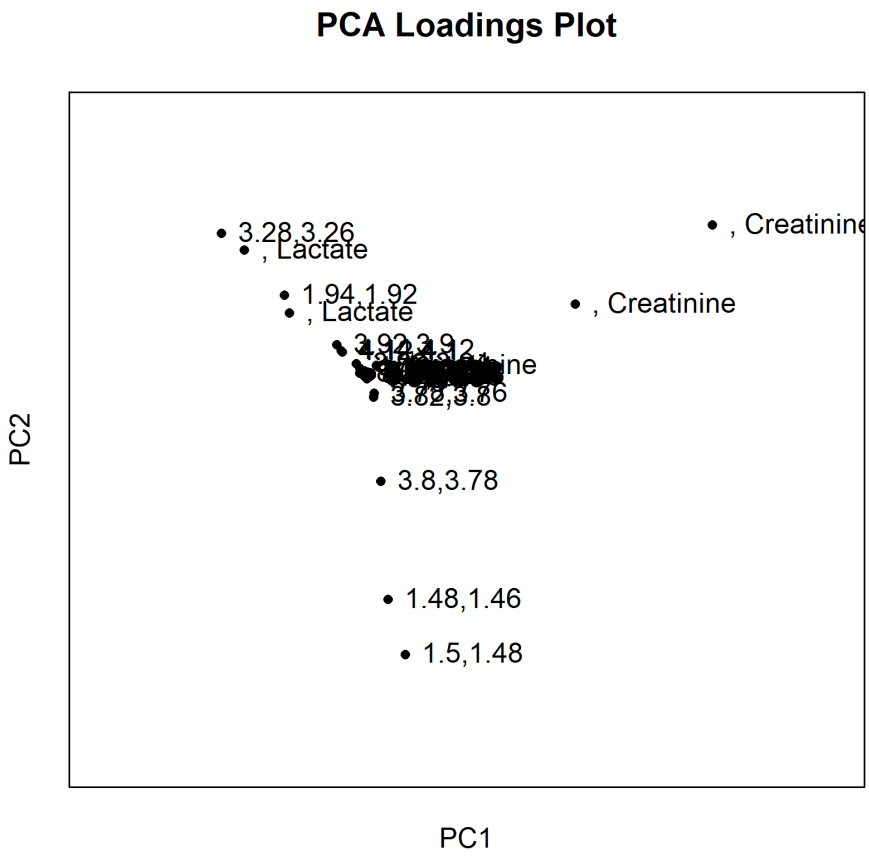


The results can be annotated with potential metabolite identities using the command metadatamrbin(), or in the command line as follows:

```
> mrbinResults2D<-editmetabolitesmrbin(mrbinResults2D,borders=matrix(c(
+       1.346,1.324,21,23,
+       3.052,3.043,30.5,33.5,
+       4.066,4.059,57,59.5
+     ),ncol=4,byrow=TRUE),metabolitenames=c(
+     "Lactate",
+     "Creatinine",
+     "Creatinine"
+     ))
```

The loadings plot of the PCA can be displayed as follows:

```
> plotPCA(mrbinResults2D,loadings=TRUE,annotate=TRUE)
```



2.8 Example: Lipid Data Analysis

Analyzing lipid NMR signals can be accomplished using custom bin lists rather than a regular grid of bins. This can be done as in the following code:

```
> results <- mrbin(parameters=list(dimension="1D",binMethod="Custom bin list",
+  specialBinList=matrix(c(5.45,5.2,0,160,
+                          2.9,2.74,0,160,
+                          2.14,1.93,0,160,
+                          1.41,1.2,0,160,
+                          0.94,0.8,0,160,
+                          2.44,2.2,0,160,
+                          4.325,4.26,0,160
+                          ),ncol=4,byrow=TRUE,dimnames=list(c(
+                          "-CH=CH- Methene",
+                          "=CH-CH2-CH= Diallylic",
+                          "-CH2-CH=CH- Allylic",
+                          "-CH2- Methylene",
+                          "-CH3 Methyl",
+                          "COO-CH2-CH2- Methylene_to_carboxyl",
+                          "Glycerol"
+                          ),NULL)),
```

```
+  referenceScaling="Yes",reference1D=c(0.03,-0.03),removeSolvent="No",
+  noiseRemoval="No",PQNScaling="No",fixNegatives="Yes",logTrafo="No",
+  NMRfolders=c(system.file("extdata/1/10/pdata/10",package="mrbin"),
+               system.file("extdata/2/10/pdata/10",package="mrbin"),
+               system.file("extdata/3/10/pdata/10",package="mrbin"))
+  ))
```

When using custom bin lists, each spectral data point may be part of multiple bins. For rectangular bin lists, each data point will only be counted once. The lipid signal areas used in this example are based on Klein et al, 2011. Please note that the data used in this example does not include lipid signals but signals of small molecules.

2.9 Recreating Data and Parameters

In order to create reproducible results, mrbin will save the used parameters to a text file. Please keep this file. You may want to share this file in a data repository when publishing your findings.

While it is fine to view the parameter text file in a text editor, please do not change its contents, as this may break its formatting.

In order to recreate a previous data set, or to reload previously used parameters, use:

```
> mrbinResults<-mrbin()
```

and select ""Reload from file" when asked "Set parameters or use existing parameters?". This will restore all parameters that were previously used. If the file was created using an older version of mrbin, this may cause inconsistencies. Missing parameters will be added using standard parameters. Ideally, download the older mrbin version at CRAN and use the old version to recreate the data in an exact way.

Please be aware that bins will have to be recalculated in this case, so the original NMR spectra will have to be present to do this.

2.10 mrbin Workflow

The sequence of data processing is as follows:

- Gathering initial parameters from user
- Creating a set with coordinates of each bin
- Optional: Removing solvent region
- Optional: Removing additional regions
- Optional: Cropping of HSQC spectra to the region along the diagonal
- Optional: Merging regions containing peaks with unstable positions such as citric acid
- Reading NMR spectral data
- Optional: Scaling to reference region
- Binning
- Quality control plot, including PCA

- Optional: Removal of bins containing mostly noise
- Optional: Replacement of negative values (atnv transform)
- Optional: PQN transform
- Optional: Log transform
- Quality control plot, including PCA
- Returning final bin data and parameters as an mrbin object

3 atnv: Affine Transformation of Negative Values

The function atnv replaces (column-wise) negative values by a small positive number. The number is calculated as an affine transformation to the range of the lowest positive number to 0,01*the lowest positive number (of this column). Ranks stay unchanged. Positive numbers are not altered.

If sample-wise noise levels are available, the median noise level of samples with negative values is calculated and replaces the lowest positive number in case it is smaller. If no noise data is available, the 1% percentile of all positive values in the data set is used as an estimate.

It is recommended to use this function AFTER noise removal and other data clean-up methods, as it may alter (reduce) the noise level of the binned data. If no NMR data and noise levels are provided as arguments, the function will use NMR data and noise levels from the global variables mrbin.env$bins and mrbin.env$mrbinTMP.

To use own (user provided) data:

```
> atnv(NMRdataMatrix,noiseLevelVector)
```

To use current mrbin data from the internal memory, use atnv()) without parameters. This requires data loaded using mrbin(). This is usually not necessary as it is included in the mrbin work flow.

4 mrplot: NMR Plotting

The mrbin package contains basic plotting commands for 1D and 2D NMR data.

Most convenient is using the mrplot function, which is menu-based:

```
> mrplot()
```

In mrplot, multiple spectra shown overlaid. If both 1D and 2D spectra are added, both are shown in region-matched plots, e.g. for metabolite identification purposes.

To use the more basic commands, you need to first load one NMR spectrum:

```
> readBruker(dimension="1D",
+   folder=system.file("extdata/1/10/pdata/10",package="mrbin"))
> plotNMR()
```

The system.file command loads example data from the mrbin package. To use your own spectra, please adjust as follows:

```
> readBruker(dimension="1D",
+   folder="C:/Bruker/TopSpin3.6.1/data/guest/nmr/sample_1/12/pdata/10")
> plotNMR()
```

For 2D data, this code reads as follows:

```
> readBruker(dimension="2D",
+   folder=system.file("extdata/1/12/pdata/10",package="mrbin"))
> plotNMR()
```

There are multiple commands for editing the plot:

```
> zoom(left=4.6, right=2, top=10, bottom=150) #Exact zoom
> zoomIn() #Zoom in
> zoomOut() #Zoom out
> intPlus() #Increase intensity
> intMin() #Decrease intensity
> left() #Move spectrum to the left
> right() #Move spectrum to the right
```

For 2D data, you can additionally use the following commands:

```
> contMin() #Decrease minimum contour level (show more small peaks)
> contPlus() #Increase minimum contour level (remove small peaks)
> up() #Move spectrum up
> down() #Move spectrum down
```

5 fia: Feature Impact Assessment of Artificial Neural Networks

The function fia() finds features that can change the outcomes of a model's prediction. For example, fia=1.00 means single compound found in all, or 100 percent of samples. fia=2.45 would indicate that this compound is found in pairs in all but 45 percent of tested samples.
You will need a trained Artificial Neural Network (ANN) model available in R to use fia(), e.g. using packages keras and/or tensorflow. A function named predict needs to be present for fia() to work. If the function name of the prediction function is different, the function name has to be provided in the parameter functionNamePredict. Please make sure to have loaded all required packages before starting fia().

As an example, we will a logit model instead of ANN, as ANN training would require additional programs to be installed.

```
> #First, define group membership and create the example feature data
> group<-factor(c(rep("Group1",4),rep("Group2",5)))
> names(group)<-paste("Sample",1:9,sep="")
> dataset<-data.frame(
+     Feature1=c(5.1,5.0,6.0,2.9,4.8,4.6,4.9,3.8,5.1),
+     Feature2=c(2.6,4.0,3.2,1.2,3.1,2.1,4.5,6.1,1.3),
+     Feature3=c(3.1,6.1,5.8,5.1,3.8,6.1,3.4,4.0,4.4),
+     Feature4=c(5.3,5.2,3.1,2.7,3.2,2.8,5.9,5.8,3.1),
+     Feature5=c(3.2,4.4,4.8,4.9,6.0,3.6,6.1,3.9,3.5),
+     Feature6=c(6.8,6.7,7.2,7.0,7.3,7.1,7.2,6.9,6.8)
+ )
> rownames(dataset)<-names(group)
```

```
> #train the logit model
> mod<-glm(group~Feature1+Feature2+Feature3+Feature4+Feature5+Feature6,
+    data=data.frame(group=group,dataset),family="binomial")
> fiaresults<-fia(model=mod,dataSet=dataset,factors=group,
+   parameterNameData="newdata",firstLevel=0,type="response")
> fiaresults$scores
```

6 Known Issues

6.1 Firewall Warnings

If parallel computing is turned on and the package parallel is installed, mrbin will try to use the socket approach for computing. This requires establishing network connections to the local cluster, which might trigger the firewall. It is safe to unblock these connections.

6.2 Pop-Up Windows

mrbin is set up to ask for user input through pop-up windows. This requires graphics support, otherwise the user input will be asked through command line menus, which is less user friendly but still offers the full functionality.

6.3 Apple/Mac Computers And RStudio

In some cases, running mrbin from within RStudio on Apple computers will not generate pop-up windows. To enable pop-up windows, it might be helpful to install the newest version of xquartz from https://www.xquartz.org.

6.4 Spectra are Missing

If a Bruker spectrum is not shown during browsing, please make sure a file with filename title is present in the PROCNO folder of that spectrum. You can create a title file by opening the spectrum in Bruker Topspin, selecting the Title tab, entering a title and clicking the disk symbol for saving.

7 License

8 Citation

If you are using mrbin in a publication, please cite the following manuscript:

Klein, M.S. (2021): Affine Transformation of Negative Values for NMR Metabolomics Using the mrbin R Package. J. Proteome Res. 20(2):1397-1404, DOI: 10.1021/acs.jproteome.0c00684

9 References

Klein MS, Dorn C, Saugspier M, Hellerbrand C, Oefner PJ & Gronwald W (2011): Discrimination of Steatosis and NASH in Mice Using Nuclear Magnetic Resonance Spectroscopy. *Metabolomics* 7:237-246