A mixed model for normalization of metabolomics data with applications to correlation analysis

Alexandra Jauhiainen, Basetti Madhu, John Griffiths, and Simon Tavaré

November 12, 2014

Abstract

In metabolomics the goal is to identify and measure the concentrations of different metabolites (small molecules) in a cell or a biological system. The metabolites form an important layer in the complex metabolic network, and the interactions between different metabolites are often of interest. It is crucial to perform proper normalization of metabolomics data and in [**?**], a normalization approach is proposed that is based on a mixed model, with simultaneous estimation of a correlation matrix. The methodology is implemented in R, and this document introduces how to apply the method to real NMR example data.

1 Data

We use two versions of a real NMR dataset that stems from six cohorts of human diploid fibroblasts (HDFs). For each cohort, a number of batches were extracted, which in turn gave rise to a number of samples (52 in total). The design is illustrated in Figure 2 in [**?**].

In one version of this data set, named `Ystd`, TSP was used for chemical shift calibration and metabolite quantitation. In total 28 metabolites were uniquely characterized. In the other version of the dataset, `Yraw`, TSP was not used for quantitation. In this set, 26 spectral features could be identified, some of which some are sums of signals for several metabolites.

The two datasets are not directly comparable, since, in addition to the issues involving standardization, the same metabolite might not correspond to the same peaks in both sets. The different metabolites are named `mr1,mr2,...` in `Yraw`, and `ms1,ms2,...` in `Ystd`.

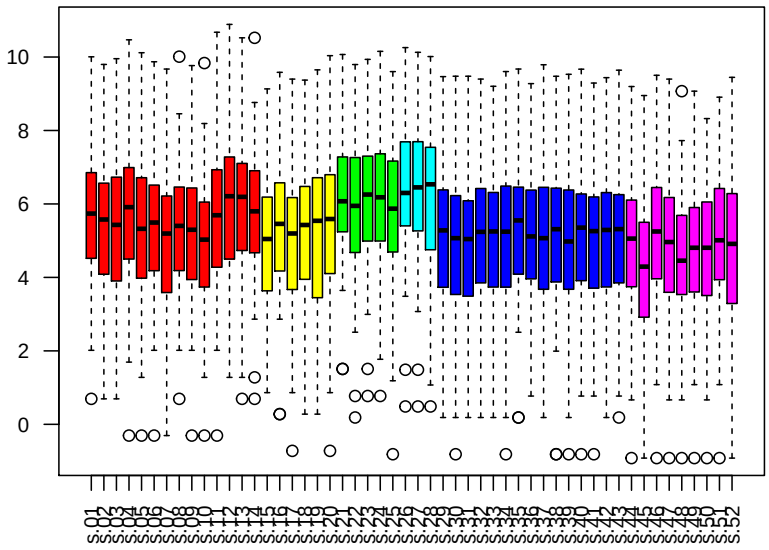
The data has been log-transformed, and can be loaded with the command illustrated below. Necessary packages are also loaded before further analysis:

```
> library(gplots)
> library(hglm)
> source("normfunctions.R")
> load("NMR.RData")
```

The cohorts and batches are given in the objects `cohort` and `batch`, respectively.

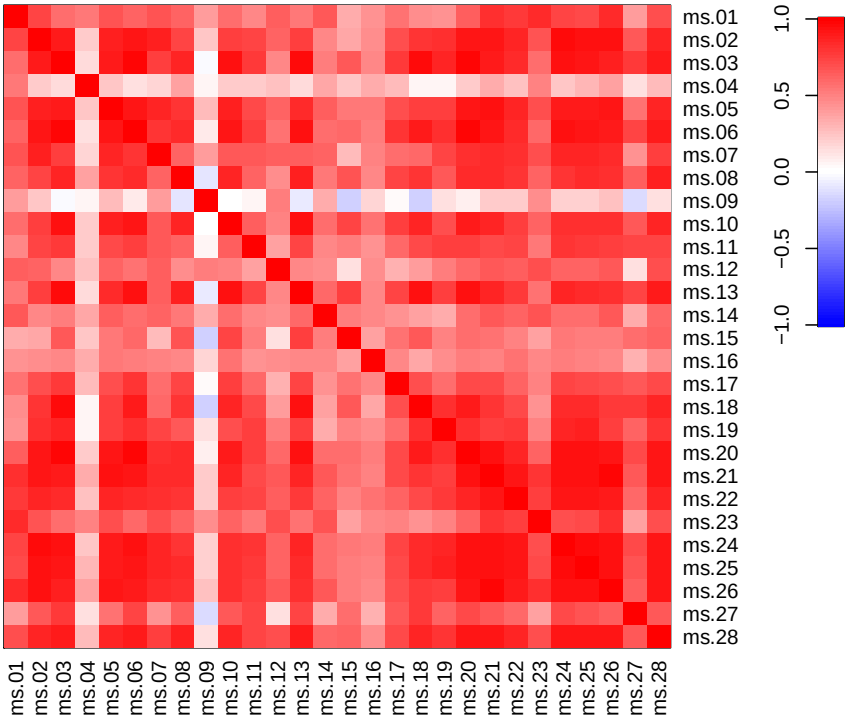
We can take a preliminary look at the data by using boxplots with colors indicating batches or cohorts. Here we plot the `Ystd` data with cohorts indicated.

```
> col.cohort <- rep(rainbow(6),times=table(cohort))
> boxplot(Ystd,col=col.cohort,las=2,cex.axis=0.8)
```



We can also look at correlation heatmaps of the data. First we estimate the correlation matrix by Pearson correlations and plot a heatmap. The code to generate the heatmap is provided with the other functions. We are using `Ystd` here, and we see that the correlations are mostly positive for these non-normalized data.

```
> cor.std <- cor(t(Ystd),method="pearson")
> heatmap(cor.std)
```



2 Normalization

As indicated in the previous paragraph, these data are in need of normalization. The technical artifacts and standardization contribute to large positive correlations. We apply the mixed model to normalize the data by the following commands.

```
> YstdN <- normalizeMixed(Ystd,cohort=cohort,batch=batch)
> YrawN <- normalizeMixed(Yraw,cohort=cohort,batch=batch)
```

As the default, the function returns the normalized data only, although it also simultaneously estimates a covariance matrix for the metabolites. The covariance matrix can be returned by setting the option `return.cov=TRUE`, which will give the normalized data and the covariance matrix as a list. Another option is simply to estimate the covariance/correlation matrix from the normalized data, as is done below.

During the normalization, the covariance matrix is estimated using maximum likelihood according to $\frac{1}{L}\sum_{l=1}^{L} RV_l RV_l^T$ where L is the number of samples, and RV_l the residuals (see Section 3 in [**?**]). Due to the iteration this is a sensible choice, since the mean (giving rise to the residuals) is considered fixed when the covariance is estimated. A sample covariance matrix is usually estimated by $\frac{1}{L-1}\sum_{l=1}^{L} RV_l RV_l^T$, which differs slightly from our estimate.

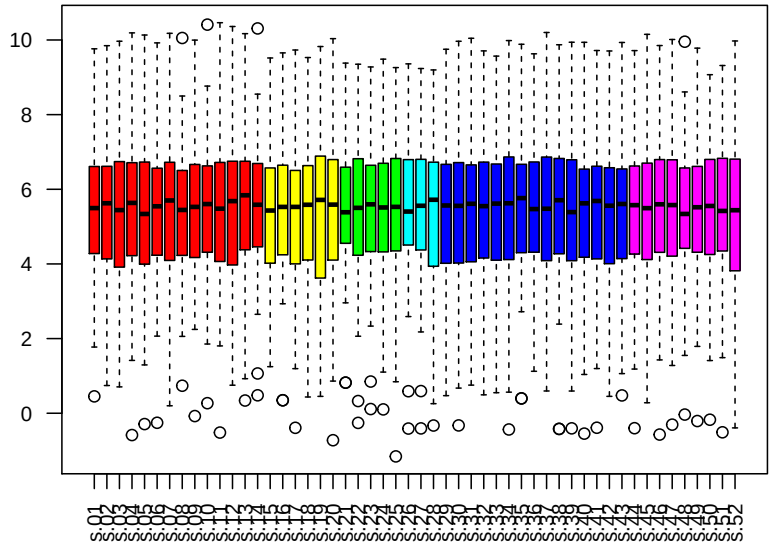
Hence, using the option `return.cov=TRUE` will give rise to a slightly different estimate of the covariance matrix, compared to when we use the normalized values to estimate the sample covariance. However, the differences between the two estimates are generally very small, and when

the covariances are translated into correlations these differences will just manifest themselves as small numerical deviations (order of magnitude $<$ 1e-8 in our case).

The `normalizeMixed` function can also be called with the option `cohort=NULL`, which will fit a model with only a batch and a sample-within-batch effect.

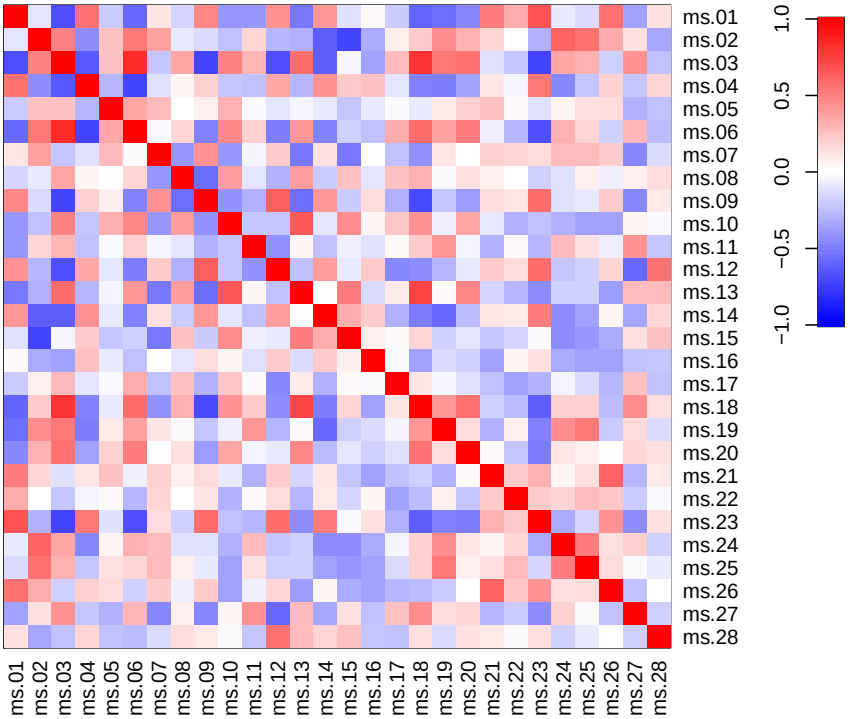
We can now take a look at the normalized data. Below is a boxplot for the normalized `Ystd` data. The samples show more similarity than prior to normalization.

```
> boxplot(YstdN,col=col.cohort,las=2,cex.axis=0.8)
```



We can also look at a new correlation heatmap, which seems more reasonable from a biological point of view.

```
>  cor.stdN <- cor(t(YstdN),method="pearson")
>  heatmap(cor.stdN)
```



Additional functionality exists in the `normalizeMixed` function. If the batches included have large differences in variance, this can be correct for using an initial model fit, by selecting the option `scale.bv=TRUE`. The option is set to `FALSE` by default. This option requires that the `nlme` package is installed.

It is also possible to run the mixed model normalization without the simultaneous estimation of a covariance matrix (suitable in situations where $p > n$). This is governed by selecting the option `iterative=F` when calling the `normalizeMixed` function.

References

[1] A. Jauhiainen, B. Madhu, J. Griffiths, and S. Tavaré. Normalization of metabolomics data with applications to correlation maps. *Bioinformatics*, 2014; 30(15):2155-61