Fusing prior knowledge with microbial metabolomics
Verouden, M.P.H.

Citation for published version (APA):
Verouden, M. P. H. (2012). Fusing prior knowledge with microbial metabolomics

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Maximum likelihood scaling (MALS)
3. Maximum likelihood scaling

Abstract

A filtering procedure is introduced for multivariate data that does not suffer from noise amplification by scaling. A maximum likelihood principal component analysis (MLPCA) step is used as a filter that partly removes noise. This filtering can be used prior to any subsequent scaling and multivariate analysis of the data and is especially useful for data with moderate and low signal-to-noise ratio’s, such as metabolomics, proteomics and transcriptomics data.

KEYWORDS: scaling, preprocessing, omics-data, noise, filtering.
3.1 Introduction

In many multivariate data analysis methods scaling is often performed as a pre-processing step. There are often good reasons to perform such a scaling step, for example prior to a principal component analysis (PCA) of the data \[58, 57\]. In many cases, the correlation between variables is more important than the covariance between these variables. In that case, autoscaling (AS) can be performed on the data to analyze the correlations. The problem with this type of scaling is that it does not take any noise characteristics into consideration. AS can easily result in noise amplification and an increase in heteroscedasticity of the data. This could lead to poor results for autoscaled PCA \[102\]. Also other data analysis methods, such as clustering and classification, suffer from noise amplification after scaling.

The noise amplification is especially a problem in data with a moderate or low signal-to-noise ratio. Even in data with a high signal-to-noise ratio AS has to be performed carefully in order not to amplify the noise in signal scarce domains. In the emerging post-genomic field, metabolomics, proteomics and transcriptomics data are becoming available. Such data usually has large variation in the dynamic range for the different variables with a moderate to low signal-to-noise ratio and we expect our method to perform well in these areas. In the literature, there are several papers \[59, 99, 103-105\] on how to deal with noisy data in a PCA context. The methods described in these papers calculate the most likely PCA solution taking the distribution of the noise into account: the maximum likelihood solution. These methods, however, focus on modeling whereas our procedure focuses on filtering. An example that a component model is used as a noise filter can be found in Chen et al. \[106\].

The method we propose here is a two-step method. It starts with calculating the above mentioned maximum likelihood solution, in the second step this solution is scaled. Subsequently the data can be analyzed with any data analysis tool. The advantage of this procedure is that in the first step a part of the noise is removed and the scaling step can be performed on data with reduced noise.

We shall demonstrate the superiority of MALS over ordinary scaling using two artificial test cases with different noise characteristics. Furthermore, MALS will be used on a real life metabolomics data set \[58\]. From this example, it will become clear how to use MALS in practice. As an analysis tool for the data we use PCA being one of the most used techniques in multivariate data analysis. Although in the literature there are many papers on scaling and many papers on noise in multivariate data analysis, this paper introduces a new method that deals with the problems arising from scaling noisy data.

3.2 Methods

In the following section we will first explain briefly how AS works and subsequently describe the MALS approach. PCA is chosen as an example of a much used multivariate analysis method, but the results carry over to other methods as well. Matrices are written as boldface uppercase characters; vectors as boldface
3. Maximum likelihood scaling

Figure 3.1: MALS, the first step filters the data and the second step provides the user with the desired view on the data.

3.2.1 Autoscaling (AS)

Let $X$ denote the matrix of $J$ variables obtained at $I$ experiments. AS comes down to mean-center the data and scale each column by its standard deviation:

$$x_{\text{scaled}}^{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

(3.1)

where $\bar{x}_j$ represents the mean of the $j$th column of $X$ and $s_j$ is the standard deviation of that column ($x_{ij}$ is the typical element of $X$). The matrix $X_{\text{auto}}$ contains the numbers $x_{\text{scaled}}^{ij}$.

3.2.2 Maximum likelihood scaling (MALS)

The first step of MALS is to partly remove noise from the original data. This is achieved by a MLPCA. After this first step the model $\hat{X}$ is obtained. Next, $\hat{X}$ can be scaled with an appropriate scaling method, for example AS. The MALS approach is depicted in the Figure 3.1. Subsequently, the data can be analyzed with an appropriate analysis method, for example PCA.

3.2.3 Maximum likelihood principal component analysis (MLPCA)

Standard PCA, see Jollife [96], minimizes the following objective function: $g(T, P) = ||X - TP^T||^2$ for preset dimensions of the matrices $T$ and $P$; where the matrix $X$ contains the measurements. MLPCA can use knowledge of error information of the variables and the error co-variation between variables. If there is no co-variation between the error of the variables then the problem can be simplified into a weighted PCA (WPCA) formulation: $gw(T, P) = ||W \circ (X - TP^T)||^2$, where $\circ$ denotes a Hadamard (element wise) product. Here the matrix $W$ has the same dimensions as the data and contains the reciprocal measurement uncertainty.
Several algorithms exist for MLPCA and WPCA [99,103,104,107,108]. In this paper we do not consider correlated error structures but only consider cases that can be solved using WPCA. The proposed strategy also holds for the general case. For solving the minimization problem a strategy is adopted similar to Schuermans et al. [107] The objective function is written in such a form that the minimization can be performed by the non-linear solver of MATLAB (lsqnonlin). The MATLAB m-files are available on the web: http://www.bdagroup.nl. We found this approach much faster than the available software for this type of problems, a conclusion that is in line with the results from Schuermans et al. [107] The minimization stops at a point where either the residuals or the solution do not change more than predefined stop criteria. In all simulations performed in this paper the criteria are reduced until the solution is stable.

3.3 Test cases

3.3.1 Artificial data

For illustration we start with the use of artificial data to compare our approach with straightforward AS the original (non-filtered) data. Artificial data can be made in an infinite numbers of ways. The first condition we put on our artificial data is that the values of variables vary over orders of magnitude. If the importance of the small variables is comparable to the importance of the large values it is logical to perform some type of scaling. Secondly, the number of experiments is made smaller than the number of variables. These are characteristics that are also found in for example metabolomics data.

The data matrix \( \mathbf{X}_{\text{true}} \) is constructed by defining two smaller matrices \( \mathbf{A} \) and \( \mathbf{B} \). The size of \( \mathbf{A} \) is chosen to be 6 by 2 and the elements of \( \mathbf{A} \) are drawn from a standard normal distribution. The size of \( \mathbf{B} \) is chosen to be 10 by 2 and the elements of \( \mathbf{B} \) are drawn from a uniform distribution between 0 and 1. To introduce the differences mentioned in the paragraph above the values of the rows 6 to 8 of \( \mathbf{B} \) are multiplied with \( 10^3 \) and rows 9 and 10 are multiplied with \( 10^5 \).

The matrices \( \mathbf{A} \) and \( \mathbf{B} \) that were obtained following the procedure described above were:

\[
\mathbf{A} = \begin{pmatrix}
-2.4634e-001 & +4.8530e-001 \\
+6.6302e-001 & -5.9549e-001 \\
-8.5420e-001 & -1.4967e-001 \\
-1.2013e+000 & -4.3475e-001 \\
-1.1987e-001 & -7.9330e-002 \\
-6.5294e-002 & +1.5352e+000
\end{pmatrix}
\]
3. Maximum likelihood scaling

\[
B = \begin{pmatrix}
9.0161e-001 & 9.8299e-001 \\
5.5839e-003 & 5.5267e-001 \\
2.9741e-001 & 4.0007e-001 \\
4.9162e-002 & 1.9879e-001 \\
6.9318e-001 & 6.2520e-001 \\
3.7589e+002 & 7.5367e+002 \\
9.8765e+000 & 7.9387e+002 \\
8.4472e+004 & 6.2080e+004 \\
3.6775e+004 & 7.3128e+004
\end{pmatrix}
\]

With these matrices the artificial data are created:

\[
X_{\text{true}} = AB^T
\]  (3.2)

The absolute values in \(X_{\text{true}}\) range from \(1.2429e-002\) to \(1.2847e+005\). The autoscaled \(X_{\text{true}}\) (\(X_{\text{true,auto}}\)), is the target (noiseless) data. The results of AS the noisy data and the MALS approach are compared with the true results. Two cases are considered, one with homoscedastic noise and the other with heteroscedastic noise.

\[
X = X_{\text{true}} + N_i
\]  (3.3)

Where \(i = 1\) indicates the homoscedastic case and \(i = 2\) is the heteroscedastic case.

3.3.2 Homoscedastic case

Here \(N_1\) consists of i.i.d. noise from a normal distribution with zero mean and a variance of 0.2. This leads to a signal-to-noise ratio in \(X\) ranging from \(10^{-6}\) to 10, which is in the order of magnitude of types of omics data. The autoscaled matrix \(X\) is denoted by \(\hat{X}_{\text{auto}}\). The difference between \(\hat{X}_{\text{auto}}\) and \(X_{\text{true,auto}}\) is the error using the traditional approach:

\[
\text{error} = \| \hat{X}_{\text{auto}} - X_{\text{true,auto}} \|^2
\]  (3.4)

For MALS the MLPCA for \(X\) is calculated. The number of principal components is two and the non-linear solver algorithm was used. The result is then autoscaled to obtain the matrix \(\hat{X}_{\text{MALS}}\). The difference between the true value and the MALS model is then given by:

\[
\text{error}_{\text{MALS}} = \| \hat{X}_{\text{MALS}} - X_{\text{true,auto}} \|^2
\]  (3.5)

The noise for \(X\) is generated 20000 times and both the MALS and the AS errors are calculated. The average error for the AS approach is 2.3761 while for the MALS approach this is 1.2690. Figure 3.2 shows the distribution of the errors, for MALS a solid line is used and for AS a dotted line is used. The error in the MALS approach is 97.835% of the time smaller than the AS approach.
3.3. Test cases

Figure 3.2: The error distribution of the homoscedastic test case, for the autoscaling (dotted line) and the MALS (solid line) approach.

Figure 3.3: The error distribution of the heteroscedastic test case, for the autoscaling (dotted line) and the MALS (solid line) approach.
3.3.3 Heteroscedastic case

In this case the matrix $N_2$ is constructed from $N$ that is a matrix with i.i.d. noise from a normal distribution with zero mean and a variance of 0.2 and from $X_{\text{true}}$:

$$N_2 = N \circ X_{\text{true}} \quad (3.6)$$

Again $\circ$ denotes a Hadamard product.

The error made by the AS approach is calculated using Equation 3.4 and the MALS error is obtained from Equation 3.5. The $X$ data are again generated 20000 times and both errors from AS and MALS are calculated. The average error is 0.91989 for AS and 0.55148 for MALS. Figure 3.3 shows the error distribution: the error in the MALS approach is in 96.7% of the time smaller than the AS approach. From Figures 3.2 and 3.3 it can be seen that MALS is superior to AS.

To illustrate the effect of the MALS approach on a multivariate analysis we performed a PCA on $X_{\text{true auto}}$, $\hat{X}_{\text{MALS}}$ and on $\hat{X}_{\text{auto}}$. We did this for ten homoscedastic noise realization as described in subsection 3.3.2 Homoscedastic case. As can be seen from the score plot, Figure 3.4, the $+$ from the MALS approach are much closer to the diamonds that represent the true solution than the $\bullet$ from the AS approach. Thus using MALS as a preprocessing tool gives better results than AS.

This is no surprise because Figure 3.2 already showed that MALS is on average better than AS. The loadings plot in Figure 3.5 shows a similar result. The MALS preprocessing result resemble the true results much better than the AS approach. The true results and the MALS follow the well-known circular trajectory of a 2 PC model fully describing the data. Some of the results from AS preprocessing are quite far from the true values thus they could lead to some false conclusions.

3.3.4 Real data

To demonstrate the possibility of performing MALS on a real metabolomics dataset we use the data described in van den Berg et al. [58]. These data consist of measurements of a part of the metabolome of *Pseudomonas putida* S12 grown on four different carbon sources. In this paper we used 124 metabolites that were measured with GC-MS. Experimental details can be found in van der Werf et al. [12]. In total we consider 13 batch fermentations. Three experiments had as carbon source fructose, five experiments ran on glucose, three on gluconate and two on succinate. In the remainder of this paper the fructose experiment are denoted by the numbers 1, 2 and 3, the glucose experiments by the numbers 4, 5, 6, 7 and 8. The gluconate experiments are the numbers 9, 10 and 11 and finally the succinate experiments are denoted by 12 and 13. The experiments on glucose, gluconate and succinate contain an analytical duplo. The duplos are respectively 4, 5 and 9, 10 and 12, 13.

3.3.5 Determination of the weight matrix $W$

In a complicated data set there are usually several sources of variation. The variation that is induced by the experimental design is usually the interesting variation.
3.3. Test cases

Figure 3.4: The scores of PC 1 plotted against the scores of PC 2 of 10 noise realizations for homoscedastic noise, the + indicate the MALS approach and the • indicate the AS results.

Figure 3.5: The loading of PC 1 plotted against the loading of PC 2 of 10 noise realizations for homoscedastic noise, the + indicate the MALS approach and the • indicate the AS results.
Other sources of variation, for example biological, sampling and analytical are usually unwanted. A filtering method should diminish the unwanted variation to make the effects due to the experimental design more clear. The weighing matrix $W$ should reflect on the type of variation that should be filtered. The most common weighing procedures are based on analytical error [105, 109]. In these two papers it is argued that the analytical error for larger values is proportional to the signal strength of the analytical device and for small values it is constant. We follow the same approach. We take the elements in $W$ inversely proportional to their values. For the zero values this obviously would lead to problems. Therefore, we introduce a cutoff value of $10^{-4}$, this means all values smaller than $10^{-4}$ are given a weight of $10^4$. A sensitivity analysis of the exact cutoff value showed that this is not a sensitive parameter.

### 3.3.6 Determination of the number of principal components

Determining the number of principal components in MALS is not a trivial task. The approach we take here is just one of the possibilities. A first consideration is the number of PC’s in the ultimate model. The number of PC’s in MALS should be larger than the number in the ultimate model. From a scree plot of the autoscaled data it becomes clear that we need at least 3 PC’s (Figure 3.6).

In the dataset there are three analytical duplos. The weights in the matrix $W$
are chosen in such a way that the analytical error should be filtered out. Thus the analytical duplos should be closer to each other after the filtering procedure. So for all possible number of PC’s the norm of the difference between the duplos is calculated. The results are presented in Figure 3.7.

From Figure 3.7 it can be seen that the filtering for the glucose and succinate does not reduce the difference between the duplos much. Note that the point for 13 PC’s is equal to that of the original data because 13 PC’s describe the data completely. For gluconate the filter brings the duplos closer together for especially 5 and 6 PC’s. The minimum of the average curve in Figure 3.7 is at 6 PC’s. We therefore took 6 PC’s for the MALS filter. To show the effect on a subsequent performed AS and PCA in Figure 3.8 we present the score plots for AS and MALS. From this score plot it can be concluded that MALS with PCA give a different picture compared to AS with PCA. For both methods the group structure is present in Figure 3.8. But the grouping 1,2,3 and 4,5,6,7,8 and 9,10,11 and 12, 13 is tighter for the MALS approach. Figure 3.8 is also consistent with Figure 3.7, the distance between duplos 4, 5 and 12, 13 are not altered much but the duplos 9, 10 is much closer. Because we do not know the true solution it is impossible to say which method is closer to the truth. But because of theory and the results of the simulated data for which the true solution is known it is more likely that MALS is less influenced by analytical errors.
3. Maximum likelihood scaling

3.4 Discussion

On the basis of theoretical arguments, MALS should improve the results of a subsequent scaling step due to its noise filtering capability. The artificial test cases indeed show that MALS yields superior results. In more than 97% of the simulations for both cases MALS performs better and the overall mean error is about halved. The rationale behind MALS is that it acts as a filter that reduces the noise that corrupts the data. From this filtered data, a scaling step can be performed to obtain the desired view. This scaling step is not limited to AS but the arguments put forward in favor of MALS should hold for any kind of scaling, like for example level scaling or range scaling. Moreover, MALS can be combined with any subsequent multivariate analysis method. Hence, MALS is a general preprocessing tool.

We would like to make some comments on practical issues concerning MALS. First, in order to perform MALS, estimates of the unwanted variation are needed. In our opinion the best way to get such error estimates is to perform repeated measurements. If these cannot be performed, then a priori knowledge of the instrument

Figure 3.8: The score plots for AS and MALS. + Indicate MALS. • Indicate AS; 1,2,3: fructose; 4,5,6,7,8: glucose; 9,10,11: gluconate; 12,13: succinate. The duplos in the data are 4,5 and 9,10 and 12,13.
on which the measurements were performed should generate such estimates. Secondly, the number of PC’s in the MALS filtering should be such that the filtering is effective in the sense that the unwanted noise is reduced. Thirdly, if the measurement error is homoscedastic, a normal PCA is equivalent to a MLPCA. Hence, the problem can be solved with easier and faster software.

It is shown in this paper that it is possible to perform MALS on a real metabolomics dataset. The only drawback that is present using MALS is CPU time. While for an ordinary AS the computing time is in the order of seconds, the CPU time for MALS is in the order of hours for a dataset with the dimensions of the metabolomics data.

The MATLAB m-files that have been used to perform the calculation in this paper are available at http://www.bdagroup.nl. To execute these files the MATLAB optimization toolbox is required.

Acknowledgements

We thank Mariët J. van der Werf and Robert A. van den Berg, TNO Quality of Life, Zeist the Netherlands for the metabolomics data.