

---

# Active and passive learning to explore a complex metabolism data set

Susan J. Simmons<sup>1</sup>, Xiaodong Lin<sup>2</sup>, Chris Beecher<sup>3</sup>, and Young Truong<sup>4</sup> and S. Stanley Young<sup>5</sup>

<sup>1</sup> University of North Carolina at Wilmington, Wilmington, NC 28403  
[simmonssj@uncw.edu](mailto:simmonssj@uncw.edu)

<sup>2</sup> Statistical and Applied Mathematical Sciences Institute, RTP, NC 27709  
University of Cincinnati, Cincinnati, OH 45221 [linxd@samsi.info](mailto:linxd@samsi.info)

<sup>3</sup> Metabolon, Research Triangle Park, NC 27709 [cbeecher@metabolon.com](mailto:cbeecher@metabolon.com)

<sup>4</sup> University of North Carolina, Chapel Hill, NC 27599 [truong@bios.unc.edu](mailto:truong@bios.unc.edu)

<sup>5</sup> National Institute of Statistical Sciences, Research Triangle Park, NC 27709  
[young@niss.org](mailto:young@niss.org)

**Summary.** Metabolomics is the omics science concerning biochemistry. Metabolomic datasets include the quantitative measurements of all small molecules, known as metabolites, in a biological sample. These dataset are rich in information regarding dynamic metabolic (or biochemical) networks that are unattainable using classical methods and has great potential, conjointly with other omic data, in understanding the functionality of genes. Herein, we explore a complex metabolomic dataset with the goal of using the metabolic information to correctly classify individuals into different classes. Unfortunately, these datasets incur many statistical challenges: the number of samples is less than the number of metabolites; there is missing data and non-normal data; there are high correlations among the metabolites. Thus, we investigate the use of robust singular value decomposition, rSVD, and recursive partitioning to understand this metabolomic data set. The dataset consists of 63 samples, in which we know the status of the individuals (disease or normal and for the diseased individuals, if they are on drug or not). Clustering using only the metabolite data is only modestly successful. Using distances generated from multiple tree recursive partitioning was more successful.

## 1 Introduction

Complex data sets are becoming available in the post-genomic era; there is an increasing interest in the analysis of genetic information (genomics), and their transcription (transcriptomics) and subsequent translation into protein (proteomics). The recently emerging science of metabolomics completes the primary omics ladder of DNA, RNA, proteins and metabolites. As such, it provides a prototypic means of understanding of cellular function. Metabolomics

is the biochemical profiling of all small molecules, biochemicals or metabolites, in an organism. Even though the dynamic nature of metabolites makes them difficult to measure, recent advances in technology allows robust quantification of the concentrations of hundreds of metabolites from a biological sample [1]. This new omics science offers insight into the dynamic interactions in metabolic pathways and an unambiguous representation of cellular physiology [2]. Furthermore, patterns of metabolites can be used to identify biomarkers of specific disease, understand pathological development, and propose targets for drug intervention.

Data obtained through a metabolomic experiment poses a number of challenges to statistical modeling. The number of metabolites measured ( $=p$ ), usually in the hundreds is much larger than the number of biological samples ( $=n$ ):  $n \ll p$ . Additionally, there may be severe distributional difficulties such as non-normal distributions, outliers (unusual data values), missing values, and high correlations among metabolites. Common objectives are finding patterns in the data, in particular, clustering of the biological samples (rows) into groups with similar metabolic expression profiles; and clustering the metabolic concentrations (columns) into groups where the level of metabolic expression is similar in the samples.

Due to the challenges of modeling this dataset, many clustering techniques may produce erroneous results. For example, most clustering techniques are influenced by outliers and can not accommodate missing values. Further disadvantages of hierarchical clustering techniques are that the dendrograms produce orderings of rows (biological samples) that are not unique [3]. Thus, we investigate two methods that have appealing features and overcome these challenges. The first method is rSVD that was proposed by Liu et al. [3]. This technique is by-product of 'ordination', which involves finding suitable permutations of the rows and columns that lead to a steady progression of data values going down the rows and across the columns. The clusters are determined by placing vertical and possibly horizontal lines in the dataset that divide it into homogeneous blocks. The second method we examine that is useful in overcoming these statistical challenges is recursive partitioning, RP, which provides a powerful classification algorithm resulting in a tree diagram. The tree diagram identifies the metabolites that are useful in partitioning the sample into smaller, more homogeneous groups. RP uses information on the classification of the objects whereas rSVD uses only the information on the metabolites.

We explore these two techniques on a metabolomic dataset. The dataset contains 63 biological samples in which there are four different groups. There are two primary groups, diseased and healthy individuals. We will attempt to find other groups within the data set. The samples were analyzed using an ESA 16-channel Coularray detector. This detector has a technical robustness that is comparable, possibly even better than conventional mRNA [4]. The data contains blanks where the metabolic concentrations were below detection limits. The goal of this study is to use rSVD and recursive partitioning

in an exploratory manner to identify metabolites that cluster samples into consistent groups based on their biological function.

## 2 Robust Singular Value Decomposition

The robust clustering technique described by Liu et al. [3] is used to cluster the metabolomic data by biological samples and metabolic concentrations. The method involves a systematic approach of ordering the rows and columns so the underlying homogeneous clusters may be found. The technique involves approximating the dataset, which we will denote as  $X$ , with a bilinear form. The array  $X$  is viewed as an  $n$  by  $p$  array with  $n$  representing the number of biological samples (rows) and  $p$  representing the metabolic concentrations (columns). Thus, the approximate bilinear form is

$$x_{ij} = r_i c_j + e_{ij}, \quad (1)$$

where  $r_i$  is a parameter corresponding to the  $i$ th biological sample,  $c_j$  corresponds to the  $j$ th concentration, and  $e_{ij}$  is a 'residual'. Estimating the array with this bilinear form allows the dataset to be permuted by ordering the  $r_i$  values of the rows and the  $c_j$  values of the columns, which results in an array with high and low values in the corners and medium values in the middle, leading to an informative display [3].

Subsequently, grouping values of  $r_i$  that are similar will result in clusters of biological samples, and grouping similar values of  $c_j$  will give clusters of metabolites. If the residuals are small so that the  $r_i c_j$  captures all the important structure of the data matrix, then the ordination and ensuing clustering using the  $r$  or  $c$  values is essentially unique.

The method proposed by Liu et al [3] focuses on an algorithm proposed by Gabriel-Zamir [5] that uses an alternating least squares (ALS) approach to reconcile the problem of missing values. This algorithm begins with an initial estimate of the column factors  $c_j$  which are used to provide a matching scaling for the rows. Viewing

$$x_{ij} = r_i c_j + e_{ij} \quad (2)$$

as a regression of the  $i$ th row of  $X$  on the column factors identifies  $r_i$  as the coefficient of a no-intercept regression. The regression is fit row by row using all non-empty cells, which results in an estimate of the row factors  $r_i$ . The algorithm proceeds by switching the roles of the rows and columns so that bilinear form is now regarded as a regression of the  $j$ th column of  $X$  on the row factors. The regression is fit column by column using all non-empty cells in exactly the same way to calculate fresh estimates of the column factors  $c_j$ . This approach uses all the observed data, and does not require imputation of missing data.

The ALS algorithm is effective in solving the missing information problem; however, it does not address the issue of sensitivity to outliers. Using a robust regression method instead of ordinary least squares (OLS) in the alternating regressions can solve this problem. Various forms of outlier-resistant regression methods are possible, such as L1 [10], weighted L1 [7], least trimmed squares [8], or an M-estimation method. In this analysis, we choose to use the L1 method proposed by Hawkins et al. [10].

Thus, the resulting alternating robust fitting, ARF, algorithm handles missing information smoothly, without requiring a separate 'fill-in' step. And it is impervious to a minority of outlier cells. Outliers will, of course, create a problem for the ARF, as with almost any conceivable method, if they constitute the majority of the elements of any row or column.

Finding  $k$  clusters based on biological samples may be found by sorting the rows by their  $r_i$  values and finding 'breakpoints'  $b(0) = 0 < b(1) < b(2) < \dots < b(k-1) < b(k) = n$  and allocating to cluster  $h$  those metabolites which, in the reordering, have index  $b(h-1) < i \leq b(h)$ . The breakpoints need to be chosen so that the biological samples within each cluster have  $r_i$  values as similar as possible. This can be made operational by the criterion that the pooled sum of squared deviations of the  $r_i$  broken down into the  $k$  clusters should be a minimum [3]. Exact algorithms for finding breakpoints to attain this minimum are given by Steel and Venter [9], and by Hawkins [10]. Similarly, applying the optimal segmentation algorithm to the column factors  $c_j$  clusters the metabolites into any specified number of clusters such that the metabolites within clusters have  $c_j$  values as similar as possible.

Figure (1) illustrates the log transform of the raw data. Rows correspond to biological samples and the columns correspond to metabolites. In order to create the following illustration, many outliers need to be truncated. We will proceed with the rSVD analysis by permuting the rows and columns of this matrix and we will display the smoothed result by plotting the outer product of the first eigenvectors.

Figure (2) illustrates the outer product of the sorted first eigenvector pairs. Notice the shaded cluster around the lower left corner of the figure; this is the strong response outlier group, including the four samples X44, X20, X43 and X48, Group 4. The other disease groups, however, are not as clear even using the second and third eigenvector pairs. To further identify the biological clusters of the samples, we investigate recursive partitioning.

### 3 Recursive Partitioning, Results and Discussion

Recursive partitioning is an algorithm that searches through all of the variables (metabolites) and identifies the single variable produces the best 'split' among groups (samples). The test criterion for the best split is based on t-tests for two-way splits and F-tests for multi-way splits, and thus it quickly

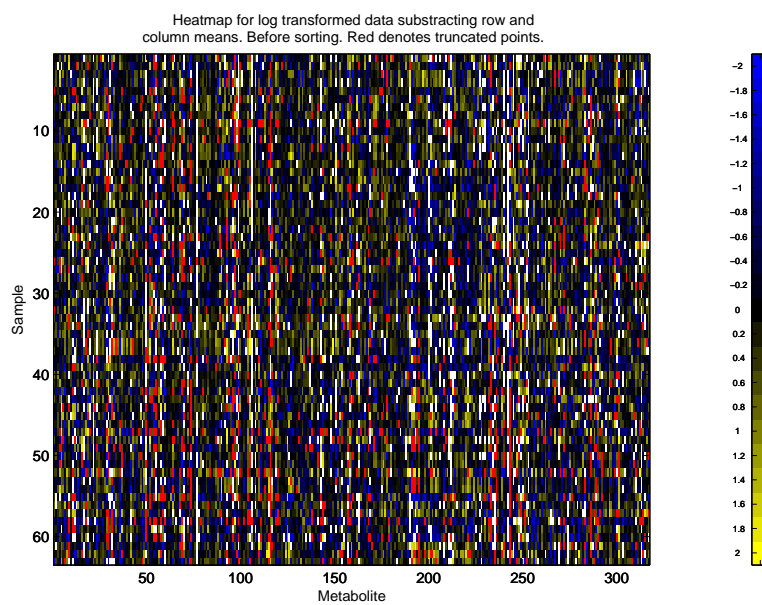


Fig. 1. Heat map of the assay values for 63 samples and 317 metabolites.

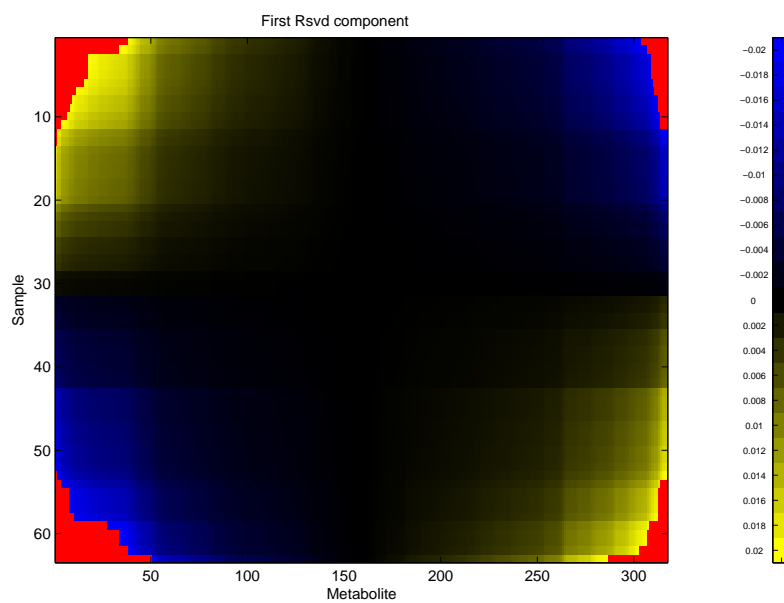


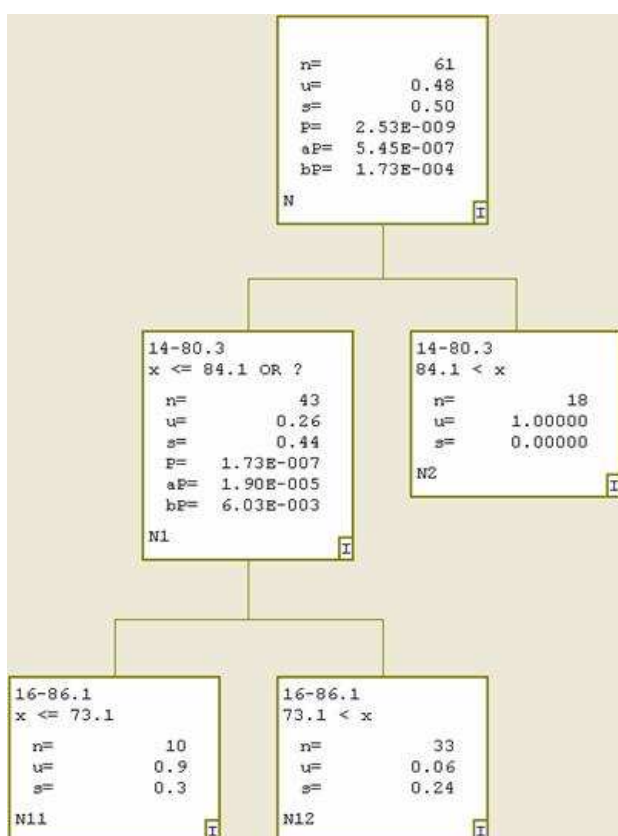
Fig. 2. Heatmap constructed from the outer product of the first row and the first column eigenvectors.

and efficiently decides the best split. When a split occurs, the group is broken into two or more daughter nodes, which are more homogeneous than the starting group. This process is repeated until no significant splits are possible. The result of this algorithm is a tree diagram that identifies the metabolites used to create each split, and the resulting groups or daughter nodes. We use the commercially available software FIRMPPlus™ [11] to perform the RP analysis.

In Figure 3 we give one tree from a multiple-tree analysis of the metabolomic data set. Node N contains 61 observations (duplicate observations, X26 and X26b were removed as outliers for this analysis). The samples are classified as normal, 0, or diseased, 1, so the average for Node N ( $u=$ ) is the probability that the sample is from a diseased person. The first step of the algorithm splits the data set into two groups depending on metabolite 14-80.3. If the level of this metabolite is greater than 84.1, the sample is placed into Node  $N2$ . Notice that this Node contains 18 individuals and all of them have the disease. If the level of the metabolite is less than or equal to 84.1, then the sample is placed into Node  $N1$ . In this node, 26 percent of the people have the disease. In Node  $N1$ , the header 14-80.3 indicates the metabolite used for the split. The second line of the header is  $x \leq 84.1$  or ?, which indicates that this node constitutes biological samples with values less than or equal to 84.1 for metabolite 14-80.3 and any samples with missing values for metabolite 14-80.3. The individuals with missing values for metabolite 14-80.3 are placed in the daughter group they are most similar to,  $N1$  or  $N2$ . If they were statistically distinct from those groups, then they would be put into a new group. This treatment of missing values allows them to be predictive in classifying the samples.

Three p-values are used to judge the validity of a split; the unadjusted p-value ( $p=$ ), computed without regard to multiple testing; the aP-value, computed to adjust for the segmentation; and the bP-value, which is a Bonferroni adjustment reflecting the number of predictors under consideration. In this dataset, there are 317 continuous predictors. We see that all the p-values in Node N are suitably small. As Node  $N2$  is pure, it is not split. Node  $N1$  is examined and after looking at all predictors and split points, the algorithm selects metabolite 16-86.1 as the split variable and a split point of 73.1. The p-values for this split are suitably small. 9/10 individuals in Node  $N11$  are diseased and 2/33 in Node  $N12$  are diseased. Of the 61 samples, three are misclassified. The algorithm examines both Nodes  $N11$  and  $N12$  looking for additional splits and does not find any more significant splits ( $bP < 0.05$  was chosen as the significance level).

The first method (ALS) attempts to identify clusters of sample without using their class labels; the second method (RP) is very different as it uses the disease indicator (0=control, 1=disease). In Figure 2, the last 8 samples shown in the lower left corner of the heat map are (from bottom up): X43, X48, X44, X20, X23, X4, X31 and X26. The first four of these constitute a clinically recognized sub-set of the disease population. The next three are



**Fig. 3.** A single recursive partitioning tree from a 100-tree analysis.

arguably within this class but were not clinically recognized. Case X26 was ultimately considered an outlier.

The corresponding first 20 metabolite from left to right are: 5-80.2, 5-34.8, 12-7.9, 12-59.1, 9-67.9, 4-63.9, 14-20.1, 3-77.8, 4-44.1, 12-21.6, 14-64.1, 5-43.5, 6-29.8, 16-8.0, 3-78.2, 16-34.9, 7-70.0, 5-77.0, 14-29.0, 5-65.2. Biochemical experts had recognized some of these as significant markers for the disease population; RP found previously recognized metabolites and found new important metabolites.

For data sets with a relatively small number of observations, it is advantageous to build many trees. FIRMPPlus builds multiple trees by randomizing the split variable used at each split point over a specified number of statistically significant variables. We use the default of 10. We build 100 trees. From the forest of trees there are a number of things that can be learned about the predictor variables and the samples. Variables that are used often over the forest of trees are important. Variables that are seldom or never used

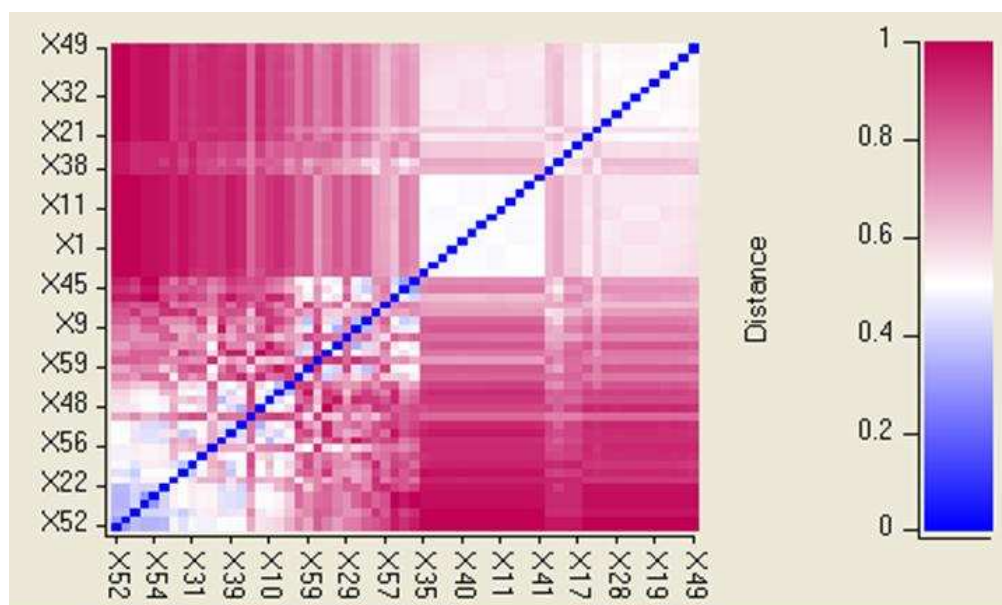
together in a tree are likely to be correlated. Variables that are used in the same tree are likely to be synergistic. Figure 4 gives the frequency that a variable is used in the forest of trees along the diagonal. We see that variable 16 – 86.1 is used often (The statistic given is the fraction of the data in a tree controlled by the variable). For example, once 16 – 86.1 is used, 3 – 44.5 is never used. If 5 – 80.2 is used, then 16 – 86.1 is often used, i.e. the variables are synergistic. The upper part of the matrix gives the joint incidence of two variables being used together in a tree and the lower triangle gives the joint incidence in standard deviation units. We can see a number of variable pairs, e.g. 4 – 37.9 and 9 – 7.1, 3 – 44.5 and 5 – 65.2.



Fig. 4. Joint incidence of variables used in 100 TP trees.

Next we can judge how similar two samples by using a RP tree-based distance. Distance is computed as the proportion of the total sample of the smallest node where two observations occur together. If they only occur together in the parent node the proportion, distance, is 1.00. If they always occur

together in a small terminal node, then the proportion, distance, is small. If they are immediately sent to different daughter nodes on the first split, then they are dissimilar. If they always occur together in a terminal node, then they are similar. Figure 5 gives a heat map of the similarity of samples.



**Fig. 5.** RP-based distance matrix of the 61 samples.

Each sample is perfectly similar to itself so we have shaded dots through the diagonal. We also see interesting patterns of three or four groups. The upper right hand group is of normal individuals. The group in the center of Figure 5 is of diseased, drug-treated individuals. The lower left hand group is of diseased, non-drug-treated individuals. The small shaded group in the lower left hand of Figure 5 is of individuals that appear to have a somewhat different manifestation of the disease. There is an ongoing effort to better identify the nature of these individuals.

## References

1. Stitt, M. and Fernie, A. R. (2003). From measurements of metabolites to metabolomics: an 'on the fly' perspective illustrated by recent studies of carbon-nitrogen interactions. *Current Opinion in Biotechnology*, **14**, 136-144.
2. Beecher, C. (1995). Metabolomics: The Newest of the 'omics' Sciences. *Innovations of Pharmaceutical Technology*, 57-64.
3. Liu, L., Hawkins, D. M., Ghosh, S., Young, S. S. (2003). Robust singular value decomposition analysis of microarray data. *PNAS*, **23**, 13167-13172.
4. Weckwerth, W. (2003). Metabolomics in systems biology. *Annual Review of Plant Biology*, **54**, 669-689.
5. Gabriel, K. R., Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights *Technometrics*, **21**, 489-498.
6. Hawkins, D. M., Liu, L., Young, S. S. (2001). Robust Singular Value Decomposition. *NISS Technical Report*, **122**, [www.niss.org/downloadabletechreports.html](http://www.niss.org/downloadabletechreports.html).
7. Croux, C., Filzmoser, P., Pison, G. and Rousseeuw, P.J. (2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, **13**(1), 23-26.
8. Ukkelberg, A and Borgen, O. (1993). Outlier detection by robust alternating regression. *Analytica Chimica Acta*, **277**, 489-494.
9. Venter, J. H. and Steel, S. J. (1996). Finding multiple abrupt change points. *Computational Statistics and Data Analysis*, **22**, 481-504.
10. Hawkins, D.M. (2001). Fitting multiple change-points to data *Computational Statistics and Data Analysis*, **37**, 323-341.
11. *FIRMPlus<sup>TM</sup>*: Golden Helix Inc, Bozeman, MT, USA. <http://www.goldenhelix.com/>.