Metabolomic Data Analysis with MetaboAnalyst 5.0

Name: guest15444671027745781

October 22, 2021

1 Data Upload and Integrity Checking

1.1 Upload your data

For statistical analysis involving complex metadata, MetaboAnalyst accepts data table and **metadata table** uploaded as two comma separated values (.csv) files. Samples can be in rows or columns for data file. The metadata table must have the same sample names. For time-series data, the time points group must be named as **Time** and **Subject**. Users need to specify the data types when uploading their data in order for MetaboAnalyst to select the correct algorithm to process them. Table 1 summarizes the result of the data checking steps.

```
Samples are in rows and features in columns The uploaded data file contains 45 (samples) by 180 (compounds) data matrix.
```

Table 1: Summary of data processing results

	Features (positive)	Missing/Zero	Features (processed)
Bleached_Hot_Day37_2	180	0	180
Bleached_Hot_Day37_3	180	0	180
Bleached_Hot_Day37_4	180	0	180
Bleached_Hot_Day37_5	180	0	180
Bleached_Hot_Day52_1	180	0	180
Bleached_Hot_Day52_2	180	0	180
Control_Ambient_Day37_2	180	0	180
Control_Ambient_Day37_3	180	0	180
Control_Ambient_Day37_4	180	0	180
Mortality_Hot_Day0_4	180	0	180
Mortality_Hot_Day0_5	180	0	180
Mortality_Hot_Day37_1	180	0	180
Mortality_Hot_Day52_3	180	0	180
Mortality_Hot_Day52_4	180	0	180
Mortality_Hot_Day52_5	180	0	180
Bleached_Hot_Day0_4	180	0	180
Bleached_Hot_Day0_5	180	0	180
Bleached_Hot_Day37_1	180	0	180
Bleached_Hot_Day52_3	180	0	180
Bleached_Hot_Day52_4	180	0	180
Bleached_Hot_Day52_5	180	0	180
Control_Ambient_Day37_5	180	0	180
Control_Ambient_Day52_1	180	0	180
Control_Ambient_Day52_2	180	0	180
Mortality_Hot_Day37_2	180	0	180
Mortality_Hot_Day37_3	180	0	180
Mortality_Hot_Day37_4	180	0	180
Mortality_Hot_Day37_5	180	0	180
Mortality_Hot_Day52_1	180	0	180
Mortality_Hot_Day52_2	180	0	180
Bleached_Hot_Day0_1	180	0	180
Bleached_Hot_Day0_2	180	0	180
Bleached_Hot_Day0_3	180	0	180
Control_Ambient_Day0_1	180	0	180
Control_Ambient_Day0_2	180	0	180
Control_Ambient_Day0_3	180	0	180
Control_Ambient_Day0_4	180	0	180
Control_Ambient_Day0_5	180	0	180
Control_Ambient_Day37_1	180	0	180
Control_Ambient_Day52_3	180	0	180
Control_Ambient_Day52_4	180	0	180
Control_Ambient_Day52_5	180	0	180
Mortality_Hot_Day0_1	180	0	180
Mortality_Hot_Day0_2	180	0	180
Mortality_Hot_Day0_3	180	0	180

1.2 Data Integrity Check

Before data analysis, a data integrity check is performed to make sure that all the necessary information has been collected. Compound concentration or peak intensity values should all be non-negative numbers. By default, all missing values, zeros and negative values will be replaced by the half of the minimum positive value found within the data (detection limits).

```
Samples are in rows and features in columns The uploaded data file contains 45 (samples) by 180 (compounds) data matrix. The data is time-series data. 3 groups were detected from primary meta-data factor: Treatment. Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed. Other special characters or punctuations (if any) will be stripped off. All data values are numeric. A total of 0 (0%) missing values were detected. By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables
```

1.3 Data Normalization

The data is stored as a table with one sample per row and one variable (bin/peak/metabolite) per column. The normalization procedures implemented below are grouped into four categories. Sample specific normalization allows users to manually adjust concentrations based on biological inputs (i.e. volume, mass); row-wise normalization allows general-purpose adjustment for differences among samples; data transformation and scaling are two different approaches to make features more comparable. You can use one or combine both to achieve better results.

The normalization consists of the following options:

1. Row-wise procedures:
    - Sample specific normalization (i.e. normalize by dry weight, volume)
    - Normalization by the sum
    - Normalization by the sample median
    - Normalization by a reference sample (probabilistic quotient normalization)[1]
    - Normalization by a pooled or average sample from a particular group
    - Normalization by a reference feature (i.e. creatinine, internal control)
    - Quantile normalization
2. Data transformation :
    - Log transformation (base 10)
    - Square root transformation
    - Cube root transformation
3. Data scaling:
    - Mean centering (mean-centered only)
    - Auto scaling (mean-centered and divided by standard deviation of each variable)
    - Pareto scaling (mean-centered and divided by the square root of standard deviation of each variable)
    - Range scaling (mean-centered and divided by the value range of each variable)

Figure 1 shows the effects before and after normalization.

[1] Dieterle F, Ross A, Schlotterbeck G, Senn H. *Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics*, 2006, Anal Chem 78 (13);4281 - 4290

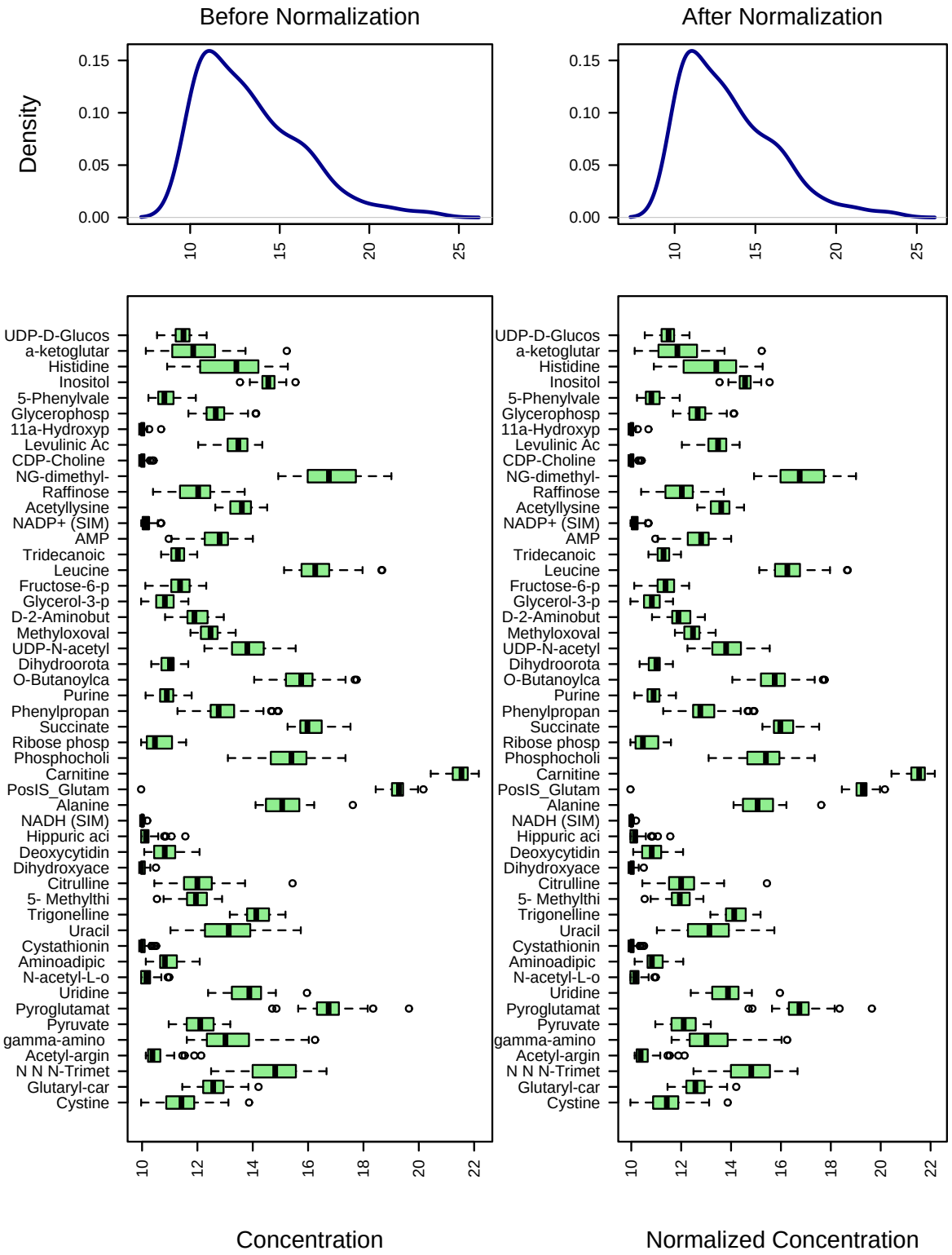

Figure 1: Box plots and kernel density plots before and after normalization. The boxplots show at most 50 features due to space limit. The density plots are based on all samples. Selected methods : Row-wise normalization: N/A; Data transformation: N/A; Data scaling: N/A.

2 Statistical and Machine Learning Data Analysis

For two-factor and time-series data, MetaboAnalyst offers several carefully selected methods for general two-factor and time-series data analysis. They include:

- Data overview:
  - Interactive Principal Component Analysis (iPCA)
  - Two-way Heatmap clustering and visualization
- Univariate method:
  - Two-way between/within-subjects ANOVA
- Multivariate approaches
  - ANOVA-Simultaneous Component Analysis (ASCA)
  - Multivariate Empirical Bayes Analysis (MEBA)

`Please note: MEBA is only applicable to time-series data analyais.`

2.1 Two-way Heatmap Visualization

The heatmap provides direct visualization of all data points in the form of colors squares. The color spectrum intuitively indicates the higher or lower values. Users can choose different clustering algorithms or distance measures to cluster the variables. The samples are ordered by the two factors with default the first factor used for primary ordering. Users can choose to switch the order.

Figure 2 shows the clustering result in the form of a heatmap.

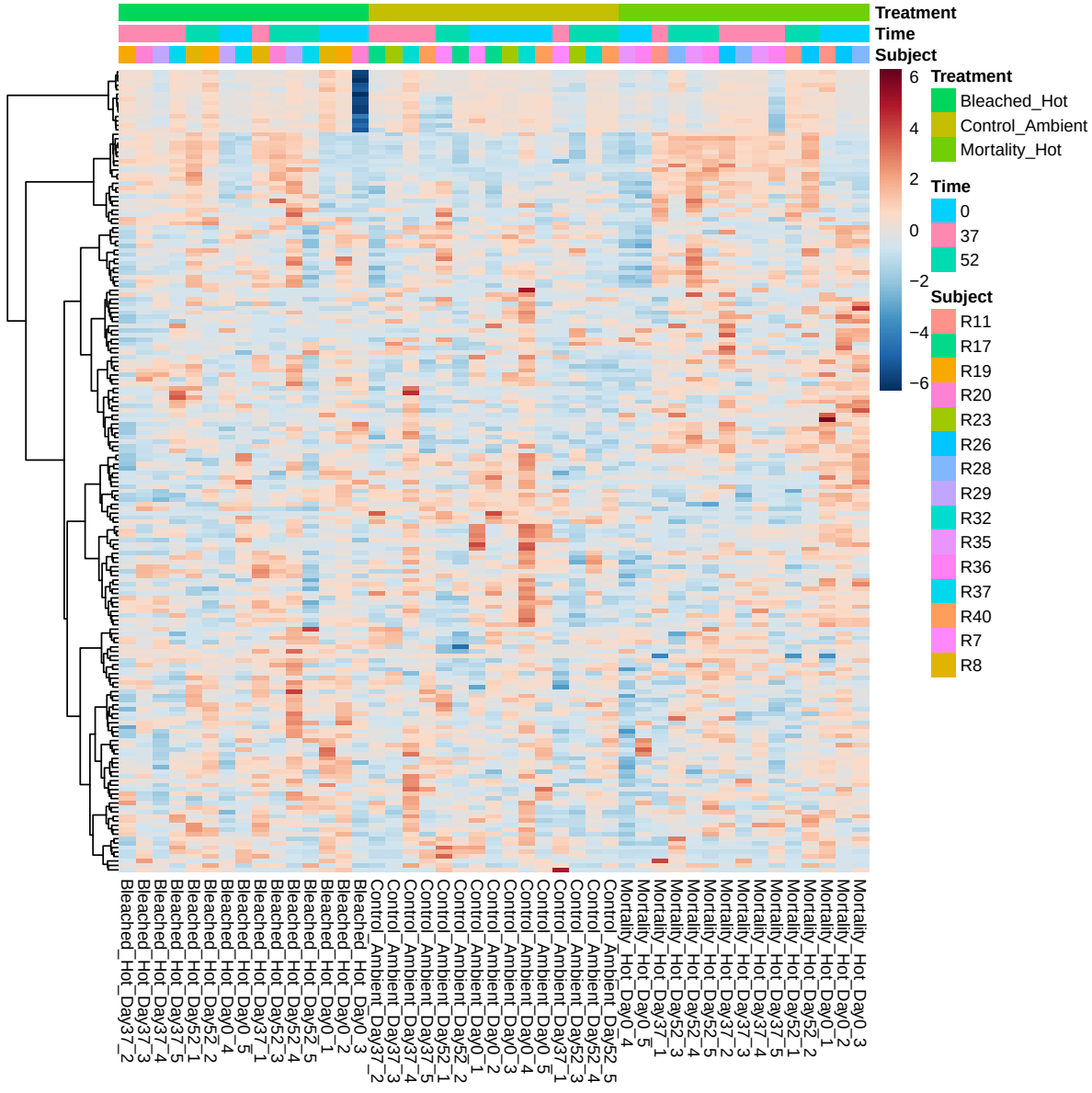

Figure 2: Clustering result shown as heatmap (distance measure using `euclidean`, and clustering algorithm using `ward.D`).

2.2 Univariate Analysis

Univariate analysis methods are the most common methods used for exploratory data analysis. For two-factor data, the basic approach is two-way ANOVA. There are two options - between-subjects ANOVA and within-subjects ANOVA. When samples are all from independent subjects (i.e. general two-way ANOVA), the between-subjects option should be selected. However, time series data contains samples measured from the same subjects from different time points. Therefore within-subjects ANOVA should be used.

Figure 3 shows the important features identified by ANOVA analysis. Table 2 shows the details of these features;

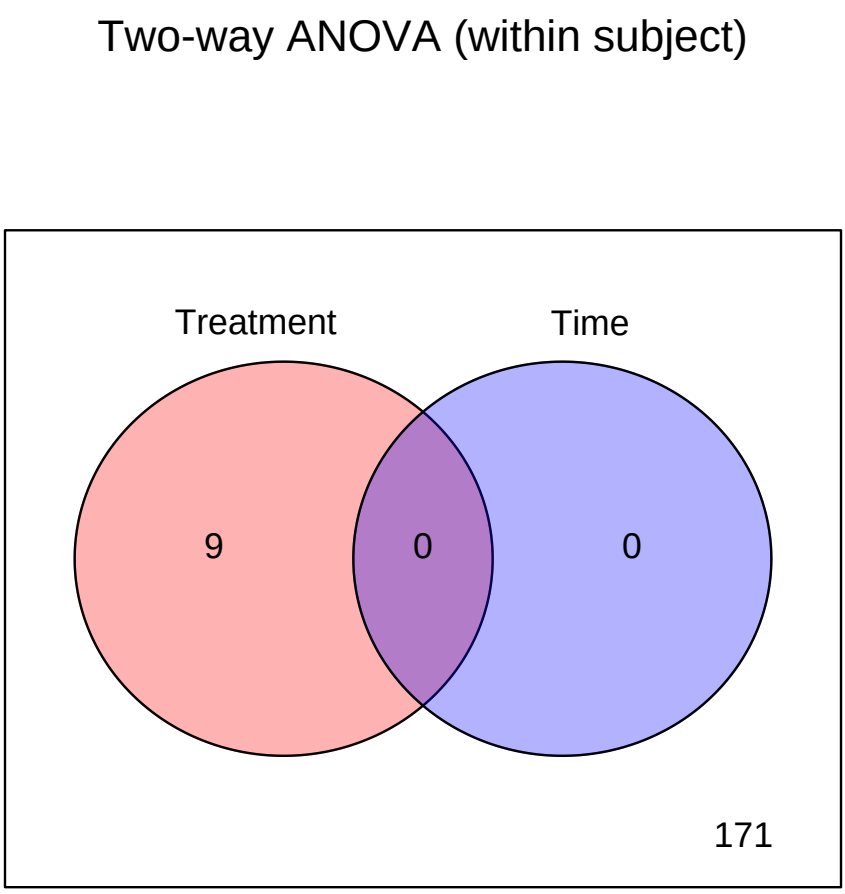

Figure 3: Plot of important features selected by two-way ANOVA.

Table 2: Important features identified by Significant features identified by advanced ANOVA

	Compounds	Treatment(F.val)	Treatment(raw.p)	Treatment(adj.p)	Time(F.val)	Time(raw.p)	Time(adj.p)
1	Adenine	10.564	0.00020727	0.037308	0.261	0.77158	0.99791
2	4-Guanidinobutanoic acid	8.6544	0.00075294	0.047274	0.13271	0.87611	0.99791
3	Acetyl proline	8.5894	0.00078794	0.047274	0.20282	0.81726	0.99791
4	O-Phosphorylethanolamine	7.6557	0.0015308	0.047274	0.23815	0.78919	0.99791
5	NADPH (SIM)	7.526	0.0016816	0.047274	0.63193	0.53679	0.99791
6	N-acetyl-glutamine	7.4036	0.0018385	0.047274	1.1508	0.32665	0.99791
7	L-Palmitoylcarnitine	7.2796	0.002013	0.047274	0.073741	0.92904	0.99791
8	Glucuronic acid	7.2038	0.0021282	0.047274	2.3233	0.11103	0.99791
9	Arginine	7.0614	0.0023637	0.047274	0.26418	0.76917	0.99791

2.3 ANOVA - Simultaneous Component Analysis (ASCA)

ASCA is a multivariate extension of univariate ANOVA approach. It is designed to identify the major patterns associated with each factor. This implementation supports ASCA model for two factors with one interaction effect. The algorithm first partitions the overall data variance (X) into individual variances induced by each factor (A and B), as well as by the interactions (AB). The formula is shown below with (E) indicates the residual Errors:

$$\mathbf{X = A + B + AB + E}$$

The SCA part applies PCA to A, B, AB to summarize major variations in each partition. Users then detect the major pattern by visualizing the PCA scores plot. MetaboAnalyst also provides model validation to test the significance of the effects associated with main effects. It is based on the Manly's unrestricted permutation of observation then calculate the permuted variation associated with each factor Finally, the permuted values are compared with the original variations The significant variables are identified based on the leverage and the Squared Prediction Errors (SPE) associated with each variables Variables with low SPE and higher leverage are modeled well after the major patterns.

Figure 4 shows the scree plots for each effect model. Figure 5 shows the major patterns associated with factor A. Figure 6 shows the major patterns associated with factor B. Figure 7 shows the major patterns associated with interaction. Figure 8 shows the results of model validations through permutations. Figure 9 shows the important features associated with factor A. Figure 10 shows the important features associated with factor B. Figure 11 shows the features that are important in the interaction.

Table 3 shows features well-modelled by Treatment. Table 4 shows features well-modelled by Time. Table 5 shows features well-modelled by Interaction model. The other details are available as .csv documents in your downloaded zip file.

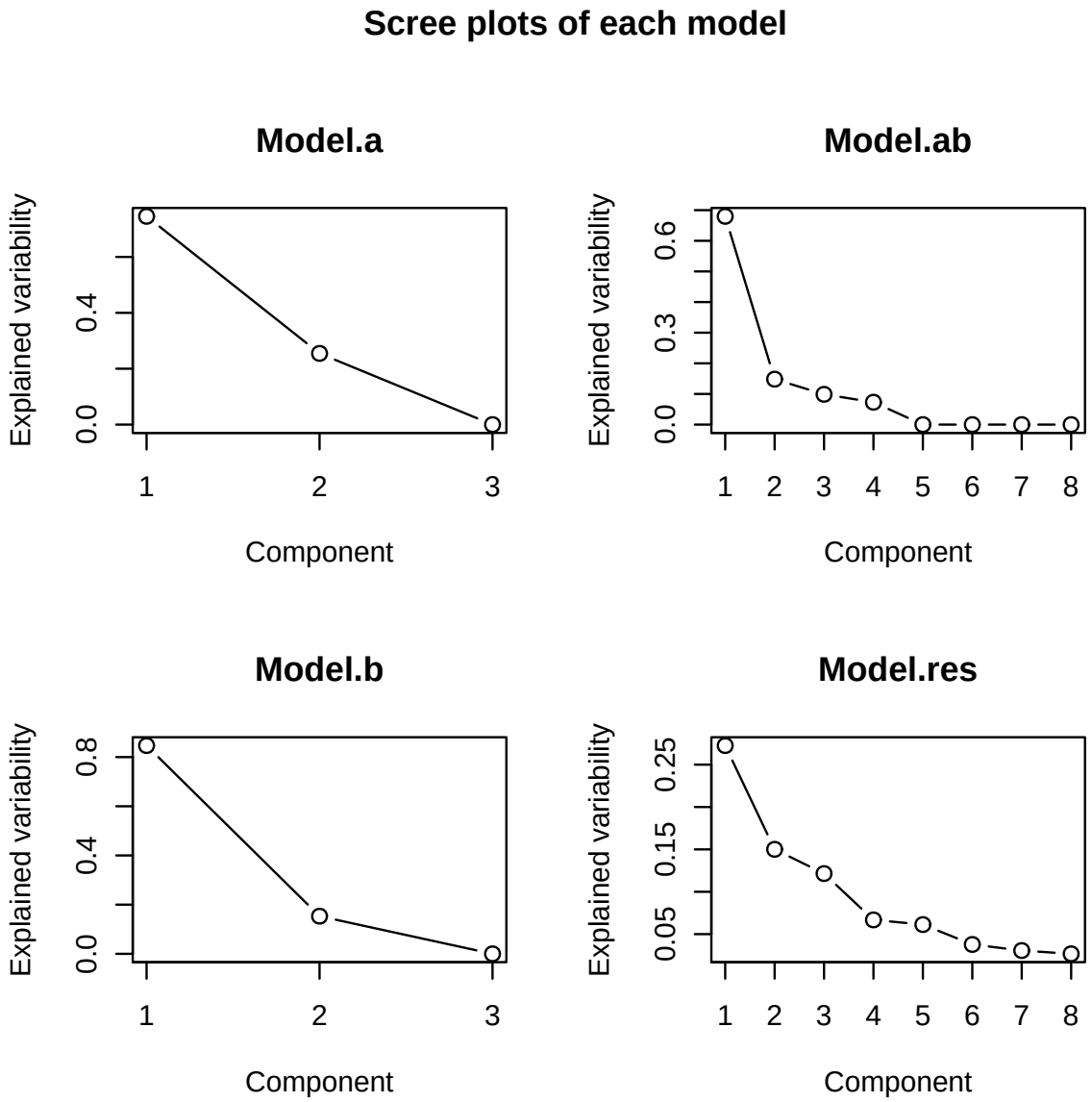

Figure 4: Scree plots for each sub model.

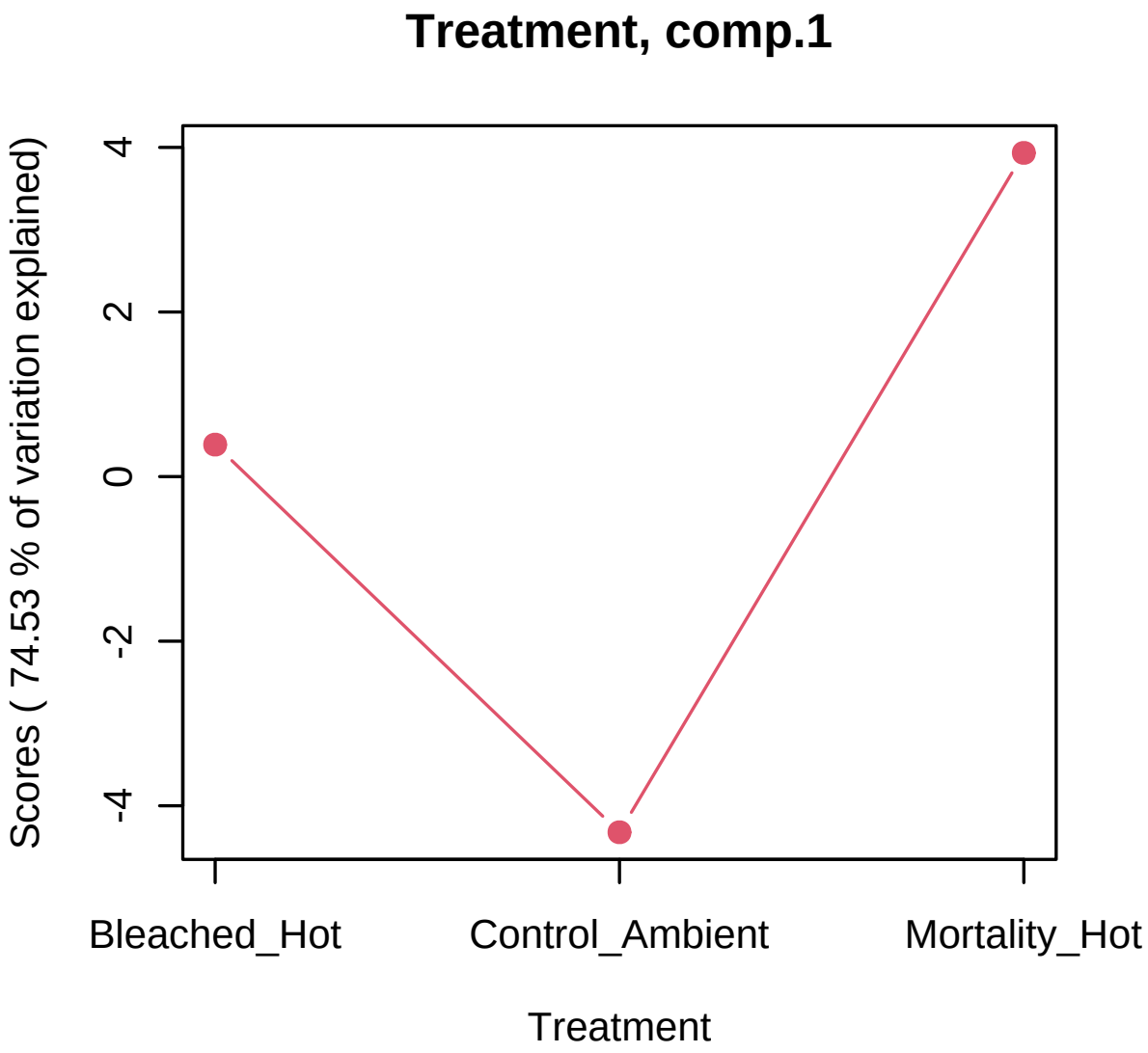

Figure 5: Major patterns associated with Treatment

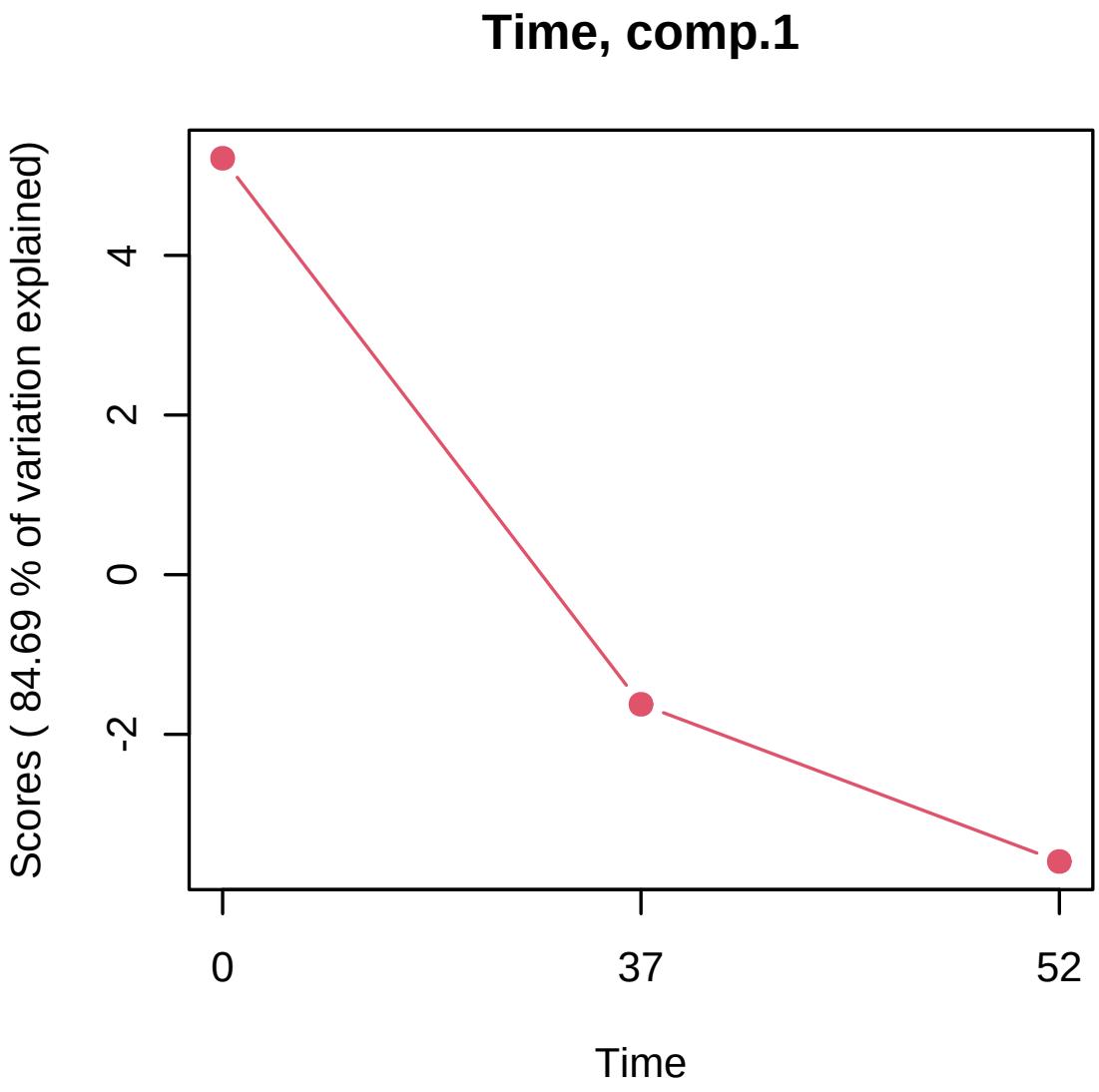

Figure 6: Major patterns associated with Time

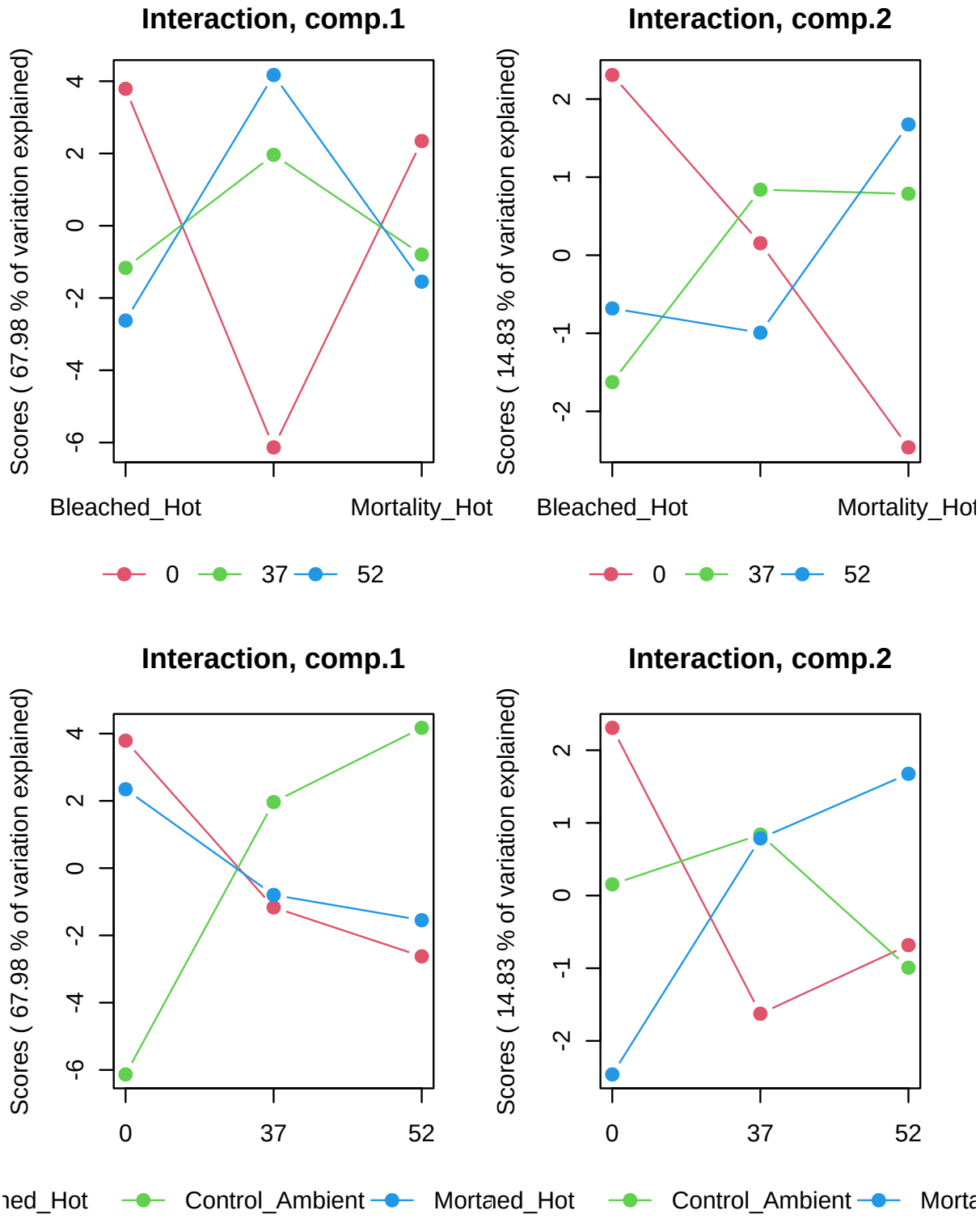

Figure 7: Major patterns associated with the Interaction between the two factors.

Figure 8: Model validation through permutations

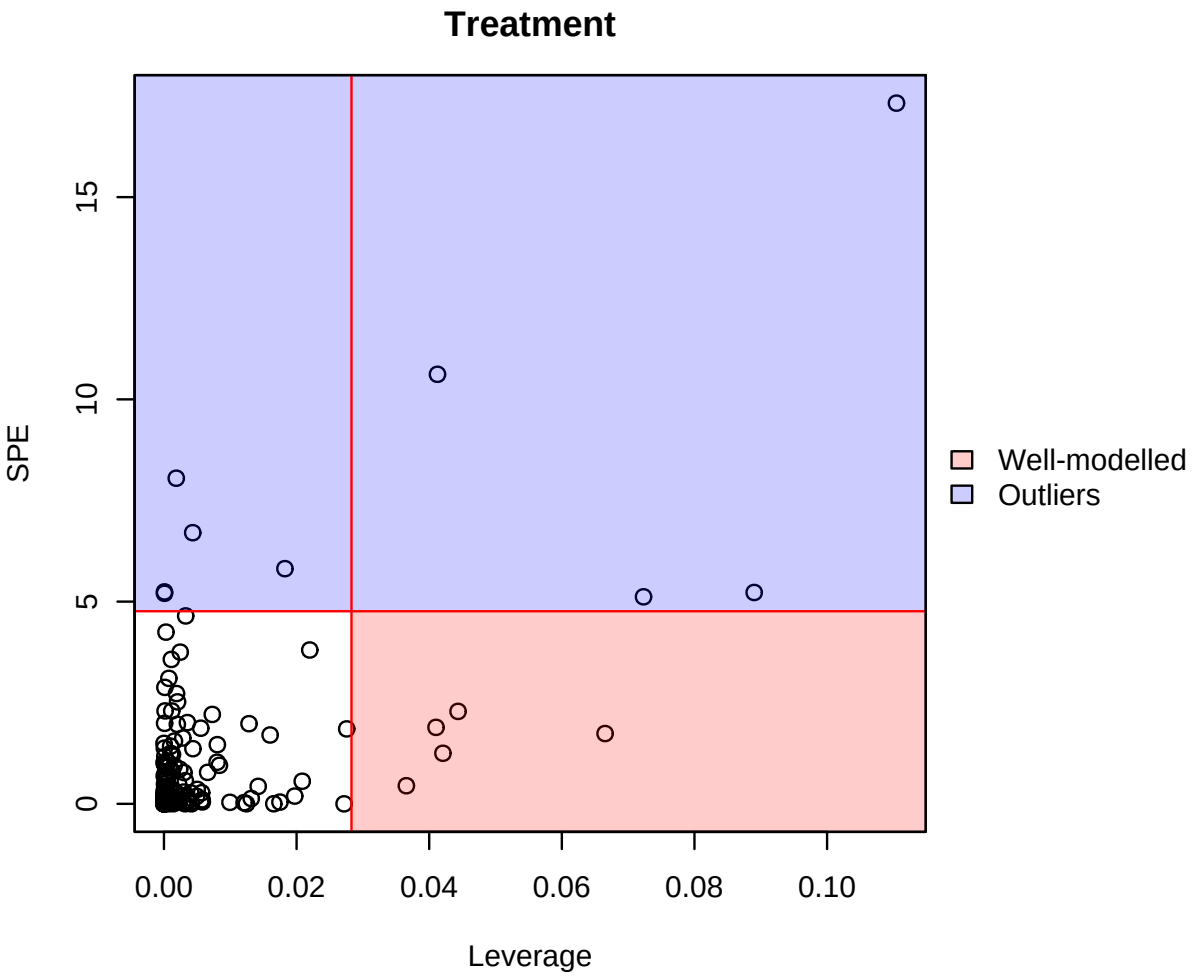

Figure 9: Important variables associated with Treatment

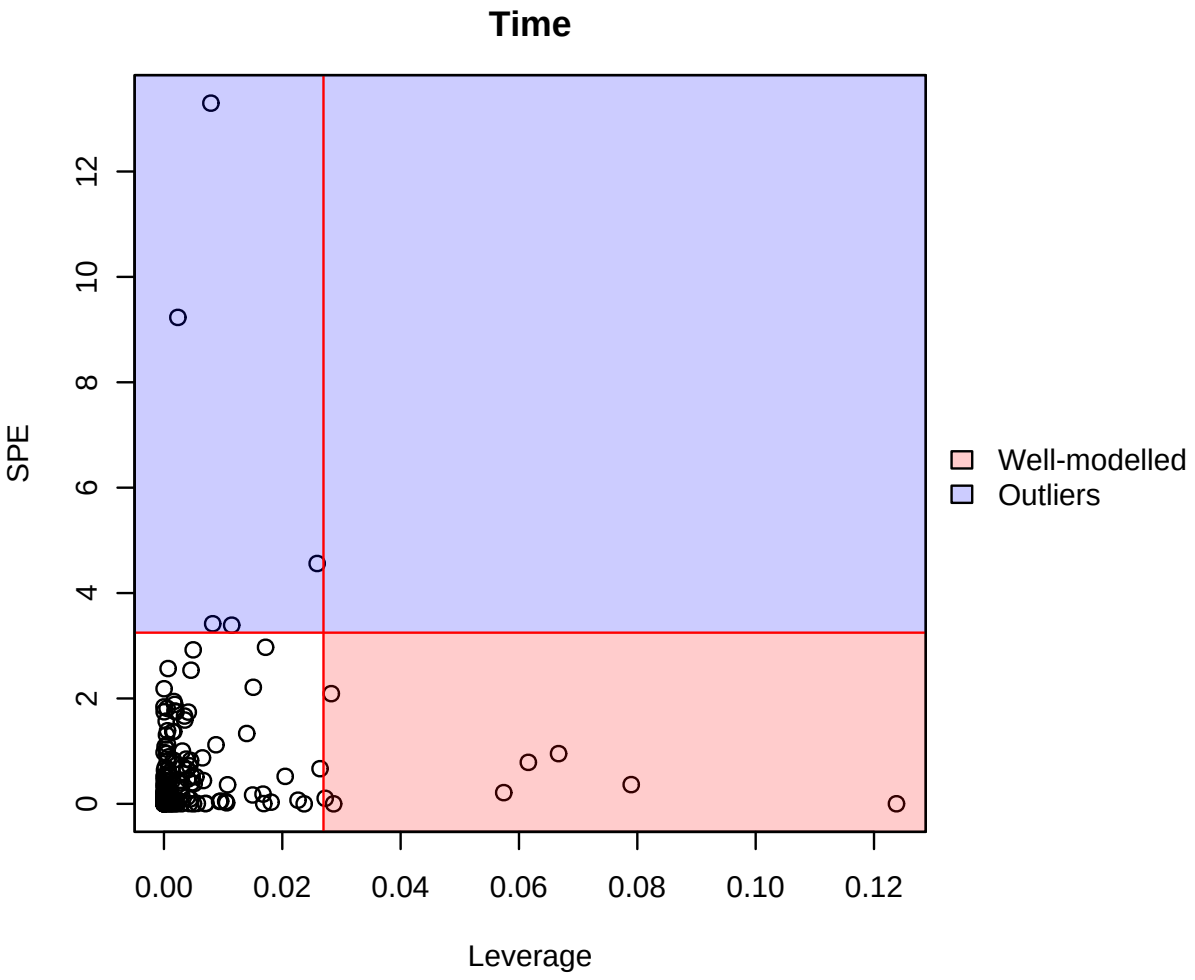

Figure 10: Important variables associated with Time

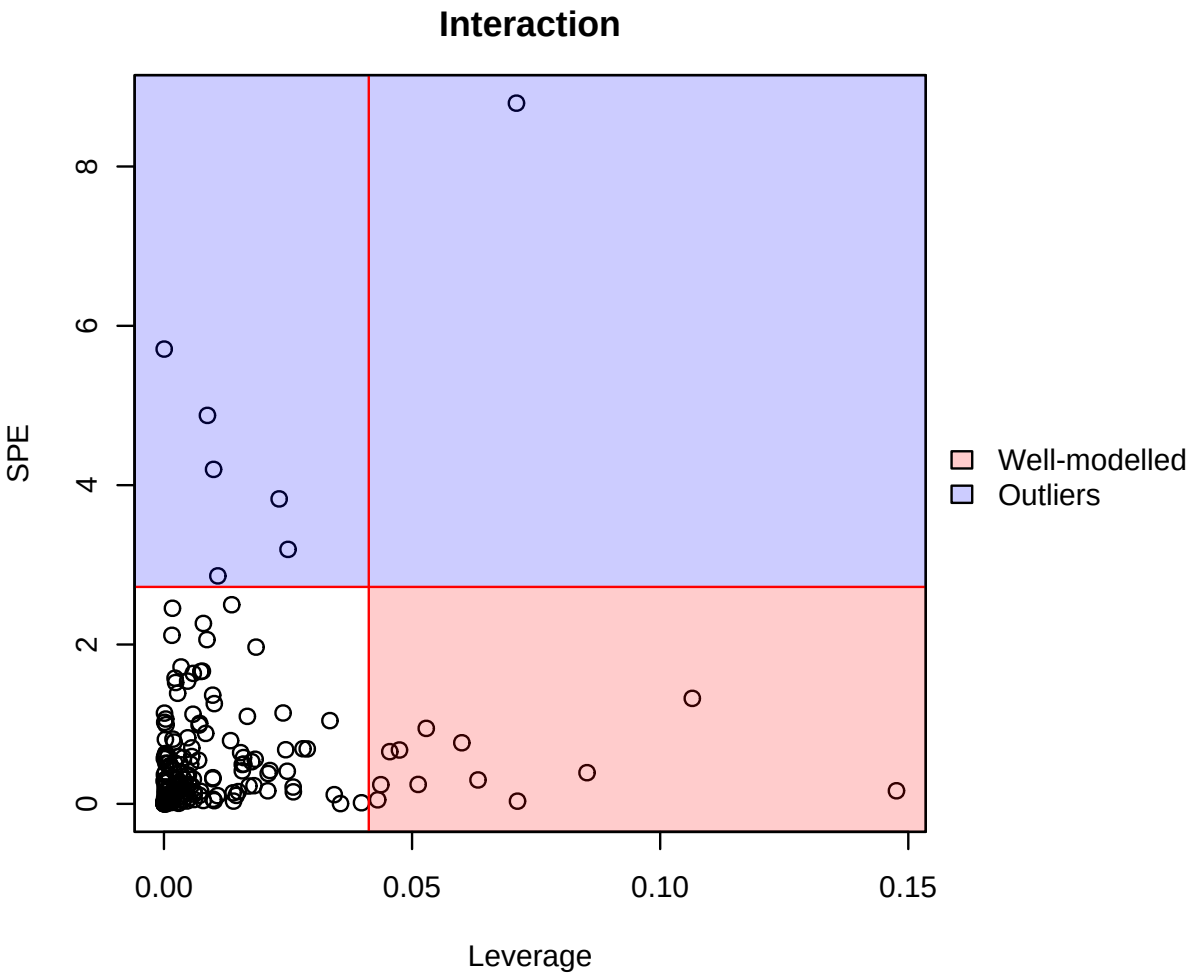

Figure 11: Variables important in interaction between the two factors

Table 3: Important features identified by ASCA. The table shows features that are well modelled by main effect Treatment.

	Compounds	Leverage	SPE
1	Aconitate	0.0665236537344261	1.73811774792659
2	Ornithine	0.0443515838717322	2.28564763456866
3	Arginine	0.042069737127083	1.25023044532387
4	Sucrose	0.0410283569717612	1.89095605669022
5	Malate	0.0365533990715317	0.447989760640942

Table 4: Important features identified by ASCA. The table shows features that are well modelled by main effect Time.

	Compounds	Leverage	SPE
1	2-Aminoethylphosphonate	0.123794834995307	0.00150912790504735
2	4-Guanidinobutanoic acid	0.0789676220474744	0.364883495955567
3	Arginine	0.0666913159053564	0.952793381337887
4	Ornithine	0.061580734304461	0.787258505716871
5	O-Phosphorylethanolamine	0.0574229182476641	0.21284988748454
6	Aconitate	0.0286732988898238	0.000454559821909018
7	Nicotinamide riboside	0.0282817887103597	2.09107601402772
8	Taurine	0.0272507624127915	0.104649777585074

Table 5: Important features identified by ASCA. The table shows features that are well modelled by interaction effect between Treatment and Time.

	Compounds	Leverage	SPE
1	2-Aminoethylphosphonate	0.147615864468603	0.164150321878781
2	4-Guanidinobutanoic acid	0.106471427479168	1.32390933322465
3	Ornithine	0.0852598454470563	0.391255326630026
4	Arginine	0.0712480685491368	0.0336930589657246
5	Aconitate	0.0632898894145795	0.300079258785085
6	NegIS_Glucose-2H7-13C6	0.0600265191871096	0.766764309527342
7	NegIS_Alanine-2H4-15N	0.0528597592694732	0.947193293240982
8	Histidine	0.0512271184843566	0.24360497127042
9	PosIS_Glutamine-2H5-15N2	0.0474601088308875	0.675932366757839
10	NegIS_Glutamine-2H5-15N2	0.0455121423364952	0.655279486611603
11	O-Phosphorylethanolamine	0.0437030557151655	0.240666683665394
12	Lysine	0.0430732479230685	0.0505216198617118

3 Appendix: R Command History

```
 [1] "mSet<-InitDataObjects(\"conc\", \"ts\", FALSE)"
 [2] "mSet<-SetDesignType(mSet, \"time\")"
 [3] "mSet<-Read.TextDataTs(mSet, \"Replacing_with_your_file_path\", \"rowts\");"
 [4] "mSet<-ReadMetaData(mSet, Replacing_with_your_file_path);"
 [5] "mSet<-SanityCheckData(mSet)"
 [6] "mSet<-ReplaceMin(mSet);"
 [7] "mSet<-SanityCheckMeta(mSet, 1)"
 [8] "mSet<-SetDataTypeOfMeta(mSet);"
 [9] "mSet<-PreparePrenormData(mSet)"
[10] "mSet<-Normalization(mSet, \"NULL\", \"NULL\", \"NULL\", ratio=FALSE, ratioNum=20)"
[11] "mSet<-PlotNormSummary(mSet, \"norm_2_\", \"png\", 72, width=NA)"
[12] "mSet<-PlotSampleNormSummary(mSet, \"snorm_2_\", \"png\", 72, width=NA)"
[13] "mSet<-ANOVA2.Anal(mSet, 0.05, \"fdr\", \"time\", 1, 0)"
[14] "mSet<-PlotANOVA2(mSet, \"aov2_1_\", \"png\", 72, width=NA)"
[15] "mSet<-CovariateScatter.Anal(mSet, \"covariate_plot_0_dpi72.png\", \"png\", 72, \"default\", \"
[16] "mSet<-PlotCmpdSummary(mSet, \"N-acetyl-glutamate\",\"NA\", 0, \"png\", 72, width=NA)"
[17] "mSet<-PlotCmpdSummary(mSet, \"Nicotinate\",\"NA\", 0, \"png\", 72, width=NA)"
[18] "mSet<-PlotCmpdSummary(mSet, \"NG-dimethyl-L-arginine\",\"NA\", 0, \"png\", 72, width=NA)"
[19] "mSet<-CovariateScatter.Anal(mSet, \"covariate_plot_1_dpi72.png\", \"png\", 72, \"default\", \"
[20] "mSet<-PlotCmpdSummary(mSet, \"N-acetyl-glutamine\",\"NA\", 0, \"png\", 72, width=NA)"
[21] "mSet<-PlotCmpdSummary(mSet, \"Sucrose\",\"NA\", 0, \"png\", 72, width=NA)"
[22] "mSet<-PlotCmpdSummary(mSet, \"Acetyllysine\",\"NA\", 0, \"png\", 72, width=NA)"
[23] "mSet<-CovariateScatter.Anal(mSet, \"covariate_plot_2_dpi72.png\", \"png\", 72, \"default\", \"
[24] "mSet<-PlotCmpdSummary(mSet, \"Homocysteine\",\"NA\", 0, \"png\", 72, width=NA)"
[25] "mSet<-CovariateScatter.Anal(mSet, \"covariate_plot_3_dpi72.png\", \"png\", 72, \"default\", \"
[26] "mSet<-PlotCmpdSummary(mSet, \"Mandelic acid\",\"NA\", 0, \"png\", 72, width=NA)"
[27] "mSet<-PlotCmpdSummary(mSet, \"Dihydroorotate\",\"NA\", 0, \"png\", 72, width=NA)"
[28] "mSet<-PlotMetaCorrHeatmap(mSet, \"pearson\", \"metaCorrHeatmap_0_\", \"png\", 72, width=NA)"
[29] "mSet<-PlotMetaHeatmap(mSet, \"overview\", \"both\", \"euclidean\", \"ward.D\", \"bwm\", F, T,
[30] "mSet<-PCA.Anal(mSet)"
[31] "mSet<-PlotPCAPairSummaryMeta(mSet, \"pca_pair_meta_0_\", \"png\", 72, width=NA, 5, \"Treatment
[32] "mSet<-iPCA.Anal(mSet, \"ipca_3d_0_.json\")"
[33] "mSet<-PlotHeatMap2(mSet, \"heatmap2_0_\", \"norm\", \"row\", \"png\", 72, width=NA, \"euclidea
[34] "mSet<-Match.PatternMeta(mSet, \"pearson\", \"Treatment\", \"feature\")"
[35] "mSet<-PlotCorr(mSet, \"ptn_1_\", \"feature\", \"png\", 72, width=NA)"
[36] "mSet<-Match.PatternMeta(mSet, \"pearson\", \"Time\", \"feature\")"
[37] "mSet<-PlotCorr(mSet, \"ptn_2_\", \"feature\", \"png\", 72, width=NA)"
[38] "mSet<-Match.PatternMeta(mSet, \"pearson\", \"Subject\", \"feature\")"
[39] "mSet<-PlotCorr(mSet, \"ptn_3_\", \"feature\", \"png\", 72, width=NA)"
[40] "mSet<-Perform.ASCA(mSet, 1, 1, 2, 2)"
[41] "mSet<-PlotModelScree(mSet, \"asca_scree_0_\", \"png\", 72, width=NA)"
[42] "mSet<-CalculateImpVarCutoff(mSet, 0.05, 0.9)"
[43] "mSet<-PlotAscaImpVar(mSet, \"asca_impa_0_\", \"png\", 72, width=NA, \"a\")"
[44] "mSet<-PlotAscaImpVar(mSet, \"asca_impb_0_\", \"png\", 72, width=NA, \"b\")"
[45] "mSet<-PlotAscaImpVar(mSet, \"asca_impab_0_\", \"png\", 72, width=NA, \"ab\")"
[46] "mSet<-PlotASCAModel(mSet, \"asca_fa_0_\", \"png\", 72, width=NA, \"a\",FALSE)"
[47] "mSet<-PlotASCAModel(mSet, \"asca_fb_0_\", \"png\", 72, width=NA, \"b\",FALSE)"
[48] "mSet<-PlotInteraction(mSet, \"asca_fab_0_\", \"png\", 72,FALSE, width=NA)"
[49] "mSet<-Perform.ASCA(mSet, 1, 1, 2, 2)"
[50] "mSet<-CalculateImpVarCutoff(mSet, 0.05, 0.9)"
[51] "mSet<-PlotAscaImpVar(mSet, \"asca_impa_0_\", \"png\", 72, width=NA, \"a\")"
[52] "mSet<-PlotAscaImpVar(mSet, \"asca_impb_0_\", \"png\", 72, width=NA, \"b\")"
[53] "mSet<-PlotAscaImpVar(mSet, \"asca_impab_0_\", \"png\", 72, width=NA, \"ab\")"
[54] "mSet<-PlotASCAModel(mSet, \"asca_fa_0_\", \"png\", 72, width=NA, \"a\",FALSE)"
[55] "mSet<-PlotASCAModel(mSet, \"asca_fb_0_\", \"png\", 72, width=NA, \"b\",FALSE)"
[56] "mSet<-PlotInteraction(mSet, \"asca_fab_0_\", \"png\", 72,FALSE, width=NA)"
```

```
[57] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 0, \"png\", 72, width=NA)"
[58] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 1, \"png\", 72, width=NA)"
[59] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 2, \"png\", 72, width=NA)"
[60] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 3, \"png\", 72, width=NA)"
[61] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 4, \"png\", 72, width=NA)"
[62] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 5, \"png\", 72, width=NA)"
[63] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 6, \"png\", 72, width=NA)"
[64] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 7, \"png\", 72, width=NA)"
[65] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 8, \"png\", 72, width=NA)"
[66] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 9, \"png\", 72, width=NA)"
[67] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 10, \"png\", 72, width=NA)"
[68] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 11, \"png\", 72, width=NA)"
[69] "mSet<-PlotCmpdSummary(mSet, \"2-Aminoethylphosphonate\",\"NA\", 12, \"png\", 72, width=NA)"
[70] "mSet<-PlotCmpdSummary(mSet, \"Ornithine\",\"NA\", 13, \"png\", 72, width=NA)"
[71] "mSet<-PlotCmpdSummary(mSet, \"Ornithine\",\"NA\", 0, \"png\", 72, width=NA)"
[72] "mSet<-PlotCmpdSummary(mSet, \"Ornithine\",\"NA\", 1, \"png\", 72, width=NA)"
[73] "mSet<-PlotCmpdSummary(mSet, \"Ornithine\",\"NA\", 2, \"png\", 72, width=NA)"
[74] "mSet<-PlotCmpdSummary(mSet, \"Ornithine\",\"NA\", 3, \"png\", 72, width=NA)"
[75] "mSet<-PlotCmpdSummary(mSet, \"Ornithine\",\"NA\", 4, \"png\", 72, width=NA)"
[76] "mSet<-PlotCmpdSummary(mSet, \"Ornithine\",\"NA\", 5, \"png\", 72, width=NA)"
[77] "mSet<-PlotCmpdSummary(mSet, \"Aconitate\",\"NA\", 6, \"png\", 72, width=NA)"
[78] "mSet<-PlotCmpdSummary(mSet, \"Aconitate\",\"NA\", 7, \"png\", 72, width=NA)"
[79] "mSet<-RF.AnalMeta(mSet, 500,7,1, \"Treatment\")"
[80] "mSet<-PlotRF.ClassifyMeta(mSet, \"rf_cls_1_\", \"png\", 72, width=NA)"
[81] "mSet<-PlotRF.VIPMeta(mSet, \"rf_imp_1_\", \"png\", 72, width=NA)"
[82] "mSet<-PlotRF.Outlier(mSet, \"rf_outlier_1_\", \"png\", 72, width=NA)"
[83] "mSet<-PlotCmpdSummary(mSet, \"Sucrose\",\"NA\", 0, \"png\", 72, width=NA)"
[84] "mSet<-PlotCmpdSummary(mSet, \"Sucrose\",\"NA\", 1, \"png\", 72, width=NA)"
[85] "mSet<-SaveTransformedData(mSet)"
[86] "mSet<-PreparePDFReport(mSet, \"guest15444671027745781\")\n"
```

The report was generated on Fri Oct 22 15:59:02 2021 with R version 4.1.1 (2021-08-10).