**User Guide for Baseline Correction of Metabolic Dataset (Removal of Confounding Effects with ICA, ICA-RCE)**

**2015-5-5**

**Run with “testdata1.csv”**

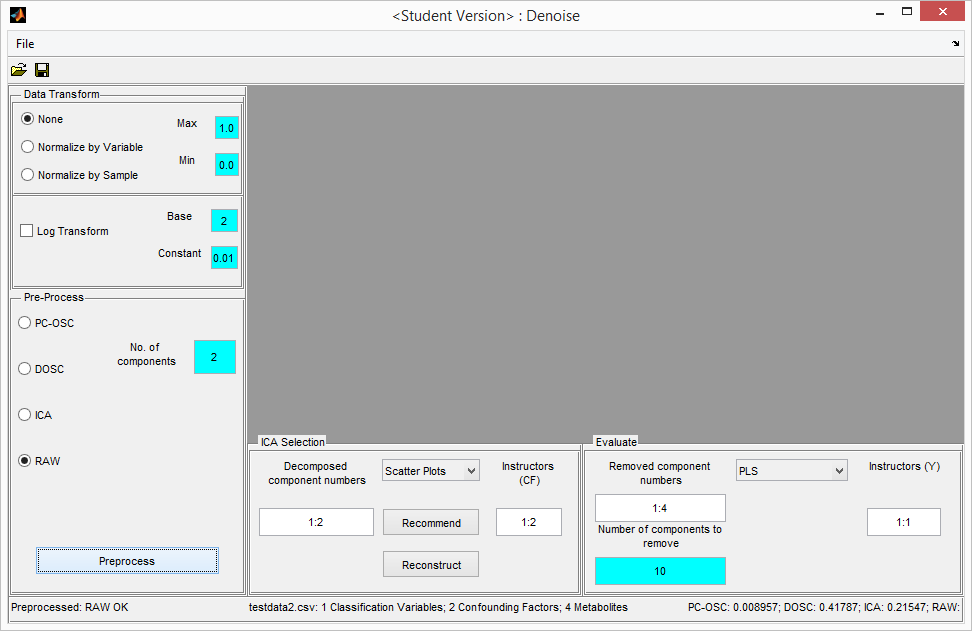
To start the program, enter the name of the main script file on Matlab:

>> main

To close the GUI, simply close the GUI, and on the command window, enter:

>> clear con

To open a data file, click on the File menu. The data file must be in a certain format in csv or xls. For now, use “testdata1.cvs” and “testdata2.csv” in the “data” subfolder.



Select “Data Transform” and “Pre-Process” methods. Then click the “Preprocess” button.

**Run with Test Data “testdata1.csv”**

Open “testdata1.csv”. Click “None” for Data Transform and “RAW” for Pre-Process. Then click on the “Preprocess” button.

Obviously, there will be no decomposition done for this option. Note, however, that in this case, the metabolite variables will be used as the decomposed components . The intention is to skip over the decomposition methods and test the accuracy of the selection and evaluation methods.

**View scatter plot of decomposed components**

Go to “ICA Selection” Panel.

Ensure that the edit box for the Decomposed Component Indices is “1:4” or its Matlab equivalent, e.g. “1,2,3,4” or “1 2 3 4”. Then select “Scatter Plots” from the combo box.

You may play with any other combinations, e.g. “1 3” or “2:4”.



**View “Area under ROC” of decomposed components (S3-1)**

As before, go to “ICA Selection” Panel. Ensure that the edit box for the Decomposed Component Indices is “1:4”.

Change the edit box for “Instructors (CF)” to 1. This means that the AUC will be calculated as instructed by the first confounding factor (“gender” in our case here).

Select “AUC” from the combo box.

****

To see the actual values, type:

>> con.objModel.aSelectAUC

ans =

av: [0.5916 0.6648 0.5770 0.5620]

X: [1 2 3 4]

Y: 1

This means that the first decomposed component has an AUC of 0.5916 when instructed by the classification variable . Similarly, has an AUC of 0.6648.

You may now change the instructor to “2” (“BMI” in our case). It will display a blank figure. This is because AUC makes decision of “detect” or “miss” by comparing binary values in the instructor. Since “BMI” is continuous rather than binary, it cannot be used for AUC.

**View “p-value of Mann-Whitney U Test” of decomposed components (S3-2)**

As for“AUC”, Mann-Whitney must be instructed by a binary variable. Thus, please ensure that the edit box for “Instructors (CF)” to 1 (“gender” in our case here) and not “2” (“BMI”in our case).



To see the actual values, type:

>> con.objModel.pSelectMannWhitney

ans =

pv: [0.1722 0.0139 0.2560 0.3591]

X: [1 2 3 4]

Y: 1

**View “p-value of Anova” of decomposed components (S3-3)**

Unlike “AUC”and “Mann-Whitney U”, the instructors for ANOVA can be continuous.

Enter “1:2” into the edit box for “Instructors (CF)” to include both “gender” and “BMI”.



>> con.objModel.pSelectAnova

ans =

pv: [2x4 double]

X: [1 2 3 4]

Y: [1 2]

View the p-values instructed by “gender” as

>> con.objModel.pSelectAnova.pv(1,:)

ans =

0.1100 0.0046 0.4897 0.5713

View the p-values instructed by “BMI” as

>> con.objModel.pSelectAnova.pv(2,:)

ans =

0.3373 0.1807 0.5719 0.4329

**View “correlation coefficients” of decomposed components with confounding factors (S3-4)**

Ensure that it is “1:4” for the edit box for “Decomposed Component Indices”, and “1:2” for the edit box for “Instructors (CF)”. Then select “Correlation” from the combo box on the “ICA Selection” panel.



>> con.objModel.rSelectCorrelation

ans =

rv: [2x4 double]

X: [1 2 3 4]

Y: [1 2]

>> con.objModel.rSelectCorrelation.rv(1,:)

ans =

-0.1582 -0.2846 -0.1318 -0.1065

>> con.objModel.rSelectCorrelation.rv(2,:)

ans =

0.2400 0.1280 0.3147 0.2888

**Recommended components to be removed**

The program is still not robust. To get the recommended components, please go through the above steps (S3-1, S3-2, S3-3, S3-4) to generate values foe the 4 bar plots. Then click on the “Recommend” button.

You will see that the edit box for “Removed component indices” become “2,1”. This means that and are recommended for removal.

Only 2 components are recommended because the edit box for “Number of components to

Remove” is 2.

**Reconstruct**

Click on “reconstruct” button”. The following will be computed:

The values are based on the edit box for “Removed component indices”. You may change it as you wish.

That’s all for the effect of this button…no graph or anything else will happen. The values of , which was earlier initialised to, will now be the corrected values. We can now proceed to evaluate the performance of .

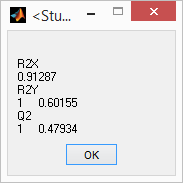
**View “PLS Score” of corrected matrix *Z* (S4-1)**

For the purpose of testing with “testdata1.csv”, say we do not reconstruct the data, or reconstruct with empty string (i.e. nothing) in the edit box for “Removed component indices”. Then will continue to be just .

Select “PLS” on the combo box on the “Evaluation” panel.







>> con.objModel.dEvaluatePLS

ans =

T: [1x1 struct]

R2X: 0.9129

R2Y: [1 0.6015]

Q2: [1 0.4793]

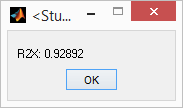
permute: [1x1 struct]

selectedCVs: 1

The R2Y and Q2 values above are 0.6015 and 0.4793 respectively. The R2X value is 0.9129.

**View “PCA Score” of corrected matrix *Z* (S4-2)**





>> con.objModel.evaluatePCAR2X

ans =

0.9289

**View “Logistic Regression” of corrected matrix *Z* with respect to Y (S4-3)**

Select “Logistic Regression” on the combo box on the “Evaluation” panel (bottom right).



>> con.objModel. pEvaluateLogisticRegression

ans =

ORcrude: [0.7957 3.4360e-04 0.0189 2.7598e-06]

pcrude: [1.0458e-04 0.0514 7.6332e-04 3.3854e-04]

CIcrude: [1x4 struct]

ORadjusted: [0.7177 1.5847e-05 0.0141 4.3705e-08]

padjusted: [1.6590e-07 0.0050 1.1348e-04 1.3430e-04]

CIadjusted: [1x4 struct]

Y: 1

**View “p-value ofMann-Whitney” of removed components in T with respect to Y (S4-3)**

Ensure that the edit box for “Removed component indices” is “1:4”, and that the edit box for “Instructor (Y)” is “1”.

Select “p-values, Mann-Whitney” on the combo box.



>> con.objModel.pEvaluateMannWhitney

ans =

pv: [2.1573e-10 5.7749e-05 5.8956e-09 1.2056e-09]

X: [1 2 3 4]

Y: 1

stats: [1x1 struct]

Similarly, select “p-values, Anova” on the combo box.



>> con.objModel.pEvaluateAnova

ans =

pv: [1.3539e-13 2.3140e-05 2.2454e-11 2.9508e-11]

X: [1 2 3 4]

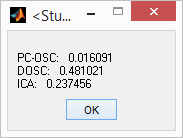
Y: 1

stats: [1x1 struct]

**View “Correlation” of removed components in T with respect to Y (S4-5)**

On the Preprocess panel (left-hand side), select each of the 3 methods “PC-OSC”, “DOSC” and “ICA” and click on the “Preprocess” button. The program makes use of “tic”, “toc” commands in Matlab to capture the actual CPU time.

Select “CPU Time” on the combo box on the “Evaluation” panel (bottom right).



>> con.objModel.timeCPU

ans =

PCOSC: 0.0161

DOSC: 0.4810

ICA: 0.2375

RAW: 6.6711e-05

**View “Correlation” of removed components in T with respect to Y (S4-5)**

Ensure that the edit box for “Removed component indices” is “1:4”, and that the edit box for “Instructor (Y)” is “1”.

Select “Correlation” on the combo box.



>> con.objModel.rEvaluateCorrelation

ans =

rv: [-0.1582 -0.2846 -0.1318 -0.1065]

pv: [0.1722 0.0127 0.2565 0.3599]

crv: [5x5 double]

cpv: [5x5 double]

X: [1 2 3 4]

Y: 1

stats: [1x1 struct]

If the edit box for “Instructor (Y)” is changed to “2” and the combo box selected for correlation, the graph will be:



>> con.objModel.rEvaluateCorrelation

ans =

rv: [0.2400 0.1280 0.3147 0.2888]

pv: [0.0368 0.2704 0.0056 0.0114]

crv: [5x5 double]

cpv: [5x5 double]

X: [1 2 3 4]

Y: 2

stats: [1x1 struct]

**Background**

When identifying the factors that influence a disease, the question is often: “Is the set of metabolites associated with the disease outcome , after accounting for the controlling variables ?”

For to be considered a confounding factor, it has to affect both and , as shown below. The choice of is based on past knowledge, but common “environmental” variables such as age, gender, and BMI are often confounding factors.

Figure : C, the confounding factors, affects both M (predictor variables) and Y (outcomes)

There are two broad methods for adjusting for confounding factors: direct method and modelling method. Direct method adjusts the data in the stratified groups according to a common distribution, whereas modelling method corrects for the confounding effect based on certain assumptions made on the data.

**Proposed Method**

The proposed method assumes that the metabolites can be decomposed into its components , and that components in that are related to the confounding factors should be removed from so as to form the corrected data set for use in further analysis.

We will break down the proposed method into the following steps:

1. Pre-processing – normalization/data transformation etc.
2. Decomposition– ICA etc.
3. Selection of Components for Removal
4. Evaluation of Corrected Data Set

Steps 3 and 4 should be done iteratively until the best outcome is obtained.

decomposition

Figure : Step 3 Selection of Components for Removal - Compare decomposed components in with confounding factors , and select the ones with the largest relation for removal from *M*.

correction

Figure : Step4 Evaluation of Corrected Data Set - Compare corrected data set Z or the removed components *Y* with the outcome Y or confounding factors *C*.

In Step 3, the vectors in are compared to the vectors in , based on the following methods:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S/N | Selection Method | Input data set | Instructor | Display |
| S3-1 | Area values under ROC | User-specified vectors in | User-specified vectors in | bar plot in descending order (x axis is the indices of the vectors in ;y axis is the AUC values) |
| S3-2 | P values of Mann-Whitney U test | User-specified vectors in | User-specified vectors in  ( is limited to categorical data that are binary) | bar plot in descending order (x axis is the indices of the vectors in ;y axis is the AUC values) |
| S3-3 | P values of ANOVA | User-specified vectors in | User-specified vectors in | bar plot in descending order (x axis is the indices of the vectors in ;y axis is the p values) |
| S3-4 | Spearman correlation coefficients | User-specified vectors in | User-specified vectors in | bar plot in descending order (x axis is the indices of the vectors in ;y axis is the correlation coefficients) |

The indices of the vectors in are sorted. The sorting order is descending for AUC and correlation (large is better), and descending for p values for Mann-Whitney and ANOVA (small is better).

The program will recommend for removal the intersection of the top few components by every criterion. The removed components, denoted as , will be displayed on the GUI. Final decision should be done by user.

The corrected data set is obtained with the following formula:

In Step 4, the corrected data set and/or the removed components in are compared to the vectors in :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S/N | Selection Method | Input data set | Instructor | Display |
| S4-1 | PLS | and | N.A | 2D Score Plot, R2X, R2Y, Q2, and 200 times permutation plot. |
| S4-2 | PCA |  | N.A | 2D Scores Plot, R2X |
| S4-3 | Logistic Regression | , with and without confounding factors . | User-specified vectors in | P, OR, and CIs of each variable in with and without confounding factors as covariates (crude and adjusted)  Box Plot of p-values with and without confounding factors as covariates. |
| S4-4 | P values from Mann-Whitney U test or ANOVA |  | User-specified vectors in | mean, median, max, min, ...?  Box Plot of p-values. |
| S4-5 | Correlation | Removed vectors in and confounding factors . | N.A | Spearman correlation coefficients of every confounding factor-removed component pair |
| S4-6 | CPU Time of Decomposition |  |  |  |