An Introduction to the *metabomxtr* package

Michael Nodzenski, Anna C. Reisetter, Denise M. Scholtens

October 26, 2021

1 Introduction

High-throughput metabolomics profiling has surged in popularity, but the frequent occurrence of missing data remains a common challenge to analyzing non-targeted output. In practice, complete case analysis, imputation, and adaptations of classic dimension reduction tools to allow for missing data have been used. A more elegant approach for metabolite-by-metabolite analysis is the Bernoulli/lognormal mixture-model proposed by Moulton and Halsey (1995), which simultaneously estimates parameters modeling the probability of non-missing response and the mean of observed values. The *metabomxtr* package has been developed to automate the process of mixture model analysis.

2 Sample Mixture Model Analysis

The following commands demonstrate typical usage of *metabomxtr*. First, load the package.

```
>   library(metabomxtr)
```

Next, load metabdata, the sample dataset. Metabdata contains metabolite levels and phenotype data of 115 of pregnant women. Columns 1:10 contain phenotype data and columns 11:59 contain log transformed metabolite levels, with missing values indicated by NA. Users should note that while metabdata is a data frame, *metabomxtr* can also accommodate matrix and *ExpressionSet* objects, and the function use is the same. For *ExpressionSets*, metabolites should be in rows of the *exprs* section, and phenotype data in columns of the *pData* section.

```
>   data(metabdata)
> dim(metabdata)

[1] 115  59
```

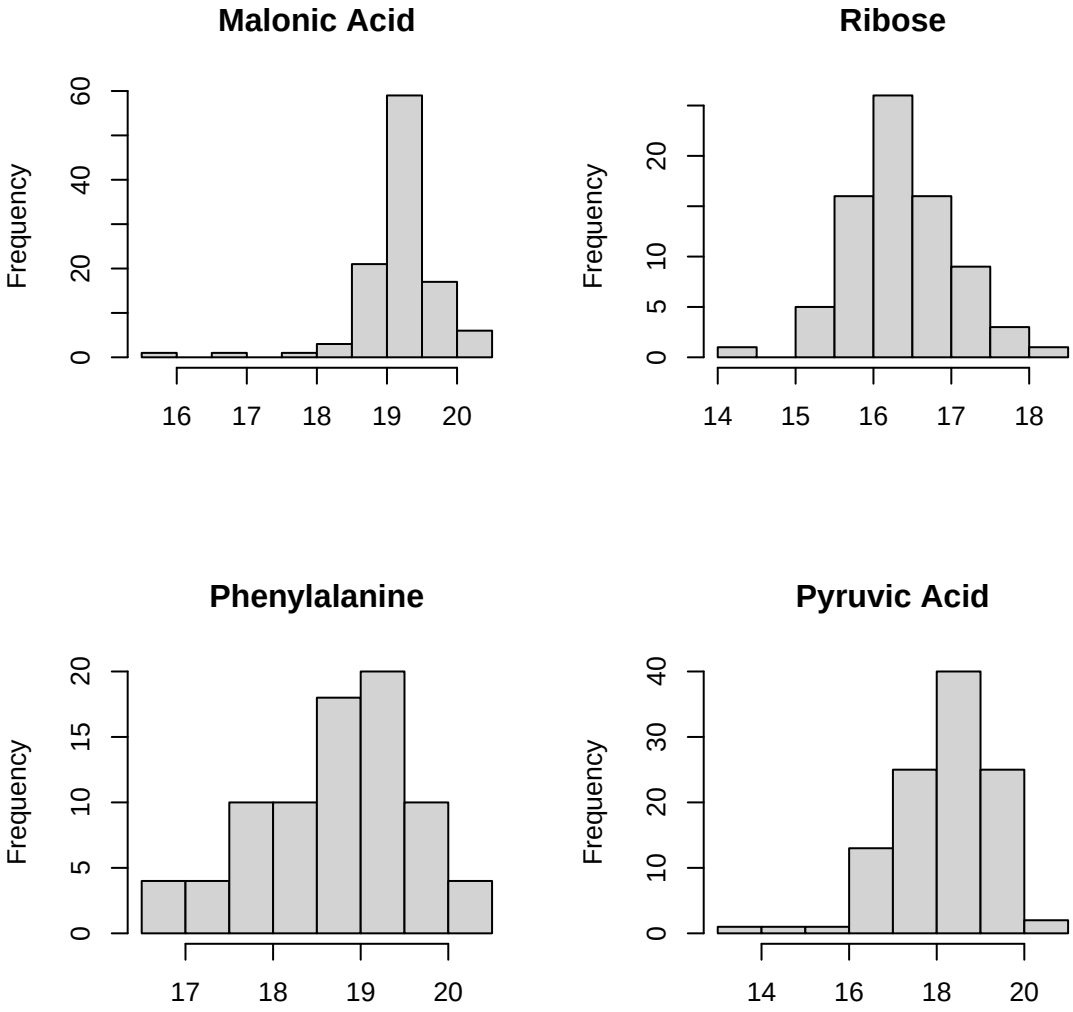
For this analysis, malonic acid, ribose, phenylalanine, and pyruvic acid are the metabolites of interest. These variables are in columns 24:27 of the dataset. We'll define a character vector of the corresponding column names for use later on, and check the number of missing values in each column.

```
>   yvars<-colnames(metabdata)[24:27]
> apply(metabdata[,yvars],2,function(x){sum(is.na(x))})

 malonic.acid        ribose phenylalanine  pyruvic.acid
            6            38            35             7
```

We'll also check the distributions of the metabolites.

```
>   par(mfrow = c(2, 2))
> hist(metabdata$malonic.acid,main="Malonic Acid",xlab=NULL)
> hist(metabdata$ribose,main="Ribose",xlab=NULL)
> hist(metabdata$phenylalanine,main="Phenylalanine",xlab=NULL)
> hist(metabdata$pyruvic.acid,main="Pyruvic Acid",xlab=NULL)
```



Each of the metabolites contains missing values and the data look fairly normal, so mixture model analysis seems appropriate.

The woman's phenotype (variable PHENO) will be the predictor of interest. This variable indicates whether the woman had high (MomHighFPG) or low (MomLowFPG) fasting plasma glucose measurements. Our goal is to determine if women with high FPG have significantly different metabolite levels than women with low FPG. To do this, we need to set the MomLowFPG group as the reference level of PHENO.

```
> levels(metabdata$PHENO)

[1] "MomHighFPG" "MomLowFPG"

> metabdata$PHENO<-relevel(metabdata$PHENO,ref="MomLowFPG")
```

Also, determine how many subjects are in each group.

```
> table(metabdata$PHENO)

 MomLowFPG MomHighFPG
        48         67
```

Next, we need to specify the full mixture model. This should be of the form ~ x1+x2...|z1+z2..., where x's represent covariates modeling metabolite presence/absence (discrete portion), and z's are covariates modeling the mean of observed values (continuous portion). The predictor of interest should be included in both the discrete and continuous portions of the full model. In addition, we'll control for the woman's age, gestational age, sample storage time, and parity in the continuous portion.

```
> fullModel<-~PHENO|PHENO+age_ogtt_mc+ga_ogtt_wks_mc+storageTimesYears_mc+parity12
```

In addition, we need to specify a reduced model. Because our goal is to evaluate the significance of the contribution of phenotype to *both* the continuous and discrete portions of the mixture model, we'll remove PHENO from both portions.

```
> reducedModel<-~1|age_ogtt_mc+ga_ogtt_wks_mc+storageTimesYears_mc+parity12
```

The *mxtrmod* function can be used to run the full model on each of the 4 metabolites of interest.

```
> fullModelResults<-mxtrmod(ynames=yvars,mxtrModel=fullModel,data=metabdata)
> fullModelResults

            .id      xInt x_PHENOMomHighFPG     zInt z_PHENOMomHighFPG
1  malonic.acid 3.8502000        -1.3325431 19.17930         0.2035945
2        ribose 0.9918867        -0.4702836 16.63081        -0.1777151
3 phenylalanine 0.4705377         0.7424348 18.60274         0.1970021
4  pyruvic.acid 2.1520649         1.3292331 17.76212         0.8376551
  z_age_ogtt_mc z_ga_ogtt_wks_mc z_storageTimesYears_mc z_parity12     sigma
1   0.005365227      0.018777421             0.09621857 -0.3438284 0.5618018
2   0.022315594      0.054468010             0.15368387 -0.4729735 0.5876773
3  -0.036657861      0.024705843            -0.04221968 -0.1819805 0.8745071
4  -0.012048400     -0.007359868             0.04422338 -0.2181451 1.0927978
  method conv    negLL nObs
1   BFGS    0 114.4586  115
2   BFGS    0 140.3466  115
3   BFGS    0 167.5460  115
4   BFGS    0 187.8470  115
```

In the output data frame, the .id column indicates metabolite, columns beginning with x's are parameter estimates for the discrete portion of the model, columns beginning with z's are parameter estimates for the continuous portion, sigma is the variance of observed values, method is the optimization algorithm used, conv indicates whether the model converged (0=convergence), negLL is the negative log likelihood, and nObs is the number of observations used.

Then, we will use *mxtrmod* to run the reduced models. Users should note the importance of specifying the *fullModel* parameter when running reduced models, which ensures that if model covariates have missing values, both full and reduced model results are based on the same set of observations.

```
> reducedModelResults<-mxtrmod(ynames=yvars,mxtrModel=reducedModel,data=metabdata,fullModel=fullModel)
> reducedModelResults

            .id      xInt     zInt z_age_ogtt_mc z_ga_ogtt_wks_mc
1  malonic.acid 2.8995698 19.28683   0.007914551      0.021036428
2        ribose 0.7085350 16.54123   0.020150030      0.049071421
3 phenylalanine 0.8705433 18.71830  -0.033505881      0.024024013
4  pyruvic.acid 2.7364050 18.21829  -0.002850798      0.007869146
  z_storageTimesYears_mc  z_parity12     sigma method conv    negLL nObs
1             0.10560160 -0.32228370 0.5702182   BFGS    0 116.9946  115
2             0.15377361 -0.49431283 0.5941046   BFGS    0 141.8404  115
3            -0.04002210 -0.13303210 0.8754378   BFGS    0 169.5777  115
4             0.07870144 -0.07151664 1.1619165   BFGS    0 195.8161  115
```

Finally, the significance of full vs. reduced models can be examined using nested likelihood ratio χ^2 tests via the *mxtrmodLRT* function. Required parameters include the output data frames from *mxtrmod* for full (parameter *fullmod*) and reduced models (parameter *redmod*). Optionally, the user may use the *adj* parameter to specify method of adjustment for multiple testing.

```
> finalResult<-mxtrmodLRT(fullmod=fullModelResults,redmod=reducedModelResults,adj="BH")
> finalResult

            .id negLLFull negLLRed     chisq df            p        adjP
1  malonic.acid  114.4586 116.9946  5.072063  2 0.0791799843 0.158359969
2        ribose  140.3466 141.8404  2.987680  2 0.2245089078 0.224508908
3 phenylalanine  167.5460 169.5777  4.063284  2 0.1311200288 0.174826705
4  pyruvic.acid  187.8470 195.8161 15.938041  2 0.0003460177 0.001384071
```

Similar to *mxtrmod* output, .id indicates metabolite, negLLFull is the negative log likelihood of the full model, negLLRed is the negative log likelihood of the reduced model, chisq is the test statistic, df are the degrees of freedom, p is the unadjusted p-value, and adjP is the adjusted p-value. Based on the FDR adjusted p-values, pyruvic acid levels are significantly different in women with high compared to low fasting plasma glucose (p=0.0014). Levels of the other three metabolites did not vary significantly between FPG groups.

As a last step, we'll put together a results table. First, calculate the estimated proportion of metabolites present for high and low FPG women.

```
> HighFPG.Prop<-round(exp(fullModelResults$xInt+fullModelResults$x_PHENOMomHighFPG)/
+ (1+exp(fullModelResults$xInt+fullModelResults$x_PHENOMomHighFPG)),digits=2)
> LowFPG.Prop<-round(exp(fullModelResults$xInt)/(1+exp(fullModelResults$xInt)),digits=2)
```

Next, calculate the estimated mean metabolite levels by FPG status, and estimated mean difference.

```
> HighFPG.Mean<-round(fullModelResults$zInt+fullModelResults$z_PHENOMomHighFPG,digits=2)
> LowFPG.Mean<-round(fullModelResults$zInt,digits=2)
> FPG.MeanDiff<-round(fullModelResults$z_PHENOMomHighFPG,digits=2)
```

Then combine with metabolite names and FDR adjusted p-values.

```
> finalResultTable<-data.frame(Metabolite=fullModelResults$.id,HighFPG.Prop=HighFPG.Prop,
+ LowFPG.Prop=LowFPG.Prop,HighFPG.Mean=HighFPG.Mean,
+ LowFPG.Mean=LowFPG.Mean,Mean.Difference=FPG.MeanDiff,
+ FDR.Adj.P=round(finalResult$adjP,digits=4))
```

Below are the final results.

	Metabolite	HighFPG.Prop	LowFPG.Prop	HighFPG.Mean	LowFPG.Mean	Mean.Difference	FDR.Adj.P
1	malonic.acid	0.93	0.98	19.38	19.18	0.20	0.1584
2	ribose	0.63	0.73	16.45	16.63	-0.18	0.2245
3	phenylalanine	0.77	0.62	18.80	18.60	0.20	0.1748
4	pyruvic.acid	0.97	0.90	18.60	17.76	0.84	0.0014

3 Session Information

- R version 4.1.1 (2021-08-10), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Running under: `Windows Server x64 (build 17763)`
- Matrix products: default
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: Biobase 2.54.0, BiocGenerics 0.40.0, metabomxtr 1.28.0, xtable 1.8-4
- Loaded via a namespace (and not attached): BiocParallel 1.28.0, DBI 1.1.1, Formula 1.2-4, MASS 7.3-54, Matrix 1.3-4, R6 2.5.1, Rcpp 1.0.7, assertthat 0.2.1, colorspace 2.0-2, compiler 4.1.1, crayon 1.4.1, dplyr 1.0.7, ellipsis 0.3.2, fansi 0.5.0, generics 0.1.1, ggplot2 3.3.5, glue 1.4.2, grid 4.1.1, gtable 0.3.0, lattice 0.20-45, lifecycle 1.0.1, magrittr 2.0.1, multtest 2.50.0, munsell 0.5.0, numDeriv 2016.8-1.1, optimx 2021-10.12, parallel 4.1.1, pillar 1.6.4, pkgconfig 2.0.3, plyr 1.8.6, purrr 0.3.4, rlang 0.4.12, scales 1.1.1, snow 0.4-3, splines 4.1.1, stats4 4.1.1, survival 3.2-13, tibble 3.1.5, tidyselect 1.1.1, tools 4.1.1, utf8 1.2.2, vctrs 0.3.8

4 References

Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. Biometrics. 1995 Dec;51(4):1570-8.