MetSizeR manual

Gift Nyamundanda, Isobel Claire Gormley, Yue Fan, William M Gallagher, and Lorraine Brennan
University College Dublin, Dublin, Ireland
gnyamundanda@gmail.com

Background

Determining sample sizes for metabolomic studies is important but due to the complexities of these experiments, currently there are no standard methods for sample size estimation in metabolomics. Since pilot studies are rarely done in metabolomics, sample size estimation approaches for high dimensional microarray experiments requiring real pilot data can not be applied. MetSizeR is a tool written in the R statistical software (R Development Core Team, 2009) for determining sample sizes for metabolomic experiments even when experimental pilot data is not available. MetSizeR can also be used for sample size estimation for both targeted and NMR metabolomic experiments. The statistical theory underlying MetSizeR is discussed in Nyamundanda et al. (2012b). Effort was put in designing the interface of MetSizeR for an easy interaction with the user. A graphic user interface (GUI) was used to design MetSizeR to encourage its widespread use, regardless of previous knowledge of R.

Availability and requirements

MetSizeR is a freely available software and is implemented within the R environment. MetSizeR requires:

1. Operating system(s): platform independent.
2. The latest version of R: it can be downloaded from the R project webpage http://www.r-project.org/.
3. It depends on few other R packages: MetabolAnalyze, gWidgets (Verzani, 2009), MASS, gtools, pscl, cairodevice, caTools, ellipse, gplots, and gdata.

Installation

To install R on your computer, go to the R webpage http://www.r-project.org/ and follow the download instructions. You must have administrative privileges to install R on your computer.

Prior to the installation of MetSizeR package, the GTK library and the R package gWidgets should be installed properly. Quick Installation of GTK library for windows user:

> install.packages("gWidgets") and choose the option RGtk2.

For Linux and Mac OS X user, for detailed download and installation information please go to the GTK website (http://www.gtk.org/download /index.php).

MetSizeR can be downloaded from the R webpage or can be installed by using command:

> install.packages("MetSizeR")

This line of command will install MetSizeR with all its R dependencies.

Load the package

After installing MetSizeR in your R environment, it can be loaded by typing command:

> library(MetSizeR)

This will take some few seconds since R needs time to load all MetSizeR dependent packages. All MetSizeR functions can be accessed via a GUI that can be activated by typing the command:

Figure 1: MetSizeR main GUI window.

```
> MetSizeR()
```

and you will be greeted by the following message.

```
Welcome to MetSizeR!!
```

The MetSizeR GUI main window will be launched as shown in Figure 1.

Estimating sample size without experimental pilot data

The main advantage of MetSizeR is that it can estimate the number of samples required for a metabolomic experiment even when experimental pilot data is not available. Let us suppose we want to determine the number of samples required for a metabolomic experiment using NMR. In the MetSizeR GUI, under the "Sample size" menu select "no pilot data" and choose the option "NMR analysis" as shown in Figure 3. The option "NMR analysis" is for the number of samples required for NMR experiment. Otherwise, if the number of samples required is for targeted metabolomic analysis, then the option "Targeted analysis" is selected.

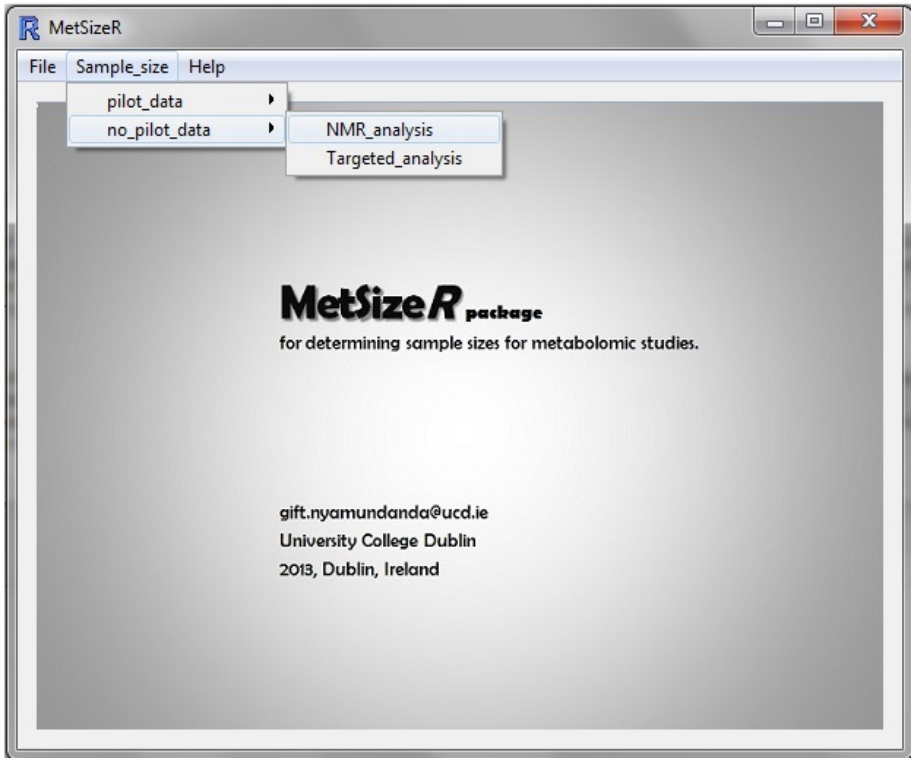

Figure 2: Selecting the type of experiment.

MetSizeR will display a window titled "MetSizeR without pilot nmr data" as shown in Figure 4. The first argument in this window is "Spectral bins", which allows the user to specify the approximate number of spectral bins from the NMR experiment. The default value is set at 200 bins. The second argument is "Proportion of significant bins". This is the proportion of spectral bins which are expected to be significant. The default value is 0.2 and it can be increased by clicking the arrow up or reduced by clicking the arrow down.

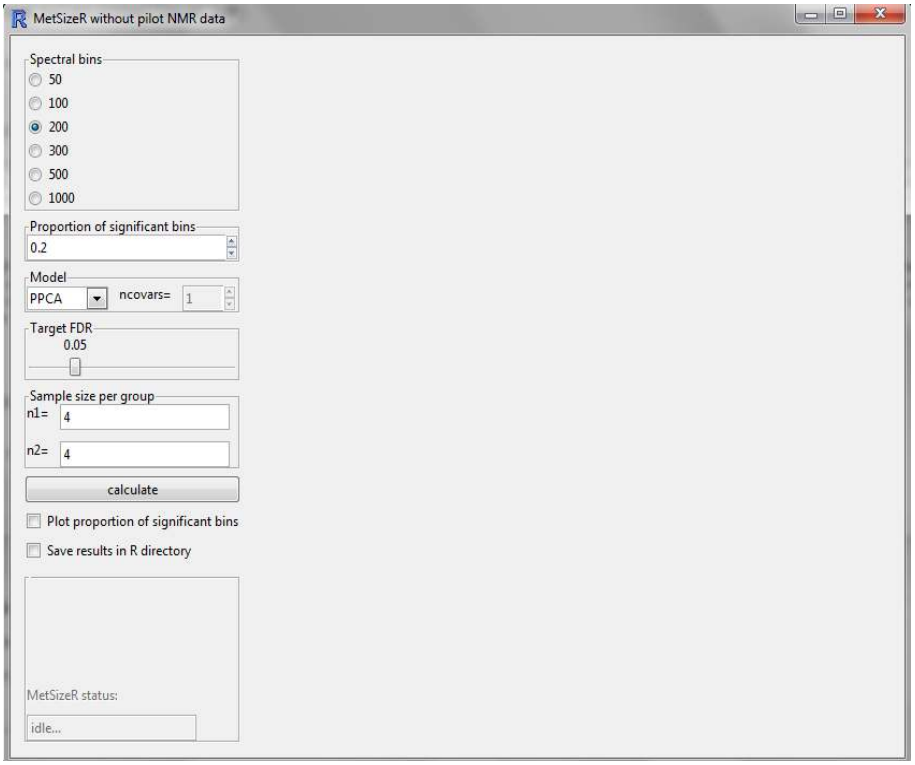

Figure 3: MetSizeR GUI window for estimating sample size for a NMR experiment with no experimental pilot data available.

The third argument in this window is "Model". This is a drop down list of the three different types of models (PPCA (Tipping and Bishop, 1999; Nyamundanda et al., 2010), PPCCA (Nyamundanda et al., 2010), and DPPCA model (Nyamundanda et al., 2012a)) which can be used to analyze the resultant data from the experiment. The PPCA model is used when the probabilistic version of PCA is going to be used for data analysis. The PPCCA model is used when covariates information is going to be used with the probabilistic PCA model and the argument "ncovars=" is used to specify the number of covariates. The DPPCA model is used to estimate the number of samples required for a longitudinal metabolomic experiment.

The argument "Target FDR" is the level of control over type I errors using false discovery rate (FDR). It is usually set at 5%, increasing it will result in a decrease in the estimated sample size which reduces the 'power' of the study. The argument "sample size per group" is the smallest or the

initial sample size to be considered by MetSizeR. Most importantly, it allows the user to provide MetSizeR the ratio of the number of samples in one group to the other group in case of unbalanced design.

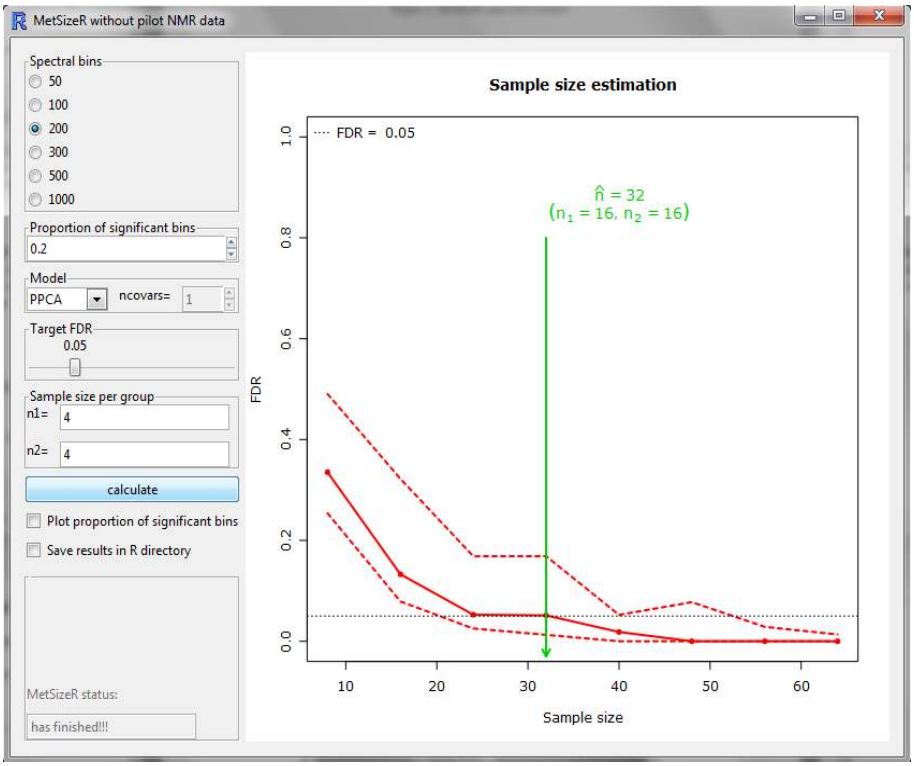

Figure 4: MetSizeR GUI results window for estimating sample size for an NMR experiment with no experimental pilot data available using PPCA model. The estimated FDR is denoted by solid red lines and the 10^{th} and 90^{th} percentiles by dashed red lines. A horizontal dashed black line is the targeted FDR at 5%. The sample size $\hat{n}$ is estimated at 32 with 16 in each treatment group.

In the first example, we want to determine the number of samples required for an NMR experiment with 300 spectral bins. About 20% of these spectral bins are expected to be significant and the PPCA model is to be used to analyze the data. Controlling FDR at 5%, click 'calculate' to estimate the number of samples. In this example the experiment has is balanced as the ratio of the number samples in one group to the other group is 1:1 with four samples in each treatment group ($n_1 = 4$ and $n_2 = 4$). Figure 4 shows that the estimated sample size is 32 with 16 samples in each group. Unfortunately, the plots in the MetSizeR GUI can not be copied or saved but the user can click "Save results in R directory" to save them in the R working directory. The results can be accessed in the folder called Results.

The expected proportion of significant spectral bins impacts on the estimated number of samples required for a metabolomic experiment. Click "Plot proportion of significant bins" to assess the effect of varying the proportion of spectral bins on different sample sizes (see Figure 5). The last

argument "MetSizeR status" indicates when MetSizeR is idle, running or when it has finished.

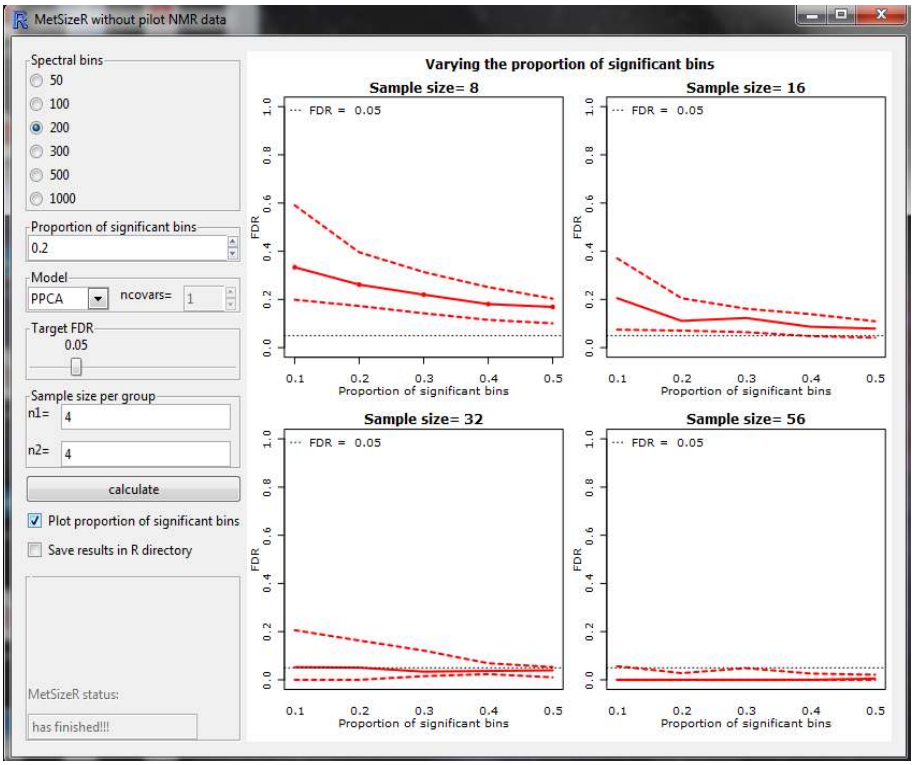

Figure 5: MetSizeR GUI results window for estimating sample size for an NMR experiment with no experimental pilot data available using PPCA model. The figure show the effect of varying the proportion of significant bins. The FDR improves when the data sample size is increased.

The approach developed here for sample size estimation is not limited to NMR data. The method has been developed to accept data from targeted metabolomic analysis using MS,thus ensuring its applicability across the metabolomics community. Figure 6 shows the estimated sample size for a targeted metabolomic experiment of 100 metabolites with one covariate. Twenty percent of metabolites are expected to be significant and the FDR is controlled at 5%. Since the sample size required is for a study in which two covariates are to be measured, the PPCCA model which allow for covariates is used.

Estimating sample size with experimental pilot data

If the experimental pilot data are available then it is recommended to estimate the number of samples required for an experiment using the data from the pilot study.

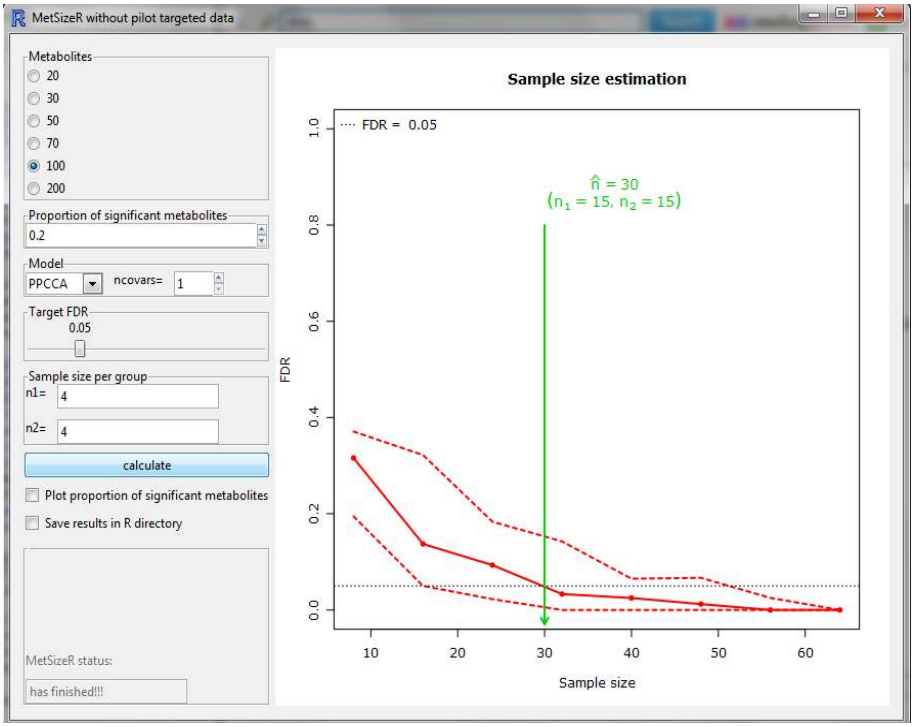

Figure 6: MetSizeR GUI results window for sample size estimation of a targeted metabolomic experiment with no experimental pilot data using the PPCCA model with one covariate. The sample size $\hat{n}$ is estimated at 30 with 15 in each treatment group.

Data format

In order for the experimental pilot data to be uploaded into MetSizeR it should be in tab separated text (txt) format. The data can be transformed into this format by entering it into a spreadsheet, storing samples in rows and metabolites or spectral bins in columns. The first row contains the names of the metabolites or spectral bins. The data are then saved as "tab separated" txt format. Figure 7 shows an example of the data format required by MetSizeR.

	A	B	C	D	E	F	G	H
1	Cystathionione	Lalanine	Sarcosine	Glycine	Ethanolamine	Lleucine	Lisoleucine1	yAminonbutyricacid
2	167.684	218.0255	61.7001	600.8361	331.2262	46.3046	3.8512	7.5035
3	144.118	154.792	95.6593	466.241	779.874	122.164	10.0568	38.7561
4	192.8645	243.5075	71.5076	770.0907	385.6113	72.6932	12.3161	12.4037
5	312.492	394.527	83.6223	186.493	327.154	69.5292	6.055	7.3614
6	267.3125	392.2191	83.1012	778.5596	627.2959	82.3141	11.2923	12.2987
7	168.2221	205.4416	76.289	612.3892	498.9038	80.3872	8.7414	19.5544
8	253.4226	370.7706	95.7844	536.0027	733.9862	71.4972	9.7399	10.9559
9	169.2267	194.7404	55.1105	412.8206	360.7942	41.3592	3.4399	6.7022
10	243.0896	299.5186	92.3215	903.9245	411.055	94.1733	18.4655	19.4575
11	153.964	206.966	101.613	919.567	650.422	89.1006	11.163	21.8174
12	250.4559	308.5949	95.1191	931.3161	423.5112	97.027	19.025	20.0471
13	251.0322	326.3961	92.3684	691.9112	604.7119	69.3205	5.7655	11.2332

Figure 7: Data format for MetSizeR: Samples are stored in the rows and metabolites in columns.

If during the pilot study some covariates were measured they can be uploaded into MetSizeR separately from the actual spectra. The data from covariates should also be in tab separated text (.txt) format with the first row containing the names of the covariates.

Uploading pilot data

The pilot data (spectra) can be uploaded into the MetSizeR by selecting "File -> Open". If the pilot data contains covariates they can be uploaded through "File -> covariates" as in Figure 8. If the pilot data is uploaded into MetSizeR successfully, the path for the pilot data will appear in the R console,

"C:/BioInfo/bmi.txt".

Figure 8: Uploading pilot data and covariates in MetSizeR.

Demo pilot data

A demonstration experimental pilot data set already in MetSizeR can be loaded by clicking the option "demo nmr pilot data" under the 'File' menu as shown in Figure 8. The demo pilot data is an extract from a metabolomic study of epilepsy. Urine samples of 18 animals in two treatment groups were collected. The goal was to identify a set of metabolites that discriminate samples from the two treatment groups. The NMR spectra consists of 189 spectral bins with nine samples from

each treatment group. The weights of animals were also measured. For more information about this experiment see Carmody and Brennan (2010). Now suppose we are interested in performing a similar experiment and we want to find out the appropriate sample size to use. Controlling FDR at 5% and setting the expected proportion of significant bins at 20%, the PPCCA model with weight as a covariate is used to estimate the sample size. Figure 9 shows the estimated sample size using the demo experimental pilot data.

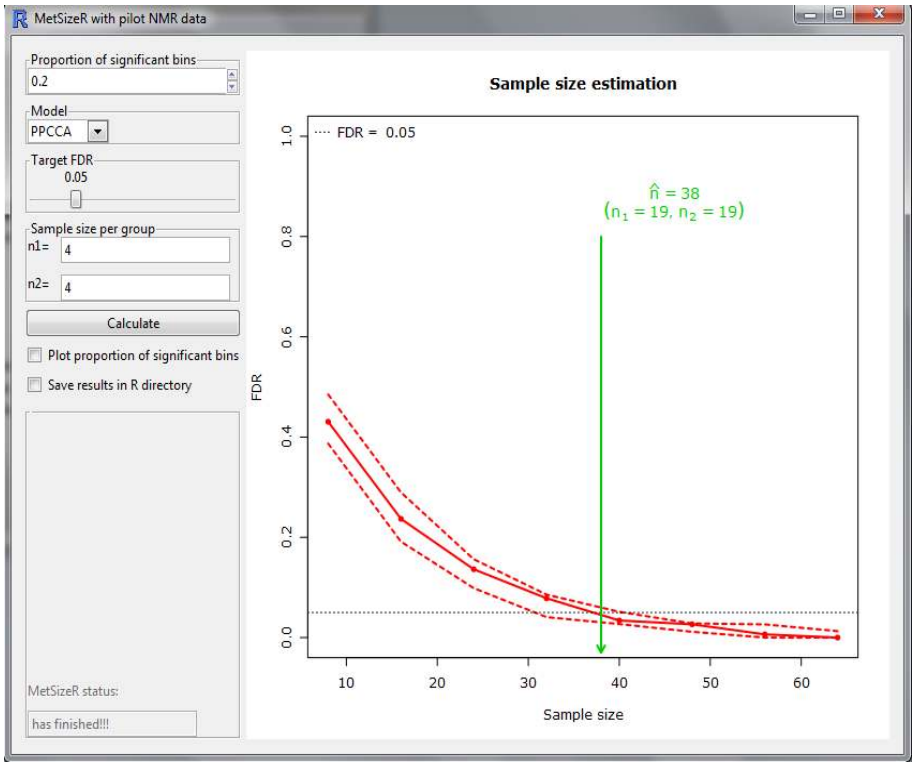

Figure 9: MetSizeR GUI results window for sample size estimation of the demo pilot NMR data using a PPCCA model with weight as a covariate. The sample size is estimated to be 40 with 20 in each treatment group.

The final example is on sample size calculation for a longitudinal metabolomic experiment with experimental pilot data from a targeted metabolomic experiment. Controlling FDR at 5% and setting the expected proportion of significant bins at 20%, the DPPCA model is used to estimate the sample size. Figure 10 shows the expected number of samples required for a longitudinal study using experimental pilot data from targeted metabolomic experiment.

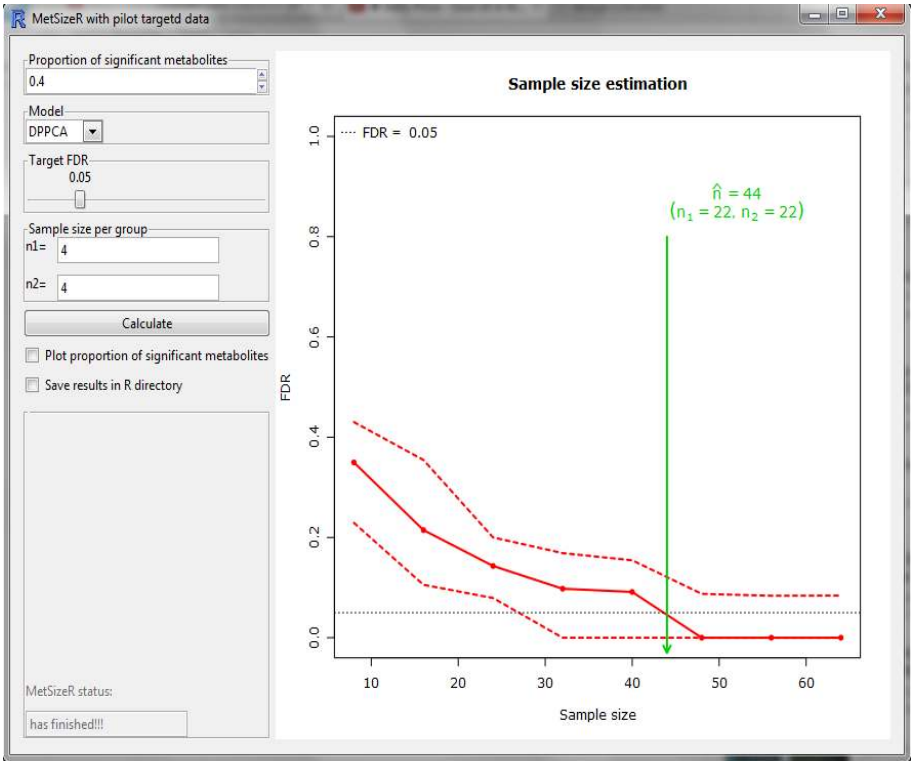

Figure 10: MetSizeR GUI window for estimating sample size for a targeted longitudinal metabolomic experiment with pilot data using a DPPCA model.

References

Carmody, S., Brennan, L., 2010. Effects of pentylenetetrazole-induced seizures on metabolomic profiles of rat brain. Neurochemistry International 56, 340–344.

Nyamundanda, G., Gormley, I.C., Brennan, L., 2010. Probabilistic principal component analysis for metabolomic data. BMC Bioinformatics 11.

Nyamundanda, G., Gormley, I.C., Brennan, L., 2012a. A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data. Technical Report. University College Dublin.

Nyamundanda, G., Gormley, I.C., Fan, Y., Gallagher, W.M., Brennan, L., 2012b. MetSizeR: selecting the optimal sample size for metabolomic studies using an analysis based approach. Technical Report. University College Dublin.

R Development Core Team, 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Tipping, M.E., Bishop, C.M., 1999. Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61, 611–622.

Verzani, J., 2009. gWidgets:gWidgets API for building toolkit-independent, interactive GUIs. R package version 0.0-35.

Appendix

File	
open	upload the txt file containing the spectrum.
covariates	upload the txt file containing the covariates.
demo nmr pilot data	load the demonstration NMR data already in MetSizeR.
quit	close the GUI.
Sample size	
pilot data	sample size estimation with experimental pilot data.
no pilot data	sample size estimation with no experimental pilot data available.
NMR data	sample size estimation for a NMR experiment.
Targeted data	sample size estimation for a targeted experiment.
Spectral bins	number of spectral bins in the NMR experiment.
Metabolites	number of metabolites for targeted analysis.
Proportion of significant bins	proportion of spectral bins expected to be significant.
Proportion of significant metabolites	proportion of metabolites expected to be significant.
Models	different types of models available.
ncovars	number of covariates for the PPCCA model.
Target FDR	level of control over type I errors.
Sample size per group	smallest sample size to be considered in each treatment group.
Save results in R directory	save results in the R working directory.
calculate	estimate the sample size.
Plot proportion of significant bins	assess the effect of varying the expected roportion of significant bins.
Plot proportion of significant metabolites	assess the effect of varying the expected proportion of significant metabolites.
MetSizeR status	displays if MetSize has finished estimating the sample size.
Help	
manual	manual for MetSizeR.