MetSizeR: selecting the optimal sample size for metabolomic studies using an analysis based approach.

Gift Nyamundanda [1] , Isobel Claire Gormley [*1] , Yue Fan[2] , William M Gallagher[2] , Lorraine Brennan[3]

[1]School of Mathematical Sciences, University College Dublin, Ireland.
[2]School of Biomolecular and Biomedical Science, University College Dublin, Ireland.
[3]School of Agriculture and Food Science, Conway Institute, University College Dublin, Ireland.

Email: Gift Nyamundanda - gift.nyamundanda@ucd.ie; Isobel Claire Gormley*- claire.gormley@ucd.ie; Yue Fan - yue.fan@ucd.ie; William M Gallagher - william.gallagher@ucd.ie; Lorraine Brennan - lorraine.brennan@ucd.ie;

*Corresponding author

Abstract

Background: Determining sample sizes for metabolomic experiments is important but due to the complexity of these experiments, there are currently no standard methods for sample size estimation in metabolomics. Since pilot studies are rarely done in metabolomics, currently existing sample size estimation approaches which rely on pilot data can not be applied.

Results: In this article, an analysis based approach called MetSizeR is developed to estimate sample size for metabolomic experiments even when experimental pilot data are not available. The key motivation for MetSizeR is that it considers the type of analysis the researcher intends to use for data analysis when estimating sample size. MetSizeR uses information about the data analysis technique and prior expert knowledge of the metabolomic experiment to simulate pilot data from a statistical model. Permutation based techniques are then applied to the simulated pilot data to estimate the required sample size.

Conclusions: The MetSizeR methodology, and a publicly available software package which implements the approach, are illustrated through real metabolomic applications. Sample size estimates, informed by the intended statistical analysis technique, and the associated uncertainty are provided.

Background

In many metabolomic experiments, one of the most important objectives is to discover the set of metabolites that plays a significant role in distinguishing samples from two different groups or populations and thus, in the identification of novel biomarkers [1]. As in any experiment, designing the experiment is critical if reliable statistically significant metabolites are to be obtained. Since metabolomic experiments are expensive, it is crucial to determine the optimal sample size $\hat{n}$ to attain the desired power to identify discriminating metabolites without wasting resources or unnecessarily sampling many subjects. However, metabolomic data are typically high dimensional and correlated meaning sample size estimation using classical statistical methods is not straight forward.

Currently, in the metabolomics literature, there is no standard method for the determination of sample size when designing a metabolomic experiment. Several methods currently exist in the literature for sample size selection in the high-dimensional data setting [2–5]. However, none of these methods are suitable for metabolomic experiments since many either assume variables have equal variance or are independent. More importantly, these methods rely on the existence of some experimental pilot data on which the actual sample size selection is then based, and are not based on the method to be used to analyze the data. In metabolomic studies, experimental pilot data are rarely available, making such sample size selection approaches redundant.

In this article, we propose a method known as MetSizeR for sample size estimation for metabolomic experiments that addresses some of these limitations. MetSizeR is founded on the idea that the method for selecting sample size firmly depends on the type of data analysis the researcher intends to employ. In a situation where experimental pilot data are not available, pseudo-metabolomic data are simulated from a statistical model. The specific statistical model from which the pseudo-metabolomic data are simulated depends on the type of statistical analysis that the metabolomic scientist intends to use. In its current form the MetSizeR approach assumes the user intends to employ one of the following three statistical analysis techniques on completion of their experiment:

1. Probabilistic Principal Components Analysis (PPCA) [6, 7].
2. Probabilistic Principal Components and Covariates Analysis (PPCCA) [7].
3. Dynamic Probabilistic Principal Components Analysis (DPPCA) [8].

Intuitively the MetSizeR method can be naturally extended to include other analysis approaches, assuming they are based on a statistical model rather than being non-parametric in nature.

MetSizeR draws on two currently existing methods (see [2] and [3]) for sample size calculation in high-dimensional data settings. While the approach in [3] is based on the existence of an experimental pilot data set, the approach detailed in [2] simulates pilot data from a statistical model. Further, while independence in the data is assumed in [2], the approach in [3] uses permutation methods to account for the correlation in the experimental pilot data. MetSizeR combines these ideas of prior simulation and permutation based techniques to estimate the sample size for metabolomic experiments. The main advantage of the developed approach is its ability to determine sample size without experimental pilot data and without assuming variable independence.

A graphic user interface (GUI) software called MetSizeR was developed to implement this approach to estimating sample sizes in R [9]. Effort was focused on designing the interface of MetSizeR to encourage its wide use in the metabolomics community regardless of previous knowledge of R. The software is available through the **R** statistical software environment `www.r-project.org`.

Methods

Metabolomic data sets are typically acquired using analytical technologies such as nuclear magnetic resonance spectroscopy (NMR) [10] and mass spectrometry (MS) [11]. The spectrum resulting from NMR spectroscopy is usually divided into spectral bins (representing variables) and the signal intensities within the bins are related to the relative abundances of metabolites. MS is typically used for targeted metabolomics in which a specified list of metabolites is measured [12]. The following section describes how the number of samples required for either an NMR or an MS metabolomic experiment can be determined under the MetSizeR approach.

Sample size estimation

Let $\bar{x}_{jg}$ be the estimate of the average signal intensity μ_{jg} for metabolite j in samples from the treatment group g which has corresponding sample size n_g, where $g = 1, 2$. Often in metabolomics, the goal of discovering a set of metabolites that discriminates between samples from two treatment groups is achieved by testing the hypothesis $\mathrm{H}_{oj} : \mu_{j1} - \mu_{j2} = 0$, on each metabolite j, where $j = 1, \ldots, p$. The aim of discovering discriminating metabolites can be framed as a multiple testing problem as there are p hypotheses to be tested and the probability of falsely declaring a metabolite as significant increases with p. It is therefore important to estimate sample size while controlling an error rate to improve the power of the test for identifying significant metabolites. MetSizeR focuses on controlling the false discovery rate

(FDR, [13]). Here, the FDR is the expected number of metabolites incorrectly deemed to be significantly different between the two treatment groups, as a proportion of the total number of metabolites declared to be significant.

The test statistic and its distribution

A test statistic widely used to identify discriminating metabolites is a two sample t-statistic. The t-statistic TS is evaluated for all metabolites, $j = 1, \ldots, p$, under the assumption that the null hypothesis of no difference $\mu_{j1} = \mu_{j2}$ is true:

$$\begin{aligned} TS_j &= \frac{(\bar{x}_{j1} - \bar{x}_{j2})}{S_j + cf}, \\ \text{where} \quad S_j &= \left\{ \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \frac{(n_1 - 1)(s_{j1})^2 + (n_2 - 1)(s_{j2})^2}{n_1 + n_2 - 2} \right\}^{\frac{1}{2}}, \end{aligned}$$

where S_j is the estimate of the pooled standard error for metabolite j. The corresponding within treatment variability estimate is $s_{jg}^2 = (n_g - 1)^{-1} \sum_{i=1}^{n_g} (x_{(jg)i} - \bar{x}_{jg})^2$ for $g = 1, 2$ where $x_{(jg)i}$ denotes the signal intensity for metabolite j in sample i from the treatment group g. A correction factor cf is a small positive value added to the standard error of each metabolite to prevent some metabolites with signal intensity near zero from having large test statistic TS_j; such a metabolite may have $TS_j \approx 0/0$.

The typical assumption about the null distribution (i.e. the distribution under the null hypotheses) of the test statistic TS_j is the t-distribution with $n_1 + n_2 - 2$ degrees of freedom. However, when the data violate such an assumption, misleading sample size estimates would result. Hence, as in [3], MetSizeR estimates the null distribution of TS_j using a permutation technique. This is a non-parametric method based on the assumption that under the null hypothesis of no difference, the distribution of the test statistic does not change no matter how the group labels of the pilot data are permuted. The data generated using this approach maintains the between subject variability and the amount of noise in the data. The null distribution of the test statistic TS is estimated by randomly permuting the group labels of pilot data and calculating the test statistic for each metabolite, TS_{jt}, where $t = 1, \ldots, T$ permutations.

Analysis based data simulation

Unfortunately, in most cases, experimental pilot data are not readily available in metabolomics since pilot studies are rarely done. Therefore, MetSizeR uses the intended statistical analysis model to simulate pilot data. The simulated pilot data are then used to learn about the null distribution of the relevant test statistic for estimating sample size. This simulation approach is similar to that in [2] in which pilot data

are simulated from the marginal model:

$$p(\mathbf{x}) \quad = \quad \int p(\mathbf{x}|\mathbf{u},\theta)dp(\mathbf{u},\theta),$$

where $\mathbf{x}$ is the $n \times p$ data matrix, $\mathbf{u}$ denotes the latent variables, and θ is a vector of unknown model parameters. Simulating from the marginal model is achieved by first generating values of the parameters and the latent variables from the prior distribution $p(\mathbf{u},\theta)$, and then simulating the data from the assumed model $p(\mathbf{x}|\mathbf{u},\theta)$ given the simulated values of $\mathbf{u}$ and θ.

Currently, MetSizeR assumes the metabolomic practitioner will use one of three different statistical models $p(\mathbf{x}|\mathbf{u},\theta)$ to analyse the data from their metabolomic experiment – either the PPCA, PPCCA or DPPCA model. Simulation of the parameters of these models is based on the model assumptions and on prior expert knowledge of metabolomic data properties. As PPCA is equivalent to the widely used Principal Components Analysis (PCA) method, simulating from the PPCA model is discussed here; details of the simulation of pilot data from the closely related PPCCA and DPPCA models are provided in the Additional File. Specifically, PPCA is a probabilistic formulation of PCA based on a Gaussian latent variable model [6, 7]. PPCA models the high dimensional spectrum $\underline{x}_i^T = (x_{i1}, \ldots, x_{ip})$ of subject i ($i = 1, \ldots, n$ where $n = n_1 + n_2$) as a linear function of the corresponding low dimensional latent variable $\underline{u}_i^T = (u_{i1}, \ldots, u_{iq})$, where ($q \ll p$). The PPCA model can be expressed as follows

$$\underline{x}_i \quad = \quad \mathbf{W}\underline{u}_i + \underline{\mu} + \underline{\epsilon}_i$$

where $\mathbf{W}$ is a $p \times q$ loadings matrix, $\underline{\mu}$ is a mean vector and $\underline{\epsilon}_i$ is multivariate Gaussian noise for observation i, i.e. $p(\underline{\epsilon}_i) = \text{MVN}_p(\underline{0}, \sigma^2\mathbf{I})$ where $\mathbf{I}$ denotes the identity matrix. The latent variable $\underline{u}_i$ is also multivariate Gaussian distributed, $p(\underline{u}_i) = \text{MVN}_q(\underline{0}, \mathbf{I})$. The maximum likelihood estimates of the loadings matrix $\mathbf{W}$ and the latent variable $\mathbf{u}$ in the PPCA model are equivalent to the traditional PCA loadings matrix and principal component scores. For a given sample size n, pilot data $\mathbf{x}$ can be simulated from the PPCA model as follows:

1. Generate parameter values from their prior distributions:

$$\begin{aligned} p(\underline{u}_i) &= \text{MVN}_q(\underline{0}, \mathbf{I}) \text{ for } i = 1, \ldots n. \\ p(\underline{w}_j) &= \text{MVN}_q(\underline{\mu}_W, \mathbf{\Sigma}_W) \text{ for } j = 1, \ldots p. \\ p(\sigma^2) &= \text{IG}(\alpha_1, \alpha_2) \end{aligned}$$

2. Given the generated model parameters and latent variables the pilot data $\mathbf{x}$ are then simulated from the PPCA model:

$$p(\underline{x}_i|\underline{u}_i, \mathbf{W}, \sigma^2) \quad = \quad \text{MVN}_p(\mathbf{W}\underline{u}_i, \sigma^2\mathbf{I}) \text{ for } i = 1, \ldots, n.$$

Estimating sample size based on pilot data simulated in this way ensures the estimated sample size is firmly dependent on the type of model being used to analyse the real experimental metabolomic data. Hence, MetSizeR represents an analysis based approach to sample size estimation for metabolomic studies. The specific steps involved in the MetsizeR algorithm are detailed in the next section.

The MetSizeR Algorithm

The MetSizeR procedure for sample size estimation starts with a number $ntry$ of different sample sizes and a user-specified FDR (denoted by $target.fdr$). It then searches for the optimal sample size $\hat{n}$ by estimating the FDR for each of the $ntry$ sample sizes. In order to estimate FDR for each sample size, the null distribution of the test statistics of all metabolites is estimated and then a shift constant is added to the test statistics of some p_o metabolites to allow them to be truly significant. The null distribution is estimated by calculating the test statistics of the permuted pilot data. After obtaining the critical values of the null distribution, the FDR is estimated. The optimal sample size $\hat{n}$ is then set to be the sample size with FDR equal to $target.fdr$.

In summary, the MetSizeR sample size estimation method proceeds as follows:

1. Specify the input parameters which include the desired level of FDR ($target.fdr$), the expected proportion m of significant metabolites and the model to be used when analyzing the observed metabolomic data.

2. Simulate pilot data of sample size n_k from the assumed analysis model, where $k = 1, \ldots, ntry$. Pilot data simulation from the PPCA model is detailed in the previous Section; the Additional File details pilot data simulation from the PPCCA and DPPCA models.

3. Estimate the null distribution for all metabolites by randomly permuting the group labels of the simulated pilot data and computing the test statistic TS_{jt} for each metabolite j and each permuted data set t for T permutations.

4. Estimate the FDR for each permuted data set $t = 1, \ldots, T$:

(a) Consider the corresponding p-vector of the test statistics $\underline{TS}_t = (TS_{1t}, TS_{2t}, \ldots, TS_{pt})$ for all metabolites on permutation t.

(b) Randomly sample $p_o = m \times p$ of the test statistics $\underline{TS}_t$ and add $\frac{\delta}{\varrho_{jt}(\sqrt{\frac{1}{n_1}+\frac{1}{n_2}})}$ to their intensities. This allows p_o metabolites to be truly significant. Here, δ is the effect size, and ϱ_{jt} is the true within group standard deviation estimated by $\frac{S_{jt}}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$.

(c) A cut off point $crit$ is set to be the p_o^{th} largest absolute value of the test statistics $\underline{TS}_t$. All metabolites with $|TS_{jt}| > crit$ are declared as significant. The FDR for permutation t can then be calculated.

5. Estimate the FDR for data simulation s by taking the 50^{th} percentile of the FDR values of $1, \ldots, T$ permutations.

6. Repeat steps 2 to 5 for $s = 1, \ldots, SIM$ simulations and report the $10^{th}, 50^{th}$ and 90^{th} percentiles of the FDR values for sample size n_k.

7. Repeat steps 2 to 6 for $k = 1, \ldots, ntry$ different sample sizes and select the optimal sample size $\hat{n}$ as the n_k with FDR $= target.fdr$.

The total number of permutations T used to estimate the sampling distribution of the test statistics TS was chosen to be twenty. In the `samr` R package [3] 20 permutations were used to estimate the null distribution and they give accurate estimates of the FDR. Here, the value of the effect size δ is chosen based on the variance of the underlying model. The optimal sample size $\hat{n}$ is estimated by predicting the sample size at $target.fdr$ using a simple linear regression model on values of FDR above and below the $target.fdr$ with their corresponding sample sizes n_k. The estimated sample size by MetSizeR ensures that the power or the confidence level in statistical tests reaches (1-$target.fdr$).

Parameter Specification: details and guidelines.

The MetSizeR algorithm requires the specification of several parameters; some are parameters relevant to the intended analysis model, and some are parameters relevant to the sample size estimation procedure itself.

In terms of the MetSizeR GUI which has been developed, the user is requested to specify parameters specific to the sample size estimation procedure i.e. the number of bins in the NMR or MS spectrum, the expected proportion of significant bins, the target FDR and the minimum sample size they wish to be

considered. The default settings of these parameters are 200 spectral bins, 20% significantly different bins, a target FDR of 5% and a minimum sample size of 4. The choice of the number and proportion of significantly different spectral bins will naturally be informed by the metabolomic practitioner's knowledge, as will the minimum sample size choice. For the target FDR, again this depends on the conservatism of the metabolomic practitioner and/or the research question of interest, but a FDR of 5% is indicative of typical statistical practice. The user can easily re-run the MetSizeR algorithm for different settings of these parameters to ascertain the influence of their particular specifications. However, within the MetSizeR GUI the user has the option of requesting plots of the expected proportion of significant bins versus the FDR, over different sample sizes, giving insight to the influence of this particularly influential parameter on sample size estimation. Regarding the specification of parameters relevant to the intended analysis model, in the MetSizeR GUI, the user is only required to specify the intended analysis model (PPCA, PPCCA or DPPCA), and in the case of PPCCA, the number of covariates to be included. Both of these decisions are again practitioner informed, depending on the particular experiment under consideration. The MetSizeR manual, available through the developed MetSizeR GUI, guides the user through these parameter specification steps using a number of illustrative examples.

The remaining parameters in the MetSizeR algorithm have been fixed within the **R** code underlying the MetSizeR GUI, but given the open source nature of **R**, these can be changed by the user if desired. In the context of the PPCA model discussed above the hyperparameters of the prior distributions of the loadings matrix $\mathbf{W}$ and the variance σ^2 are based on previous estimates of $\mathbf{W}$ and σ^2 from applications of PPCA to metabolomic data (eg. [7,8]). Each row of the loadings matrix $\mathbf{W}$ is simulated from a standard multivariate Gaussian distribution $\text{MVN}_q(\underline{0}, \mathbf{I})$ and the noise variance σ^2 is simulated from an inverse gamma distribution with shape parameter $\alpha_1 = 3$ and scale parameter $\alpha_2 = 4$. Hyperparameter settings for the PPCCA and DPPCA models are detailed in the Additional File. Within the MetSizeR algorithm four final parameters are specified: the effect size δ (fixed at 2.3, the 99th quantile of the assumed prior distribution of the loadings), the correction factor cf (fixed as the fifth percentile of the estimated standard errors of all metabolites), the number of permutations T (set to 20) and the number of simulations SIM (set at 20). These specifications are based on the choices in [3, 5, 14] in similar sample size estimation settings.

Results

This section illustrates the application of MetSizeR to different metabolomic experimental settings. In the first section, MetSizeR is employed to estimate sample size in the setting where experimental pilot data are

not available; the second section considers the case where experimental pilot data are available.

Sample size estimation using simulated pilot data

Here the MetSizeR approach to sample size estimation is illustrated in the setting where experimental pilot data are not available and it is assumed that the user has indicated that a PPCA model will be used to analyze the observed experimental data. Further, it is assumed that the user has specified that the spectra will consist of 300 spectral bins, the target FDR is 5% and the expected proportion of significant spectral bins is 20%. In this example, the user has also specified that they wish to consider a minimum sample size of ten, with five in each treatment group (i.e. $n_1 = 5$ and $n_2 = 5$). All other MetSizeR parameters are set at their default values, as detailed in the previous section. The MetSizeR method was then applied, and the $10^{th}, 50^{th}$ and 90^{th} percentiles of the FDR were calculated across a range of sample sizes and are shown in Figure 1. The sample size at which the target FDR of 5% was achieved was estimated to be 30 with 15 in each treatment group as shown in Figure 1(A).

The expected proportion of significant spectral bins specified by the user impacts on the estimated number of samples required for the metabolomic experiment. Figures 1(B), 1(C) and 1(D) demonstrate the effect on FDR of varying the expected proportion of significant spectral bins for three different sample sizes. The figures show that, increasing the expected proportion of significant spectral bins reduces the FDR.

A second example which demonstrates the applicability of MetSizeR is based on an experimental paradigm where additional information is available in the form of covariates. In this instance, the PPCCA model will be used to analyze the acquired data and thus was used to simulate pilot data with 300 spectral bins, five samples from each treatment group and two covariates. Fixing the target FDR at 5% and the expected proportion at 20%, Figure 2(A) demonstrates that when two covariates are included in the PPCCA model, the total number of samples required for such an experiment increases to 36 with 18 samples in each treatment group.

Figure 2(B) illustrates a third example of the setting where no experimental pilot data are available and the practitioner aims to conduct a longitudinal metabolomic experiment. The pilot data for this example are simulated from the DPPCA model; the data are simulated by only focusing on the first time point of the experiment as it is expected that the same number of subjects will be followed over time and that, while there may be dropouts, the largest number of subjects will be present at the first time point. Figure 2(B) shows that the expected number of samples required for a longitudinal study of 300 spectral bins with 20% significant bins and a target FDR of 5%, is 24 with 12 samples from each treatment group.

Sample size estimation with experimental pilot data

In a situation where experimental pilot data are available, parameter estimates used for simulations are based on fitting the underlying model to the experimental pilot data. Here, the application of MetSizeR is illustrated using real metabolomic data sets as experimental pilot data.

The first experimental pilot data set is from a longitudinal metabolomic animal study. Urine samples of 18 animals in two treatment groups were collected over a 15 day period and the animals' weights were measured. Details of this study have been previously detailed in [15]. Data from day 10 of the study were used as experimental pilot data here; the NMR spectra consist of 189 spectral bins with nine samples in each treatment group. The PPCCA model was fitted to the experimental pilot data, with weight as a covariate and the maximum likelihood parameter estimates from fitting this model are used for data simulations in MetSizeR. Controlling the target FDR at 5% and setting the expected proportion of significant bins at 20%, the MetSizeR method was employed. Figure 3(A) depicts that the sample size estimate is 40, with 20 samples in each treatment group. It is interesting to note that, the 10% and 90% curves in Figure 3(A) are much narrower than in the previous examples in which MetSizeR was used to estimate sample size with no experimental pilot data (Figures 1 and 2). This reduction in uncertainty is due to the fact that MetSizeR simulations are now based on fixed parameter values not on draws from prior distributions as used when experimental pilot data are not available.

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. Setting MetSizeR specifications as in the previous examples, the PPCA model was fitted to a targeted metabolomic MS pilot data set and under the MetSizR algorithm, the estimated sample size is shown in Figure 3(B).

Conclusions

Determining sample sizes in metabolomics is important but due to the complexity of these experiments, there are currently no standard methods for sample size estimation in metabolomics. Moreover, since pilot studies are rarely done in metabolomics, sample size estimation approaches for high dimensional data studies requiring experimental pilot data, cannot be applied.

The method presented in this article is a straight forward approach for determining sample sizes for metabolomic experiments whilst controlling the FDR. The main advantage of the developed approach is its ability to determine sample size even when experimental pilot data are not available. Another key

advantage is that it takes the type of analysis the researcher intends to use into consideration when estimating sample size and this can improve the power of the study. Also, since MetSizeR employs permutation techniques to estimate sample size, it accounts for correlation between metabolites and effect size variability. The method has been developed to accept both NMR and targeted MS data which will ensure wide applicability in the metabolomics community. Further, a software package facilitates easy implementation of the MetSizeR approach.

Areas of future work are multiple and varied. MetSizeR is currently designed to estimate the number of samples required for metabolomic experiments which involve two groups; modifications to the MetSizeR approach are possible to accommodate different metabolomic experimental designs. Alternatives to the permutation approach employed in MetSizeR could be examined – bootstrap sampling would provide an interesting alternative. Proof of concept metabolomic experiments are currently underway to validate the MetSizeR approach.

Availability and requirements

The package `MetSizeR` has been developed for the **R** statistical environment (`www.r-project.org`) and is freely available at `cran.r-project.org`. The package is accompanied by documentation files to facilitate its use.

Project name: `MetSizeR`

Project home page: `cran.r-project.org/web/packages/MetSizeR/`

Operating system(s): Platform independent.

Programming language: **R** platform.

Other requirements: No.

License: GPL ($\geq$ 2)

Any restrictions to use: It is available for free download.

Competing interests

The authors declare that they have no competing interests.

Authors contributions

NG was involved in algorithm development, software development and manuscript writing. ICG was involved in algorithm development, software interpretation and manuscript writing. YF and WMG were

involved in algorithm and software development. LB was involved in the study hypothesis, software interpretation and manuscript writing. All authors read and approved the final manuscript.

Authors information

Nyamundanda Gift is a PhD candidate in the PhD in Bioinformatics and Systems Biology programme in University College Dublin. Dr. Isobel Claire Gormley is a lecturer in Statistics in the School of Mathematical Sciences, University College Dublin. Dr. Yue Fan is a Postdoctoral Researcher in the School of Biomolecular and Biomedical Science in University College Dublin. Prof. William Gallagher is an Associate Professor in the School Of Biomolecular and Biomedical Science in University College Dublin. Dr. Lorraine Brennan is a lecturer in Nutritional Biochemistry in the School of Agriculture and Food Science, Conway Institute, University College Dublin.

Acknowledgements

This research was supported by the Irish Research Council for Science Engineering and Technology (IRCSET) funded PhD Programme in Bioinformatics and Systems Biology in University College Dublin, Ireland (`bioinformatics.ucd.ie/PhD/`) and by HRB Ireland (RP/2006/117).

References

1. Berk M, Ebbels T, Montana G: **A statistical framework for biomarker discovery in metabolomic time course data.** *Bioinformatics* 2011, **27**(14):1979–1985.
2. Muller P, Parmigiani G, Robert C, Rousseau J: **Optimal sample size for multiple testing.** *Journal of the American Statistical Association* 2004, **99**(468):990–100.
3. Tibshirani R: **A simple method for assessing sample sizes in microarray experiments.** *BMC Bioinformatics* 2006, **7**(106).
4. Liu P, Hwanga JTG: **Quick calculation for sample size while controlling false discovery rate with application to microarray analysis.** *Bioinformatics* 2007, **23**(6):739–746.
5. Lin WJ, Hsueh HM, Chen JJ: **Power and sample size estimation in microarray studies.** *BMC Bioinformatics* 2010, **11**(48).
6. Tipping ME, Bishop CM: **Probabilistic principal component analysis.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1999, **61**(3):611–622.
7. Nyamundanda G, Gormley IC, Brennan L: **Probabilistic principal component analysis for metabolomic data.** *BMC Bioinformatics* 2010, **11**(571).
8. Nyamundanda G, Gormley IC, Brennan L: **A dynamic probabilistic principal components model for the analysis of longitudinal metabolomics data.** Tech. rep., University College Dublin 2012.
9. R Development Core Team: *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria 2009, [http://www.R-project.org].
10. Reo NV: **Metabonomics based on NMR spectroscopy.** *Drug and Chemical Toxicology* 2002, **25**(4):375–382.

11. Dettmer K, Aronov PA, Hammock BD: **Mass spectrometry-based metabolomics.** *Mass Spectrometry Reviews* 2007, **26**:51–78.

12. Patti GJ, Yanes O, Siuzdak G: **Innovation: Metabolomics: the apogee of the omics trilogy.** *Nature Reviews Molecular Cell Biology* 2012, **13**(263–269).

13. Benjamini Y, Hochberg Y: **Controlling false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society* 1995, **57**:289–300.

14. Hwang D, Schmitt WA, Stephanopoulos G, Stephanopoulos G: **Determination of minimum sample size and discriminatory expression patterns in microarray data.** *Bioinformatics* 2002, **18**:1184–1193.

15. Carmody S, Brennan L: **Effects of pentylenetetrazole-induced seizures on metabolomic profiles of rat brain.** *Neurochemistry International* 2010, **56**(2):340–344.

Figures

Figure 1 – Sample size estimation without experimental pilot data using the PPCA model.

In each panel is the estimated FDR (solid red lines) as well as the 10^{th} and 90^{th} percentiles (dashed red lines). A horizontal dashed black line is the target FDR at 5%. **(A)** The sample size $\hat{n}$ is estimated to be 30 with 15 samples in each treatment group. **(B-D)** show the effect of varying the proportion of significant bins over a range of sample sizes.

Figure 2 – Sample size estimation without experimental pilot data using the PPCCA and DPPCA models.

(A) The estimated sample size using the PPCCA model with two covariates. **(B)** The estimated sample size for a longitudinal study using the DPPCA model.

Figure 3 – Sample size estimation with pilot data.

(A) The estimated sample size using the PPCCA model on NMR pilot data with weights of subjects as a covariate. **(B)** The estimated sample size using the PPCA model with targeted MS metabolomic pilot data.

Additional Files

Additional File — Simulating from the PPCCA and DPPCA models.