Author: Iterson, Maarten van
Title: The power of high-dimensional data in genomics research
Issue Date: 2013-11-28

General power and sample size calculations for high-dimensional genomic data

M. van Iterson[1], M. van de Wiel[2], J.M. Boer[3,4] and R.X. Menezes[2,4]

[1] *Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands,*

[2] *Department Epidemiology and Biostatistics, VU Medical Center, Amsterdam, the Netherlands,*

[3] *Laboratory of Pediatrics Erasmus Medical Center, Sophia Children's Hospital, Rotterdam, the Netherlands,*

[4] *Netherlands Bioinformatics Centre, Nijmegen, the Netherlands*

***Statistical Applications in Genetics and Molecular Biology*, 12(4):449–467, 2013.**

Abstract

In the design of microarray or next-generation sequencing experiments it is crucial to choose the appropriate number of biological replicates. As often the number of differentially expressed genes and their effect sizes are small and too few replicates will lead to insufficient power to detect these. On the other hand, too many replicates unnecessary leads to high experimental costs. Power and sample size analysis can guide experimentalist in choosing the appropriate number of biological replicates. Several methods for power and sample size analysis have recently been proposed for microarray data. However, most of these are restricted to two group comparisons and require user-defined effect sizes.

Here we propose a pilot-data based method for power and sample size analysis which can handle more general experimental designs and uses pilot-data to obtain estimates of the effect sizes. The method can also handle χ^2 distributed test statistics which enables power and sample size calculations for a much wider class of models, including high-dimensional generalized linear models which are used e.g. for RNA-seq data analysis.

The performance of the method is evaluated using simulated and experimental data from several microarray and next-generation sequencing experiments. Furthermore, we compare our proposed method for estimation of the density of effect sizes from pilot data with a recent proposed method specific for two group comparisons.

*keywords:*Density of effect-sizes; Discrete inverse problem; High-dimensional generalized linear models; Non-negative Conjugate Gradients algorithm.

4.1 Introduction

Sample size determination is concerned with the question of determining the minimum number of samples necessary to demonstrate the existence (or absence) of a difference between two or more populations of interest. The number of samples should be sufficient in that the statistical test will reject the null hypothesis, when there really exists a difference, with high probability or power. Traditional methods for sample size determination apply only to a single hypothesis test and cannot directly be applied to microarray or next-generation sequencing data.

Lee and Whitmore [2002] were one of the first who describe power and sample size analysis for microarray studies. The multiple testing problem was controlled using the easy to apply family-wise error rate. However, controlling the family-wise error rate is often too conservative for microarray data. Consequently, methods were proposed that control the multiple testing problem using the false discovery rate e.g. [Pawitan et al., 2005, Liu and Hwang, 2007, Tong and Zhao, 2008]. Recently, those methods were further adapted for using pilot data in order to obtain more realistic estimates [Ferreira and Zwinderman, 2006a, Ruppert et al., 2007, Jørstad et al., 2008], albeit restricted to two group comparisons.

One important difference between the current methods is how effect sizes, i.e., the fold-changes of the differentially expressed genes, are obtained and incorporated. For example, some authors considered a single fixed effect size for the differentially expressed genes [Tsai et al., 2004, Pawitan et al., 2005, Shao and Tseng, 2007], while others [Jung, 2005, Tong and Zhao, 2008] took the largest effect sizes of a pilot study (e.g. largest observed two-sample t test-statistics). The method by Liu and Hwang [2007] can handle a given distribution of effect sizes, parametric or non-parametric, although they do not incorporate its estimation in their procedure. A few methods estimate the distribution of effect sizes from pilot data [Ferreira and Zwinderman, 2006b, Ruppert et al., 2007, Jørstad et al., 2008, Efron, 2009, Matsui and Noma, 2011, Long et al., 2012]. However, these methods are only applicable to two-group comparisons using test statistics that can be approximated by the Student's t or the normal distribution, so methods for more general comparisons are still lacking.

Here we present a general pilot data-based method for power and sample size determination for high-dimensional data, controlling the FDR by using the adaptive version of the Benjamini-Hochberg method [Benjamini and Hochberg, 2000]. The method is based on the framework provided by Ferreira and Zwinderman [2006b], extended to handle test statistic distributions other than the normal. Our method can be used to estimate power of studies involving not only two-group comparisons, but also multiple-group comparisons, tested using an F statistic. Furthermore, our method can be used for power estimation of RNA-seq studies, where data analysis use a generalized linear model framework leading to χ^2 statistic.

The article is organized as follows. Section 4.2 describes estimation of the proportion of truly null hypotheses and the density of effect sizes. In Section 4.3 the performance of the power and sample size method is evaluated using simulated data, two microarray data sets and one data set derived from a RNA-seq experiment. We conclude with a discussion in Section 4.4.

4.2 Method

The inverse problem

Consider a set of test statistics $T_1, \cdots, T_m$. For example, these could be obtained from a microarray or RNA-seq experiment investigating some biological hypothesis of interest on m genes, possibly involving two or multiple conditions. Suppose also that a user-defined statistical model was chosen that led to the test statistic T, i.e. Student's $t-$, $F-$ or χ^2-test. Each individual test statistic will either follow the null hypothesis or the alternative hypothesis. So let us represent by F_0 and F_a the null- and alternative cumulative distribution functions of the test statistics, which follow from the chosen statistical model. Then each test statistic T follows the mixture distribution given by

$$F(t) = \pi_0 F_0(t) + (1 - \pi_0) \int F_a(t; \theta N) f_\Theta(\theta) d\theta, \tag{4.1}$$

where F_0 and F_a are known, whilst the mixing coefficient or proportion of truly null hypotheses π_0 and f_Θ the density of effect sizes are unknown. The integral is over the support of the test statistic T.

Ferreira and Zwinderman [2006a] have shown, under quite general conditions, how the power can be calculated while controlling the FDR at a desired level, within this mixture framework. The average power of the adaptive Benjamini-Hochberg approach is given by $G_a(u^*; N)$, where u^* is the unique positive solution of the equation

$$G_a(u; N) = u \frac{\pi_0(1 - \alpha)}{\alpha(1 - \pi_0)}, \tag{4.2}$$

with α representing the desired FDR level and $G_a()$ is the distribution of the p-values under the alternative hypothesis. Moreover, they have shown, given estimates of π_0 and f_Θ, how power and sample size determination can be performed when test statistics approximately follow a normal distribution, i.e. $F_a(t; \theta N) = \Phi(t - \theta N)$, with $\Phi()$ representing the cumulative distribution function of a standard normal [Ferreira and Zwinderman, 2006c]. For other sample sizes than the pilot-data, $\hat{G}_a(u^*; N')$ gives us an estimate of the average power achieved with sample size N' via

$$\hat{G}_a(u; N') = \int \Gamma(u, \theta; N') \hat{f}_\Theta(\theta) = u \frac{\hat{\pi}_0(1 - \alpha)}{\alpha(1 - \hat{\pi}_0)}, \tag{4.3}$$

where $\Gamma(\cdot, \theta; N')$ is the distribution of a p-value under the alternative hypotheses with effect size θ and sample size N'.

In this approach, it is thus essential to assume that the effect size θ is independent of the sample size N. We show that this assumption holds under more general settings, i.e. that the departure from the null distribution can be factored as an effect size and a sample size, for the commonly used test statistics: Student's t, F and the χ^2 statistics—see Appendix B.1.

To obtain estimates of π_0 and f_Θ, equation (4.1) must be solved. Solving equation (4.1) for f_Θ is called an inverse problem. Specifically, equation (4.1) is a Fredholm integral equation of the first kind. These problems are known to be ill-posed as the discretized versions of these equations lead to ill-conditioned systems of equations. Regularization methods have been proposed for solving these kind of problems [Tikhonov, 1963, O'Sullivan, 1986, Hansen, 2010].

The integral of equation (4.1) represents a convolution if F_a belongs to a location-family of distributions, e.g. $F_a(t; \theta N) = \Phi(t - \theta N)$, then explicit solutions exist [Delaigle and Gijbels, 2007]. This led to the proposal for a kernel deconvolution estimator for estimation of the density of effect sizes by Ferreira and Zwinderman [2006a]. Nevertheless, numerical implementations are not straightforward as numerical integration of highly oscillating functions are involved. Implementations based on the Fast Fourier Transform lead to fast and stable solutions [Delaigle and Gijbels, 2007, van Iterson et al., 2009].

Unfortunately the integral of equation (4.1) does not represent a convolution in general. Furthermore, the presence of π_0 and the constraints on the solution that, f_Θ should be a valid density meaning everywhere non-negative and integrate to one, make equation (4.1) an integral equation with a twist.

Discretization of the integral equation can be done in many ways: approximating the integral numerically or by first representing the density of effect sizes as a sum of basis functions and then using numerical integration. See Appendices B.2 and B.3 for an example of the former approach.

The discretized inverse problem can then be written as a linear system of equations, such as $\mathbf{Ax} = \mathbf{b}$. The matrix $\mathbf{A}$ is typically ill-conditioned. For example, when the integral is approximated on a fine grid two consecutive columns of $\mathbf{A}$ become nearly identical, therefore, $\mathbf{A}$ will not be of full column rank. In the next two sections we described two different approaches for solving the discrete inverse problem.

Tikhonov solution of the discrete inverse problem

Ridge regression was proposed by Hoerl and Kennard [1970] to solve regression problems with multicollinearity. Ridge regression is similar to the earlier proposed regularization method for solving discrete ill-posed problems of Tikhonov [1963]. Tikhonov proposed to force the L_2-norm of the solution not to exceed a certain value thereby the solution is stabilized or regularized. Solving a discretized version of equation (4.1) using Tikhonov regularization leads to the following minimization problem:

$$\min_{\mathbf{x}} ||\mathbf{b} - \mathbf{A}\mathbf{x}||_2^2 + \lambda^2 \, ||\mathbf{x}||_2^2. \tag{4.4}$$

The first term represents the usual least-squares term and the second term the constraint on the size of the solution. The weight of the constraint is balanced by λ, a positive value. Equation (4.4) can be reformulated as an ordinary least-squares problem

$$\min_{\mathbf{x}} \left|\left| \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbf{A} \\ \lambda \mathbf{I} \end{pmatrix} \mathbf{x} \right|\right|_2^2, \tag{4.5}$$

whose solution is given by

$$\mathbf{x}_\lambda = (\mathbf{A}^T\mathbf{A} + \lambda^2\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b}. \tag{4.6}$$

Several methods have been proposed for finding the optimal regularization parameter in the regression literature; such as cross-validation, generalized cross-validation or Akaike's information criterion [Hastie et al., 2001, Ruppert et al., 2003]. However, it turns out that these methods do not always work for discrete inverse problems as can be seen by a plot of the generalized cross-validation error as a function of λ shown in the upper-left panel of Figure 4.1. This plot is useless for finding the optimal value of λ. When plotted on a double-logarithmic scale, shown in the lower left panel of Figure 4.1, the generalized cross-validation error as function of λ represents an S-shaped curve, but still has no clear minimum (Varah [1983] made the same observation).

Hansen and O'Leary [1993] advocated the L-curve for selecting the regularization parameter for ill-posed problems. The L-curve is a $\log - \log$ plot of the solution norm versus the residual norm. The general shape is that of an 'L' where the vertical part represent under-smoothing and the horizontal part over-smoothing. Based on heuristic arguments they propose to use the λ that reflects the corner of the curve (see upper-right panel of Figure 4.1). By plotting or calculating the curvature of the curve the corner can be located.

In a similar way, using an estimate of the curvature, we could locate the first corner of the S-curve. The first corner reflects the transition from slow to rapid increase of the generalized cross-validation error when reducing the model complexity (see lower-left panel of Figure 4.1).

Figure 4.1 suggests that there is a close relation between the generalized cross-validation error on the log-scale and Akaike's information criterion; compare Figure 4.1 lower-left and -right panels. In Appendix B.4 we give a verification of this observation.

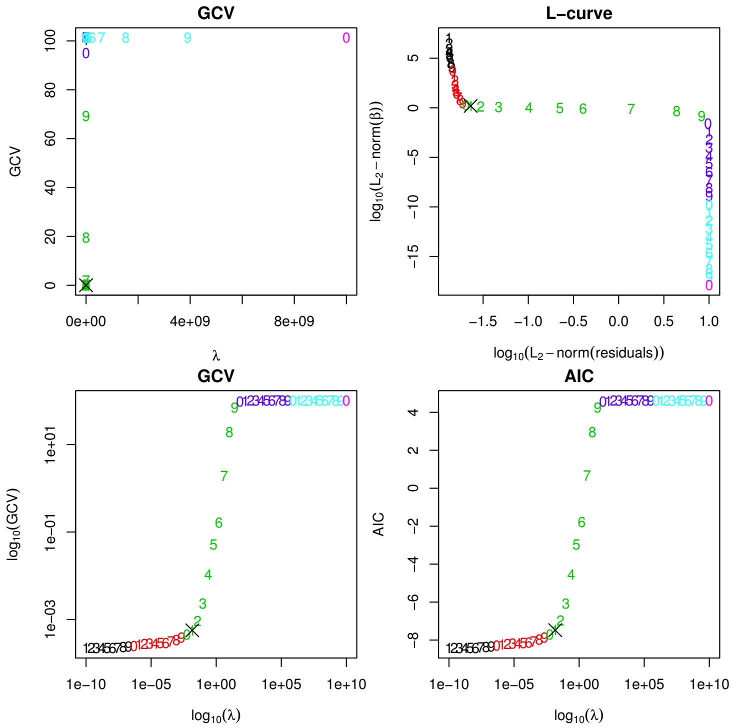

Figure 4.1: Upper-left panel: Generalized cross-validation error as function of λ. Lower-left panel: Generalized cross-validation error on double logarithmic scale. Upper-right and lower-right panel: L-curve and Akaike's information criterion as function of λ, respectively. For λ an equidistant grid of 50 points from 10^{-10} to 10^{+10} was used. A × indicates the optimal penalty parameter automatically selected as described in the text. These plots were generated using the simulated data, described in Section 4.3 "Example 1: Two group", using method B with grid-size 250.

One disadvantage of the linear Tikhonov approach is that it can lead to a solution that contains negative values, which in turn leads to a non-valid density of effect sizes (see left-panels of Figure 4.2, more detail on the data is given in Section 4.3 "Example 1: Two group"). To overcome this discrepancy one could cast the linear problem to a nonlinear one by the transformation $\mathbf{x} = \exp(\mathbf{z})$ (where exp is taken element wise) and solve for $\mathbf{z}$. Solving such a nonlinear discrete inverse problem leads necessarily to an iterative approach [Hanke et al., 2000, Calvetti et al., 2004]. Unfortunately, solving the nonlinear discrete inverse problem as proposed by Hanke et al. [2000], equation (4.5) with $\mathbf{x} = \exp(\mathbf{z})$, has not yet produced satisfying results. For this reason, we will take a different route to obtain a valid estimate of the density of effect sizes.

Constrained solution of the discrete inverse problem

Projection methods like the Conjugate Gradients method of Hestenes and Stiefel [1952] are another popular way of solving discrete inverse problems, especially the Conjugate Gradients Normal equation Residual Method (CGNR) [Golub and van Loan, 1996] is very well suited for our problem. Interestingly, the Conjugate Gradients method or partial least-squares, as it is called in the chemometrics literature, is related to ridge regression [Frank and Friedman, 1993]. For example, Phatak and de Hoog [2002] show that the partial least-squares method has similar shrinkage or regularizing properties as ridge regression.

The CGNR method is a Conjugate Gradients method applied to the normal equations $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$. The k^{th} iterative solution $\mathbf{x}^{[k]}$ of the CGNR method is defined as the solution to the minimization problem

$$\mathbf{x}^{[k]} = \min_{\mathbf{x}} ||\mathbf{A}\mathbf{x} - \mathbf{b}||_2^2 \quad s.t. \quad \mathbf{x} \in \mathbb{K}_k, \tag{4.7}$$

where the Krylov space $\mathbb{K}_k$ is defined as the span of powers of the matrix $\mathbf{A}^T\mathbf{A}$ applied to $\mathbf{A}^T\mathbf{b}$:

$$\mathbb{K}_k \equiv \text{span}\{\mathbf{A}^T\mathbf{b}, (\mathbf{A}^T\mathbf{A})\mathbf{A}^T\mathbf{b}, \ldots, (\mathbf{A}^T\mathbf{A})^{k-1}\mathbf{A}^T\mathbf{b}\}. \tag{4.8}$$

The Conjugate Gradients method is a very efficient algorithm that generates solutions that lie in a Krylov space where the number of iterations plays the role of the regularization parameter [Phatak and de Hoog, 2002] (see Appendix B.5 for a derivation that shows why the solution of a least-squares problem lies in the Krylov space).

Like the linear Tikhonov method, the CGNR method does not constrain the solution to be non-negative. However, using the zero-vector as starting point the first iterations all gave non-negative solutions. This led us to suggest to stop the CGNR method just before the first negative values are introduced, leading to a non-negative solution. The Conjugate Gradient method applied to discrete inverse problems is known to show semiconvergence, the optimal solution is reached before convergence. This further supports our early stopping rule. The performance of our stopping rule was evaluated using simulated data, described in Section 4.3 “Example 1: Two groups”, by calculating the relative error, defined as

$$\frac{\left|\left|\mathbf{x}^{\text{true}} - \mathbf{x}^{[k]}\right|\right|_2}{\left|\left|\mathbf{x}^{\text{true}}\right|\right|_2}, \tag{4.9}$$

where $\mathbf{x}^{[k]}$ represents the regularized solution e.g. at the kth iterate of the CGNR method. The upper panel of Supplementary Figure B.1 shows the relative error for 25 simulated data sets using the CGNR method, the × indicates solution $\mathbf{x}^{[k]}$ with all elements ≥ 0. According to our stopping rule, the majority of the selected solutions have relative error near the lowest point. For comparison, the lower panel shows the relative error for Tikhonov regularization, where the optimal solution is obtained using the L-curve method (some of these solutions do not lead to valid densities).

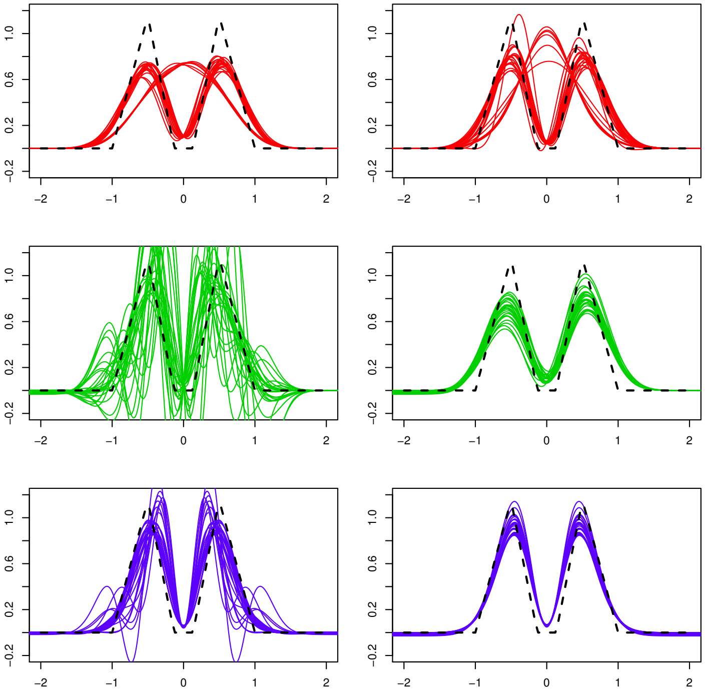

Figure 4.2: Example 1: Estimated density of effect sizes on a grid of 750 points. Left panels using the Tikhonov regularization and right-panels CGNR method. In the upper panels method A, an estimate of the density of test-statistics, is used in the middle panel method B and lower panel method C.

4.3 Examples

Simulated Data

Example 1: Two groups

In Appendix B.3 we described one method, using an estimated density of the test statistics, for estimation of both π_0 and the density of effect sizes. Our approach also allows estimation by using an empirical distribution of the test statistics or transformations thereof, such as an estimated distribution of the p-values. In this example we will show how accurate π_0, the power and the density of effect sizes are estimated using these three approaches which we will refer to as method A: estimated density of the test statistics; method B: estimated cumulative distribution of the test statistics; and method C: estimated cumulative distribution of the p-values.

The simulation study represents a two group microarray experiment with $m = 5000$ features and $N = 10$ samples per group. Let R_{ij} and S_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, N$ be random variables where $S_{ij} \sim N(-{}^{\mu_i}/_2, 1)$ and $R_{ij} \sim N({}^{\mu_i}/_2, 1)$ with the first $\mu_i = 0$ for $i = 1, \ldots, m_0$ and the remaining μ_i for $i = m_0 + 1, \ldots, m$ were drawn from a symmetric bi-triangular distribution as described by Langaas et al. [2005]. A total of 25 data sets were simulated with the proportion of non-differentially expressed features fixed at $\pi_0 = 0.8$. Additional technical noise is added to the model using $X \sim \log(e^R + e^\epsilon)$ and $Y \sim \log(e^S + e^\epsilon)$ with $\epsilon \sim N(0, 0.1^2)$, so that the noise is positive and additive.

A simple t test was applied to each of the 5000 features to obtain a vector of test statistics and corresponding p-values. The FDR was controlled at 0.1 using the adaptive Benjamini-Hochberg method. We also evaluated the effect of the numerical integration error by using different grid sizes: 100, 250, 500 and 750 points.

The observed power, calculated as the mean proportion of correctly declared significant features out of all alternative features (1000) over 25 simulated data sets is slightly above 0.3, indicating that the effect sizes are small.

Using method A the mean power is mostly overestimated while π_0 is estimated more accurately (Figure 4.3). Method B tends to overestimate both power and π_0. Finally, method C yields the most accurate average power. All three methods have comparable estimates of the density of effect sizes (Figure 4.2). Note that there is little gain in using grid size 750 compared to 500 as it does not improve π_0 estimates. Increasing the grid size used for the numerical integration may help by approximating the integral more accurately, but it will also increase the ill-conditionedness of the inverse problem. Other integration methods can then be considered as alternatives, e.g. the trapezoidal or Simpson's method, which can have similar accuracy with a smaller grid size.

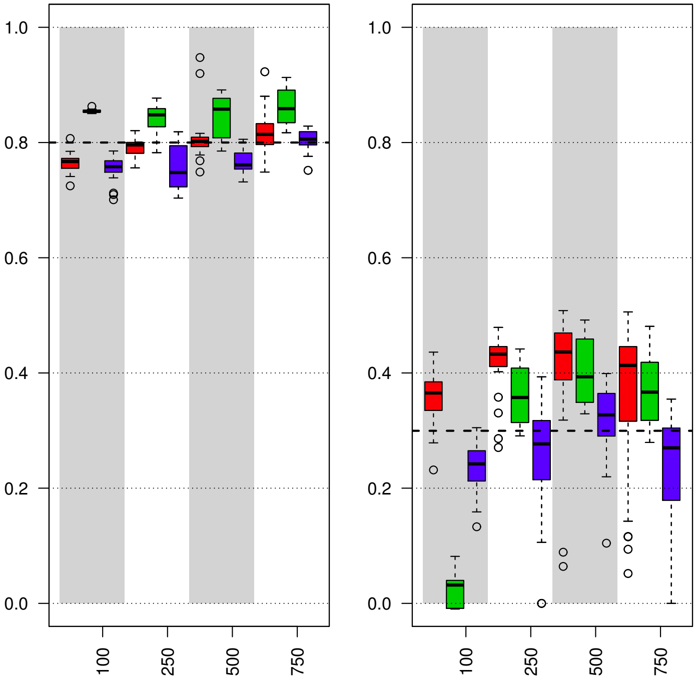

Figure 4.3: Example 1: Estimates of the proportion of non-differentially expressed genes and the average power. Dashed horizontal lines are indicating a) the true π_0, left-panel, and b) estimated expected power, right-panel. The x-axis indicates the different grid sizes used for the numerical integration and the use of different estimation methods for the density of effect sizes, where in each rectangular region the first box-plot indicates the method A, second method B and third C.

Example 2: Power prediction

In this example we examine the quality of power predictions based on pilot-data are using t- and F-test statistics. The same simulation set-up as described in Section 4.3 “Example 1: Two group” was used. We only show results using method A with a grid-size of 500 and the proportion of non-differentially expressed genes of $\pi_0 = 0.7$. We considered three pilot-data sample sizes; 5, 10 and 15 per group and calculated t-test statistics and p-values for each. For the F-test statistics we just square the t-test statistics, so each gene under the null hypothesis follows a central F with $(1, N-2)$ degrees of freedom. The power was predicted for per-group sample sizes of 15, 20, 25 and 30. In order to assess how good the predictions are, the true positive rate of simulated data sets with these larger sample sizes are calculated using the adaptive Benjamini-Hochberg FDR at 0.1. Since, in a simulation study, it is known apriori which genes are differentially expressed the true positive rate can be calculated as the number of declared differentially expressed genes that are truly differentially expressed over the total number of truly differentially expressed genes. The whole procedure was repeated 25 times (see Figure 4.3).

The predicted power increases most for the smaller sample sizes, leveling off for larger sample sizes, as expected. Power predicted using the F-statistics is a little higher than the t-statistics, and is in better agreement with the true positive rate. For the t-statistics the pilot-data of sample size 5 per group has the largest variation, whereas for the F-statistics the pilot-data set with sample size 15 per group. This is probably, because for larger sample sizes the density of effect sizes is better estimated then for smaller sample sizes were the density of effect sizes is estimated to have larger effect sizes. As the power is increasing faster for the smaller effect sizes this could give rise the large variability observed for the larger sample sizes.

Example 3: Comparing estimation of the density of effect sizes

Recently, Long et al. [2012] proposed an improvement of the method by Ruppert et al. [2007] for estimation of the density of effect sizes, in case test statistics were derived from a two-group comparison. They assume the density of effect sizes can be represented by a normal density or convex combination of normal densities which yields a fast estimation method. This assumption seems to favour estimating effect size densities that are unimodal. In order to illustrate this, we run two simulation studies. The first one was based on an unimodal effect size density, as assumed by Long et al. [2012], and showed slightly better results using Long's method compared to ours, both in terms of estimated π_0 and in terms of the Hellinger distance between the estimated and true effect size densities (Supplementary Figures B.2 and B.3).

The second simulation study uses the bitriangular density of effect sizes proposed by Langaas et al. [2005], as in our example 1, and showed better results using our method compared to Long's (Supplementary Figures B.4 and B.5). Thus, when the true effect size density is bimodal, our method yields a better fit to the data compared to Long's method. Our method does not make specific assumptions about the effect size density shape and is, therefore, more flexible.

Our method estimates the density of effect sizes and π_0 simultaneously. π_0 estimates of both simulations are similar to those of the methods proposed by [Storey, 2003, Langaas et al., 2005] (Supplementary Figures B.2 and B.5).

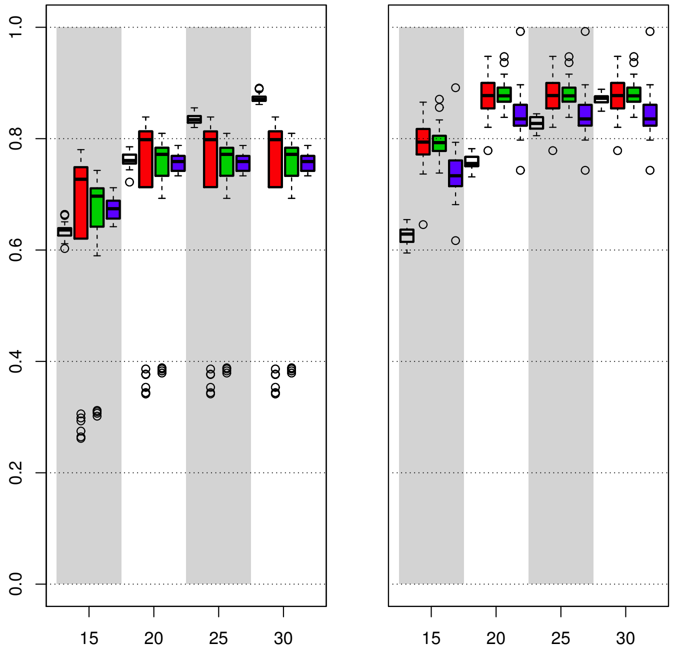

Figure 4.4: Example 2: Left panel predicted power using t-statistics and right panel using the F-statistics for sample sizes ranging from 15-30. The white box indicates the true positive rate over 25 simulated data sets for the sample sizes ranging from 15-30. The red, green and blue boxes are predicted power based on pilot-data of size 5, 10 and 15.

Experimental Data

Example 4: Two group

In this example we carry out power and sample size analysis on experimental microarray data with a two-group design, containing 38 arrays. We take subsets of the data and do power predictions for the full-sized data set. Thus the objective is to compare the predicted power with those for the full-sized data set.

The data was taken from the `multtest` (2.10.0) package. Briefly, the data contained 38 tumor samples divided in two groups; 27 acute lymphoblastic leukemia (ALL) samples and 11 acute myeloid leukemia (AML) samples. Randomly, 25 subsets were constructed containing 15 ALL and 5 AML samples. Again, simple *t* tests were applied to obtain test-statistics and corresponding p-values.

From Table 4.1, it can be seen that the mean predicted power is close to that of the full sized data for all three methods (Supplementary Figure B.6). These results show that the predicted power is within an acceptable range of the observed power of data with same sample size as was used for the prediction of the power.

Table 4.1: Example 4: Estimated average power for the full sized data and subsets of data.

Method	Full data	Subsets
A	0.78	$0.78^1(0.04^2)$
B	0.64	0.73 (0.05)
C	0.79	0.77 (0.06)

[1] mean, [2] standard deviation

Example 5: Factorial design

In this example we will illustrate how a power and sample size analysis can be performed for a factorial design. The data for this example is taken from the `estrogen` (1.8.7) package from BioConductor [Gentleman et al., 2004]. The data gives results from a 2×2 factorial experiment on MCF7 breast cancer cells, involving as factors estrogen (present or absent) and length of exposure (10 or 48 hours). The objective was to identify genes that respond to an estrogen treatment and to classify these genes as early or late responders, which can be done in terms of an F-test statistic. We followed the analysis as presented in the `limma` users guide to obtain the set of F-test statistics and corresponding p-values.

We will estimate the average power again using the three estimation methods, A, B and C, as described earlier. Estimated density of effect sizes do not differ much between the three approaches (upper panel of Figure 4.5). As a result, the predicted power varies between the three approaches only for the smallest sample size (lower panel of Figure 4.5). Indeed, for the observed total sample size 8 the power is estimated to lie in the range $[0.45, 0.65]$. With the sample size of the pilot data only half of the time the differentially expressed genes are detected. Larger sample sizes yield power estimates above 0.8, i.e. now the differrentially expressed genes are detected with 80% probability. From Figure 4.5 we conclude that, to guarantee an average power of 80%, the study should include at least 15 patients per group.

Some authors suggest leaving the region near zero out of the density of effect sizes when computing power predictions. The reason is that small effect are very hard to validate and thus are not typically of interest. In this example, leaving out a region $[0, 0.5]$ gives estimated power curves that are even closer to each other (Supplementary Figure B.7).

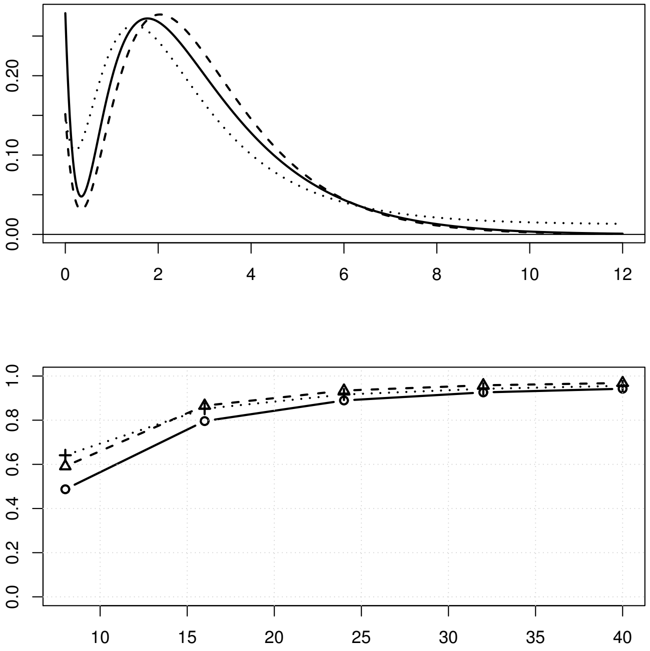

Figure 4.5: Example 5: Upper panel: Estimated density of effect sizes using estimation method A (solid), B (dashed) and C (dotted). Lower panel: estimate power curves for sample sizes up to 5× larger than the pilot data.

Example 5: RNA-seq data

In this example we will show how our method can be applied to count data of an RNA-seq experiment. We will use the data described by 't Hoen et al. [2008] comparing gene expression profiles in the hippocampi of transgenic δC-doublecortin-like kinase mice with that of wild-type mice. Data was downloaded from GEO [Edgar et al., 2002] (GSE10782) and analyzed using the generalized linear model approach implemented in `edgeR` (2.4.3) [McCarthy et al., 2012, Robinson et al., 2010]. A tag-wise dispersion parameter was estimated using the Cox-Reid adjusted profile likelihood approach for generalized linear models as implemented in `edgeR`. Using a treatment contrast, we tested per tag if there was a genotype effect using the likelihood ratio statistic. This test statistic follows asymptotically a χ^2 distribution with one degree of freedom.

The π_0 estimates vary between methods (0.69, 0.69 and 0.73, using respectively methods A, B and C), as do corresponding effect size densities (see upper panel of Figure 4.6). Nevertheless, predicted power curves generated by the three methods are fairly similar (lower panel Figure 4.6), with the small difference observed driven mostly by the mass of effect sizes estimated to be on the tail. Indeed, using Method B the effect size density has the heaviest tail, and the estimated power is correspondingly the largest. On the other hand, Method C yields an effect size density with the lightest tail, leading to the smallest predicted power. In practice, researchers should choose the method that most closely resembles the approach used for data analysis. Thus, if in this case the intention is to analyse the data using edgeR-derived test statistics and their χ^2-square asymptotic densities, then method A should be chosen. However, if on the other hand p-values will be extracted for further comparisons with other methods, then method C might be the most appropriate.

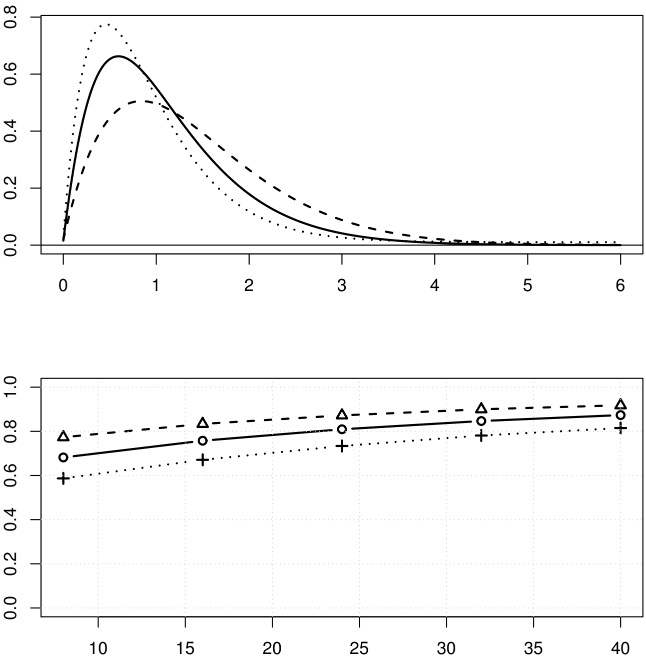

Figure 4.6: Example 5: Upper panels shows the estimated density of effect sizes using the three methods. Lower panel the predicted power curves, from 8, the total sample size of the pilot-data to 40. Method A (solid), B (dashed) and C (dotted).

4.4 Discussion

Most methods so far proposed for pilot data-based power and sample size calculations using high-dimensional data focused on two-group comparisons, and so far have assumed that the test statistics used follow either a normal or a Student-t distribution [Ferreira and Zwinderman, 2006c, Scheinin et al., 2010, Ruppert et al., 2007, Jørstad et al., 2008, Matsui and Noma, 2011]. Our approach is thus much more general, as it can be used for a large class of test statistics including those generated from generalized linear models applied to a wide variety of high-dimensional data. Types of data that can be handled include microarray and next-generating sequencing. Our method is general in the sense that it can handle any kind of high-dimensional data, where parametric forms for the null and alternative distribution of test statistics (or transformations thereof, like p-values) are known, and a set of pilot data is available. In particular, our method can be used for multiple-group comparisons, as well as deviance-based testing used to compare nested models.

The method relies on the adaptive version of the Benjamini-Hochberg FDR method, which does not explicitly take into account correlation among features. Some authors [Tibshirani, 2006, Lin et al., 2010] suggested to use resampling-based FDR in the power and sample size analysis, but neither π_0 nor the density of effect sizes were then estimated. Our approach could be extended to use a resampling-based estimate for the null distribution of test statistics (or p-values), making it nonparametric. In this case, the nonparametric alternative distribution could be based on a shifted version of the nonparametric null distribution.

We have considered two methods to solve the system of equations involved in our approach. One of them, the nonlinear Tikhonov optimization problem, described in section 4.2, did not produced satisfactory results (data not shown), as it led occasionally to estimated densities that are not non-negative everywhere. Another approach would be to follow the suggestion of [Eilers and Marx, 2010] and use a combination of first- and second-order penalties, so that positivity of estimated coefficients can be imposed. This is an interesting direction for future research. Other authors have used constrained estimation methods, requiring quadratic programming or an expectation-maximization algorithm [van de Wiel and Kim, 2007, Ruppert et al., 2007, Jørstad et al., 2008, Matsui and Noma, 2011] and, thus, leading to computationally intensive algorithms.

We should point out that, when there is not enough information about the effect size density in the pilot data, estimates can be unstable. This is especially the case when π_0 is close to one, a large fraction of effect sizes is small, and the pilot data sample size is small (data not shown). This observation was also reported by Long et al. [2012]. Such situations may especially arise when using sequencing data, which is a very sensitive technology that may, as a result, require larger pilot data studies for reliable power calculations.

Our method assumes that test statistics, with known null and alternative distributions, are produced by the data analysis. So any type of test statistic distribution family can be handled. Our package, SSPA (available from http://www.bioconductor.org, see Appendix F for the vignette of the package), can handle any given family of distribution, so long as both null and alternative are known, at least empirically. One natural further generalization would be to attempt to solve the problem in terms of p-values, independently of the test statistics' distribution. This would require an alternative distribution that depends explicitly on effect size and sample size, in such a way that the effect size density is sample-size invariant. A parametric distribution family that could be considered here is the beta, provided one of its parameters can be factorized as a product of sample size and effect size. Non-parametric estimates for the p-values null and alternative distributions could also be considered, via resampling say, provided the same factorization can be obtained. A further alternative is to look for a functional form that can be used approximately, e.g. by a p-value transformation leading to approximately normally distributed random variables, both under the null and alternative hypothesis.

To conclude, our method for power and sample size calculation is flexible, and can be used for power studies in a wide variety of statistical models used for hypothesis testing in high-dimensional data studies.

Acknowledgments

The work was conducted within the Centre for Medical Systems Biology (CMSB), established by the Netherlands Genomics initiative/Netherlands Organisation for Scientific Research (NGI/NWO) and part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). *Conflict of Interest*: None declared.

4.5 Bibliography

M.L.T. Lee and G.A. Whitmore. Power and sample size for DNA microarray studies. *Statistics in Medicine*, 21:3543–3570, 2002.

Y. Pawitan, S. Michiels, S. Koscielny, A. Gusnanto, and A. Ploner. False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, 21(13):3017–3024, 2005.

P. Liu and J.T.G. Hwang. Quick calculation for sample size while controlling false discovery rate with application to microarrays. *Bioinformatics*, 26(6):739–746, 2007.

T. Tong and H. Zhao. Practical guidelines for assessing power and false discovery rate for fixed sample size in microarray experiments. *Statistics in Medicine*, 27:1960–1972, 2008.

J.A. Ferreira and A. Zwinderman. Approximate Sample Size Calculations with Microarray Data: An Illustration. *Statistical Applications in Genetics and Molecular Biology*, 5(1):1, 2006a.

D. Ruppert, D. Nettleton, and J.T.G. Hwang. Exploring the information in p-values for the analysis and planning of multiple-test experiments. *Biometrics*, 63(2):483–495, 2007.

T.S. Jørstad, H. Midelfart, and A.M. Bones. A mixture model approach to sample size estimation in two-sample comparative microarray experiments. *BMC Bioinformatics*, 9 (117):1, 2008.

C. Tsai, S. Wang, D. Chen, and J. Chen. Sample size for gene expression microarray experiments. *Bioinformatics*, 21(8):1502–1508, 2004.

Y. Shao and C. Tseng. Sample size calculation with dependence adjustment for fdr-control in microarray studies. *Statistics in Medicine*, 26:4219–4237, 2007.

S.H. Jung. Sample size for FDR-control in microarray data analysis. *Bioinformatics*, 21(14): 3079–3104, 2005.

J.A. Ferreira and A. Zwinderman. On the Benjamini-Hochberg Method. *Annals of Statistics*, 34(4):1827–1849, 2006b.

B. Efron. Empirical bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association*, 104:1015–1028, 2009.

S. Matsui and H. Noma. Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments. *Biometrics*, 2011.

Q. Long, D. Nettleton, and J.C.M. Dekkers. Improved estimation of the noncentrality parameter distribution from a large number of t-statistics, with applications to false discovery rate estimation in microarray data analysis. *Biometrics*, 68(4):1178–1187, 2012.

Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with indenpendent statistics. *Journal of Educational and Behavioral Statistics*, 25: 60–83, 2000.

J.A. Ferreira and A. Zwinderman. Approximate Power and Sample Size Calculations with the Benjamini-Hochberg Method. *The International Journal of Biostatistics*, 2(1):1, 2006c.

A.N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet mathematics doklady*, 151:501–504, 1963.

F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1 (4):502–518, 1986.

P.C. Hansen. *Discrete Inverse Problems: Insight and Algorithms*. SIAM: Fundamentals of algorithms series, 2010.

A. Delaigle and I. Gijbels. Frequent problems in calculating integrals and optimizing objective functions: A case study in density deconvolution. *Statistical Computations*, 2:7349–355, 2007.

M. van Iterson, P.A. 't Hoen, P. Pedotti, G.J. Hooiveld, J.T. den Dunnen, G.J. van Ommen, J.M. Boer, and R.X. Menezes. Relative power and sample size analysis on gene expression profiling data. *BMC Genomics*, 10:439–449, 2009. doi: 19758461.

A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learing*. New York:Springer-Verlag, 2001.

D Ruppert, M.P. Wand, and R.J. Carroll. *Semiparametric regression.* New York: Cambridge University Press., 2003.

J.M. Varah. Pitfalls in the numerical solution off linear ill-posed problems. *SIAM: Journal on Scientific and Statistical Computing*, 4:164–176, 1983.

P.C. Hansen and D.P. O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM Journal Scientific Computation*, 14:1487–1503, 1993.

M. Hanke, J.G. Nagy, and C. Vogel. Quasi-newton approach to nonnegative image restorations. *Linear Algebra and its Applications*, 316:223–236, 2000.

D. Calvetti, B. Lewis, L. Reichel, and F. Sgallari. Tikhonov regularization with nonnegativity constraint. *Electronic Transactions on Numerical Analysis*, 18:153–173, 2004.

M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49:409–436, 1952.

G.H. Golub and C.F. van Loan. *Matrix computations.* John Hopkins university press, 1996.

I.E. Frank and J.H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

A. Phatak and F. de Hoog. Exploiting the connection between PLS, Lanczos methods and conjugate gradients: alternative proofs of some properties of PLS. *Journal of Chemometrics*, 16:361–367, 2002.

M. Langaas, B.H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society Series B*, 67(4):555–572, 2005.

J. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of Statistics*, 31:2013–2035, 2003.

R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y.H. Yang, and J. Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. URL http://genomebiology.com/2004/5/10/R80.

P.A.C. 't Hoen, Y. Ariyurek, H.H. Thygesen, E. Vreugdenhil, R.H.A.M. Vossen, R.X. de Menezes, J.M. Boer, G.B. van Ommen, and J.T. den Dunnen. Deep Sequencing-based Expression analysis shows Major Advances in Robustness, Resolution and Inter-lab Portability over Five Microarray Platforms. *Nucleic Acids Research*, pages 1–11, 2008.

R. Edgar, M. Domrachev, and A.E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

D.J. McCarthy, Y. Chen, and G.K. Smyth. Differential expression analysis of multifactor rna-seq expriments with respect to biological variation. *Nucleic Acids Research*, 40(10): 4288–4297, 2012.

M.D. Robinson, D.J. McCarthy, and G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.

I. Scheinin, J.A. Ferreira, S. Knuutila, G.A. Meijer, M.A. van de Wiel, and B. Ylstra. Cghpower: exploring sample size calculations for chromosomal copy number experiments. *BMC Bioinformatics*, 11:331–341, 2010.

R. Tibshirani. A simple method for assessing sample size in microarray experiments. *BMC Bionformatics*, 7:106–112, 2006.

W. Lin, H. Hsueh, and J. Chen. Power and sample size estimation in microarray studies. *BMC Bioinformatics*, 11:48–57, 2010.

P.H.C. Eilers and B.D. Marx. Splines, knots and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, n/a, doi:10.1002/wics.125:1, 2010.

M. van de Wiel and K. Kim. Estimating the false discovery rate using nonparametric deconvolution. *Biometrics*, 63(3):806–815, 2007.