29

Sample Size Choice for Microarray Experiments

Peter Müller,
M.D. Anderson Cancer Center

Christian Robert and Judith Rousseau
CREST, Paris

Abstract

We review Bayesian sample size arguments for microarray experiments, focusing on a decision theoretic approach. We start by introducing a choice based on minimizing expected loss as theoretical ideal. Practical limitations of this approach quickly lead us to consider a compromise solution that combines this idealized solution with a sensitivity argument. The finally proposed approach relies on conditional expected loss, conditional on an assumed true level of differential expression to be discovered. The expression for expected loss can be interpreted as a version of power, thus providing for ease of interpretation and communication.

29.1 Introduction

We discuss approaches for a Bayesian sample size argument in microarray experiments. As is the case for most sample size calculations in clinical trials and other biomedical applications the nature of the sample size calculation is to provide the investigator with decision support, and allow an informed sample size choice, rather than providing a black-box method to deliver an optimal sample size.

Several classical approaches for microarray sample size choices have been proposed in the recent literature. Pan et al. (2002) develop a traditional power argument, using a finite mixture of normal sampling model for difference scores in a group comparison microarray experiment. Zien et al. (2002) propose to plot ROC-type curves to show achievable combinations of false-negative and false-positive rates. Mukherjee et al. (2003) use a machine learning perspective. They consider a parametric learning curve for the empirical error rate as a function of the sample size,

and proceed to estimate the unknown parameters in the learning curve. Lee and Whitmore (2002) set up an ANOVA model, and reduce the sample size choice to a traditional power analysis in the ANOVA model. Bickel (2003) proposes an approach based on a formal loss function with terms corresponding to a payoff for correct discoveries and a penalty for false discoveries. The loss function is equivalent to the loss L introduced below.

An interesting sequential approach is developed in Fu et al. (2005). After each microarray, or batch of arrays, they compute the posterior predictive probability of mis-classification for the next sample. Sampling continues until this probability achieves some pre-specified threshold.

In Müller et al. (2004) we develop a Bayesian decision theoretic approach to sample size selection for group comparison microarray experiments. We assume that each array reports expression for n genes. Also, we assume that the sample size choice is about multiple arrays with independent biologic samples recorded on each array (excluding, among others, technical repeats based on the same biologic sample).

Main Features of the Proposed Approach. Before introducing the formal setup and approach we provide a brief summary of the results discussed in more detail later. This will help to motivate and focus the following formal discussion. Let J denote the sample size, i.e., the number of microarrays that we recommend to be carried out. In a decision theoretic approach, we define a criterion for the sample size recommendation by stating how much a specific sample size would be worth for a hypothetical outcome y of the experiment, and an assumed hypothetical truth, i.e., true values of all relevant parameters θ. This function of decision, data and parameters is known as the utility function. Alternatively, flipping signs we get the loss function. Of course, at the time of the sample size selection the future data y is not known, and the parameters θ will never be known. One can argue (DeGroot, 1970; Robert, 2001) that a rational decision maker should then choose a sample size based on expected loss, taking the expectation with respect to the relevant probability distribution on parameters and future data. The relevant distribution is the posterior predictive distribution conditional on any data available at the time of making the decision. In the absence of any data this is the prior predictive distribution. Some complications arise when the nature of the decision is sequential. See below.

Figure 29.1 shows expected loss for a microarray sample size selection. The loss function is $L(J, y, \theta) = \mathrm{FD} + c\mathrm{FN} - k \cdot J$, where FD denotes the

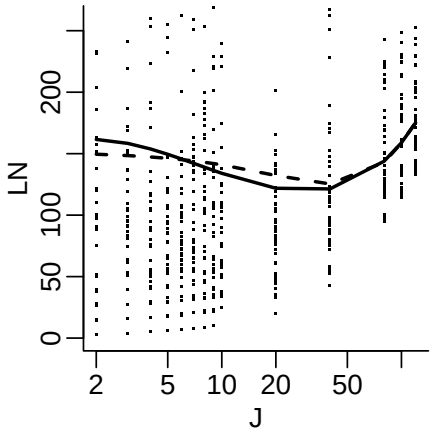


Fig. 29.1. Expected loss as a function of sample size J for a two-group comparison microarray experiment. Evaluating the expectation involves a large scale Monte Carlo simulation. See the text for details. The dots show realized losses for simulated experiments. The solid line plots an estimate of expected loss based on a parametric fit of the dots. The dashed line shows the same using a spline fit. Sample size J is plotted on a logarithmic scale. Note the relatively flat nature of the expected loss, rendering a sample size recommendation difficult.

number of false positives (truly not differentially expressed genes that are flagged), and FN the number of false negatives (truly differentially expressed genes that are not discovered). See Section 29.2.1 for a formal definition of FD and FN. The function includes two trade-off parameters, c and k. See the following sections for more details about the choice of the loss function, the nature of the expectation, including complications that arise from a sequential decision setup, details of the probability model, and the Monte Carlo simulation used to evaluate expected loss. The relatively flat nature of the expected loss hinders a decisive sample size recommendation based on expected loss alone. To be of practical use, the minimum is too sensitive to technical, arbitrary choices of details in the loss function and probability model. We will therefore proceed with a closer look at important features of the expected loss function. In particular, we will consider expected loss conditional on an assumed true level of differential expression for one gene, marginalizing with respect to future data and all other parameters as before. This adds an additional

dimension to the plot in Figure 29.1. Let ρ_i denote the assumed true level of differential expression for gene i. We assume that ρ_i is defined such that $\rho_i = 0$ is interpreted as non-differential expression, and $\rho_i > 0$ as differential expression. We consider expected loss as a function of J and ρ_i. Focusing on only the change in expected utility across J and ρ_i, and dropping the deterministic sampling cost $k \cdot J$, we argue that the plot can be interpreted as a variation of power. Details are discussed in the next section. See Figure 29.2 for an example.

The rest of this chapter is organized as follows. In Section 29.2 we cast sample size choice as a decision problem. In 29.2.1 we argue that sample size choice should be considered as a sequential decision problem. Solving the sequential decision problem we start in Section 29.2.2 with the terminal decision of selecting a list of differentially expressed genes, and proceed in 29.2.3 to address the sample size problem. In Section 29.3 we develop a Monte Carlo scheme to evaluate expected losses. In Section 29.4 we introduce a specific probability model. Section 29.5 discusses the use of pilot data. Finally, section 29.6 demonstrates the proposed approach in an example.

29.2 Optimal Sample Size as a Decision Problem

A decision problem is specified by a set of possible actions $d \in D$; a set of relevant unknown quantities, typically parameters θ and data y; a probability model $p_d(\theta, y)$; and a loss function $L(d, y, \theta)$ that formalizes the relative preferences over decisions for assumed hypothetical values of y and θ. The probability model for data and parameters can depend on the decision d. See, for example, Berger (1993) for a general description. In the application to microarray sample size choice the decision d includes the sample size J, the data y are the gene expressions that will be recorded in the J microarray experiments, and θ typically includes indicators for true differential expression for each of the n genes under the biologic conditions of interest.

The optimal decision is the action d^* that minimizes the loss in expectation, $d^* = \arg\min E\{L(d, \theta, y)\}$. The expectation is with respect to the relevant probability model. In so-called non-sequential problems, the relevant probability model is $p_d(\theta, y)$. In general, the calculation of expected utility might involve more steps. As we will argue, this is the case for the sample size problem.

29.2.1 The Decision Problem

Approaching sample size choice as a decision problem it is important to recognize the sequential nature of the decision. In words, optimal sample size is always defined in the context of the intended inference or decision that will be carried out eventually, once all data is collected (terminal decision). Different inference goals might lead to different sample size recommendations. We therefore need to consider the entire sequence of (i) the sample size decision, (ii) the observation of gene expressions for the chosen number of arrays, and (iii) the terminal decision about differentially expressed genes. Let J denote the sample size choice, let n denote the number of genes that are recorded on each of the J arrays, let $y^J = (y_1, \ldots, y_J)$ denote the data for J arrays, and let $\delta = (\delta_1, \ldots, \delta_n)$ denote the terminal decision, with $\delta_i = 1$ if gene i is reported as differentially expressed, and $\delta_i = 0$ otherwise. The problem involves two decisions, $d = (J, \delta)$. The terminal decision δ is made *after* observing the data. We thus condition on y^J, and the expected loss integral is only with respect to the unknown parameters θ. In contrast, the sample size is chosen *before* observing the data. We thus marginalize with respect to both, data y and parameters θ, substituting the optimal terminal decision δ^*. Decision problems with such multi-step structure are known as sequential decision problems. The optimal decision δ^* is defined as before, with the expected loss taken w.r.t. the posterior distribution, $\delta^*(y^J) = \arg\min \int L(d, \theta, y) \mathrm{d}p(\theta \mid y^J)$. We include an argument y^J in the notation for δ^* to highlight the dependence on the observed data. The optimal sample size choice is defined by

$$J^* = \arg\min \int L(\delta^*(y^J), \theta) \, \mathrm{d}p_d(\theta, y^J). \tag{29.1}$$

The conditions for J^* and $\delta^*(y^J)$ define an ideal solution, following from first principles about rational decision making (Robert, 2001). In practice, several compromises are made when implementing Bayesian optimal design.

An attractive feature of the proposed approach is that the nature of the optimal decision does not depend on details of the probability model. The only required assumption is that the probability model include indicators $r_i \in \{0, 1\}$ for true differential expression of gene i. Except for this minimal assumption, we can discuss the optimal decision before defining a specific probability model, requiring only a loss function to complete the formal description of the decision problem.

We define a loss function that defines a tradeoff of false negative and

false rejection counts. Let $\text{FN} = \sum_i (1 - \delta_i)\ r_i$ denote the number of false negatives, and let $\text{FD} = \sum_i \delta_i (1 - r_i)$ denote the false rejections (discoveries). The counts FN and FD are functions of the parameters r_i and the data y^J, implicitly through $\delta_i(y^J)$. We use the loss function

$$L(J, \delta, \theta, y^J) = \text{FD} + c\,\text{FN}.$$

The loss function does not include a term representing sampling cost. See the discussion below, when we consider the optimal decision sample size choice.

29.2.2 The Terminal Decision δ^*

The decision about the optimal sample size in any experiment is always relative to the intended data analysis after carrying out the experiment. This is formalized in the definition (29.1) by requiring to plug in the optimal rule δ^* about reporting genes as differentially expressed. It is therefore natural to first discuss δ^* before we consider the original sample size question.

Let $\overline{r}_i = Pr(r_i = 1 \mid y^J)$ denote the marginal posterior probability of gene being differentially expressed. It can be easily shown (Müller et al., 2004) that under L the optimal decision δ_i^* is of the form

$$\delta_i^*(y^J) = I(\overline{r}_i > t),$$

i.e., flag all genes with marginal posterior probability of differential expression beyond a certain threshold. The threshold is $t = c/(c + 1)$. The optimal rule is very intuitive and similar to a popularly used methods to control (frequentist) expected false discovery rate (Benjamini and Hochberg, 1995; Storey, 2003), with the critical difference that the rule is defined as a cutoff for marginal probabilities instead of nominal p-values. See also Genovese and Wasserman (2003) for more discussion of Bayesian variations of the Benjamini and Hochberg rule.

29.2.3 Sample Size Choice

We now use the optimal decision δ^* to substitute in definition (29.1). First we note that $L(J, \delta, \theta, y^J)$ does not include any sampling cost. To define an optimal sample size J^* we could add a deterministic sampling cost, say kJ. However, the choice of the tradeoff k is problematic. We therefore prefer to use a goal programming approach, plotting expected loss as a function of J, and allowing the investigator to make an informed

choice by, for example, selecting the minimum sample size to achieve expected loss below a certain target.

Doing so we run into an additional complication. Let $\overline{L}(J)$ denote the expected loss

$$\overline{L}(J) = \int L(\delta^*(y^J), \theta)\, \mathrm{d}p_d(\theta, y^J). \tag{29.2}$$

For relevant sample sizes the expected loss $\overline{L}(J)$ is far too flat to allow a conclusive sample size choice. In fact, in Müller et al. (2004) we show that the prior expectation of FN, plugging in the optimal rule δ^*, decreases asymptotically as $O_P(\sqrt{\log J/J})$.

The flat nature of the expected loss surface is a typical feature for decision problems in many applications. A common solution to address this problem is to consider sensitivity analysis of the expected loss with respect to some relevant features of the probability model. In particular, we assume that for each gene the probability model includes a parameter $\rho_i \geq 0$ that can be interpreted as level of differential expression, with $\rho_i = 0$ for non-differentially expressed genes. Assuming a gene with true $\rho_i > 0$, we explore the change in expected loss as a function of ρ_i and J. In other words, we consider the integral (29.2), but conditioning on an assumed true value for ρ_i, instead of including it in the integration. Assuming a large number of genes, fixing one ρ_i leaves the inference for all other genes approximately unchanged, impacting the loss function only when the i-th gene is (wrongly) not flagged as differentially expressed and adds to FN. Thus, for $\rho_i > 0$, the only effected term in the loss function is the i-th term in the definition of FN, i.e., $(1 - \delta_i)\, r_i$. We are lead to consider

$$\beta_i(J, \rho_i) \equiv Pr(\delta_i = 0 \mid y^J, \rho_i) = Pr(\overline{r}_i > t \mid y^J, \rho_i) \tag{29.3}$$

The probability includes the marginalizations over all other genes, and the application of the optimal terminal rule δ^*. Assuming that the probability model is exchangeable over genes, we can drop the index from β_i. The expression $1 - \beta(J, \rho)$ has a convenient interpretation as power, albeit marginalizing over all unknowns except for ρ_i. We refer to $\beta(J, \rho)$ as predictive power.

29.3 Monte Carlo Evaluation of Predictive Power

Evaluation of $\beta(J, \rho)$ is most conveniently carried out by Monte Carlo simulation. Let J_0 and J_1 denote minimum and maximum sample sizes

under consideration. We first describe the algorithm in words. Simulate many, say M, possible experiments $(\theta^m, y^m_{J_1})$, $m = 1, \ldots, M$, simulating responses for a maximum number J_1 of arrays. For a grid of sample sizes, from J_1 down to J_0, compute $\overline{r}_i$ for each gene i, each simulation m, and each sample size J on the grid. Record the triples $(J, \rho_i, \overline{r}_i)$ across m, i and J. Plot $\delta_i = I(\overline{r}_i > t)$ against J and ρ_i. Finally, fitting a smooth surface through δ_i as a function of (J, ρ_i) we estimate $\beta(J, \rho)$. The algorithm is summarized by the following steps. To simplify notation we drop the i index from $\overline{r}_i$, ρ_i and δ_i.

(i) Simulate experiments $(\theta^m, y^m_{J_1}) \sim p(\theta)\, p(y_{J_1} \mid \theta)$, $m = 1, \ldots, M$.
(ii) Compute $\overline{r}$ across all genes $i = 1, \ldots, n$, simulations $m = 1, \ldots, M$, and for all samples sizes J on a given grid. Record all triples $(J, \rho, \overline{r})$.
(iii) Let $\delta = I(\overline{r} > t)$ and Fit a smooth surface $\hat{\beta}(J, \rho)$ through δ as a function of (J, ρ).

Note 1: Most probability models for microarray data assume that y_j are independent given the parameters θ. This allows easy simulation from the joint probability model.

Note 2: Evaluating posterior probabilities $\overline{r}_i$ usually involves posterior MCMC. However, the MCMC requires no burn-in since $p(\theta)p(y^J \mid \theta) = p(y^J)\, p(\theta \mid y^J)$. In words, the prior draw θ generated in step 1 is a draw from the posterior distribution given y^J. It can be used to initialize the posterior MCMC.

The plot of $\hat{\beta}(J, \rho)$ is used for an informed sample choice, in the same way as power curves are used in sample size arguments under a frequentist paradigm.

29.4 The Probability Model

29.4.1 A Hierarchical Mixture of Gamma/Gamma Model

The proposed approach builds on the model introduced in (Newton et al., 2001; Newton and Kendziorski, 2003). Let X_{ij} and Y_{ij} denote appropriately normalized intensity measurements for gene i on slide j under the two biologic conditions of interest, i.e., $y^J = (X_{ij}, Y_{ij},\ i = 1, \ldots, n \text{ and } j = 1, \ldots, J)$ We assume conditionally independent measurements given gene specific scale parameters $(\theta_{0i}, \theta_{1i})$:

$$X_{ij} \sim \text{G}amma(a, \theta_{0i}) \text{ and } Y_{ij} \sim \text{G}amma(a, \theta_{1i}).$$

We define a hierarchical prior probability model, including a positive prior probability for a tie between θ_{i0} and θ_{i1}, corresponding to non-differential expression across the two conditions. We introduce a parameter $r_i \in \{0, 1\}$ as latent indicator for $\theta_{0i} = \theta_{1i}$, and assume

$$\theta_{0i} \sim \text{G}amma(a_0, \nu)$$

and

$$p(\theta_{1i} \mid r_i, \theta_{0i}) = \begin{cases} I(\theta_{1i} = \theta_{0i}) & \text{if } r_i = 0 \\ \text{G}amma(a_0, \nu) & \text{if } r_i = 1 \end{cases}$$

with $Pr(r_i = 0) = p_0$. The model is completed with a prior for the parameters $(a, a_0, p) \sim \pi(a, a_0, p)$, and fixed ν. We assume *a priori* independence and use marginal gamma priors for a_0 and a, and a conjugate beta prior for p. As in Newton et al. (2001) , the above model leads to a closed form marginal likelihood after integrating out θ_{1i}, θ_{0i}, but still conditional on $\eta = (p, a, a_0)$. Let $X_i = (X_{ij},\ j = 1, \ldots, J)$ and $Y_i = (Y_{ij},\ j = 1, \ldots, J)$. We find

$$p(X_i, Y_i | r_i = 0, \eta) = \left\{ \frac{\Gamma(2Ja + a_0)}{\Gamma(a)^{2J}\Gamma(a_0)} \right\} \frac{(\nu)^{a_0} \left(\prod_j X_{ij} \prod_j Y_{ij}\right)^{a-1}}{[(\sum_j X_i + \sum_j Y_i + \nu)]^{2a+a_0}}$$

and

$$p(X_i, Y_i | r_i = 1, \eta) = \left\{ \frac{\Gamma(aJ + a_0)}{\Gamma(a)^J \Gamma(a_0)} \right\}^2 \frac{(\nu\nu)^{a_0} \left(\prod_j X_{ij}\ \prod_j Y_{ij}\right)^{a-1}}{[(\sum_j X_{ij} + \nu)(\sum_j Y_{ij} + \nu)]^{a+a_0}},$$

and thus the marginal distribution is

$$p(X_i, Y_i | \eta) = p_0\, p(X_i, Y_i \mid r_i = 0, \eta) + (1 - p_0)\, p(X_i, Y_i \mid r_i = 1, \eta) \quad (29.4)$$

Availability of the closed form expression for the marginal likelihood greatly simplifies posterior simulation. Marginalizing with respect to the random effects reduces the model to the 3-dimensional marginal posterior $p(\eta \mid y) \propto p(\eta) \prod_i p(X_i, Y_i | \eta)$. Conditional on currently imputed values for η we can at any time augment the parameter vector by generating $r_i \sim p(r_i \mid \eta, X_i, Y_i)$ as simple independent Bernoulli draws, if desired.

29.4.2 A Mixture of Gamma/Gamma Model

One limitation of a parametric model like this hierarchical Gamma/Gamma model is the need to fix specific model assumptions. The investigator has to select hyper-parameters that reflect the relevant experimental

conditions. Also, the investigator has to assume that the sampling distribution for observed gene expressions can adequately be approximated by the assumed model. To mitigate problems related with these requirements we consider a model extension that still maintains the computational simplicity of the basic model, but allows for additional flexibility.

A computationally convenient implementation is a mixture extension of the basic model. In particular, we replace the Gamma distributions for $p(X_{ij}|\theta_{0i})$ and $p(Y_{ij}|\theta_{1i})$ by scale mixtures of Gamma distributions

$$X_{ij} \sim \int Ga(a, \theta_{0i}\, q_{ij})\ \mathrm{d}p(q_{ij}|w,m) \text{ and}$$
$$Y_{ij} \sim \int Ga(a, \theta_{1i}\, s_{ij})\ \mathrm{d}p(s_{ij}|w,m) \tag{29.5}$$

where $p(q \mid w,m)$ is a discrete mixing measure with $P(q = m_k) = w_k$ $(k = 1,\ldots,K)$. Locations $m = (m_1,\ldots,m_K)$ and weights $w = (w_1,\ldots,w_K)$ parameterize the mixture. To center the mixture model at the basic model, we fix $m_1 = 1.0$ and assume high prior probability for large weight w_1. We use the same mixture for s_{jk}, $P(s_{jk} = m_h) = w_h$. The model is completed with $m_k \sim Ga(b,b)$, $k > 1$ and a Dirichlet prior $w \sim Dir_K(M \cdot W, W, \ldots, W)$. Selecting a large factor M in the Dirichlet prior assigns high prior probability for large w_1, as desired. By assuming a dominating term with $m_1 = 1.0$ and $E(m_k) = 1$, $k > 1$, we allocate large prior probability for the basic model and maintain the interpretation of θ_{0i}/θ_{1i} as level of differential expression.

Model (29.5) replaces the Gamma sampling distribution with a scale mixture of Gamma distributions. This is important in the context of microarray data experiments, where technical details in the data collection process typically introduce noise beyond simple sampling variability due to the biological process. A concern related to microarray data experiments prompts us to introduce a further generalization to allow for occasional slides that are outliers compared to the other arrays in the experiment. This happens for reasons unrelated to the biologic effect of interest but needs to be accounted for in the modeling. We achieve this by adding a second mixture to (29.5)

$$(X_{ij}|q_{ij}, g_j) \sim Ga(a, \theta_{0i}\, g_j\, q_{ij}) \text{ and } (Y_{ij}|s_{ij}, g_j) \sim Ga(a, \theta_{1i}\, g_j\, s_{ij}), \tag{29.6}$$

with an additional slide specific scale factor g_j. Paralleling the definition of $p(q_{ij}|w,m)$ we use a finite discrete mixture $P(g_j = m_{gk}) = w_{gk}$, $k = 1,\ldots,L$ with a Dirichlet prior $(w_{g1},\ldots m_{gL}) \sim Dir_L(M_g \cdot W_g, W_g, \ldots, W_g)$, $m_{gk} \sim Ga(b_g, b_g)$ for $k > 1$ and $m_{g1} \equiv 1$.

29.4.3 Posterior MCMC

Posterior inference is implemented by Markov chain Monte Carlo (MCMC) posterior simulation. See, for example, Tierney (1994), for a review of MCMC methods. MCMC simulation proceeds by iterating over the following transition probabilities. We use notation like $[x \mid y, z]$ to indicate that x is being updated, conditional on the known or currently imputed values of y and z. We generically use θ^- to indicate all parameters, except the parameter on the left side of the conditioning bar.

(i) $[q_{ij} \mid \theta^-, X_i]$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$.
(ii) $[s_{ij} \mid \theta^-, Y_i]$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$.
(iii) $[g_j \mid \theta^-, X, Y]$, $j = 1, \ldots, J$.
(iv) $[a \mid \theta^-, X, Y]$
(v) $[a_0 \mid \theta^-, X, Y]$
(vi) $[m_h \mid \theta^-, X, Y]$, $h = 1, \ldots, K$
(vii) $[w \mid \theta^-, X, Y]$, $w = (w_1, \ldots, w_K)$
(viii) $[m_g \mid \theta^-, X, Y]$, $g = 1, \ldots, L$
(ix) $[w_g \mid \theta^-, X, Y]$, $w_g = (w_{g1}, \ldots, w_{gL})$.
(x) $[K \mid \theta^-, X, Y]$
(xi) $[L \mid \theta^-, X, Y]$

All but steps (x) and (xi) are standard MCMC transition probabilities. Changing K and L we use reversible jump MCMC (Green, 1995). See Richardson and Green (1997) for a description of RJMCMC specifically for mixture models. Our reversible jump implementation includes a merge move to combine two terms in the current mixture, a matching split move, a birth move and a matching death move. Details are similar to Richardson and Green (1997), with the mixture of gammas replacing the mixture of normals. Inference is based on a geometric prior on the number of terms K and L in both mixtures.

29.5 Pilot Data

The flexible mixture model allows to use pilot data to learn about details of the sampling distribution. We envision a process where the investigator either uses available data from similar previous experiments, or collects preliminary data to allow estimation of the mixture model parameters before proceeding with the sample size argument. The pilot data might not include samples under both biologic conditions. Pilot data is often available only for control tissue. For such data a reduced

version of the model, using only the parts of the model relevant for X_{ij} is used. This is sufficient to estimate the mixture model parameters.

In summary, we proceed in two stages. In a first stage the pilot data is used to fit the mixture model. Let X^o_{ij}, $j = 1, \ldots, J^o$, denote the pilot data. We will use posterior MCMC simulation to estimate the posterior mean model. This is done once, before starting the optimal design. We then fix the mixture model at the posterior modes $\widehat{K}$ and $\widehat{L}$, and the posterior means $(\bar{w}, \bar{m}, \bar{w}_g, \bar{m}_g) = E(w, m, w_g, m_g \mid X^o, \widehat{K}, \widehat{L})$. We proceed with the optimal sample size approach, using model (29.5) with the fixed mixtures. The

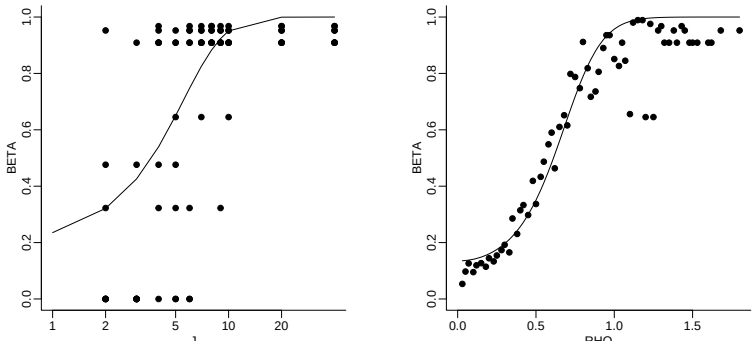
29.6 Example

For illustration we consider the data reported in Richmond et al. (1999), and used in Newton et al. (2001). We use the control data as pilot data to plan the sample size for a hypothetical future experiment. Estimating the mixture model we find a posterior mode $\hat{K} = 3$ and $\hat{L} = 2$.

We now fix K and L at the posterior mode, and the remaining mixture parameters (m, w, m_g, w_g) at their conditional posterior means, conditional on $K - 3$ and $L = 2$. We then use the mixture Gamma/Gamma model with fixed mixture parameters to proceed with the Monte Carlo simulation to compute $\beta(J, \rho)$. In the context of the mixture of Gamma/Gamma model we define $\rho_i = \log(\theta_{0i}/\theta_{1i})$, the log ratio of scale parameters for gene i. Figure 29.2 shows the estimated predictive power curves $\beta(J, \rho)$. The left panel shows $\beta(J, \rho)$ for fixed ρ. Aiming for four-fold differential expression, the plot shows the predictive power that can be achieved with increasing sample size. The left panel of Figure 29.2 summarizes the surface $\beta(J, \rho)$ by fixing J at $J = 15$, and plotting predictive power against assumed level of differential expression ρ. For increasing level of ρ the figure shows the predictive power that can be achieved with $J = 15$ arrays. The points show the simulated true fraction of rejections for J and ρ on a grid. The estimated surface $\beta(J, \rho)$ is based on *all* simulations, across all ρ and J. But the plot only shows the simulations corresponding to the shown slice of the surface.

29.7 Conclusion

We have discussed ideas for a Bayesian decision theoretic sample size argument for microarray experiments. The strength of the approach is



(a) $\beta(J, \rho = \log(4))$ (b) $\beta(J = 15, \rho)$

Fig. 29.2. Power β (labeled BETA in the plot) against sample size for assumed four-fold over-expression, $\rho = \log(4)$ (left), and against ρ for sample size $J = 1$ (right). Power $\beta(J, \rho)$ is defined in (29.3) as the average posterior probability of discovery, conditional on the true level of differential expression $\rho_i = \log(\theta_{0i}/\theta_{1i})$.

the opportunity to use essentially arbitrarily complex probability models. The proposed mixture Gamma/Gamma model is an example. But the argument is valid for any probability model, as long as the model includes latent variables r_i that can be interpreted as indicators for a true effect for gene i, and parameters ρ_i that can be interpreted as strength of the effect. In particular, the probability model could include more complicated designs than two-sample experiments.

Limitations of the proposed approach are the assumed independence across genes, and the implicit 0-1 loss function. More general loss functions could, for example, include a weight proportional to the true ρ_i in the penalty for false negatives. More general models could explicitly allow for networks and dependence.

Bibliography

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Berger, J. O. (1993), *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag Inc.

Bickel, D. R. (2003), "Selecting an optimal rejection region for multiple testing: A decision-theoretic alternative to FDR control, with an application to microarrays," Tech. rep., Medical College of Georgia.

DeGroot, M. (1970), *Optimal Statistical Decisions*, New York: McGraw Hill.

Fu, W., Dougherty, E., Mallick, B., and Carroll, R. (2005), "How many samples are needed to build a classifier: a general sequential approach," *Bioinformatics*, 21, 63–70.

Genovese, C. and Wasserman, L. (2003), *Bayesian Statistics 7*, Oxford: Oxford University Press, chap. Bayesian and Frequentist Multiple Testing, p. to appear.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Lee, M.-L. and Whitmore, G. (2002), "Power and sample size for microarray studies," *Statistics in Medicine*, 11, 3543–3570.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T., and Mesirov, J. (2003), "Estimating Dataset Size Requirements for Classifying DNA Microarray Data," *Journal of Computational Biology*, 10, 119–142.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004), "Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays," *Journal of the American Statistical Association*, 99.

Newton, M. A. and Kendziorski, C. M. (2003), "Parametric Empirical Bayes Methods for Micorarrays," in *The analysis of gene expression data: methods and software*, New York: Springer.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001), "On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data," *Journal of Computational Biology*, 8, 37–52.

Pan, W., Lin, J., and Le, C. T. (2002), "How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach," *Genome Biology*, 3(5), research0022.1–0022.10.

Richardson, S. and Green, P. (1997), "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society, series B*, 59, 731–792.

Richmond, C. S., Glasner, J. D., Mau R., Jin, H., and Blattner, F. (1999), "Genome-wide expression profiling in Escherichia coli K-12," *Nucleic Acid Research*, 27, 3821–3835.

Robert, C. P. (2001), *The Bayesian Choice: from Decision-theoretic Foundations to Computational Implementation*, Springer-Verlag Inc.

Storey, J. D. (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the q-value," *The Annals of Statistics*, 31, 2013–2035.

Tierney, L. (1994), "Markov chains for exploring posterior distributions (with Discussion)," *Annals of Statistics*, 22, 1701–1762.

Zien, A., Fluck, J., Zimmer, R., and Lengauer, T. (2002), "Microarrays: How Many Do You Need?" Tech. rep., Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany.