Empirical Bayes estimation of a sparse vector of gene expression changes

Stephen Erickson* and Chiara Sabatti*†

Departments of *Statistics and †Human Genetics,

UCLA, Los Angeles, California 90095

UCLA Statistic Department Preprint # 413

February 2005

Running head Empirical Bayes and sparsity in microarrays

Keywords Microarrays, MCMC, FDR, thresholding, ℓ_1 norm.

Corresponding author Chiara Sabatti

Department of Human Genetics

UCLA School of Medicine

695 Charles E. Young Drive South

Los Angeles, California 90095-7088 (USA)

FAX: (310) 794-5446

Phone: (310) 794-9567

e-mail: csabatti@mednet.ucla.edu

Abstract

Gene microarray technology is often used to compare the expression of thousand of genes in two different cell lines. Typically, one does not expect measurable changes in transcription amounts for a large number of genes; furthermore, the noise level of array experiments is rather high in relation to the available number of replicates. For the purpose of statistical analysis, inference on the "population" difference in expression for genes across the two cell lines is often cast in the framework of hypothesis testing, with the null hypothesis being no change in expression. Given that thousands of genes are investigated at the same time, this requires some multiple comparison correction procedure to be in place. We argue that hypothesis testing, with its emphasis on type I error and family analogues, may not address the exploratory nature of most microarray experiments. We instead propose viewing the problem as one of estimation of a vector known to have a large number of zero components. In a Bayesian framework, we describe the prior knowledge on expression changes using mixture priors that incorporate a mass at zero and we choose a loss function that favors the selection of sparse solutions. We consider two different models applicable to the microarray problem, depending on the nature of replicates available, and show how to explore the posterior distributions of the parameters using MCMC. Simulations show an interesting connection between this Bayesian estimation framework and both false discovery rate (FDR) control, and misclassification minimizing procedures. Finally, two empirical examples illustrate the practical advantages of this Bayesian estimation paradigm.

1 Statistical analysis of microarray experiments

The development of gene expression array (microarray) technology opened the possibility to study, on a genome scale and in systematic fashion, how the information statically coded in DNA translates in the dynamic of cell life. The scientific community's sizeable interest in the opportunity opened by this line of research spurred the development of statistical methodology to analyze the

measured data appropriately. Starting from the initial contributions appearing in 2000 to date (end of 2004), there has been a growing body of statistical literature devoted to the analysis of microarray data. There are a variety of questions in this area that statistics can help to address; in the present work we limit our consideration to the analysis of datasets obtained comparing transcription levels of genes in cells grown in baseline conditions with those in cells that represent one experimental condition of interest (for example, a specific growth media, a cell line where a gene has been knocked out, or a cancerous cell line). In such experiments—which can be conducted both with one channel as well as two channel arrays—the object of inference are the changes in expression of the assayed genes.

Without any presumption of accounting for all the contributions, it appears to us that two areas of statistical methodology have emerged as particularly relevant in this context: empirical Bayes procedures and corrections for multiple comparisons. To our knowledge, the first work using an empirical Bayes procedure to analyze gene expression array data is due to Newton *et al.* [23] in 2001. Since then a number of authors have used empirical Bayes procedures for microarray data, for example Baldi and Long [3], Ibrahim *et al.* [15], Tseng *et al.* [27], and Ishwaran and Rao [16]. Indeed, in a context where the number of observations per genes is typically very limited and the number of genes very large, empirical Bayes procedures that allow "borrowing" strength across genes appear to be a very effective tool to channel the little information available on a large number of genes.

While the cited papers propose an analysis of microarray within the estimation framework, other groups adopt a test-of-hypothesis approach: the question of relevance appeared to be which genes actually undergo a change in expression and which do not, and the framework of hypothesis testing was selected. It is immediately evident that if one intends to conduct separate tests on thousand of genes, it is necessary to correct for the large number of tests. The analysis of microarray experiments offered fertile ground for the analysis of novel approaches to this problem: the usefulness of measuring global error in terms of false discovery rate (FDR) has been argued, and

multiple methods to estimate FDR, and local FDR, under a series of hypothesis on the test statistics have been proposed, for example Tusher *et al.* [28], Efron *et al.* [10], Storey and Tisbshirani [26], Sabatti *et al.* [25], and Reiner *et al*. [24]

These two trusts of the statistical analysis of microarray experiments have also been combined: the work by Efron *et al.* [10], for example, provides an interpretation of the FDR criteria in terms of nonparametric empirical Bayes procedure; moreover, a sizeable portion of the empirical Bayesian approaches focus on the problem of testing the hypothesis that each gene may not experience a change in expression in the experiment under study, while they often differ in how to deal with the problem of multiple comparisons. As an example, Newton *et al.* [23] define a Bayesian version of FDR.

In this manuscript, we suggest adopting an empirical Bayes framework for the analysis of microarray experiments, effectively abandoning the hypothesis testing framework. We do so on the grounds that type I error may not represent the best description of the loss for scientists involved in gene expression studies with microarray screens. Nevertheless, the procedure we propose, based on the estimation of a sparse vector, bears substantial resemblance to the results of test of hypothesis controlling FDR at some levels. This allows us to underline yet another connection between empirical Bayes procedures and FDR, and also to make some suggestions on the level at which it may be reasonable to control FDR in microarray experiments.

2 To test or not to test?

We are interested in the comparison of expression values for N genes across two experimental conditions. In the present paper, we assume that the measurements come from cDNA-spotted microarrays (also variously referred to as *slides* or *chips*) , but many of the methodological aspects translate directly to other types of experimental devices. For convenience, we introduce the vector parameter $\theta = (\theta_1, \ldots, \theta_N)$, representing the true population change in expression of genes $i =$

$1, \ldots, N$ between two compared conditions. Associating to each gene a single θ_i is rather arbitrary as, even if we could measure it with no error, the expression change of a given gene will be different in each cell. With θ_i, then, we refer to an abstract "population" value that is differently realized at each cell. The nature of gene expression array experiments is such that, for any pair of cell conditions that are compared, one expects only a certain fraction of the genes on the screen to change expression. This is trivially true for the large number of genes that are not expressed in neither of the two conditions. Moreover, the changes in experimental conditions experienced by the cells are such that they are expected to generate a relatively small number of changes; it would be of little scientific worth to compare radically dissimilar cellular conditions. In terms of the introduced parameter θ, this amounts to the knowledge that a large portion of the θ_i are effectively zero.

The data available to make inferences on the θ_i are measurements of fluorescent intensity at discrete locations on a microarray, each location being associated with a specific gene i. The exact functional relationship between increases in expression and increases in intensity levels is not known, and the number of replicates for a given experiment is typically rather small. Given the outlined characteristics of the parameter of interest and the data available for inferential purposes, both the scientists generating the data and the statisticians analyzing it have found that the most basic use of array data is simply addressing the question of which genes experience a change in expression; that is, for which genes i does $\theta_i \neq 0$. Most commonly, this question has been addressed casting the problem in the test of hypothesis framework: the null hypotheses are taken to be $H_i : \theta_i = 0$ (reflecting the knowledge that zero is the most common value for the parameter), a test is performed for each hypothesis, and a procedure for correcting for multiple comparisons is implemented.

Certainly, a test of hypothesis framework allows one to consider $\theta_i = 0$ as a special and common value and deflects emphasis from the estimation of a precise value for θ, which may be rather difficult in this case. By adopting this inferential framework, however, one buys into

much more: the type I error (and its family analogues) becomes the primary element of interest in the sense that a false rejection of the null hypothesis $H_i : \theta_i = 0$ is considered more serious than a false non-rejection. It is not clear to us that this is really an appropriate choice in the context of microarray. On the one hand, microarray experiments are not considered by the scientific community as means to test, but as instruments to generate, hypotheses: that a gene appears up-regulated in a condition on the basis of microarray experiments is not taken as solid evidence for this to be the case, no matter how small the p-value or strong the estimated effect. Only when other experimental techniques (such as *in situ* hybridization, etc.) are able to verify the change in expression is the hypothesis considered corroborated. And while it is clear that one does not want to try to replicate the change in expression of thousands of non-changing genes, one has to consider that the cost of investigating suspicious genes with other experimental techniques is not extremely high. On the other hand, the failure to identify from the thousand of genes in the screen the ones that change expression represents a very serious loss for the researcher. Based on this consideration, we think that other inferential frameworks may be more appropriate. In particular, we would like to point the reader's attention to three approaches that are available to the statistician and that, not surprisingly, are all quite related to testing and correcting for multiple comparisons.

1. One can describe loss as the number of misclassifications, that is, the total number of false positives and false negatives, and choose a procedure aimed at minimizing that loss. This approach is rather straightforward in a Bayesian context, but has not typically been explored in a frequentist setting, a recent exception being the work of Genovese and Wasserman [11].

2. It is possible to cast the question in a model selection framework, discussing which parameters $\theta_i \neq 0$ need to be introduced to maximize a criterion such as AIC, BIC, or MDL. The literature on model selection is extensive; particularly relevant for our setting is the work of Casella and George [6] and George and McCulloch [13], which illustrates how to carry out model selection in a Bayesian setting and use a Markov chain Monte Carlo (MCMC)

algorithm for exploration of the prior, and Ishwaran and Rao [16], who suggest using this framework for gene expression array analysis.

3. Yet another approach is to estimate θ as a sparse vector, that is, using the knowledge that most of its elements are zero. This methodology has been developed primarily in the context of wavelet thresholding. There, the choice of an appropriate basis (as wavelets) has the effect of transforming the observations in a sparse vector that it is known to have, on theoretical grounds, only a small number of non-zero components: denoising of the original observations can be achieved by thresholding the coordinates of the transformed vector that are not significantly different from zero. Donoho and Johnstone [8] provides an introduction to this area, and Abramovich, Sapatinas, and Silverman [2] illustrates how to set up a Bayesian model for this framework.

All three of the described approaches have clear links with each other and with hypothesis testing. The manuscript by Abramovich, Sapatinas, and Silverman [2] illustrates particularly well the connections between estimation of a sparse vector and test of multiple hypotheses with corrections for multiple comparisons, exploring both Bonferroni-type procedures and FDR procedures. In this manuscript, we illustrate how the framework of estimation of sparse vectors can be applied to the microarray context and what its connection to the more common multiple testing approaches are. We devote attention to this particular approach because it aims to estimate the entire vector θ, both identifying the zero components and denoising the remaining ones, thus providing the biologist with a rich set of information. Additionally, it offers a framework through which to understand and improve the thresholding rules that biologists have typically used in analyzing microarray data.

The remainder of the manuscript is organized as follows. Section 3 introduces a Bayesian framework to estimate sparse vectors by identifying appropriate priors and loss functions and describes the form of the estimator with a simple data model. Section 4 shows how the same principles can be applied to more general data model and introduces a hierarchical Bayesian framework

for the estimation of unknown parameters; connection with FDR-controlling procedures in the context of microarray are explored using a simulation study. Section 5 presents the analysis of two datasets and illustrate how our inferential framework can be applied to general models for microarray data.

3 Bayesian estimation of a sparse mean vector

We introduce into the Bayesian inferential framework the knowledge that the vector of parameters θ is sparse by 1) defining a prior distribution on θ that gives substantial weight to the zero value and 2) selecting a loss function that penalizes non-sparse estimates. In particular, we specify:

$$\theta_i \sim \omega\,\mathcal{N}(0,p) + (1-\omega)\delta_0(\theta_i), \quad \theta_i \;\; iid \tag{1}$$

$$\mathcal{L}(\widehat{\theta}) = \sum_{i=1}^{N} |\theta_i - \widehat{\theta}_i| \tag{2}$$

The prior distribution assumes independence between the θ_i (conditional on p and ω) and assigns a positive probability to $\theta_i = 0$, which indicates that there is no functional change in expression for gene i between the baseline and experimental conditions. Notice that this translates, irrespective of the probability model linking the θ_i to the observed data, to a *posterior* distribution of the θ_i that also has point mass at zero. Mixture prior distributions of this nature are typically used in a Bayesian framework when researchers want to test a point hypothesis (see Berger [4]). They have been used in regression variable selection (Mitchell and Beauchamp [21], George and McCulloch [13]) and are becoming increasingly popular in microarray analysis (Newton [22], Ishwaran and Rao [16], Ibrahim *et al.* [15]), although in these three papers the Bayesian model is used within a hypothesis testing framework. Abramovich, Sapatinas, and Silverman [2] describe their use in conjunction with the ℓ_1 loss function in the context of wavelet coefficient denoising. We will initially review their model, in order to provide some insight in the nature of these assumptions. A crucial difference between the setting of Abramovich, Sapatinas, and Silverman [2] and ours

is that ultimately we do not assume ω known, while these authors do. In this respect, our prior assumptions are closer to those of Johnstone and Silverman [17], even if our data model differs substantially.

The proposed ℓ_1 loss function has the following notable characteristics: 1) it is a global loss (that is, the error is evaluated across all the θ_i) so that it naturally leads to a control of global error, and 2) being based on absolute rather than squared difference, it does not discount small errors—which are the ones most likely to be made when estimating a parameter that is truly equal to zero. Abramovich, Sapatinas, and Silverman [2] state that ℓ_1 loss is the "natural measur[e] for spatially inhomogeneous functions"; to see why this is so, it is useful to observe how this loss function behaves with respect to sparsity. A vector with small ℓ_1 norm tends to be sparser than a vector with the same ℓ_2 norm. For example, the n-dimensional vector $(1/n, \ldots, 1/n)$ has ℓ_2 norm $1/n$, but ℓ_1 norm equal to 1. The ℓ_1 loss function, then, favors sparser error vectors and therefore behaves more like a misclassification rate than the ℓ_2 loss function does. Coupling this observation with the knowledge, embedded in our prior choice, that the parameter vector is likely to contain a sizeable number of zeroes, one gets an inferential procedure that tends to select zero as a point estimate when there is a substantial probability that this is the true value and no error is made.

This is readily apparent if one considers what happens to Bayesian inference under this prior and loss function. Under ℓ_1 loss, Bayesian risk is minimized by choosing the posterior median, not mean, as the estimator for θ_i. As stated above, the posterior distribution of each θ_i will have mass at zero, and its median will be zero whenever this mass is larger than $1/2$ or when the difference between the probabilities of the two tails is small enough. Under ℓ_2 loss, on the other hand, the Bayes estimator is the posterior mean and therefore will never equal zero. To make this point explicit and illustrate further properties, we consider a specific data model that was originally considered in Abramovich, Sapatinas, and Silverman [2], mindful that it represents an oversimplification of microarray data.

Suppose that M slides are hybridized and the observed microarray data are represented by

$\{y_{ik}\}$, $k = 1, \ldots, M$. These data are assumed (and will be for the rest of the paper) to be quality-controlled, channel-normalized, and log-transformed differences between experimental cells and control cells. Tseng *et al.* [27] and Yang *et al.* [29] describe rank-invariant normalization methods applicable to the experiments described here. Let us then assume that

$$y_{ik} = \theta_i + \varepsilon_{ik}, \tag{3}$$

where the $\varepsilon_{ik} \overset{iid}{\sim} \mathcal{N}(0, \pi)$. Throughout this paper, Gaussian distributions are specified by their mean and *precision*, defined as the reciprocal of variance. Under these conditions, an analytical solution for the posterior median can be derived:

$$\widehat{\theta}_i = \mathrm{med}(\theta_i | \{y_{ik}\}) = \mathrm{sign}(\bar{y}_{i\bullet}) \max(0, \zeta_i)$$

where

$$\zeta_i = \frac{\bar{y}_{i\bullet}}{p/\pi + 1} - \frac{1}{\pi\sqrt{p/\pi + 1}} \Phi^{-1}\left\{\frac{1 + \min(\xi_i, 1)}{2}\right\}$$

and ξ_i is the posterior odds ratio for the component at 0, namely

$$\xi_i = \frac{1-\omega}{\omega}\sqrt{\pi/p + 1}\exp\left\{-\frac{\pi\bar{y}_{i\bullet}}{2(p/\pi + 1)}\right\}.$$

This analytic expression allows us to point out some features of the ℓ_1 estimator. It is a continuous thresholding rule, with a thresholding level that depends on the noise variance, the sparsity level, and the precision of the Gaussian component of the prior. When $\widehat{\theta}_i$ is not zero, it is a shrunk version of the data average. Figure 1 shows how the threshold changes under three different levels of noise variance. With ω set to .05 and p set to 1, the noise variance $1/\pi$ is variously set at $1/4$, $1/16$, and $1/64$. Asymptotically, the function approaches a slope of $1/(p/\pi + 1)$. The right-hand graph plots $\widehat{\theta}_i$ as a function of $\bar{y}_{i\bullet}\sqrt{\pi}$, which is the mean scaled by the standard deviation: while changes are minor, it is clear that the threshold is not linear in the noise variance $1/\pi$. The relationship between noise variance, sparsity, and threshold is shown in greater detail in figure 2. In the right-hand plot,

noise variance is held constant and the threshold is plotted as a function of (constant and known) sparsity. The threshold is highest for the most sparse data, while the threshold shrinks to zero for well-populated data.

We have already discussed at length the attractiveness of thresholding estimators in the context of gene expression array studies, underscoring how they lead to inferential results that reflect practitioners' beliefs on the nature of changes in transcription levels, in particular, that the change will be zero for a large proportion of genes. Thresholding estimators lead to a result that is also connected to the one of hypothesis testing. Abramovich *et al.* [1] shows how the universal threshold of Donoho and Johnstone [8] can be related to a Bonferroni correction of Gaussian tests. In the same paper, the authors stress how FDR-controlling procedures correspond to a thresholding that is adaptive to the unknown level of sparsity. The Bayesian thresholding estimator presented above depends on the level of sparsity in the data, but so far this has been assumed known. In the following section, we will illustrate how it is possible, in a hierarchical Bayesian model, to estimate the level of sparsity and obtain a thresholding estimator that is adaptive to unknown sparsity. This property appears particularly appealing in the microarray context, where the advantage of FDR-controlling procedures has been repeatedly argued. Moreover, the connection between hierarchical Bayes models and FDR-controlling procedures, underscored by Efron *et al.* [9] in a non-parametric context, is of additional interest.

4 Hierarchical Bayes and adaptive thresholding

In this section, we start considering a more realistic model for gene expression and make more appropriate assumptions on the identity of unknown parameters. In practice, there are a variety of models for gene expression that one may consider, and the inferential framework we are describing can be applied, with comparable implications, to any such model as long as it includes a population change-in-expression parameter θ. When presenting some data analysis results in the

next section, for example, we will use the model proposed by Tseng *et al.* [27]. For the time being, we intend to focus on basic properties of our Bayesian inferential framework and illustrate them with a comparative analysis.

Our setting is as follows. The experiments involve N genes and M slides. We assume that the observed expression value for gene i in slide k is distributed as

$$y_{ik} = \theta_i + \varepsilon_{ik}, \tag{4}$$

where $\varepsilon_{ik} \overset{ind}{\sim} \mathcal{N}(0, \pi_i)$, with a different variance for each gene. On the noise precision π_i we place a gamma prior with shape a_i and rate b_i, notated $\pi_i \sim Gamma(a_i, b_i)$. Notice that not only the π_i but also the a_i and b_i are free to vary from gene to gene. The a_i and b_i are considered fixed and known, although in practice their values will be estimated from a separate set of calibration slides, in which mRNA extracted from comparable cells lines are oppositely dyed and hybridized to the same slide, or from the the comparison data y_{ik}, or from a combination of both. Using the data y_{ik} to estimate prior parameters follows an empirical Bayes paradigm. While we attempt to avoid overtapping the data, there are situations where using the data to estimate certain nuisance parameters is both appropriate and robust.

It is possible to place a gamma prior on the expression precision p, although in practice we found that it may be convenient to choose one specific value of p to mimic the average effect size of the changes in expression in the dataset. We describe here the case where p is assumed to follow the prior distribution $p \sim Gamma(\alpha, \beta)$. The weight parameter ω receives a beta prior, $\omega \sim Beta(c_0, c_1)$. The parameters α, β, c_0, and c_1 are specified to yield diffuse, essentially noninformative, distributions.

Figure 3 diagrams the described model, showing its hierarchical structure. For notational convenience, let $z_i = 1_{(\theta_i \neq 0)}$, indicating whether or not gene i experiences any expression change between experimental and control conditions. With this in mind, we can write an expression pro-

portional to the joint posterior likelihood of the parameter set $\{\omega, p, \theta, \pi\}$:

$$\begin{aligned} p(\omega, p, \theta, \pi|y) \quad \propto \quad & \omega^{c_0}(1-\omega)^{c_1} p^{\alpha-1} e^{-\beta p} \\ & \prod_{i=1}^{N} \omega^{z_i}(1-w)^{(1-z_i)} f(\theta_i|z_i) \pi_i^{a_i-1} e^{-b_i \pi_i} \\ & \prod_{i=1}^{N} \prod_{k=1}^{M} \pi_i^{1/2} e^{-\pi_i (y_{ik}-\theta_i)^2/2} \end{aligned} \tag{5}$$

Unlike what is illustrated in the previous section, it is not possible to derive analytical expression for the posterior medians of θ_i. We describe an MCMC routine for exploration of the posterior distributions. It is a collapsed Gibbs sampler (see Liu, Wong, and Kong [19]), based on the following conditional distributions:

$$\pi_i|\omega, p, \theta, \pi_{-i} \quad \sim \quad Gamma(a_i + M/2, b_i + \sum\nolimits_{k=1}^{M} (y_{ik} - \theta_i)^2/2) \tag{6}$$

$$p(z_i = 0|\omega, p, \pi) \quad \propto \quad \omega_0 \exp\left\{-\pi_i \sum\nolimits_{k=1}^{M} y_{ik}^2/2\right\} \tag{7}$$

$$p(z_i = 1|\omega, p, \pi) \quad \propto \quad \omega_1 \sqrt{\frac{p}{p + M\pi_i}} \exp\left\{\frac{M^2 \pi_i^2 \bar{y}_{i\bullet}^2}{2(p + M\pi_i)} - \pi_i \sum\nolimits_{k=1}^{M} y_{ik}^2/2\right\} \tag{8}$$

$$\theta_i|\omega, p, \pi, z_i = 0 \quad \sim \quad \delta_0 \tag{9}$$

$$\theta_i|\omega, p, \pi, z_i = 1 \quad \sim \quad \mathcal{N}(\bar{y}_{i\bullet}\pi_i M/(p + \pi_i M), p + \pi_i M) \tag{10}$$

$$\omega|p, \theta, \pi \quad \sim \quad Beta(c_0 + \Omega_0, c_1 + \Omega_1) \tag{11}$$

$$p|\omega, p, \theta, \pi \quad \sim \quad Gamma(\alpha + \Omega_1/2, \beta + \sum\nolimits_{i:z_i=1} \theta_i^2/2), \tag{12}$$

where $\Omega_j = \sum_{i=1}^{N} 1(z_i = j)$ for $j = 0, 1$.

4.1 Simulations

In order to explore the performance of our MCMC procedure and the characteristics of our inferential framework, we conducted a simulation study. To interpret results, it is useful to analyze the same simulated datasets with another procedure for benchmarking purposes. Given that FDR-controlling procedures have been repeatedly proposed for the analysis of this type of dataset and

that they are, like ours, adaptive to unknown sparsity, we decided to use them as our benchmark. In particular, because we wanted to compare the performance of the ℓ_1 estimator to those of another estimator (rather than a hypothesis testing procedure) we considered the following FDR-based thresholding estimator, as described in Abramovich *et al.* [1] We indicate with q the desired level of FDR, which represents a tuning parameter for this thresholding estimator. Let t_i be the t-statistic corresponding to gene i, resulting in a p-value p_i. We notate the ordered vector of p-values by $(p_{(1)}, \ldots, p_{(N)})$. Let k be the largest i for which $p_{(i)} < q(i/N)$, and t_k the corresponding t-statistic. Then the FDR thresholding estimator of level q sets $\widehat{\theta}_i = 0$ if $t_i \leq t_k$, otherwise $\widehat{\theta}_i = \bar{y}$.

Each simulated data set is comprised of $N = 1000$ genes, with $M = 4$ slide replicates per gene. Given ω, which specifies the degree of sparsity, ωN randomly selected genes are assigned a non-zero θ_i generated from $\mathcal{N}(0, 1)$, while for all other genes $\theta_i = 0$. In different simulations, ω varies among $1, 2.5, 5, 10, 15$, and 25%. Given the θ_i, noise variance is added according to the hierarchical model. Namely, gene-specific precisions π_i are generated from $Gamma(a, b)$, where $a = 2$ and b varies among $1/10$, $4/10$, and $16/10$ in different simulations, each increase in b representing a doubling of the average noise deviation. For both the MCMC and FDR-controlling procedures, a and b are considered fixed and known. Therefore, t-statistics for FDR control are computed using an adjusted variance estimate incorporating the prior gamma distribution,

$$\widehat{\sigma}^2_{\mathrm{FDR}i} \;=\; \frac{\widehat{\sigma}_i^2(M/2) + b}{(M/2) + a}, \tag{13}$$

where $\widehat{\sigma}_i^2$ is the usual variance estimate computed from the $M = 4$ data points for gene i. To compute p-values, the resulting t-statistics are compared to the t distribution with $M - 1 + 2a$ degrees of freedom.

The twelve plots in figure 4 compare the performance of the ℓ_1 estimator to FDR control at $q = 0.05$ and $q = 0.20$, for simulations at the various levels of sparsity and noise variance. Each point on these plots represents an average of twenty independent simulations.

The first row of plots shows the predicted proportion of zero-changers, that is, one minus the

predicted sparsity. The dashed line at unity slope represents the true value. In every instance, each of the three statistics overestimates the proportion of zeroes, increasingly so as the level of noise variance increases. This is to be expected and not really a problem in terms of the interpretability of the results as, given our model for θ, some of the true changers have θ_i very close to zero. At the lowest noise variance, the ℓ_1 estimator falls between the two FDR-controlling procedures. As noise variance increases, however, FDR control becomes increasingly conservative while ℓ_1 still discriminates between the different levels of sparsity. The next row shows mean absolute error and gives some hint for this phenomenon. As noted before, the ℓ_1 statistic minimizes absolute error loss, so it is no surprise that ℓ_1 performs at least as well as the FDR-controlling statistics in each case. The third row shows the actual false discovery rate achieved by the three estimators. FDR control does achieve its goal of staying under the two dotted lines at .05 and .20, respectively, whatever the sparsity. The ℓ_1 estimator produces increases values of FDR as the noise variance increases. Generally speaking, the ℓ_1 estimator seems to behave as an FDR-thresholding procedure that adaptively selects the level of FDR control based on noise variance. Clearly this does not control FDR at a predetermined level q, but does appear to lead to a minimization of the misclassification rate, as illustrated in figure 5. This is not surprising, given the properties of ℓ_1 loss described in the previous section, and represents an important point in favor of our proposed estimator.

5 Data analysis and more realistic models of expression data.

We now consider two datasets and present the results of their analysis with ℓ_1 estimation. The first example consists of a series of experiments aimed at elucidating the variation in gene expression resulting from the suppression of transcription of a gene known to cause Friedreich's ataxia. The second dataset was collected with the explicit intention of studying the noise level of cDNA experiments and compares the expression values of genes in *E. coli* when cells are grown in two different media.

5.1 Friedreich's ataxia

Friedreich's ataxia is one of the most common forms of autosomal recessive ataxia in humans and is caused by reduced expression of the frataxin-producing gene on chromosome 9q13. Clinical features include progressive neurological, cardiovascular, skeletal, and endocrine abnormalities. The disease has been modeled in transgenic mice in which the expression of frataxin is reduced to 25%–36% of wild-type levels. (Knocking out the entire gene is lethal for mice, see Miranda *et al.* [20]) These mice are phenotypically identical to wild types because of compensatory effect. In order to achieve a finer resolution of the phenotype, a study of the changes in expression levels for a large set of genes involved in brain function appears as a promising strategy. Indeed, the laboratory of professor Dan Geschwind is carrying out experiments along these lines [7].

We were involved in the analysis of one of the initial experiments in this study. A custom mouse cDNA microarray of roughly 10,000 genes was probed with cDNA from three brain regions (brain stem, cerebellum, and spinal cord) affected in Friedreich's ataxia, from each of four transgenic mice, two male and two female. Four wild-type mice served as controls, again two male and two female. The two male controls, however, were combined into a single male control, same with the female. Thus, although cells from the two male transgenic mice were hybridized to separate slides, they were compared to the same controls; likewise with the females. Each of the twelve comparisons, furthermore, were performed on a pair of dye-flipped slides, which were subsequently averaged to help control labeling bias. After data normalization and quality control, a data set consisting of 8,578 genes was produced for analysis.

The Friedreich's ataxia data are analyzed using the model described in the previous section. Initial analyses show that expression patterns vary significantly between brain regions, so the set of θ_i is allowed to vary independently between each brain region. In other words, the data are analyzed with N not merely equal to 8,578 but $3 \times 8{,}578 = 25{,}734$. However, because the probe design is the same regardless of brain region, noise variances $1/\pi_i$ are constrained not to vary from

brain region to brain region. Thus $\pi_i = \pi_{i+8578} = \pi_{i+17156}$ for $i = 1, \ldots, 8578$, and estimation of the π_i benefits from a three-fold increase in readings per estimate.

In addition to the twelve dye-crossed pairs of comparison slides, a dye-crossed pair of calibration slides was produced in which cerebellar tissue from two transgenic mice was oppositely dyed and hybridized. Because both tissue donors are transgenic, $\theta_i = 0$ for all genes on the calibration slides, and therefore all variation is due to biological and measurement variance, which in the single-layer model is represented by the $1/\pi_i$.

Recall that the prior distributions on the π_i are $Gamma(a_i, b_i)$. In this example, the limited amount of calibration data dictates that we constrain a_i and b_i to be constant between genes. If we represent the calibration data by x_i, then viewing the distribution of the $1/x_i^2$ can help determine reasonable values for a and b. In this example, viewing a histogram of the $1/x_i^2$ led to setting a to 0.7 and b to 0.09. As we show in a second empirical example, a larger complement of calibration data enables gene-specific estimation of one or both of these parameters.

A full complement of comparative data was not available for every gene, and restricting analysis to genes with complete data would have been excessively restrictive. Computation of conditional posteriors, therefore, required slight modification. For example, the conditional posterior distribution of π_i is $Gamma(a_i + M/2, b_i + \sum_{k=1}^{M}(y_{ik} - \theta_i)^2/2)$. For a given gene i, if not all the y_{ik} are available then M is replaced by the number of populated y_{ik}, and the summation in the rate parameter is only over those k for which y_{ik} is populated.

We wish to compare the ℓ_1 approach to FDR control as a benchmark. The computation of t-statistics and p-values reflects the prior knowledge on noise variance gleaned from the calibration slides, and adjustments are directly analogous to those described in the simulation example. Figure 6 shows the collection of p-values. On the left is a histogram of all p-values. The increased weight toward zero indicates the existence of differential expression in a significant number of genes. On the right-hand plot of ordered p-values, however, we see that FDR control at the moderate values of $q = .05$ and $q = .30$ identifies *no* differentially expressed genes. This is a somewhat disappointing

result, as it appears to be biologically very reasonable to expect a change in the expression of some genes. Only by setting q to the unusually high value of 0.55 yields a reasonable number of genes, roughly 200.

Figure 7, on the other hand, shows the ℓ_1 analysis of the spinal cord data. In both plots, the ℓ_1 estimate (that is, the posterior median) is shown on the vertical axis. The left-hand plot simply shows $\bar{y}$ on the horizontal axis, while the right-hand plot shows $\bar{y}$ scaled by a simple estimate of standard deviation. The right-hand plot exhibits a shrink-or-kill behavior that is roughly equivalent to that shown in the simpler example of figure 1. Deviations from a simple functional line are due to the gene-specific variance estimates. In other words, the right-hand plot can be seen as a random admixture of several of the plots shown in figure 1. The ℓ_1 estimator identifies 240 genes as differentially expressed. The earlier simulations indicate that this number is probably a conservative estimate, even though roughly half of the genes identified are doomed to be false discoveries.

5.2 E. coli

We use the *E. coli* data analyzed in Tseng *et al.* [27] as a second empirical example. In this experiment, *E. coli* cells were grown in contrasting growth media, namely glucose and acetate. There were $E = 3$ biological replicates, each of which was hybridized to $R = 2$ replicate slides. Additionally, there were four calibration slides, two of which compared acetate-grown to acetate-grown cells, and two of which compared glucose-grown to glucose-grown. Following channel normalization and quality filtering, a data set of $N = 2291$ genes was prepared for analysis.

Given the structure of this dataset, it is possible and meaningful to introduce a more sophisticated model for expression values. In particular, we are going to use the same hierarchical data model adopted in Tseng *et al.* [27], except using our mixture prior and ℓ_1 loss, and then compare our results to those obtained with the slightly different inferential framework of the original paper.

The hierarchical model explicitly differentiates between biological and measurement error sources of variance. Instead of simply referring to M replicates, E biological subjects are identified, with R slide replicates per subject, for a total of $R \times E$ observations per gene. Figure 8 diagrams the hierarchical structure of the model. For a given gene i, there are E *unobserved* subject-specific expression changes $\mu_{i1}, \ldots, \mu_{iE}$, with subjects indexed by $l = 1, \ldots, E$. The prior distributions on the μ_{il} are, as in the single-layer model, Gaussian with means θ_i and variances $1/\pi_i$. Similarly, the priors on the precisions π_i are gamma with shape and rate parameters allowed, in the most general form, to vary from gene to gene.

For a given subject l and gene i, there are R *observed* expression changes $y_{il1}, \ldots, y_{ilR}$, with replicates indexed by $k = 1, \ldots, R$. Given the subject-specific μ_{il}, the priors on the y_{ilk} are Gaussian with means μ_{il} and variances $1/\tau_i$. The priors on the precisions $1/\tau_i$ are again gamma with possibly gene-specific shapes A_i and rates B_i.

The collapsed Gibbs sampler procedure to explore the posterior obtained with this model is very similar to the one previously described. Below we provide the conditional distributions used in each step has to be repeated for each possible value of subscripts i, l, and k.

$$\mu_{il}|y, \theta, \pi, \tau \sim \mathcal{N}\left((\pi_i\theta_i + \bar{y}_{il\bullet}\tau_i R)/(\pi_i + \tau_i R), \pi_i + \tau_i R\right) \quad (14)$$

$$\pi_i|y, \theta \sim Gamma(a_i + E/2, b_i + \sum\nolimits_{k=1}^{E}(y_{ik} - \theta_i)^2/2) \quad (15)$$

$$\tau_i|y, \mu \sim Gamma(A_i + RE/2, B_i + \sum\nolimits_{l=1}^{E}\sum\nolimits_{k=1}^{R}(y_{ilk} - \mu_{il})^2/2) \quad (16)$$

$$p(z_i = 0|\omega, p, \pi) \propto \omega_0 \exp\left\{-\pi_i \sum\nolimits_{k=1}^{M} y_{ik}^2/2\right\} \quad (17)$$

$$p(z_i = 1|\omega, p, \pi) \propto \omega_1 \sqrt{\frac{p}{p + M\pi_i}} \exp\left\{\frac{M^2\pi_i^2\bar{y}_{i\bullet}^2}{2(p + M\pi_i)} - \pi_i \sum\nolimits_{k=1}^{M} y_{ik}^2/2\right\} \quad (18)$$

$$\theta_i|\omega, p, \pi, z_i = 0 \sim \delta_0 \quad (19)$$

$$\theta_i|\omega, p, \pi, z_i = 1 \sim \mathcal{N}(\bar{y}_{i\bullet}\pi_i M/(p + \pi_i M), p + \pi_i M) \quad (20)$$

$$\omega|p, \theta, \pi \sim Beta(c_0 + \Omega_0, c_1 + \Omega_1) \quad (21)$$

$$p|\omega, p, \theta, \pi \sim Gamma(\alpha + \Omega_1/2, \beta + \sum\nolimits_{i:z_i=1} \theta_i^2/2), \quad (22)$$

where $\Omega_j = \sum_{i=1}^{N} 1(z_i = j)$ for $j = 0, 1$.

Prior distributions for the noise variance components $1/\pi_i$ and $1/\tau_i$ were estimated from a combination of calibration and comparison data. The priors for the biological variances $1/\pi_i$ were estimated from a weighted average of gene-specific and genome-wide statistics, while the priors for slide-to-slide variances $1/\tau_i$ were computed genome-wide and were the same for each gene.

Under this more complex model structure, we no longer use FDR control as a benchmark, because it is unclear how a t-test would take advantage of the additional information yielded by the hierarchical structure and calibration data. Instead, we use as a benchmark the framework of Tseng *et al.* [27], which is the same as the Bayes structure described above except for the prior distribution on θ. In their application, they place a flat, noninformative prior on the θ_i, which leads to posteriors of θ_i that have no point mass at zero.

Because the flat-prior posteriors place no special emphasis on zero-changers, there is no implicit Bayesian test of the hypothesis $H_0 : \theta_i = 0$. In Tseng *et al.* [27], the authors report 95% confidence intervals for the θ_i and identify a gene as up- or down-regulated when its confidence interval does not include zero. Figure 9 compares mixed-prior and flat-prior results. Both graphs plot the mixed-prior posterior median against a normalized within-gene average. We recognize the characteristic shape from earlier figures and note that deviations from a single functional line are due to gene-to-gene variation in posterior distributions of biological and slide variance. On the left-hand graph, genes whose flat-prior 95% confidence interval falls outside of zero are highlighted, while the right-hand graph does the same at the 99% level. Our analysis identifies 168 differentially expressed genes, which falls in between the 75 identified by flat-prior 99% CI and the 266 identified by flat-prior 95% CI.

Figure 10 compares the two posteriors in a slightly different fashion. Posterior medians are again plotted on the vertical axis, but the horizontal axis is a simple data mean, meaning that non-zero points show more spread than in figure 9. The vertical lines span 95% confidence intervals from the flat prior model, and we note that with the exception of only one gene at the far right,

all posterior medians from the mixture prior fall within them. Confidence intervals that span zero are colored differently from those that remain above or below; this emphasizes that the ℓ_1 analysis selects more genes than selection based on the 99% CI, but fewer that one based on 95%, as shown in figure 9.

6 Discussion

We have described how to estimate the value of the population expression changes for the genes surveyed in an array experiments in a hierarchical Bayes framework under the assumption that a large number of such changes is effectively zero. Using an estimation, rather than test of hypothesis approach deflects the emphasis from type I error control. We argue this to be appropriate for exploratory investigations whose aim is to identify genes that likely change expression across two cell lines characterized by expression differences.

Our simulation study is conducted under rather simplistic hypotheses on the nature of the noise in array experiments; it is not possible to extrapolate the performance of our estimator in that setting to "real life." The characteristics of our procedure and its relation to a false discovery control that emerge from the simulation are, however, likely to be valid more generally. In particular, as the ℓ_1 loss function is related to misclassification error, our estimates will tend to provide a low misclassification. This can be described as selecting a level of FDR control adaptively on the signal-to-noise ratio.

The Friedreich's ataxia data set provides an illustration of such advantage: despite a low number of replicates compared to the experimental noise level, we were able to provide a list of genes whose transcript levels may change in the frataxin-deficient mice. While a large number of false positives is to be expected, their proportion is selected to minimize a global loss, not determined by a researcher's arbitrary choice of a "reasonable" number of genes in the final list.

The *E. coli* data illustrate another advantage of the estimation procedure we suggest. Often

the results from expression experiments are used in further analysis, such as clustering or classification and regression procedures. In such cases, practitioners often summarize the results of replicate comparisons between two cell lines with their average. Using our procedure to provide this summary allows a better use of the information in the data set not only by borrowing strength across genes, but also by passing to further analysis the fact that a given gene appeared not to have any detectable change in expression across the two studied conditions.

While we think that the description of array experiments we have provided correctly illustrates the nature of the majority of them, there are some cases where either considering θ a sparse vector or the goal of the experiment merely exploratory is less appropriate. A striking example of this are the "subtraction arrays" described in Geschwind et al. [14]: in a first step, a subtraction technique is used to identify transcripts that appear to be more abundant in one of the two cell lines under study; in a second step, the cDNA corresponding to these are spotted on an array and the amounts of their expression further quantified. In this situation one clearly expects the majority of genes to change expression and the assumption of sparsity on θ may not be appropriate. Furtermore, the experiment has a less exploratory nature and a more tight type I error control may be appropriate.

The estimation of sparse high dimensional vectors is an important topic in contemporary statistics [8, 1, 2, 17]. The application that is most often considered is perhaps thresholding of wavelets coefficients. Interesting theoretical results are obtained for cases where specific assumptions on the noise level are possible. The hierarchical Bayes framework presented here applies this methodology to a novel problem; in combination with the MCMC algorithm it allows for considerable variation in distributional assumptions, of which only a couple are illustrated here. This flexibility makes it particularly appealing in the context of microarray analysis and considerably broadens the domain of estimation procedures that rely on sparsity of the parameters.

Acknowledgments

We thank Prof. Dan Geshwind and his laboratory and Prof. James C. Liao and his laboratory for letting us use data they collected. Chiara Sabatti acknowledges support from the NSF (grants DMS 0239427 and DCC0326605), NASA/Ames (grant NCC2-1364), and NIH (Grant GM53275), and Stephen Erickson from the NIH (UCLA genomic analysis and interpretation training program).

References

[1] Abramovich, F., Y. Benjamini, D. Donoho, I. Johnstone. Adapting to Unknown Sparsity by controlling the False Discovery Rate. Technical Report, Statistics Department, Stanford University, 2000.

[2] Abramovich, F., T. Sapatinas, B. W. Silverman. Wavelet Thresholding via a Bayesian Approach. Journal of the Royal Statistical Society, Series B, 1998, 60:4, 725–749.

[3] Baldi, P., A.D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics, 2001, 17:6, 509–519.

[4] Berger J. O. Statistical decision theory and Bayesian analysis. Springer-Verlag, 2nd ed, 1985.

[5] Benjamini, Y., Y. Hochberg. The control of the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 1995, 57:1, 289–300.

[6] Casella, G., E.I. George. Explaining the Gibbs Sampler. The American Statistician, 1992, 46, 167–174.

[7] Choi, S., D. Tentler, M.M. Santos, M. Pandolfo, D.H. Geschwind. Microarray Analysis of Frataxin-Deficient Mouse Brain Identifies Compensatory Changes in Gene Expression and a Potential Role for Immune-Related Genes. Program No. 301.10. 2003. Abstract Viewer/Itinerary Planner. Washington, DC: Society for Neuroscience, 2003 Online.

[8] Donoho, D.L., I.A. Johnstone. Ideal spatial adaptation by wavelet shrinkage. Biometrika, 1994, 81, 425–455.

[9] Efron, B., R. Tibshirani. Empirical Bayes Methods and False Discovery Rates for Microarrays. Technical report, 2001, available from http://www-stat.stanford.edu/ tibs/research.html.

[10] Efron, B., R. Tibshirani, J.D. Storey, V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association, 2001, 96:456, 1151–60.

[11] Genovese, C., L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. Journal of the Royal Statistical Society, Series B, 2002, 64:3, 499–517.

[12] George, E.I., D.P. Foster. Calibration and empirical Bayes variable selection. Biometrika, 2000, 87:4, 731–747.

[13] George, E.I., R.E. McCulloch. Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association, 1993, 88:423, 881–889.

[14] Geschwind, D.H., J. Ou, M.C. Easterday, J.D. Dougherty, R.L. Jackson RL, Chen Z, Antoine H, Terskikh A, Weissman IL, Nelson SF, Kornblum HI. (2001) A genetic analysis of neural progenitor differentiation. *Neuron* 29:325–39.

[15] Ibrahim, J.G., M.-H. Chen, R.J. Gray. Bayesian Models for Gene Expression With DNA Microarray Data. Journal of the American Statistical Association, 2002, 97, 88–99.

[16] Ishwaran, H., J.S. Rao. Detecting differentially expressed genes in microarrays using Bayesian model selection. Journal of the American Statistical Association, 2003, 98:462, 438–455.

[17] Johnstone, I.M., B.W. Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. The Annals of Statistics, 2004, 32:4.

[18] Li, C., W.H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc. Natl. Acad. Sci, 2001, 98, 31–36.

[19] Liu, J.S., W.H. Wong, A. Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika, 1994, 81:1, 27–40.

[20] Miranda C.J., M.M. Santos, K. Ohshima, J. Smith, L. Li, M. Bunting, M. Cossee, M. Koenig, J. Sequeiros, J. Kaplan, M. Pandolfo. Frataxin knockin mouse. FEBS Letters, 2002, 512, 291–297.

[21] Mitchell, T. J., J. J. Beauchamp. Bayesian Variable Selection in Linear Regression. Journal of the American Statistical Association, 1988, 83:404, 1023–32.

[22] Newton, M.A., A. Noueiry, D. Sarkar, and P. Ahlquis. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics, 2004, 5:2, 155–176.

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

[24] Reiner A., D. Yekutieli, Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics, 2003, 19:3, 368–375.

[25] Sabatti C., S. Service, N. Freimer. False Discovery Rate in Linkage and Association Genome Screens for Complex Disorders. Genetics, 2003, 164, 829–833.

[26] Storey J.D., R. Tibshirani. Statistical significance for genome-wide studies. Proceedings of the National Academy of Sciences, 2003, 100, 9440–5.

[27] Tseng, G. C., M.-K. Oh, L. Rohlin, J. C. Liao, W. H. Wong. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Research, 2001, Vol. 29, No. 12 2549–57.

[28] Tusher, V., R. Tibshirani, C. Chu. Significance Analysis of Microarrays Applied to Transcriptional Responses to Ionizing Radiation. Proceedings National Academy of Science, 2001, 98, 5116–21.

[29] Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research, 2002, Vol. 30, No. 4 e15.

Figure 1: Thresholding behavior as noise variance changes

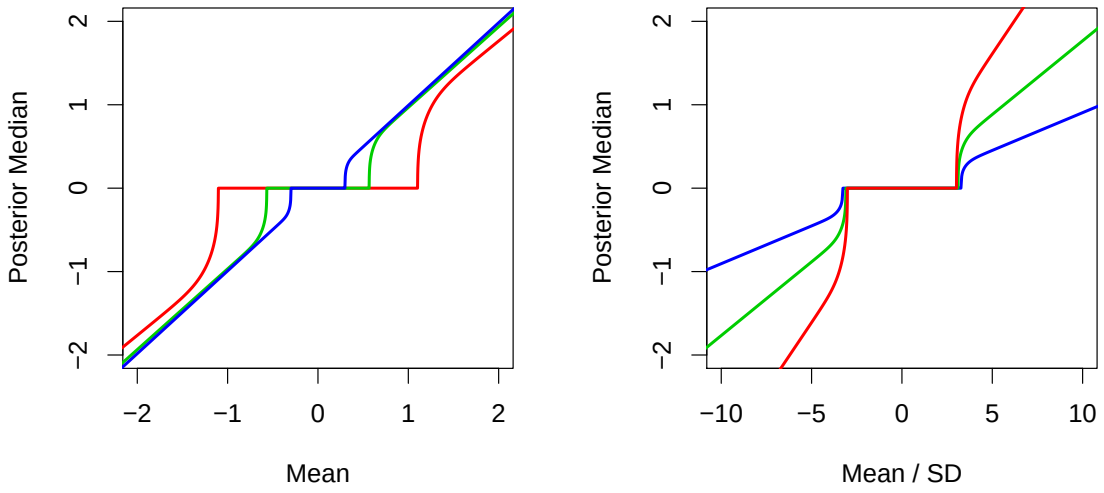


Figure 2: Threshold as a function of noise variance and sparsity

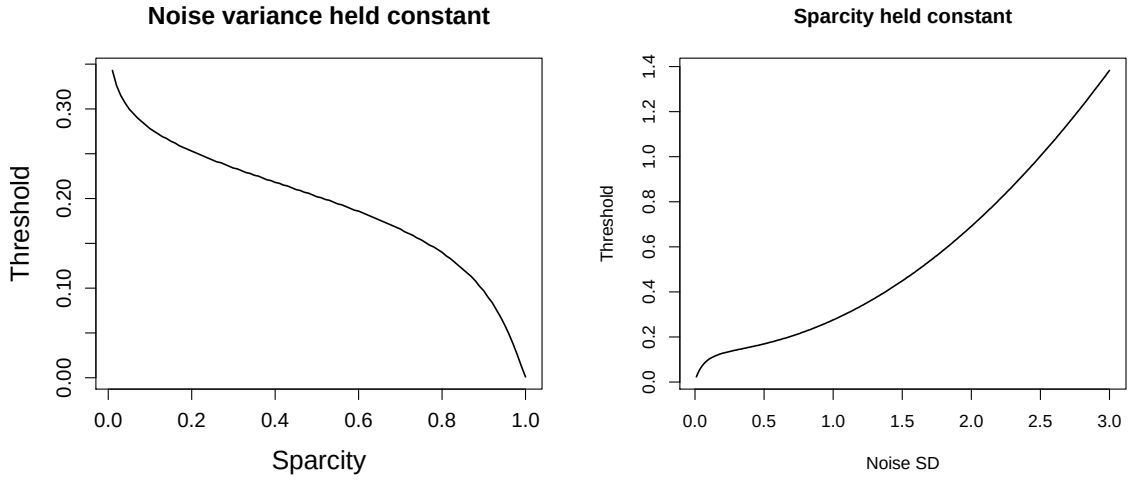
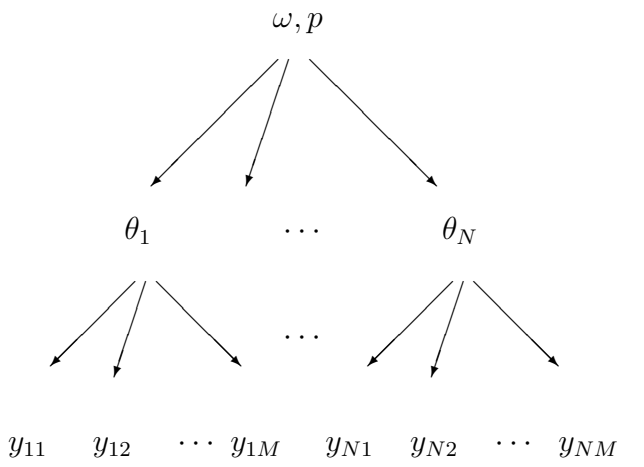


Figure 3: Single-layer model



$\omega \sim Beta(c_0, c_1), \;\; p \sim Gamma(\alpha, \beta)$

$\theta_i|\omega, p \sim \omega\, \mathcal{N}(0, p) + (1 - \omega)\delta_0(\theta_i)$

$y_{ik}|\theta_i, \pi_i \sim \mathcal{N}(\theta_i, \pi_i), \;\; \pi_i \sim Gamma(a_i, b_i)$

Figure 4: Performance of ℓ_1 estimator and FDR control at $q = 0.05$ and 0.20

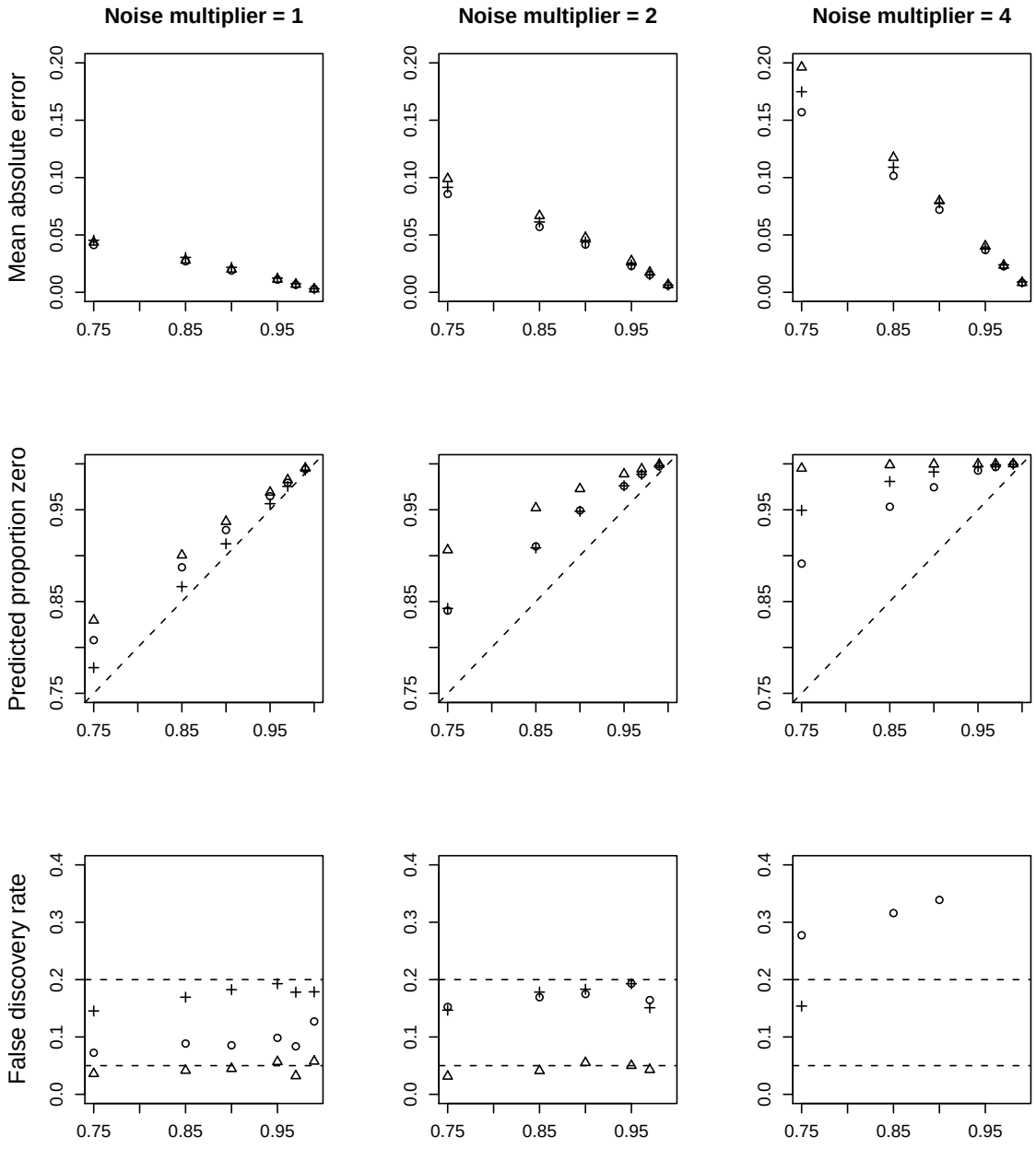


Figure 5: Comparison of overall misclassification rates

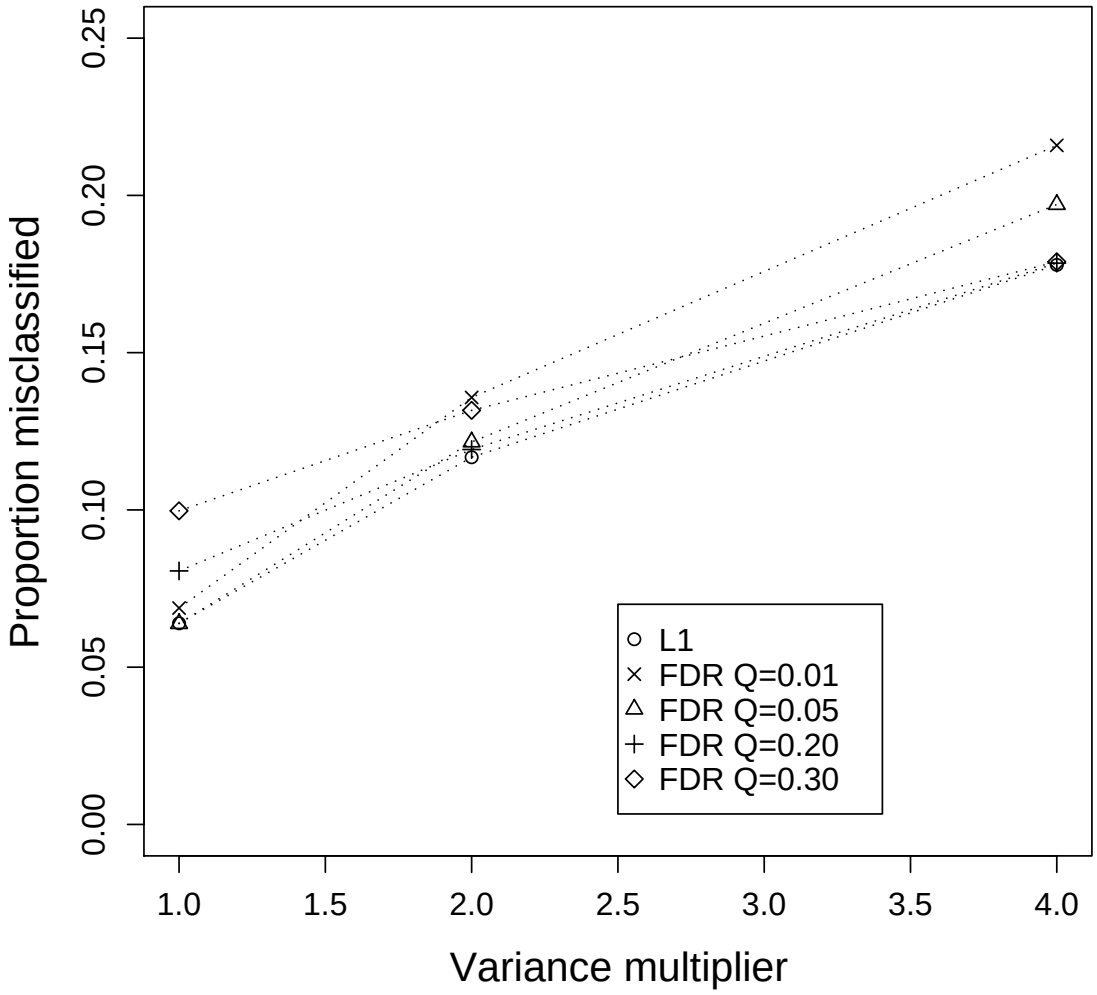


Figure 6: Friedreich's ataxia FDR-controlling analysis, spinal cord

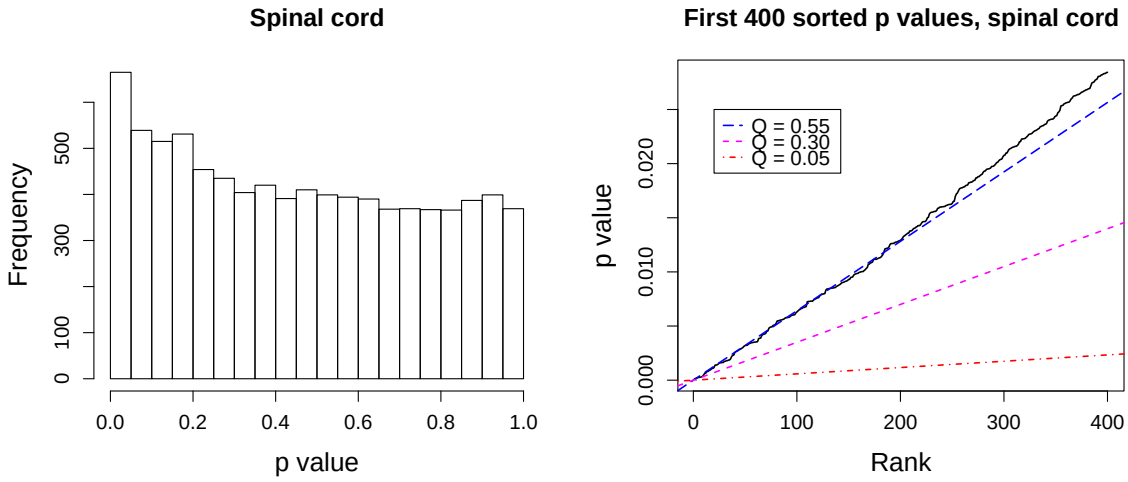


Figure 7: Friedreich's ataxia ℓ_1 analysis, spinal cord

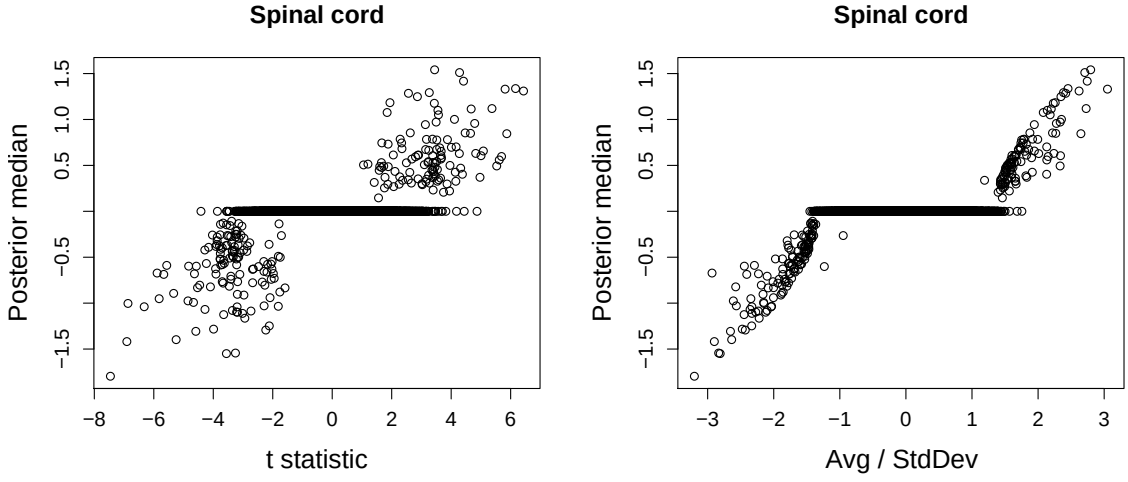
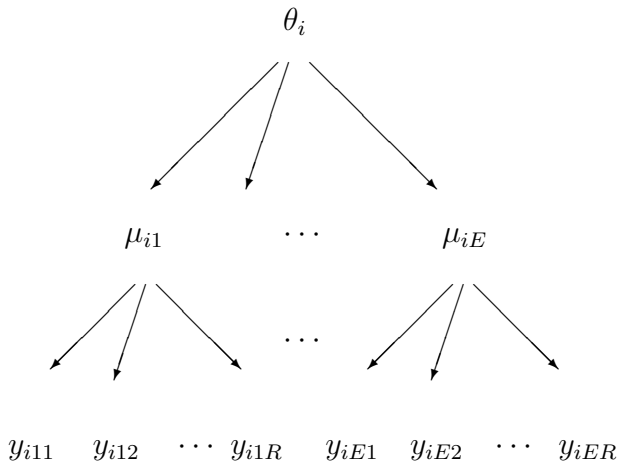


Figure 8: Two-layer model



$$\theta_i|\omega, p \sim \omega\,\mathcal{N}(0,p) + (1-\omega)\delta_0(\theta_i)$$

$$\mu_{il}|\theta_i, \pi_i \sim \mathcal{N}(\theta_i, \pi_i), \;\; \pi_i \sim Gamma(a_i, b_i)$$

$$y_{ilk}|\mu_{il}, \tau_i \sim \mathcal{N}(\mu_{il}, \tau_i), \;\; \tau_i \sim Gamma(A_i, B_i)$$

Figure 9: Comparison of ℓ_1 analysis with flat-prior confidence intervals, *E. coli* example

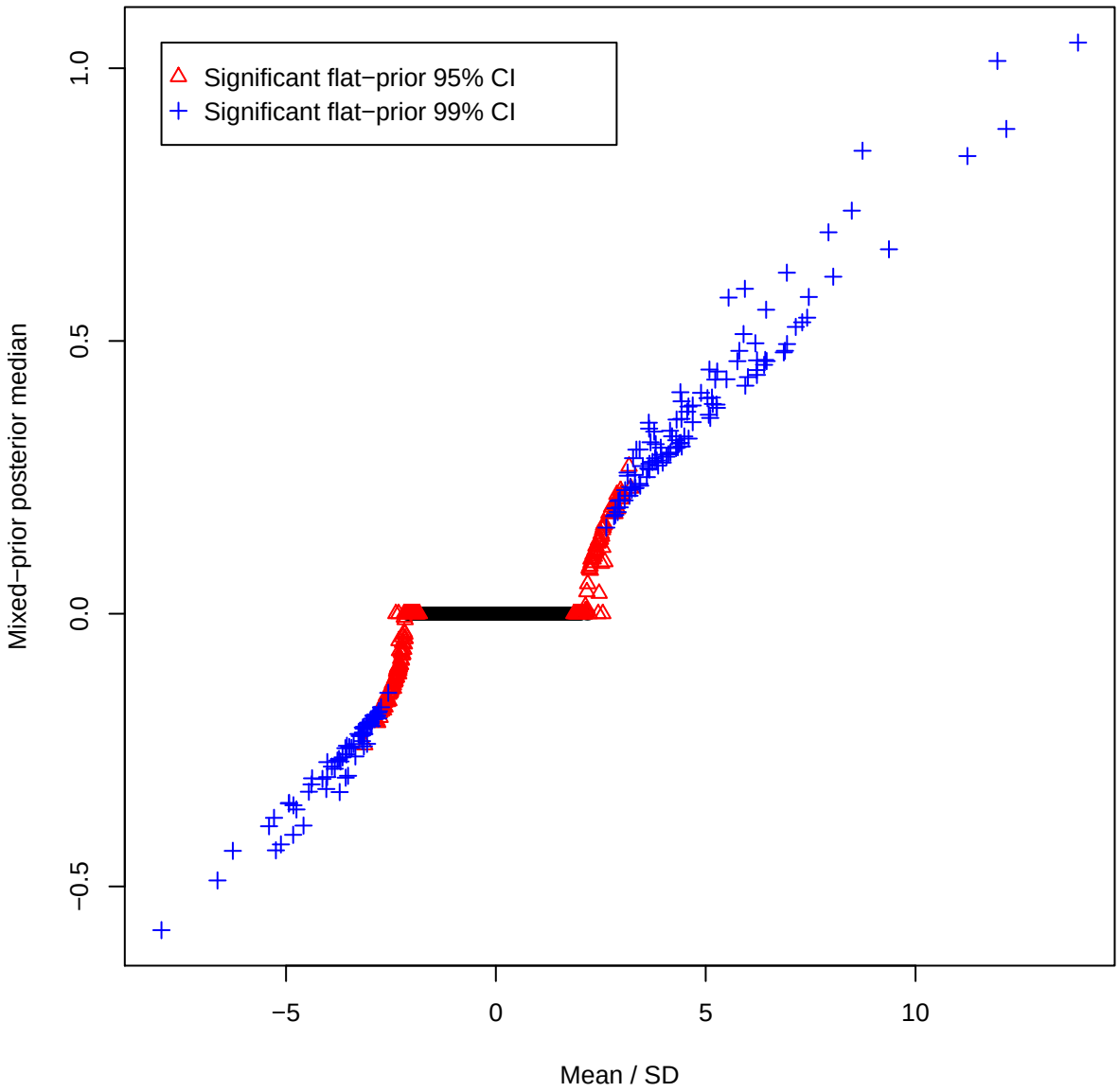


Figure 10: ℓ_1 posterior median and flat-prior 95% confidence intervals vs. data average, *E. coli* example