An Exploration of Aspects of Bayesian Multiple Testing *

James G. Scott and James O. Berger
University of Texas and Duke University

May 25, 2003

Abstract

There has been increased interest of late in the Bayesian approach to multiple testing (often called the multiple comparisons problem), motivated by the need to analyze DNA microarray data in which it is desired to learn which of potentially several thousand genes are activated by a particular stimulus. We study the issue of prior specification for such multiple tests; computation of key posterior quantities; and useful ways to display these quantities. A decision-theoretic approach is also considered.

Key words and phrases: Multiple hypothesis tests; Multiple comparisons; Hierarchical models; Hyperpriors; Importance sampling; Posterior inclusion probabilities; Decision theory.

1 Introduction

Suppose we observe $\boldsymbol{x} = (x_1, \ldots, x_M)$, where the x_i arise independently from normal densities, $N(x_i \mid \mu_i, \sigma^2)$, with variance σ^2 unknown, and that it is desired to determine

*This research was supported by the U.S. National Science Foundation, under a Research Experience for Undergraduates supplement to grant DMS-0103265. Part of the work was done while the first author was visiting ISDS at Duke University. The interest of the second author in multiple decision problems originated in problems posed by Shanti Gupta, and it is wonderful to see his enormous influence acknowledged by this volume.

which of the means, μ_i, are nonzero. For example, the x_i's might be the measured overexpression or underexpression of M different genes from a DNA microarray, where x_i is modeled as a true mean μ_i plus unknown measurement error having variance σ^2. (Often, some transformation of the data is first performed to make the normality assumption more reasonable.) It is then desired to classify the genes as either active or inactive ($\mu_i = 0$ or $\mu_i \neq 0$), and to estimate the magnitudes of their effects. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$.

The model specification is completed by defining a model index parameter $\boldsymbol{\gamma}$ to be an M-dimensional vector of 0's and 1's such that

$$\gamma_i = \begin{cases} 0 & \text{if } \mu_i = 0 \\ 1 & \text{if } \mu_i \neq 0\,. \end{cases}$$

Then the full likelihood can be written

$$f(\boldsymbol{x} \mid \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\mu}) = \prod_{j=1}^{M} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_j - \gamma_j \mu_j)^2}{2\sigma^2} \right) \right]. \tag{1}$$

We investigate the Bayesian approach to this problem. One of the attractions of the Bayesian approach is that there is no need to introduce a penalty term for performing thousands of simultaneous tests; Bayesian testing has a built-in penalty or 'Ockham's razor effect' (cf. Berger and Jefferys, 1992).

A natural Bayesian approach is to assume that there is a common prior probability p that each $\mu_i = 0$, and that the μ_i are zero or not independently. Note that we expect p to be large in the microarray context (i.e., most genes are expected to be inactive). Westfall, Johnson, and Utts (1997) emphasize that, in general, choice of the prior probabilities of hypotheses in multiple comparisons has a major effect on the posterior probabilities; our emphasis here is on letting the data themselves choose p.

The nonzero μ_i are typically modeled as arising from a common normal density $N(\mu_i \mid 0, V)$ with mean 0 and unknown variance V. The 0 mean could be replaced by a nonzero unknown mean, with only a slight complication of the analysis. (In the microarray context, the mean zero assumption is basically the statement that over and under expression are equally likely; data transformation may well be necessary for this to be reasonable.) Thus we have specified a random effects model or Bayesian hierarchical model for the μ_i. Waller and Duncan (1969) first introduced this in a multiple comparisons setting; see Hobert

(2000) for a modern introduction and many other references to hierarhical modeling.

The Bayesian specification will be completed by choosing a prior distribution $\pi(p)$ and (independently in our case) a joint prior distribution $\pi(V, \sigma^2)$. We devote considerable attention to issues involved in the choice of these prior distributions.

Of primary inferential interest will be

1. $p_i = \Pr(\gamma_i = 0 \mid \boldsymbol{x})$, the posterior probability that $\mu_i = 0$ given the data (the posterior probability that the gene is inactive in the microarray example); $1 - p_i$ is often called the *posterior inclusion probability* (cf. Barbieri and Berger, 2003).

2. $\pi(\mu_i \mid \gamma_i = 1, \boldsymbol{x})$, the marginal posterior density of μ_i, given that it is non-zero (indicating the amount of over or under expression of the gene in the microarray context);

3. $\pi(V, \sigma^2, p \mid \boldsymbol{x})$, the joint posterior distribution for the key hyperparameters.

It will be seen that posterior computations involving these quantities are straightforward using an importance sampling scheme, even if M is in the thousands.

Among the other issues that are addressed in the paper are:

- How can posterior information best be summarized?

- How sensitive are inferences to the choices of the priors $\pi(p)$ and $\pi(V, \sigma^2)$?

- Can a decision-theoretic perspective, based on an unbalanced loss function that reflects the differential costs of misclassification of the μ_i, lead to a useful classification procedure?

Berry and Hochberg (1999) give a general review and insightful discussion of Bayesian multiple comparison approaches, and places them in context of non-Bayesian approaches. Applications to microarray experiments, of multiple comparison approaches similar to that studied in this paper, include Efron, Tibshirani, Storey, and Tusher (2001) and West (2003).

In Section 2, we discuss the choice of prior distributions on model parameters, and give expressions for the key posterior quantities mentioned above. We also prove propriety of the posterior for the recommended priors. Section 3 presents an importance sampling algorithm for computing posterior quantities of interest. Section 4 discusses useful posterior

summaries, and presents illustrations. Section 5 presents a decision-theoretic procedure for deciding whether a given observation is a signal or noise, and evaluates the performance of the procedure on simulated data. Section 6 presents some conclusions.

2 The Prior and Posterior Distributions

2.1 Priors on V and σ^2

In the absence of strong prior information about V and σ^2, we suggest use of

$$\pi(V, \sigma^2) = (V + \sigma^2)^{-2}\,. \tag{2}$$

The motivation for this choice follows from writing

$$\pi(V, \sigma^2) = \pi(V \mid \sigma^2) \cdot \pi(\sigma^2) \equiv \frac{1}{\sigma^2}\left(1 + \frac{V}{\sigma^2}\right)^{-2} \cdot \frac{1}{\sigma^2}.$$

The prior $\pi(V \mid \sigma^2)$ is a proper prior for V, given σ^2, which is needed since V is a hyperparameter that does not occur in all models under consideration – in particular, the model in which all of the μ_i are zero. (For many references to this issue and other general problems encountered in Bayesian model selection, see the review papers Berger and Pericchi, 2001, and Chipman, George, and McCulloch, 2001). Note that $\pi(V \mid \sigma^2)$ is scaled by σ^2 (which is commonly done for a hypervariance), and its form is an obvious modification of the usual objective prior for a hypervariance, $(\sigma^2 + V)^{-1}$ (which cannot be used here because it is improper). The mild quadratic decrease in $\pi(V \mid \sigma^2)$ ensures that the prior is not overly influential in the analysis.

Use of $\pi(\sigma^2) = 1/\sigma^2$ would seem to be problematical because it is also improper, but this parameter is common to all models under consideration. Furthermore, if one were to integrate out all parameters in each model, except σ^2, it can be shown that the resulting models would be scale invariant models, with scale parameter σ^2. Hence it follows from Berger, Pericchi and Varshavsky (1998) that use of $\pi(\sigma^2) = 1/\sigma^2$ is justified as a 'predictive matching' prior.

For discussion of elicitation of subjective priors for normal hyperparameters in multiple comparisons, see DuMouchel (1988). Gopalan and Berry (1998) consider relaxation of the

normality assumption through use of Dirichlet process priors.

2.2 The prior on p

An obvious objective choice of this prior is $\pi(p) = 1$. Typically, however, one does have strong prior information about p and, often, this information is that p is large (i.e., that most of the μ_i are zero). A convenient functional form that represents this type of information is

$$\pi(p) = (\alpha + 1)p^{\alpha}\,, \tag{3}$$

where α is an adjustable parameter allowing one to control how much of $\pi(p)$'s mass is concentrated near 1. Note the case of $\alpha = 0$, which defaults back to the uniform prior.

One simple way to specify α would be to make a 'best guess' $\hat{p}$, for p, and interpret this guess as the prior median. Solving for α leads to the choice

$$\alpha = \frac{\log(.5)}{\log(\hat{p})} - 1\,.$$

As long as one doesn't choose $\hat{p}$ to be extremely close to 1, the resulting prior is not overly concentrated. For example, suppose the best guess for p is .9, leading to $\alpha = 5.58$. The resulting prior has first decile of .7, first quartile of .81, and third quartile of .96, reflecting a reasonable amount of variation .

2.3 The posterior distribution

Under the above modeling assumptions, the posterior density of $\Theta = (p, V, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\mu})$ is

$$\pi(\Theta \mid \boldsymbol{x}) = C_1^{-1} \cdot f(\boldsymbol{x} \mid \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\mu}) \cdot \left[\prod_{j=1}^{M} N(\mu_i \mid 0, V)\right] \cdot \pi(\gamma \mid p) \cdot \pi(V, \sigma^2) \cdot \pi(p)\,, \tag{4}$$

where $\pi(\boldsymbol{\gamma} \mid p) = \prod_{i=1}^{M} p^{1-\gamma_i}(1-p)^{\gamma_i}$ and C_1 is the normalization constant.

Lemma 1 *The posterior distribution in (4) is proper (i.e., C_1 is finite).*

Proof. From the usual formula for convolution of two normal distributions, it is immediate that

$$\sum_{\gamma_j=1}^{2} \int f(x_j \mid \mu_j, \sigma^2, \gamma_j) N(\mu_j \mid 0, V) p^{1-\gamma_j} (1-p)^{\gamma_j} \, d\mu_j$$
$$= \frac{p}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x_j^2}{2\sigma^2}\right) + \frac{(1-p)}{\sqrt{2\pi(\sigma^2+V)}} \exp\left(\frac{-x_j^2}{2(\sigma^2+V)}\right). \tag{5}$$

It follows, from (1) and (4), that C_1 can be written

$$C_1 = \int_0^1 \int_0^\infty \int_0^\infty \prod_{j=1}^{M} \left[\frac{p}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x_j^2}{2\sigma^2}\right) + \frac{(1-p)}{\sqrt{2\pi(\sigma^2+V)}} \exp\left(\frac{-x_j^2}{2(\sigma^2+V)}\right) \right] \times \pi(V, \sigma^2)\pi(p)\, dV\, d\sigma^2 dp.$$

Making the change of variables $u = (\sigma^2 + V)^{-1}$ in the inner integral, and noting that $\mathrm{d}u = (\sigma^2+V)^{-2}\mathrm{d}V$, this simplifies to

$$C_1 = \frac{1}{(2\pi)^{M/2}} \int_0^1 \int_0^\infty \int_0^{\sigma^{-2}} \prod_{j=1}^{M} \left[\frac{p}{\sigma} \exp\left(-\frac{x_j^2}{2\sigma^2}\right) + (1-p)\sqrt{u} \exp\left(-\frac{u x_j^2}{2}\right) \right] \pi(p)\, du\, d\sigma^2\, dp. \tag{6}$$

It is easy to show, for all (p, σ, u), that

$$\frac{p}{\sigma} \exp\left(-\frac{x_j^2}{2\sigma^2}\right) + (1-p)\sqrt{u} \exp\left(-\frac{u x_j^2}{2}\right) < c\,,$$

for some constant c (that will depend on the x_j). Hence we can bound the product in the integrand in (6) by

$$\frac{c^{M-1}p}{\sigma} \exp\left(-\frac{x_1^2}{2\sigma^2}\right) + (1-p)\sqrt{u} \exp\left(-\frac{u x_1^2}{2}\right) \prod_{j=2}^{M} \left[\frac{p}{\sigma} \exp\left(-\frac{x_j^2}{2\sigma^2}\right) + (1-p)\sqrt{u} \exp\left(-\frac{u x_j^2}{2}\right) \right].$$

Continuing in the same way, with subsequent terms in this last product, leads to the bound

$$\sum_{j=1}^{M} \frac{c^{M-1}p}{\sigma} \exp\left(-\frac{x_j^2}{2\sigma^2}\right) + [(1-p)\sqrt{u}]^M \exp\left(-\frac{u|\,\boldsymbol{x}\,|^2}{2}\right), \tag{7}$$

for the product in the integrand in (6).

Inserting the first expression from (7) into (6), in place of the product, and performing an easy computation shows that the resulting integral is finite. Hence we only need to establish the finiteness of

$$\int_0^1 \int_0^\infty \int_0^{\sigma^{-2}} [(1-p)\sqrt{u}]^M \exp\left(-\frac{u|\,\boldsymbol{x}\,|^2}{2}\right) \pi(p)\, du\, d\sigma^2\, dp\,.$$

Changing orders of integration allows explicit integration over σ^2, resulting in the equivalent expression

$$\int_0^1 (1-p)^M \pi(p)\, dp \int_0^\infty u^{(M-2)/2} \exp\left(-\frac{u|\,\boldsymbol{x}\,|^2}{2}\right) du\,,$$

which is clearly finite, completing the proof. □

Lemma 2 *The marginal posterior distribution of* (V, σ^2, p) *is given by*

$$\pi(V, \sigma^2, p \mid \boldsymbol{x}) = C_1^{-1} \prod_{j=1}^{M} \left[\frac{p}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x_j^2}{2\sigma^2}\right) + \frac{(1-p)}{\sqrt{2\pi(\sigma^2+V)}} \exp\left(\frac{-x_j^2}{2(\sigma^2+V)}\right)\right] \times \pi(V, \sigma^2)\pi(p). \tag{8}$$

Proof. This follows directly from (4) and (5). □

Lemma 3

$$\begin{aligned} p_i &\equiv Pr(\gamma_i = 0 \mid \boldsymbol{x}) = Pr(\mu_i = 0 \mid \boldsymbol{x}) \qquad (9)\\ &= \int_0^1 \int_0^\infty \int_0^\infty \left[1 + \frac{(1-p)}{p} \sqrt{\frac{\sigma^2}{\sigma^2+V}} \exp\left(\frac{x_i^2 V}{2\sigma^2(\sigma^2+V)}\right)\right]^{-1} \pi(V, \sigma^2, p \mid \boldsymbol{x})\, dV\, d\sigma^2\, dp\,. \end{aligned}$$

Proof. Clearly

$$p_i = \int_0^1 \int_0^\infty \int_0^\infty Pr(\mu_i = 0 \mid V, \sigma^2, p, \boldsymbol{x})\, \pi(V, \sigma^2, p \mid \boldsymbol{x})\, dV\, d\sigma^2\, dp\,,$$

and $Pr(\mu_i = 0 \mid V, \sigma^2, p, \boldsymbol{x})$ is simply the posterior probability of the hypothesis that a normal mean is zero and is given in, e.g., Berger (1985). □

Lemma 4

$$\pi(\mu_i \mid \gamma_i = 1, \boldsymbol{x}) = \int_0^1 \int_0^\infty \int_0^\infty \left(2\pi\tau^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\tau^2}(\mu_i - \rho)^2\right) \pi(V, \sigma^2, p \mid \boldsymbol{x})\, dV\, d\sigma^2\, dp\,, \tag{10}$$

where $\rho = Vx_i/(\sigma^2 + V)$ *and* $\tau^2 = V\sigma^2/(\sigma^2 + V)$.

Proof. Clearly

$$\pi(\mu_i \mid \gamma_i = 1, \boldsymbol{x}) = \int_0^1 \int_0^\infty \int_0^\infty \pi(\mu_i \mid \gamma_i = 1, V, \sigma^2, p, \boldsymbol{x})\, \pi(V, \sigma^2, p \mid \boldsymbol{x})\, dV\, d\sigma^2\, dp\,,$$

and $\pi(\mu_i \mid \gamma_i = 1, V, \sigma^2, p, \boldsymbol{x})$ is simply the posterior density of a normal mean in a conjugate prior situation, as given in, e.g., Berger (1985). □

3 Computational Implementation via Importance Sampling

From (8), it is apparent that posterior computations could be approached via mixture model analysis; see Robert (1996) and Celeux, Hurn, and Robert (2000) for discussion and other references to computation in such situations. The large magnitude of M, in typical motivating microarray examples, presents a challenge within this approach.

Luckily, the computational problem here can be attacked very simply, because of the

fact that most posterior quantities of interest can be expressed as expectations of the form

$$\int_0^1 \int_0^\infty \int_0^\infty h(V, \sigma^2, p)\pi(V, \sigma^2, p \mid \boldsymbol{x})\, dV\, d\sigma^2\, dp. \tag{11}$$

An example is p_i in (9) (with $h(V, \sigma^2, p)$ being the bracketed expression), which one wants to compute for $i = 1, \ldots, M$. When M is large, it is most efficient to compute such posterior expectations using importance sampling, because one can use the same multivariate importance sample for all computations.

To implement importance sampling (see Berger, 1985, and Robert and Casella, 1999, for introductions to importance sampling), it is convenient to first eliminate domain restrictions, by transforming to the parameters

$$\xi = \log(V), \quad \eta = \log(\sigma^2), \quad \lambda = \log\left(p/(1-p)\right),$$

so that the integral in (11) becomes, after changing variables,

$$\int_{-\infty}^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty h(e^\xi, e^\eta, (1+e^{-\lambda})^{-1})\pi^*(\xi, \eta, \lambda)\, d\xi\, d\eta\, d\lambda\,, \tag{12}$$

where $\pi^*(\xi, \eta, \lambda)$ is the transformation of $\pi(V, \sigma^2, p \mid \boldsymbol{x})$ given by

$$\pi^*(\xi, \eta, \lambda) = \pi(e^\xi, e^\eta, (1+e^{-\lambda})^{-1} \mid \boldsymbol{x})\, e^{(\xi+\eta+\lambda)}\, (1+e^\lambda)^{-2}\,.$$

Next find the posterior mode, $(\hat{\xi}, \hat{\eta}, \hat{\lambda})$, of $\pi^*(\xi, \eta, \lambda)$ and compute H, the Hessian matrix of partial derivatives at this mode. Then, as an importance function, use the multivariate Student importance density $t_3(\xi, \eta, \lambda \mid (\hat{\xi}, \hat{\eta}, \hat{\lambda}), aH^{-1})$, having mean $(\hat{\xi}, \hat{\eta}, \hat{\lambda})$, covariance matrix aH^{-1}, and 3 degrees of freedom. Values of $a > 1$ should be tried until the importance function adequately covers the major part of the support of $\pi^*(\xi, \eta, \lambda \mid \boldsymbol{x})$, and yet is not so disperse that the efficiency of the sampling algorithm is compromised. We found that choosing a in the neighborhood of 5 was typically successful, although it is always a good idea to plot the proposal distribution on the same scale as the posterior to make sure the tails are adequately sampled.

Finally, one draws a sample $(\xi_i, \eta_i, \lambda_i)$, $i = 1, \ldots, m$, from $t_3(\xi, \eta, \lambda \mid (\hat{\xi}, \hat{\eta}, \hat{\lambda}), aH^{-1})$;

defines the importance sampling weights $w_i = \pi^*(\xi_i, \eta_i, \lambda_i)/t_3(\xi_i, \eta_i, \lambda_i \mid (\hat{\xi}, \hat{\eta}, \hat{\lambda}), aH^{-1})$; and approximates the integral in (11) as

$$\frac{\sum_{i=1}^m h(e^{\xi_i}, e^{\eta_i}, (1+e^{-\lambda_i})^{-1})\, w_i}{\sum_{i=1}^m w_i}\,.$$

4 Posterior Summaries and Illustrations

We illustrate the above ideas on a simulated example based on generation of ten "signal" points from a $N(0, V)$ distribution, with V usually equal to 9 or 16, together with simulated n "noise" points from a $N(0, 1)$ distribution, for various values of n. A variety of different priors on p are also considered, to study sensitivity to this prior choice.

4.1 Posterior inclusion probabilities, $1 - p_i$

Table 1 illustrates how the Bayesian procedure automatically introduces a penalty that handles multiple testing. For the most part, the posterior inclusion probabilities, $1 - p_i$, decrease as the number, n, of "noise" observations grows, so that the same observation is viewed as implying less evidence of a nonzero mean when more tests are simultaneously considered. This happens because, as n grows, the posterior distribution of p concentrates closer to 1, which, as is clear from (9), will result in the p_i moving closer to 1. For the most extreme observations, this ordering does not hold for small and moderate n, because of the fact that V, σ^2, and p are then somewhat confounded, so that the procedure cannot easily separate signal from noise.

To check the sensitivity of the inferences to the prior on p, we considered both the 'objective' uniform prior and a prior weighted towards large values of p, namely $\pi(p) = 11p^{10}$, which has a median of about .93 (thus specifying that we think that a little over 90% of the observations are noise). As Table 1 shows, the prior on p does have a strong effect on the p_i's. While this effect tends to fade as n increases, it becomes negligible only for extremely large n. This reinforces the importance of using subjective information about p.

	10 Signal Observations									
n	−5.65	−5.56	−2.62	−1.20	−1.01	−0.90	−0.15	1.65	1.94	3.57
25	.97	.97	.71	.31	.28	.26	.20	.43	.51	.88
100	.99	.99	.47	.21	.20	.19	.16	.26	.31	.75
500	1	1	.34	.07	.06	.06	.04	.11	.15	.79
5000	1	1	.11	.02	.02	.02	.01	.03	.04	.42

Table 1: *Above:* The posterior probabilities of the 10 signal means being nonzero as n, the number of noise observations, increases. Here, the signal $\mu_i \sim N(0, 9)$, and the signal observations $x_i \sim N(\mu_i, 1)$. The prior on p is uniform.
Below: $1 - p_i$ for the same 10 signal observations with $\pi(p) = 11p^{10}$.

	10 Signal Observations									
n	−5.65	−5.56	−2.62	−1.20	−1.01	−0.90	−0.15	1.65	1.94	3.57
25	.91	.90	.37	.10	.08	.08	.06	.15	.20	.65
100	.98	.98	.23	.06	.05	.05	.04	.08	.11	.58
500	1	1	.26	.04	.04	.03	.02	.07	.10	.74
5000	1	1	.08	.01	.01	.01	.01	.02	.03	.36

4.2 Marginal posterior mean summaries

An informative way of summarizing the posterior information about the μ_i is given in Figure 1 (essentially following Mitchell and Beauchamp, 1988). For each mean,

- place a bar at zero, the height of which gives p_i, the posterior probability that the corresponding mean is zero;
- plot the density $(1 - p_i)\pi(\mu_i \mid \gamma_i = 1, \boldsymbol{x})$, which indicates where μ_i is likely to be if it is nonzero.

Note that the $\pi(\mu_i \mid \gamma_i = 1, \boldsymbol{x})$ can be computed on a grid, using (10) and the importance sampling scheme discussed in Section 3.

Both components of this posterior distribution are important. Consider, for instance, observation −2.62 in Figure 1. If one considered only the density portion, one would note that virtually all of the curve is concentrated at values less than 0, which might lead one to conclude that the observation is a signal; in fact, $p_i = .45$, so that there is a very significant chance that the mean is zero. On the other hand, reporting only $p_i = .45$ might lead one to reject the mean as uninteresting, even though the values of the density portion

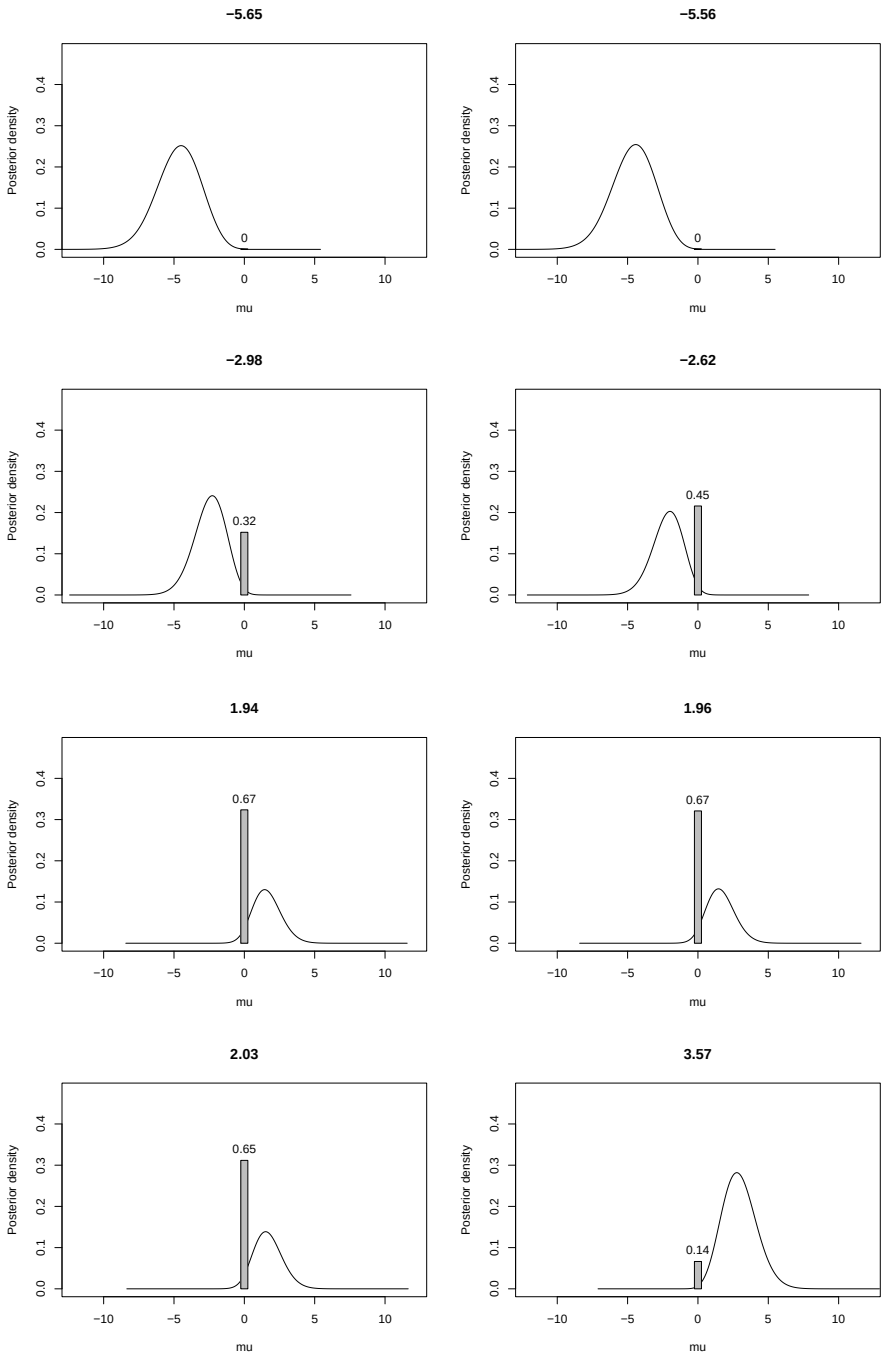

Figure 1: The marginal posterior distributions of some selected μ_i's for the case with 10 data points and 100 noise points. The bar at zero gives p_i, the posterior probability that the mean is zero, while the plotted density is $(1 - p_i)\pi(\mu_i \mid \gamma_i = 1, \boldsymbol{x})$.

are interestingly negative. One ends up concluding that it is far from clear whether -2.62 is a signal or not but, if it *is* a signal, then it has a potentially interesting mean.

4.3 Marginal posteriors for hyperparameters

Marginal posterior distributions of p, V, and σ^2 are significantly affected by the prior on p and the number of noise observations n. Figures 2 through 7 present histograms of the posteriors for various cases: the first three when there are 10 signal observations from a $N(0, 9)$ distribution and 100 noise observations, and the second three when there are 25 signal observations from a $N(0, 16)$ distribution and 500 noise observations. These histograms were also computed by importance sampling, based on choosing $h(V, \sigma^2, p)$ in (11) to be indicator functions over the ranges of each bar in the histograms in Figures 2 through 7 (with the same importance sample used for each bar in a figure).

In the second situation, one would expect considerably more precise marginal posterior distributions, both because the numbers of signal and noise observations are larger and because the signals are stronger (larger V). That is indeed the case, although the marginal posteriors of V remain rather surprisingly disperse. In the first situation, all marginal posteriors are disperse. This is an important indicator of the difficulty of accurate inference in this type of mixture situation.

The figures again show a significant effect of the choice of the prior distribution for p. The effect of $\pi(p)$ on the posterior for p is straightforward, but its effect on the marginal posteriors for V and σ^2 are not so obvious. Indeed, in the case of 10 signal points, using the more informative prior for p actually seems to *increase* uncertainty about V, as evidenced by the thicker right tail of the marginal posterior for V.

5 A Decision-Theoretic Approach

Suppose we have computed the p_i's and are now interested in deciding which of the observations are signals and which are noise. (In the microarray example, this would involve deciding which genes are active and which are not.) One could simply choose a cutoff P and say that all observations x_i with $p_i \leq P$ are signals, but it is not at all clear how to choose P.

A natural decision-theoretic approach to this problem, introduced by Duncan (1965)

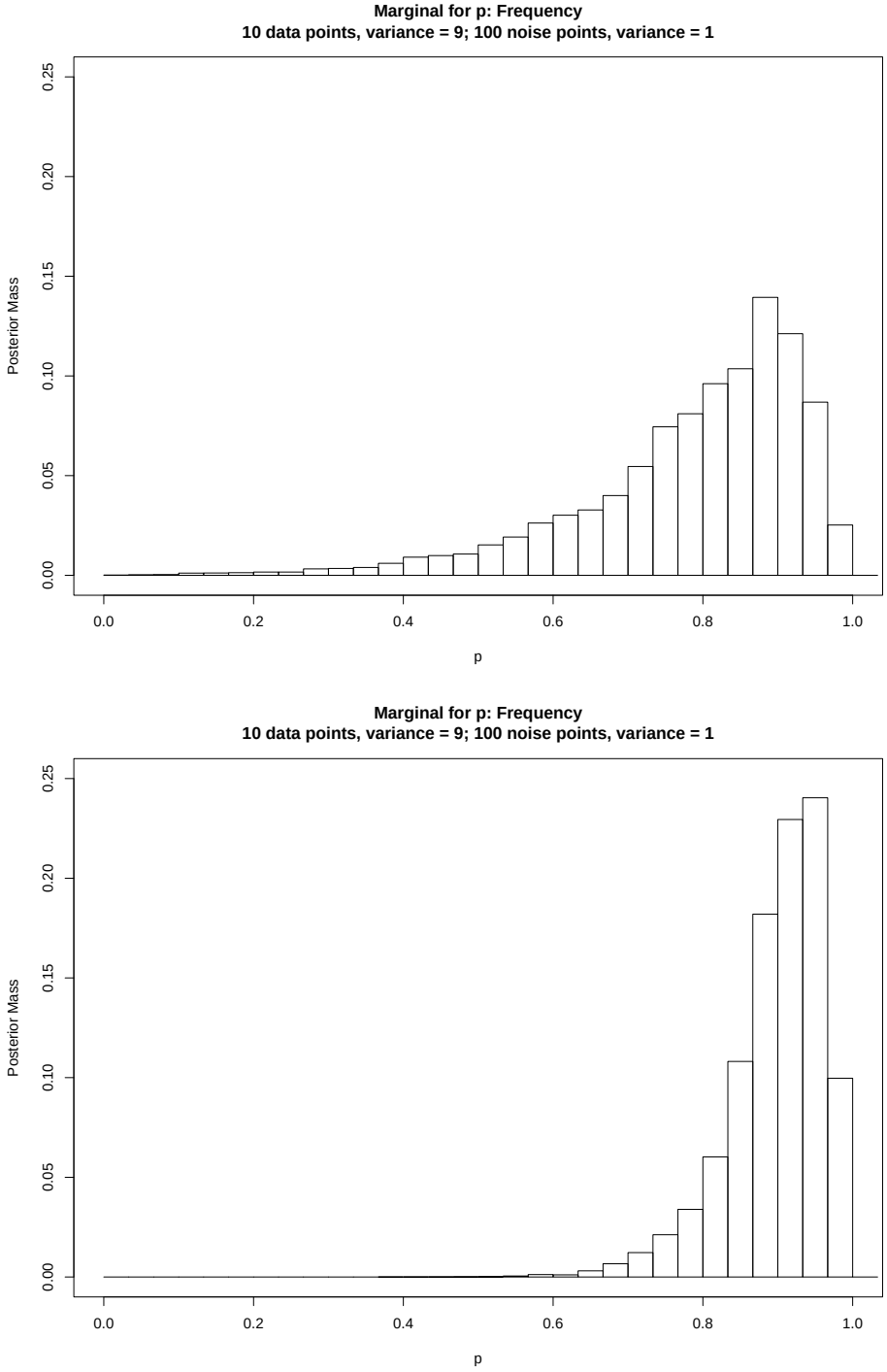

Figure 2: Marginal posteriors of p for the case with 10 data points and 100 noise points. Top: $\pi(p) = 1$. Bottom: $\pi(p) = 11p^{10}$.

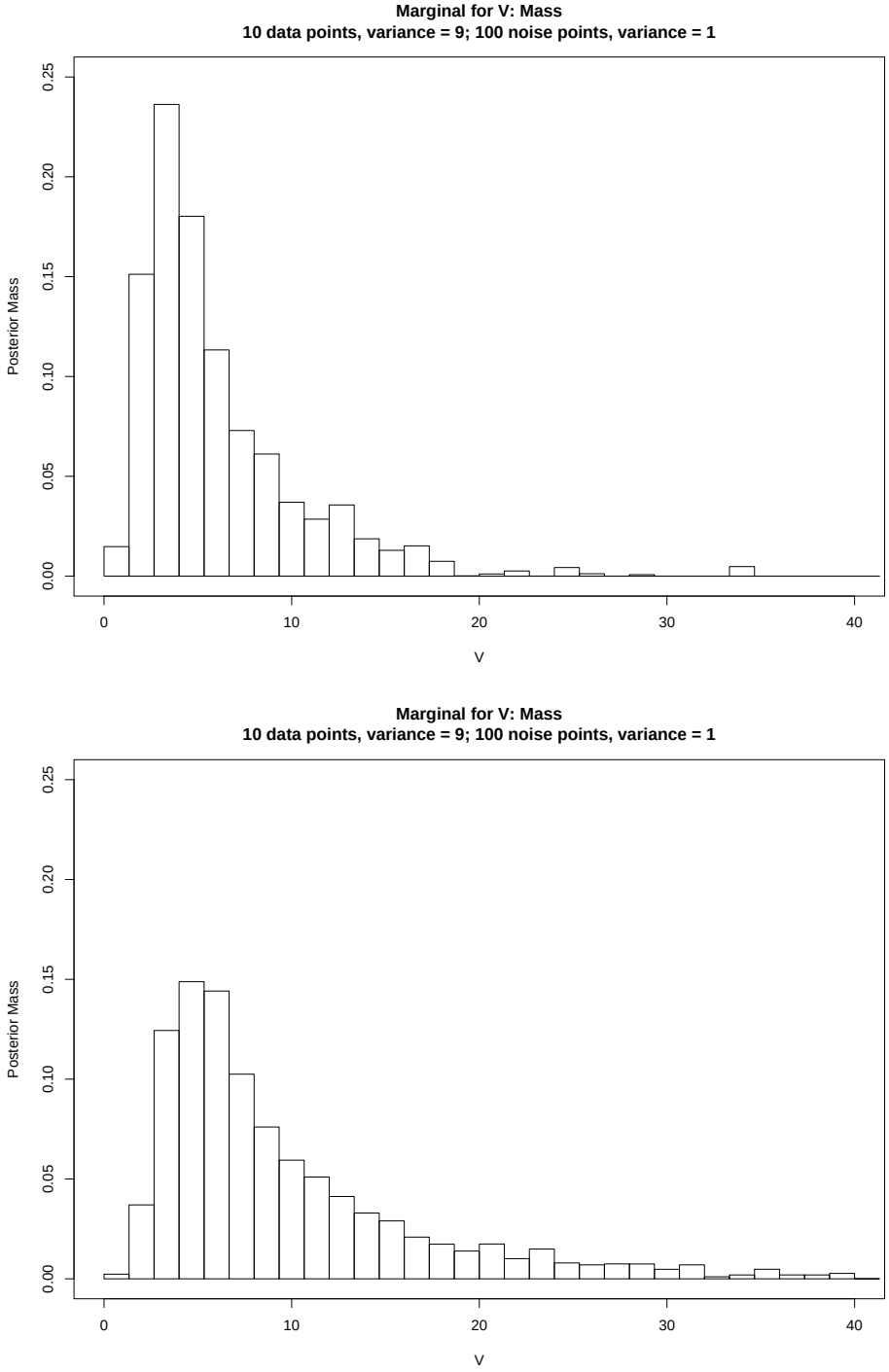

Figure 3: Marginal posteriors of V for the case with 10 data points and 100 noise points. Top: $\pi(p) = 1$. Bottom: $\pi(p) = 11p^{10}$.

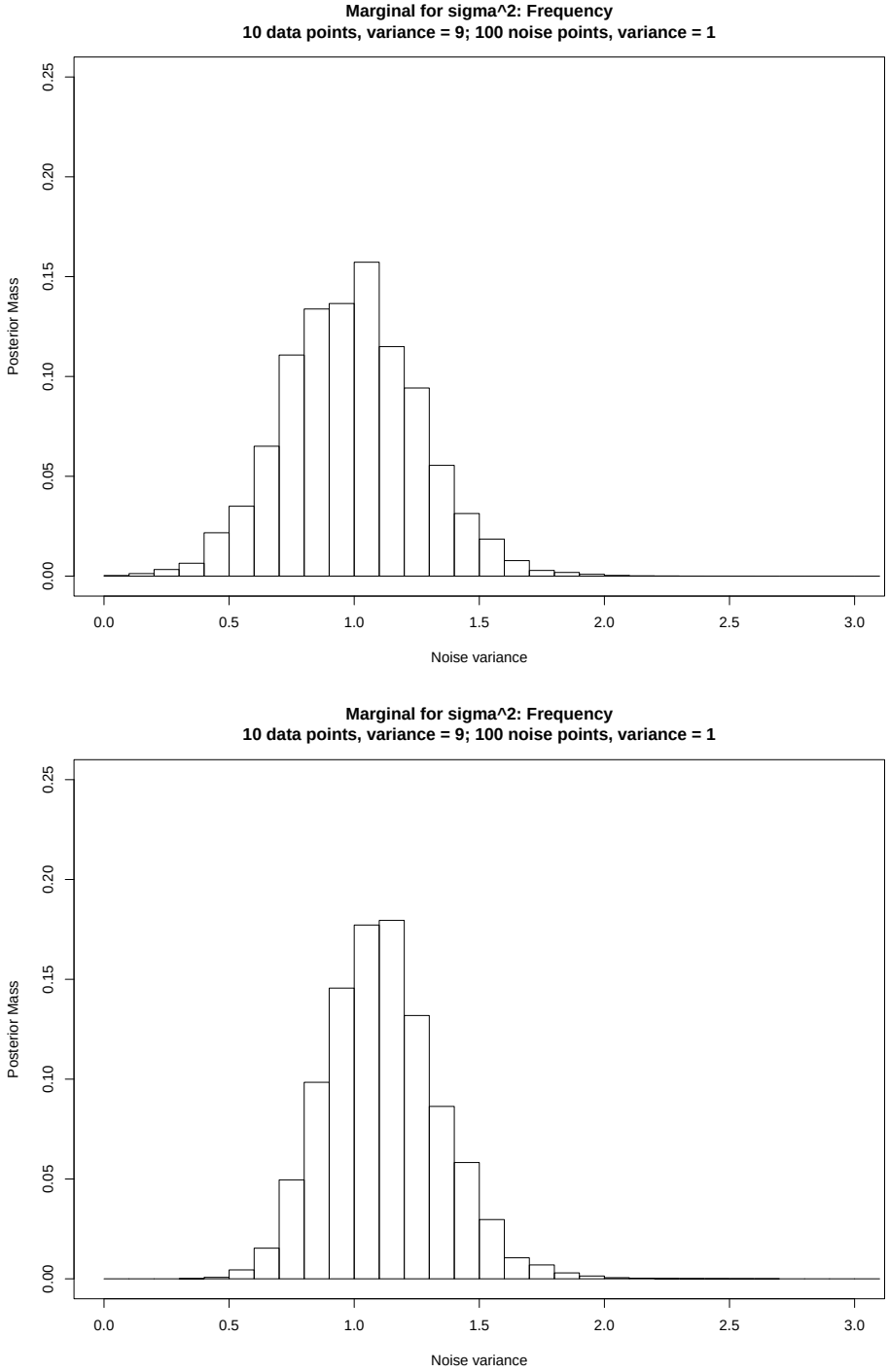

Figure 4: Marginal posteriors of σ^2 for the case with 10 data points and 100 noise points. Top: $\pi(p) = 1$. Bottom: $\pi(p) = 11p^{10}$.

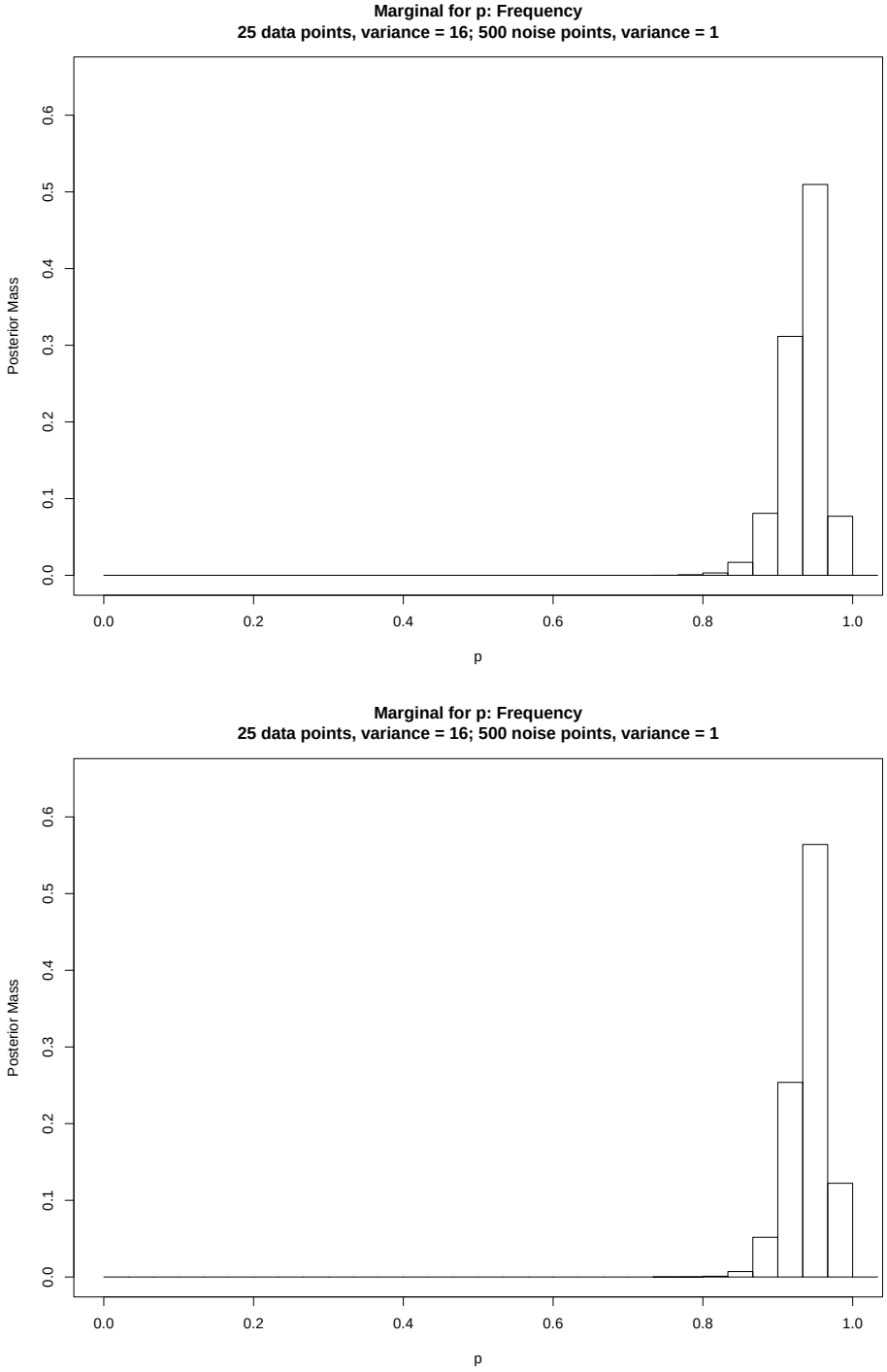

Figure 5: Marginal posteriors of p for the case with 25 data points and 500 noise points. Top: $\pi(p) = 1$. Bottom: $\pi(p) = 11p^{10}$.

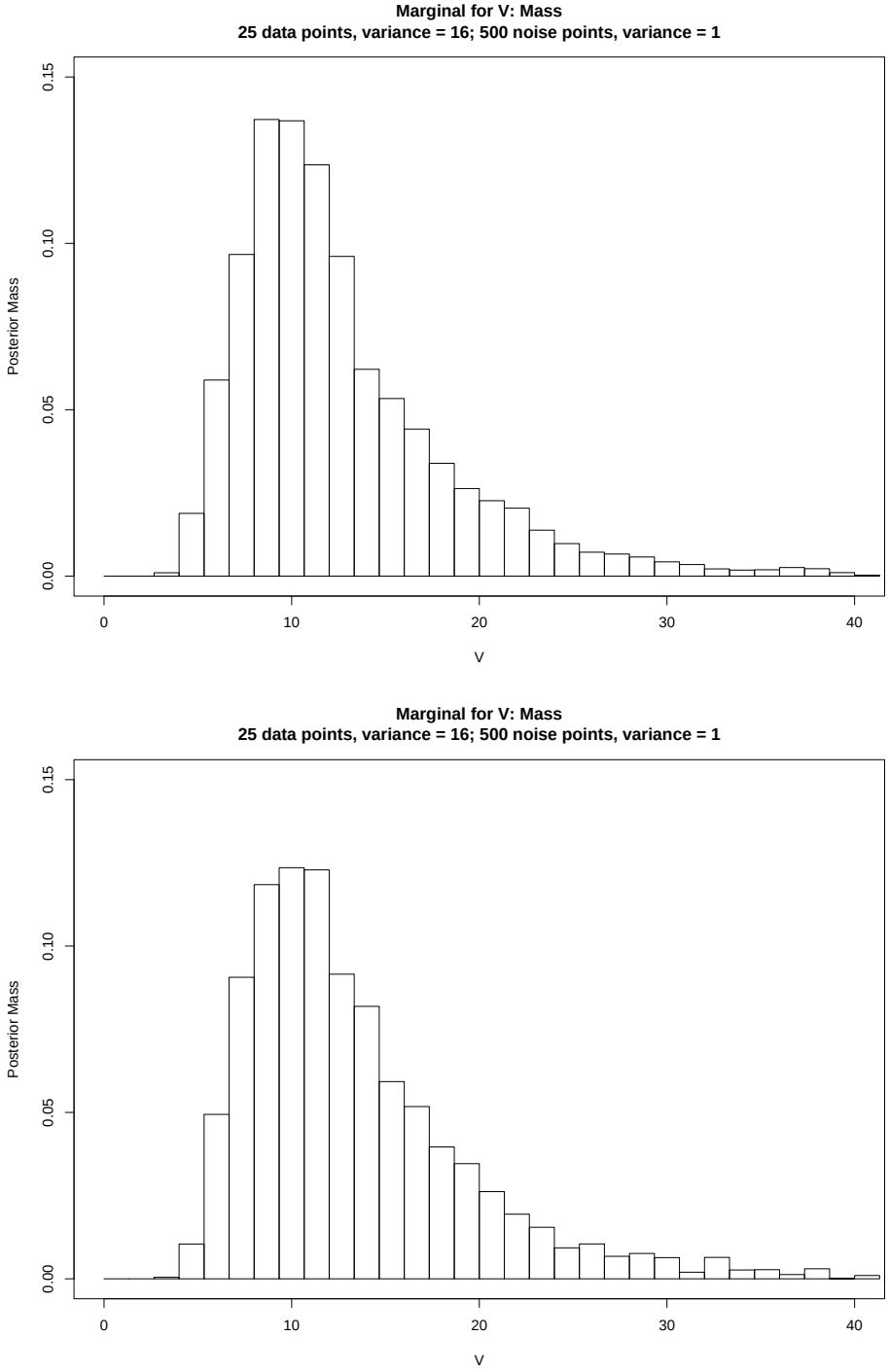

Figure 6: Marginal posteriors of V for the case with 25 data points and 500 noise points. Top: $\pi(p) = 1$. Bottom: $\pi(p) = 11p^{10}$.

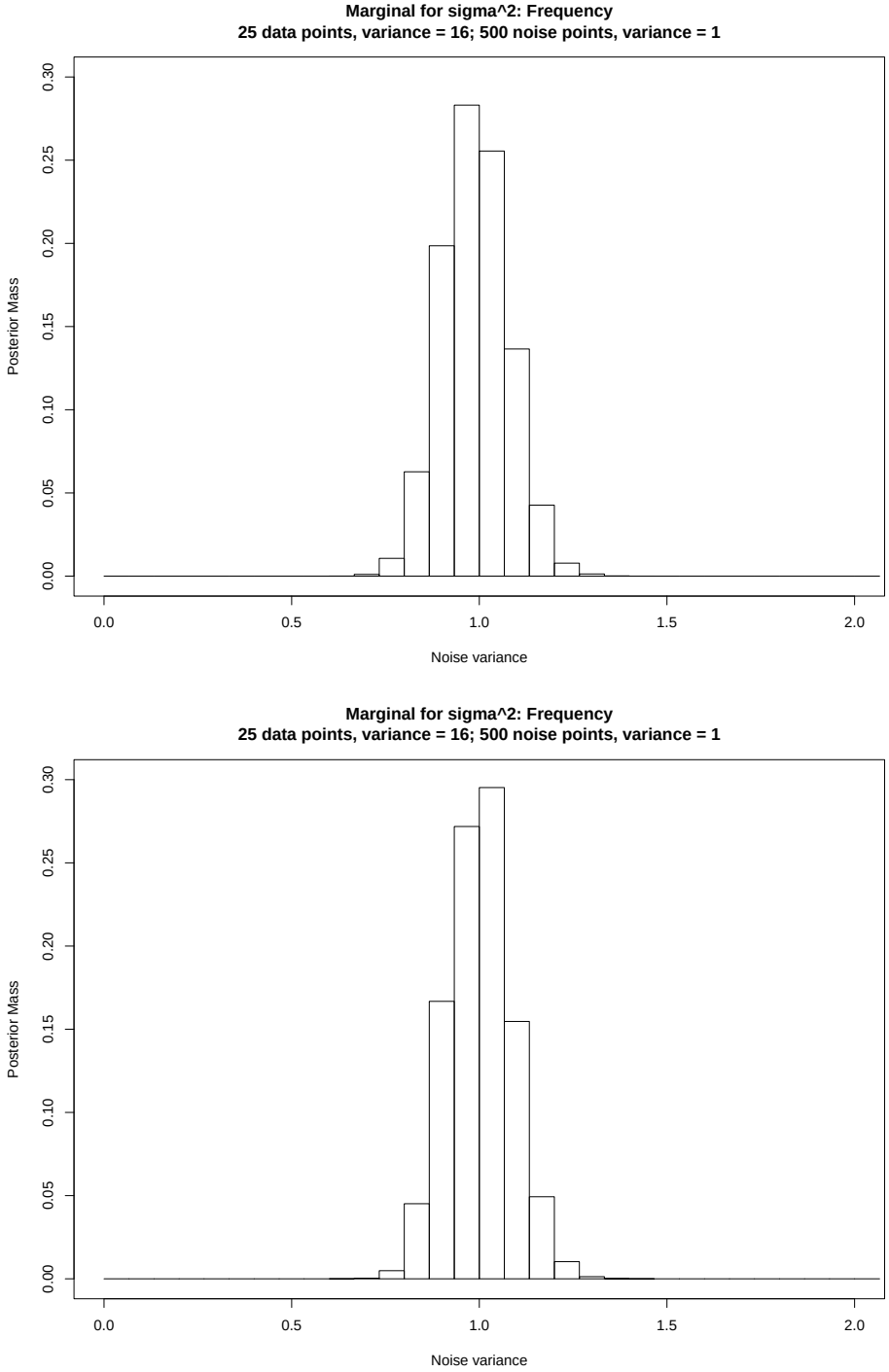

Figure 7: Marginal posteriors of σ^2 for the case with 25 data points and 500 noise points. Top: $\pi(p) = 1$. Bottom: $\pi(p) = 11p^{10}$.

and further discussed in Waller and Duncan (1969), is to separately specify the cost of a false positive (declaring μ_i to be a signal when it is zero), and the cost of missing a true signal. For instance, in the microarray example, the loss incurred by a false positive might be the fixed cost of doing a targeted experiment to verify that the gene is active (the cost being wasted when the gene is found to be inactive). On the other hand, the loss in incorrectly classifying an active gene as inactive might well be proportional to the distance of the corresponding mean from zero; thus a greater loss is incurred by failing to detect a gene whose mean level of expression is 5 than by failing to detect one whose expression level is 2.

Let S_i denote the action "Call μ_i a signal", and let N_i denote the action "Call μ_i noise." Then the loss function discussed above would be written

$$L(S_i, \mu_i) = \begin{cases} 1 & \text{if } \mu_i = 0 \\ 0 & \text{if } \mu_i \neq 0\,, \end{cases} \tag{13}$$

$$L(N_i, \mu_i) = \begin{cases} 0 & \text{if } \mu_i = 0 \\ c|\mu_i| & \text{if } \mu_i \neq 0\,, \end{cases} \tag{14}$$

where c is an adjustable parameter reflecting the relative costs of each type of error.

The posterior expected losses of each action can then be computed as

$$E[L(S_i, \mu_i) \mid \boldsymbol{x}\,] = \int L(S_i, \mu_i)\pi(\mu_i \mid \boldsymbol{x})\, d\mu_i = p_i\,,$$
$$E[L(N_i, \mu_i) \mid \boldsymbol{x}\,] = c \cdot (1 - p_i) \int_{-\infty}^{\infty} |\mu_i| \cdot \pi(\mu_i \mid \gamma_i = 1, \boldsymbol{x})\, d\mu_i.$$

Thus the posterior expected loss is minimized if we take action S_i (calling μ_i a signal) whenever $E[L(S_i, \mu_i) \mid \boldsymbol{x}\,] < E[L(N_i, \mu_i) \mid \boldsymbol{x}\,]$, i.e., whenever

$$p_i < \frac{c \cdot \int_{-\infty}^{\infty} |\mu_i| \cdot \pi(\mu_i \mid \gamma_i = 1, \boldsymbol{x})\, d\mu_i}{1 + c \cdot \int_{-\infty}^{\infty} |\mu_i| \cdot \pi(\mu_i \mid \gamma_i = 1, \boldsymbol{x})\, d\mu_i}. \tag{15}$$

Interestingly, this loss function can thus be viewed as a vehicle for allowing determination of the cutoff P discussed earlier.

It is conceivable that this cutoff could be quite high. For example, a situation with many thousands of noise points would drive down the p_i's for all observations. Yet the

posterior expectation of $|\mu_i|$ for an extreme observation could be so large that the procedure might still take action S_i, even if the posterior odds *against* x_i being a signal are quite large.

To illustrate use of the procedure, and study sensitivity to the choice of c, we considered the choices $c = 1$ and $c = 3$, the latter placing a much greater relative emphasis on the importance of discovering false nulls. We considered a variety of combinations of choices of α in the prior for p in (3), choices of the numbers of signal and noise observations, and choices of V. For each combination considered, we generated the indicated number of $N(0, 1)$ noise observations; combined them with the indicated signal observations; applied the decision procedure; and recorded the number of signal observations that were correctly classified as signals, and the number of noise observations that were incorrectly classified as signals. These results are found in Tables 2 and 3. Note that each table entry represents only a single trial, so that the results are random, depending on the generated values of the noise observations; for larger n there is naturally less randomness.

Another common method of dealing with multiple testing is the False Discovery Rate (FDR) procedure, first introduced in Benjamini and Hochberg (1995). We do not attempt to explore the relationship of this procedure to Bayesian methodology (such explorations have been undertaken in Shaffer, 1999; Efron, Tibshirani, Storey, and Tusher, 2001; Storey, 2001 and 2002; and Genovese and Wasserman, 2002), but do record the FDR of the above decision procedure for each of the cases given in the tables.

As expected, the choice $c = 3$ generally resulted in more of the signals being detected. It also tended to produce more false positives, as would be expected. In judging the performance of the procedure, it is important to note that many of the signal observations actually had values rather close to zero (since they were generated from a $N(0, V)$ distribution), and correctly classifying such observations as signals, in a background sea of noise, is simply not possible.

6 Conclusion

The suggested Bayesian procedure can be called objective, when the uniform prior on p is used. It is of considerable interest that such a procedure exists, and that it can easily be implemented computationally, even when M is huge. We saw, however, that incorporation of subjective prior information about p can be very beneficial, and so recommend its use,

Case				Performance		
# signal	V	# noise	α	Non-zero μ_i's Discovered	False Positives	FDR
10	9	25	0	6 of 10	3	.33
10	9	25	5	3 of 10	0	0
10	9	25	10	4 of 10	1	.20
10	9	100	0	4 of 10	3	.43
10	9	100	5	3 of 10	0	0
10	9	100	10	3 of 10	0	0
10	9	500	0	4 of 10	7	.64
10	9	500	5	3 of 10	2	.40
10	9	500	10	3 of 10	3	.50
10	9	5000	0	3 of 10	4	.57
10	9	5000	5	3 of 10	1	.25
10	9	5000	10	3 of 10	1	.25

Table 2: Performance of the decision theoretic procedure with $c = 1$ above and $c = 3$ below. The signal points are (-5.65, -5.56, -2.62, -1.20, -1.01, -0.90, -0.15, 1.65, 1.94, 3.57).

Case				Performance		
# signal	V	# noise	α	Non-zero μ_i's Discovered	False Positives	FDR
10	9	25	0	4 of 10	0	0
10	9	25	5	4 of 10	0	0
10	9	25	10	4 of 10	0	0
10	9	100	0	6 of 10	15	.71
10	9	100	5	4 of 10	5	.56
10	9	100	10	4 of 10	5	.56
10	9	500	0	4 of 10	11	.73
10	9	500	5	3 of 10	1	.25
10	9	500	10	3 of 10	1	.25
10	9	5000	0	2 of 10	0	0
10	9	5000	5	3 of 10	4	.57
10	9	5000	10	3 of 10	5	.625

Case				Performance		
# signal	V	# noise	α	Non-zero μ_i's Discovered	False Positives	FDR
25	16	50	0	13 of 25	1	.07
25	16	50	5	14 of 25	1	.07
25	16	50	10	10 of 25	0	0
25	16	500	0	11 of 25	1	.08
25	16	500	5	11 of 25	1	.08
25	16	500	10	13 of 25	1	.07
25	16	5000	0	9 of 25	1	.10
25	16	5000	5	10 of 25	1	.09
25	16	5000	10	10 of 25	2	.17
25	16	10000	0	10 of 25	3	.23
25	16	10000	5	9 of 25	1	.10
25	16	10000	10	8 of 25	0	0

Table 3: Performance of the decision theoretic procedure with $c = 1$ above and $c = 3$ below. The signal points are: (-7.93, -7.04, -4.46, -4.28, -3.85, -2.01, -1.81, -1.66, -1.62, -1.28, -0.57, 0.32, 0.74, 1.11, 1.82, 1.94, 2.23, 2.7, 2.82, 3.06, 3.62,4.79, 5.48, 5.62, 8.12).

Case				Performance		
# signal	V	# noise	α	Non-zero μ_i's Discovered	False Positives	FDR
25	16	50	0	16 of 25	3	.16
25	16	50	5	16 of 25	2	.11
25	16	50	10	14 of 25	1	.07
25	16	500	0	14 of 25	14	.50
25	16	500	5	13 of 25	4	.24
25	16	500	10	11 of 25	2	.15
25	16	5000	0	10 of 25	4	.29
25	16	5000	5	10 of 25	9	.47
25	16	5000	10	10 of 25	4	.29
25	16	10000	0	10 of 25	7	.41
25	16	10000	5	10 of 25	4	.29
25	16	10000	10	10 of 25	10	.50

when available.

We did not consider use of subjective information about either V or σ^2, since it is less typically available. Incorporation of such information could also be beneficial, however, especially information about V. (When there are a large number of noise observations, the data can usually do a good job of determining σ^2.)

We recommend adoption of the displays in Figure 1 for presentation of the results for interesting means. They instantly convey both the probability that the the mean is zero, and the likely magnitude of the mean, if it is not zero.

Deciding which means to classify as signal and which to classify as noise is inherently a decision problem, and we presented one reasonable formulation of the problem, in which only a single quantity c need be specified. Of course, specification of this scaling constant is itself a non-trivial problem in utility elicitation.

The most striking feature of the results presented herein is that the answers seem highly sensitive to unknown features of the problem, and rather sensitive to prior specifications. This was indicated in the subjective context by Westfall, Johnson, and Utts, (1997); that it also occurs in an attempt at an objective analysis strongly reinforces the notion that there is no magic answer available for the multiple testing problem.

References

[1] Barbieri, M., and Berger, J. (2003). Optimal predictive model selection. *Ann. Statist.*, to appear.

[2] 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,* **B 57**, 289–300.

[3] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd edition), New York: Springer-Verlag.

[4] Berger, J. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison (with Discussion). In *Model Selection*, P. Lahiri, ed., Institute of Mathematical Statistics Lecture Notes – Monograph Series, volume 38, Beachwood Ohio, 135–207.

[5] Berger, J., Pericchi, L. and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya*, **60**, 307-321.

[6] Berry, D.A., and Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference,* **82**, 215–277.

[7] Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Am. Statist. Assoc.* **95**, 957–70.

[8] Chipman, H., George, E.I., and McCulloch, R.E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*, P. Lahiri (Ed.), Institute of Mathematical Statistics Lecture Notes – Monograph Series, volume 38, Beachwood Ohio, 66–134.

[9] DuMouchel, W. (1988). A Bayesian model and graphical elicitation procedure for multiple comparisons. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), *Bayesian Statistics 3*. Oxford University Press. Oxford, England.

[10] Duncan, D.B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7**, 171-222.

[11] Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association,* **96**, 1151–1160.

[12] Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society,* **B 64**, 499–517.

[13] Gopalan, R. and Berry, D.A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *J. Amer. Statist. Assoc.* **93**, 1130-1139.

[14] Hobert, J. (2000). Hiearchical models: a current computational perspective. *J. Amer. Statist. Assoc.*, **95**, 1312–1316.

[15] Jefferys, W., and Berger, J. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64–72.

[16] Mitchell, T.J., and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.* **83**, 1023–1036.

[17] Robert, C. (1996). Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice*, Ed. W. Gilks, S. Richardson & D. Spiegelhalter, pp. 441–64. London: Chapman and Hall.

[18] Robert, C.P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

[19] Shaffer, P.J. (1999). A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. *J. Statist. Plann. Inference* **82**, 197-213.

[20] Storey J.D. (2001). The positive False Discovery Rate: A Bayesian interpretation and the q-value. *Technical Report,* Department of Statistics, Stanford University.

[21] Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society,* **B 64**, 479–498.

[22] Waller, R.A. and Duncan, D.B. (1969). A Bayes rule for the symmetric multiple comparison problem. *J. Amer. Statist. Assoc.* **64**, 1484-1503.

[23] West, M. (2003). Bayesian factor regressionm models in the "large p, small n" paradigm. In *Bayesian Statistics 7*, J.M. Bernardo, M.J. Bayarri, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West (Eds.). Oxford University Press, Oxford.

[24] Westfall, P.H., Johnson, W.O. and Utts, J.M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* **84**, 419-427.