7-18-2006

FDR and Bayesian Multiple Comparisons Rules

Peter Muller
M.D. Anderson Cancer Center

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, gp@jimmy.harvard.edu

Kenneth Rice
University of Washington

Proc. Valencia / ISBA 8th World Meeting on Bayesian Statistics
Benidorm (Alicante, Spain), June 1st–6th, 2006

FDR and Bayesian Multiple Comparisons Rules

Peter Müller, Giovanni Parmigiani & Kenneth Rice
M.D. Anderson Cancer Center, USA *Johns Hopkins University, USA*
University of Washington, Seattle, USA
pmueller@mdanderson.org gp@jhu.edu kenrice@u.washington.edu

Summary

We discuss Bayesian approaches to multiple comparison problems, using a decision theoretic perspective to critically compare competing approaches. We set up decision problems that lead to the use of FDR-based rules and generalizations. Alternative definitions of the probability model and the utility function lead to different rules and problem-specific adjustments. Using a loss function that controls realized FDR we derive an optimal Bayes rule that is a variation of the Benjamini and Hochberg (1995) procedure. The cutoff is based on increments in ordered posterior probabilities instead of ordered p-values. Throughout the discussion we take a Bayesian perspective. In particular, we focus on conditional expected FDR, conditional on the data. Variations of the probability model include explicit modeling for dependence. Variations of the utility function include weighting by the extent of a true negative and accounting for the impact in the final decision.

Keywords and Phrases: Decision problems; Multiplicities; False discovery rate.

1. INTRODUCTION

We discuss Bayesian approaches to multiple comparison problems, using a Bayesian decision theoretic perspective to critically compare competing approaches. Multiple comparison problems arise in a wide variety of research areas. Many recent discussions are specific to massive multiple comparisons arising in the analysis of high throughput gene expression data. See, for example, Storey et al. (2004) and references therein. The basic setup is a large set of comparisons. Let r_i denote the unknown truth in the i-th comparison, $r_i = 0$ (H_0) versus $r_i = 1$ (H_1), $i = 1, \ldots, n$. In the context of gene expression data a typical setup defines r_i as an indicator for gene i being differentially expressed under two biologic conditions of interest. For each gene a suitably defined difference score z_i is observed, with $z_i \sim f_0(z_i)$ if $r_i = 0$, and $z_i \sim f_1(z_i)$ if $r_i = 1$. This is the basic setup of the discussions in Benjamini and Hochberg (1995); Efron et al. (2001); Storey (2002); Efron and Tibshirani (2002); Genovese and Wasserman (2002, 2003); Storey et al. (2004); Newton et al. (2004); Cohen and Sackrowitz (2005) and many others. A traditional approach to address

the multiple comparison problem in these applications is based on controlling false discovery rates (FDR), the proportion of false rejections relative to the total number of rejections. We discuss details below. A similar setup arises in the analysis of high throughput protein expression data, for example, mass/charge spectra from MALDI-TOF experiments, as described in Baggerly et al. (2003).

Many other applications lead to similar massive multiple comparison problems. Clinical trials usually record data for an extensive list of adverse events (AE). Comparing treatments on the basis of AEs takes the form of a massive multiple comparison problem. Berry and Berry (2004) argue that the hierarchical nature of AEs, with AEs grouped into biologically different body systems, is critical for an appropriate analysis of the problem. Another interesting application of multiple comparison and FDR is in classifying regions in image data. Genovese et al. (2002) propose an FDR-based method for threshold selection in neuroimaging. Shen et al. (2002) propose an enhanced procedure that takes into account spatial dependence, specifically in a wavelet based spatial model. Another traditional application of multiple comparisons arises in record linkage problems. Consider two data sets, A and B, for example billing data and clinical data in a clinical trial. The record matching problem refers to the question of matching data records in A and B corresponding to the same person. Consider a partition of all possible pairs of data records in A and B into matches versus non-matches. A traditional summary of a given partition is the Correct Match Rate (CMR), defined as the fraction of correctly guessed matches relative to the number of true matches. See, for example, Fortini et al. (2001, 2002). Another interesting related class of problems are ranking and selection problems. Lin et al. (2004) describe the problem of constructing league tables, i.e., reporting inference on ranking a set of units (hospitals, schools, etc.). Lin et al. explicitly acknowledge the nature of the multiple comparison as a decision problem and discuss solutions under several alternative loss functions.

To simplify the remaining discussion we will assume that the multiple comparison problem arises in a microarray group comparison experiment, keeping in mind that the discussion remains valid for many other massive multiple comparison. A microarray group comparison experiment records gene expression for a large number of genes, $i = 1, \ldots, n$, under two biologic conditions of interest, for example tumor tissue and normal tissue. For each gene we are interested in the comparison of the two competing hypotheses that gene i is differentially expressed versus not differentially expressed. We will refer to a decision to report a gene as differentially expressed as a discovery (or positive, or rejection), and the opposite as a negative (fail to reject).

2. FALSE DISCOVERY RATES

Many recently proposed approaches to address massively multiple comparisons are based on controlling false discovery rates (FDR), introduced by Benjamini and Hochberg (1995). Let δ_i denote an indicator for rejecting the i-th comparison, for example flagging gene i as differentially expressed and let $D = \sum \delta_i$ denote the number of rejections. Let $r_i \in \{0, 1\}$ denote the unknown truth, for example an indicator for true differential expression of gene i. We define $\text{FDR} = (\sum(1 - r_i)\delta_i)/D$ as the fraction of false rejections, relative to the total number of rejections. The ratio defines a summary of the parameters (r_i), the decisions (δ_i) and the data (indirectly, through the decisions). As such it is neither Bayesian nor frequentist. How we proceed to estimate and/or control it depends on the chosen paradigm. Traditionally one considers the (frequentist) expectation $E(\text{FDR})$, taking an expectation

over repeated experiments. This is the definition used in Benjamini and Hochberg (1995). Applications of FDR to microarray analysis are discussed, among many others, in Efron and Tibshirani (2002). Storey (2002, 2003) introduces the positive FDR (pFDR) and the q-value and improved estimators for the FDR. In the pFDR the expectation is defined conditional on $D > 0$. Efron and Tibshirani (2002) show the connection between FDR and the empirical Bayes procedure proposed in Efron et al. (2001) and the FDR control as introduced in Benjamini and Hochberg (1995). Genovese and Wasserman (2003) discuss more cases of Bayes-frequentist agreement in controlling various aspects of FDR. Let p_i denote a p-value for testing $r_i = 1$ versus $r_i = 0$. They consider rules of the type $\delta_i = I(p_i < t)$. Controlling the posterior probability $P(r_i = 0 \mid Y, p_i = t)$ is stronger than controlling the expected FDR for a threshold t. Specifically, let $\mathrm{FDR}(t)$ denote FDR under the rule $\delta_i = (p_i < t)$, let $\hat{q}(t) \approx P(r_i = 0 \mid Y, p_i = t)$ denote an approximate evaluation of the posterior probability for the $i-$th comparison, and let $Q(t) \approx E(\mathrm{FDR}(t))$ denote an asymptotic approximation of expected FDR. Then $q(t) \leq Q(t)$. The argument assumes concavity of the c.d.f. for p-values under the alternative $r_i = 1$. Genovese and Wasserman (2002) also show an agreement of confidence intervals and credible sets for FDR. They define the realized FDR process $\mathrm{FDR}(T)$ as a function of the threshold T and call T a (c, α) confidence threshold if $P(\mathrm{FDR}(T) < c) \geq 1 - \alpha$. The probability is with respect to a sampling model that includes an (unknown) mixture of true and false hypotheses. The Bayesian equivalent is a posterior credible set, i.e., controlling $P(\mathrm{FDR}(T) \leq c \mid Y) \leq 1 - \alpha$. Genovese and Wasserman (2003) show that controlling the posterior credible interval for $\mathrm{FDR}(T)$ is asymptotically equivalent to controlling the confidence threshold.

Let z_i denote some univariate summary statistic for the i-th comparison, for example a p-value. Many discussions are in the context of an assumed i.i.d. sampling model for z_i, from a mixture model $f(\cdot)$ with terms f_0 and f_1 corresponding to subpopulations of differentially and not-differentially expressed genes, respectively:

$$z_i \sim p_0\, f_0(z_i) + (1 - p_0)\, f_1(z_i) \equiv f(z_i).$$

Using latent indicators $r_i \in \{0, 1\}$ introduced earlier, the mixture is equivalent to the hierarchical model:

$$p(z_i \mid r_i = j) = f_j(z_i) \text{ and } Pr(r_i = 0) = p_0 \tag{1}$$

Let F_0 and F denote the c.d.f. for f_0 and f. Efron and Tibshirani (2002) define FDR for rejection regions of the type $\{z_i \leq z\}$,

$$\mathrm{Fdr}(z) \equiv p_0 F_0(z)/F(z)$$

and denote it "Bayesian FDR". The Bayesian label is justified by the use of Bayes theorem to find the probability of false discovery given $\{z_i \leq z\}$, which they show is equivalent to the defined Fdr statistic. The probability statement is in the context of the assumed mixture model, for assumed known f_0, f_1 and p_0. In particular, there is no learning about p_0. However, using reasonable data-driven point estimates for the unknown quantities f_0, f_1 and p_0, the Fdr statistic provides an good approximation for what $P(r_{n+1} = 1 \mid z_{n+1} \leq z, Y)$ would be in a full Bayesian model with flexible priors on f_1 and f_0. Here and throughout this paper we use Y to generically indicate the observed data. Efron et al. (2001) introduce local FDR, as

$$\mathrm{fdr}(z) \equiv p_0 f_0(z)/f(z).$$

Under the mixture model, and conditioning on f_0, f_1 and p_0, the fdr statistic is the probability of differential expression, so $\text{fdr}(z) = Pr(r_i = 1 \mid z_i = z, Y, f_0, f_1, p_0)$. As before, one can argue that under a sufficiently flexible prior probability model on f_0, f_1, p_0, reasonable point estimates can be substituted for the unknown quantities, allowing us to interpret values of fdr as posterior probabilities, without reference to a specific prior model (subject to identifiability constraints). In the following sections we argue that posterior inference for the multiple comparison should consider more structured models. Inference should not stop with the marginal posterior probability of differential expression. Rules of the type $\delta_i = I(z_i \leq z)$ are intuitive, but not necessarily optimal. In the context of a full probability model, and assuming a reasonable utility function, it is straightforward to derive the optimal rule.

Considering frequentist expectations of FDR, i.e., expectations over repeated sampling, we need to consider expectations over a ratio of random variables. Short of uninteresting trivial decision rules, the decision $\delta_i = \delta_i(Y)$ is a function of the data and appears in both, numerator and denominator of the ratio. The discussion significantly simplifies under a Bayesian perspective. The only unknown quantity in $\text{FDR} = \sum \delta_i(1 - r_i)/D$ is the unknown r_i in the numerator. Let $v_i = P(r_i = 1 \mid Y)$ denote the marginal posterior probability of gene i being differentially expressed and define

$$\overline{\text{FDR}} = E(\text{FDR} \mid Y) = \sum (1 - v_i)\delta_i / D. \tag{2}$$

Newton et al. (2004) consider decision rules that classify a gene as differentially expressed if $v_i > \gamma^*$, fixing γ^* to achieve a certain pre-set false discovery rate, $\overline{\text{FDR}} \leq \alpha$. Newton et al. (2004) comment on the dual role of v_i in decision rules like $\delta_i = I(v_i > \gamma^*)$. It determines the decision, and at the same time already reports the probability of a false discovery as $1 - v_i$ for $\delta_i = 1$ and the probability of a false negative as v_i for $\delta_i = 0$.

3. POSTERIOR PROBABILITIES ADJUST FOR MULTIPLICITIES

"Posterior inference adjusts for multiplicities, and no further adjustment is required." The statement is only true with several caveats. First, the probability model needs to include a positive prior probability of non-differential expression for each gene i. Second, the model needs to include a hyperparameter that defines the prior probability mass for non-differential expression. For example, consider the mixture model (1), with independence across i, conditional on p_0. The statement requires that p_0 be a parameter with a hyperprior $p(p_0)$, rather than fixed. Scott and Berger (2003) discuss the nature of this adjustment and show some examples. In the context of microarray data analysis, Do et al. (2005) carry out the same simulation experiment in the context of a mixture model as in (1). Results are shown in Table 1. The table shows marginal posterior probabilities v_i for differential expression. The nature of the model is such that v_i depends on the gene only through an observed difference score z_i, making it meaningful to list v_i by observed difference score z_i. The marginal posterior probability of differential expression adjusts for the multiplicities. If there are many truly negative comparisons, as in the third row of the table, then the model reduces the marginal probabilities of differential expression. If on the other hand there are many truly positive comparisons, as in the first row, then the model appropriately increases the marginal probabilities.

The probability model need not be i.i.d. sampling. Any probability model that includes a positive prior probability for $r_i = 0$ and $r_i = 1$, i.e., any model that allows inference on how comparisons between units are true or false leads to a similar

	Observed z scores								
p_0	-5.0	-4.0	-3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
0.4	**1.0**	**1.0**	**1.0**	**0.9**	**0.5**	0.2	**0.4**	**0.9**	**1.0**
0.8	**0.9**	**0.9**	**0.8**	**0.4**	0.1	0.1	0.1	**0.4**	**0.8**
0.9	**0.5**	0.4	0.3	0.1	0.1	0.0	0.0	0.1	0.3

Table 1: Posterior probabilities of differential expression, as a function of the observed difference score z_i, under three different simulation truths, using $p_0 = 0.4$ (first row), 0.8 (second row) and 0.9 (third row) for the proportion of false comparisons. Probabilities $v_i > 0.4$ are marked in bold face.

adjustment. Berry and Hochberg (1999) discuss this perspective. An interesting probability model is proposed in Gopalan and Berry (1998). They consider the problem of comparing group means in an ANOVA setup. They introduce a prior probability model on all possible partitions of matching group means using the probability model on random partitions that is implied by sampling from a random probability measure with a Dirichlet process prior.

Berry and Berry (2004) discuss inference for adverse events (AE) in clinical trials, proposing a hierarchical model to address the multiplicity issue. The data are occurrences of a large set of adverse events reported in a two arm clinical trial comparing standard versus experimental therapy. The authors argue that the conclusion about a set of AEs with elevated occurrence under the experimental therapy should be different, depending on whether these AEs cluster in the same body system, or are scattered across different body systems. In the latter case it should be considered more likely that the reported AEs are due to random occurrence, whereas in the earlier case it seems more likely that the increased AEs are caused by the drug. Berry and Berry (2004) develop a three-level hierarchical model with levels corresponding to AEs, body systems, and the collection of all body systems. The proposed hierarchical model leads to the desired inference. Due to borrowing of strength in the hierarchical model AEs that cluster in the same body system lead to higher posterior probability of an underlying true difference than if the same AE counts were observed across different body systems.

4. DECISION THEORETIC APPROACHES

In a review of a Bayesian perspective on multiple comparisons Berry and Hochberg (1999) comment that "finding posterior distributions of parameters is only part of the Bayesian solution. The remainder involves decision analysis." Computing posterior probabilities of differential expression only estimates parameters in the probability model. It does not yet recommend a specific decision about flagging genes as differentially expressed or not. Reasonable solutions are likely to follow some notion of monotonicity. All else being equal, genes with higher marginal probability of differential expression should be more likely to be reported as differentially expressed. However, differing levels of differential expression, focused interest in some subsets of genes, and inference about dependence might lead to violations of monotonicity. More importantly, this argument, without refinement, does not provide the threshold beyond which comparisons should be rejected.

It can be shown (Müller et al., 2004) that under several loss functions that

combine false negative and false discovery counts and/or rates the optimal decision rule is of the following form. Recall that δ_i is an indicator for the decision to report gene i as differentially expressed and $v_i = Pr(r_i = 1 \mid Y)$ denotes the marginal posterior probability of differential expression for gene i. The optimal decision is to declare all genes with marginal probability beyond a threshold as differentially expressed:

$$\delta_i^* = I(v_i > t). \tag{3}$$

The choice of loss function determines the specific threshold. In Müller et al. (2004) we consider four alternative loss functions. Similar to FDR we define FD $= \sum(1 - r_i)\delta_i$ and FN $= \sum r_i(1 - \delta_i)$ as the false positive and false negative counts, and FNR $=$ FN$/(n - D)$ as the false negative ratio. We use $\overline{\text{FD}}$, $\overline{\text{FN}}$ and $\overline{\text{FNR}}$ for the posterior expectations. All are easily evaluated. For example, $\overline{\text{FN}} = \sum v_i(1 - \delta_i)$. Considering various combinations of these statistics we define alternative loss functions. Since the posterior expectation is straightforward, we specify the loss functions already as posterior expected loss. The first two loss functions are linear combinations of the false negative and positive counts and ratios. We define

$$L_N(\delta, z) = c\,\overline{\text{FD}} + \overline{\text{FN}}, \tag{4}$$

and $L_R(\delta, z) = c\,\overline{\text{FDR}} + \overline{\text{FNR}}$. The loss function L_N is a natural extension of $(0, 1, c)$ loss functions for traditional hypothesis testing problems (Lindley, 1971). From this perspective the combination of error rates in L_R seems less attractive. The loss for a false discovery and false negative depends on the total number of discoveries or negatives, respectively. Genovese and Wasserman (2002) interpret c as the Lagrange multiplier in the problem of minimizing FNR subject to a bound on FDR. They compare the Benjamini and Hochberg (1995) rule (BH) and the optimal rule under L_R and show that BH almost achieves the optimal risk, in particular for a large fraction of true nulls.

Alternatively, we consider bivariate loss functions that explicitly acknowledge the two competing goals:

$$L_{2R}(\delta, z) = (\overline{\text{FDR}}, \overline{\text{FNR}}), \quad L_{2N}(\delta, z) = (\overline{\text{FD}}, \overline{\text{FN}}).$$

We need to define the minimization of the bivariate functions. A traditional approach to select an action in multicriteria decision problems is to minimize one dimension of the loss function while enforcing a constraint on the other dimensions (Keeney et al., 1976). We thus define the optimal decisions under L_{2N} as the minimization of $\overline{\text{FN}}$ subject to $\overline{\text{FD}} \leq \alpha_N$. Similarly, under L_{2R} we minimize $\overline{\text{FNR}}$ subject to $\overline{\text{FDR}} \leq \alpha_R$.

Under all four loss functions the optimal rule is of the form (3). See Müller et al. (2004) for a statement of the optimal cutoffs t. The result is true for any probability model with non-zero prior probability for differential and non-differential expression. In particular, the probability model could include dependence across genes.

One of the assumptions underlying these loss functions is that all false negatives are equally undesirable, and all false positives are equally undesirable. This is inappropriate in most applications. A false negative for a gene that is truly differentially expressed, but with a small difference across the two biologic conditions, is surely less of a concern than a false negative for a gene that is differentially expressed with a large difference. The large difference might make it more likely that follow up experiments will lead to significant results. Assume now that the probability model

includes for each gene i a parameter m_i that can be interpreted as the level of differential expression, with $m_i = 0$ if $r_i = 0$, and $m_i > 0$ if $r_i = 1$. For example, in the hierarchical gamma/gamma model proposed in Newton et al. (2001) this could be the absolute value of the log ratio of the gamma scale parameters that index the sampling distributions under the two biologic conditions. A log ratio of $m_i = 0$ implies equal sampling distributions, i.e., no differential expression. In the mixture of Dirichlet process model of Do et al. (2005), m_i would be the absolute value of the latent variable generated from the random probability measure with Dirichlet process prior. A natural extension of the earlier loss functions is to

$$L_m(m, \delta, z) = -\sum \delta_i \, m_i + k \sum (1 - \delta_i) m_i + c \, D.$$

A similar weighting with the relative magnitude of errors is underlying Duncan's (1965) multiple comparison procedure. Since $r_i = 0$ implies $m = 0$, the summations go only over all true positives, $r_i = 1$. The loss function includes a reward proportional to m_i for a correct discovery, and a penalty proportional to m_i for each false negative. The last term encourages parsimony, without which the optimal decision would be to trivially flag all genes. Straightforward algebra shows that the optimal decision is similar to (3). Let $\overline{m}_i = E(m_i \mid Y)$ denote the posterior expected level of differential expression for gene i. The optimal rule is

$$\delta_i^* = I\{\overline{m}_i \geq c/(1 + k)\} :$$

Flag all genes with $\overline{m}_i$ greater than a fixed cutoff. The optimal rule remains essentially the same if we replace m_i in the loss function by some function of m, allowing in particular for the loss to be a non-linear function of the true level of differential expression.

$$L_f(m, \delta, z) = -\sum \delta_i f_D(m_i) + \sum (1 - \delta_i) f_N(m_i) + c \, D. \tag{5}$$

The functions $f_D(m)$ and $f_N(m)$ would naturally be S-shaped, monotone functions with a minimum at $m = 0$, and perhaps level off for large levels of m. Let $\overline{f_N}_i = E(f_N(m) \mid Y)$ denote the posterior expectation for $f_N(m_i)$, and similarly for $\overline{f_D}$. The optimal decision is

$$\delta_f^* = I\{\overline{f_D}_i + k\overline{f_N}_i > c\}.$$

Flag all genes with sufficiently large expected reward for discovery and/or penalty for a false negative. The rule δ_f^* follows from the fact that the choice of m_i in L_m was arbitrary.

The introduced loss functions are all generic in the sense of being reasonable loss functions without reference to a specific decision related to the multiple comparisons. If the goal of the inference is a very specific decision with a clearly recognizable implication, a problem-specific loss function should be used as the relevant criterion for the multiple comparison. For example, Lin et al. (2004) and Shen and Louis (1998) consider the problem of ranking units like health care providers. Ranking is a specific form of a multiple comparison problem. It could be described as all pairwise comparisons, subject to transitivity. They introduce loss functions that formalize the implications of a specific ranking, relative to the true ranks, and show the optimal rules for several such loss functions.

Example 1: Epithelial Ovarian Cancer (EOC)
Wang et al. (2004) report a study of epithelial ovarian cancer (EOC). The goal of the study is to characterize the role of the tumor microenvironment in EOC. To this end the investigators collected tissue samples from patients with benign and malignant ovarian pathology. Specimens were collected, among other sites, from peritoneum adjacent to the primary tumor. RNA was co-hybridized with reference RNA to a custom-made cDNA microarray including a combination of the Research Genetics RG_HsKG_031901 8k clone set and 9,000 clones selected from RG_Hs_seq_ver_070700. A complete list of genes is available at

`http://nciarray.nci.nih.gov/gal_files/index.shtml`

(The array is listed as custom printing Hs-CCDTM-17.5k-1px).

We focus on the comparison of 10 peritoneal samples from patients with benign ovarian pathology versus 14 samples from patients with malignant ovarian pathology. The raw data was pre-processed using BRB ArrayTool (`http://linus.nci.nih.gov/BRB-ArrayTools.html`). In particular, spots with minimum intensity less than 300 in both fluorescence channels were excluded from further analysis. See Wang et al. (2004) for a detailed description.

We computed probabilities of differential expression using the POE model proposed in Parmigiani et al. (2002). Inference is summarized by marginal probabilities of differential expression v_i. One parameter in the model is interpretable as the level of differential expression. Briefly, the basic POE model includes a trinary indicator e_{it} for gene i and sample t, with $e_{it} \in \{-1, 0, 1\}$ for under-expression, normal and over-expression. In a variation of the original POE model we use a probit prior for e_{it}. The probit prior includes a regression on an indicator for malignant ovarian pathology. We denote the corresponding coefficient in the model by m_i, and interpret it as the level of differential expression for gene i. The original model does not include a gene-specific parameter r_i that can be interpreted as differential expression for gene i. We define $r_i = I(|m_i| > \epsilon)$, using $\epsilon = 0.5$. Figure 1 shows the selected lists of reported genes under the loss functions L_N and L_m (marked L_N and L_m). To facilitate the comparison we calibrated the tradeoff parameter c in both loss functions to fix $D = 20$. The difference in the two solutions are related to the difference between statistical significance and biologic significance. Because of varying precisions, it is possible that a gene with a very small level of differential expression reports a high posterior probability of differential expression, and vice versa.

5. APPROXIMATING BENJAMINI AND HOCHBERG'S RULE

The earlier introduced loss functions and the corresponding Bayes rules control various aspects of false discovery and false negative counts and rates. While similar in spirit, the rules are different from methods that have been proposed to control frequentist expected FDR, for example the rule defined in Benjamini and Hochberg (1995), henceforth BH.

It is not possible to justify BH (applied to the sorted p-values) as an optimal Bayes rule under a loss function that includes a combination of FD(R) and FN(R). This is shown in Cohen and Sackrowitz (2005) and extended in Cohen and Sackrowitz (2006). BH can be described as a step-up procedure that compares ordered observed difference scores $z_{(i)}$ with pre-determined critical cutoffs C_i. Let j denote the smallest integer i with $z_{(i)} > C_i$. All comparisons with difference scores beyond $z_{(j)}$ are rejected. Cohen and Sackrowitz (2005) show that such rules are inadmissible. The discussion includes a simple example that makes the inadmissibility easily

interpretable. Consider a set of (ordered) p-values $p_{(i)}$ with $p_{(i)} = n\frac{\alpha}{n} - \epsilon$, i.e., all equal values. In particular, the largest p-value, $p_{(n)}$, falls below the BH boundary $(j\,\alpha)/n$. The BH rule would lead us to reject all comparisons. Now consider $p_{(i)} = i\frac{\alpha}{n} + \epsilon$. The p-values $p_{(i)}$, $i = 1, \ldots, n-1$ are substantially smaller, and $p_{(n)}$ is only slightly larger. Yet, we would be lead not to reject any comparison.

But interestingly, it is possible to still mimic the mechanics of the popular BH method as the Bayes rule under a specific loss function. The rule replaces the p-values with increments in posterior probabilities. The correspondence is not exact, and can not be in the light of Cohen and Sackrowitz' inadmissibility results.

Recall that $\delta_i(z) \in \{0, 1\}$ denotes the decision rule for the i-th comparison, $r_i \in \{0, 1\}$ is the (unknown) truth, $v_i = Pr(r_i = 1 \mid Y)$ are the marginal posterior probabilities, $\text{FD} = \sum \delta_i(1 - r_i)$ are the false discovery count, $\overline{\text{FD}} = \sum \delta_i(1 - v_i) = E(\text{FD} \mid Y)$, and $D = \sum \delta_i$ is the number of rejections. Let $w_i = 1 - v_i$ denote the marginal probability of the i-th null model.

Consider the loss function $\ell_B(\delta, z, r) = I(\text{FD} > \alpha D) - g_D$, with a monotone reward g_D for the number of discoveries. Marginalizing w.r.t. r, conditional on the data, we find the expected loss $L_B(\delta, z) = P(\text{FD} > \alpha D \mid Y) - g_D$. By Chebycheff's inequality, $P(\text{FD} > \alpha D) \leq \overline{\text{FD}}/(\alpha D)$. Using this upper bound, we define $L_B \approx$

$$L_U(\delta, z) \equiv \frac{\overline{\text{FD}}}{\alpha D} - g_D = \overline{\text{FDR}}/\alpha - g_D.$$

Without loss of generality assume $w_1 \leq w_2 \leq w_3 \ldots$ are ordered. We show that under L_U with $g_D = D/n$, the optimal decision is to use a threshold equal to the largest j with (appropriately defined) increment in posterior probability w_j less than $(j\alpha)/n$. See below for the appropriate definition of increment in w_j.

For fixed D, the optimal rule selects the D largest probabilities v_i. Let $\delta_i^D = I(i \leq D)$ denote this rule. To determine the optimal rule we still need to find the optimal $D = j$. Consider the condition $L_U(\delta^j, z) \leq L_U(\delta^{j-1}, z)$ for preferring δ^j over δ^{j-1}:

$$\frac{1}{\alpha j}\sum_{i=1}^{j} w_i - g_j \leq \frac{1}{\alpha(j-1)}\sum_{i=1}^{j-1} w_i - g_{j-1}.$$

After some simplification, and letting $\overline{w}_j = \frac{1}{j}\sum_{i=1}^{j} w_i$ denote the average across comparisons 1 through j, and $\Delta w_j = w_j - \overline{w}_{j-1}$, the condition becomes $L_U(\delta^j, z) < L_U(\delta^{j-1}, z)$ if $\Delta w_j < (g_j - g_{j-1})\alpha j$. A similar condition is true for lag k comparisons. Let $\overline{w}_{ij} = \frac{1}{j-i+1}\sum_{h=i}^{j} w_i$ and $\Delta w_{ij} = \overline{w}_{i+1,j} - \overline{w}_i$.

$$L_U(\delta^j, z) \leq L_U(\delta^{j-k}, z) \text{ if } \Delta w_{j-k,j} \leq (g_j - g_{j-k})\alpha j/k \tag{6}$$

The earlier condition was the special case for $k = 1$. For $g_j = j/n$ the condition becomes $\Delta w_{j-k,j} \leq \frac{\alpha\, j}{n}$. Condition (6) characterizes the optimal rule δ^*. Let $B(2) \equiv 1$ and $B(j) = \max_{i<j}\{i : \Delta w_{B(i),i} < \frac{\alpha\, i}{n}\}$. In words, $B(j)$ is the best rule δ^i, $i < j$. The optimal rule is δ^j for

$$j = \max_i \left\{ i : \Delta w_{B(i),i} \leq \frac{\alpha\, i}{n} \right\}.$$

This characterizes the optimal rule by an algorithm like BH, applied to the increments in posterior probabilities $\Delta w_{B(i),i}$.

An alternative justification of a BH type procedure is the following approximation. Recall that under the loss function L_B, the optimal rule must be of the type $\delta_i = I(v_i \geq v_j)$ for some optimal j. If we knew the number n_0 of true null hypotheses, then we would find $\overline{\text{FD}} \leq (1-v_j)n_0$, and thus $\overline{\text{FDR}} \leq (1-v_j)n_0/j$. Assume that the probabilities v_i are ordered. To minimize v_j, while controlling $\overline{\text{FDR}}$ we would determine the cutoff by the maximum j with $(1-v_j) \leq j\alpha/n_0$. Finally, replacing n_0 by the conservative bound n we get a BH type rule on the posterior probabilities $(1-v_j)$.

A fundamental difference of BH and the rule under L_U is the use of posterior probabilities instead of p-values. Of course, the two are not the same. The relationship of p-values and posterior probabilities is extensively discussed, for example, in Casella and Berger (1987), Sellke et al. (2001) and Bayarri and Berger (2000).

The loss function L_B serves to make the use of BH type rules plausible under an approximate expected loss minimization argument. We would not, however, recommend it for practical use. Assume we had fantastically good data, with $v_i \in \{0,1\}$, i.e., we essentially know the truth for all comparisons. The rule $\delta_i = I(v_i = 1)$ that reports the known truth is not optimal. Under L_B it can be improved by knowingly reporting false positives with $v_i = 0$. This is possible since L_B rewards for large D, and only introduces a penalty for false positives if the set threshold αD is exceeded. A similar statement applies for L_U.

6. FDR AND DEPENDENCE

In previous sections we argued for the use of posterior probabilities to account for multiplicities, and for a decision theoretic approach to multiple comparisons. The two arguments are not competing, but naturally complement each other. A structured probability model helps us to identify genes that might be of more interest than others.

In particular, the dependence structure of expression across genes might be of interest. If the goal is to develop a panel of biomarkers to classify future samples, then it is desireable to have low correlation of the expression levels for the reported set of differentially expressed genes. For other applications one might want to argue the opposite. Recall the example about inference on adverse events mentioned in the introduction.

In Müller and Parmigiani (2006) we introduce a probability model for gene expression that includes dependence for subsets of genes. The dependent subsets are typically identified as genes with a common functionality or genes corresponding to the nodes on a pathway of interest. The probability model allows us to use known pathways to formulate informative prior probability models for dependence across genes that feature in that pathway. Alternatively, for a small to moderately large set of genes the model allows us to learn about dependence starting from relatively vague prior assumptions. We briefly outline the features of the model that are relevant for the decision about reporting differentially expressed genes. Dependence is introduced not on the observed gene expressions, but on imputed trinary indicators $e_{it} \in \{-1,0,1\}$ for under- and over-expression of gene i in sample t. We build on the POE model introduced in Parmigiani et al. (2002), and already briefly mentioned earlier. In a variation of the basic POE model, in Müller and Parmigiani (2006) we represent the probabilities for the trinary outcome by a latent normal random

variable z_{it}, with

$$e_{it} = \begin{cases} -1 & \text{if } z_{it} < -1 \\ 0 & \text{if } -1 < z_{it} < 1 \\ 1 & \text{if } z_{it} > 1. \end{cases} \quad (7)$$

The latent variables z_{it} are continuous random variables that allow us to introduce the desired dependence on related genes, as well as a regression on biologic condition. Let x_t denote a sample-specific covariate vector including an indicator x_{t1} for the biologic condition of the sample t and other sample-specific covariates. For example, in the case of a two group comparison between tumor and normal tissues, x_{t1} could be a binary indicator of tumor. Also, let $\{e_{jt};\ j \in N_i\}$ denote the trinary indicators for other genes that we wish to include as possible parent nodes in the dependent prior model for z_{it}. We assume a regression

$$z_{it} = g(x_t, e_{jt},\ j \in N_i) + \epsilon_i, \quad (8)$$

with mean function $g(\cdot)$ and standard normal residuals ϵ_i. The regression on e_{jt} introduces the desired dependence, and the regression on x_t includes the regression on the biologic condition x_{t1}, as before. Let m_i denote the regression coefficient for x_{t1}, the biologic condition. Also, define Σ_1 as the correlation matrix of $\{z_{it};\ \delta_i = 1\}$, the latent scores corresponding to the reported genes. The model allows us to include a term in the loss function that penalizes the reporting of highly correlated genes. We modify L_f to

$$L_D(m, \delta, z) = -k_1\ \log(|\Sigma_1|) - k_2\ \sum \delta_i f_D(m_i) + k_3\ \sum (1-\delta_i) f_N(m_i) + k_4\, c\, D. \quad (9)$$

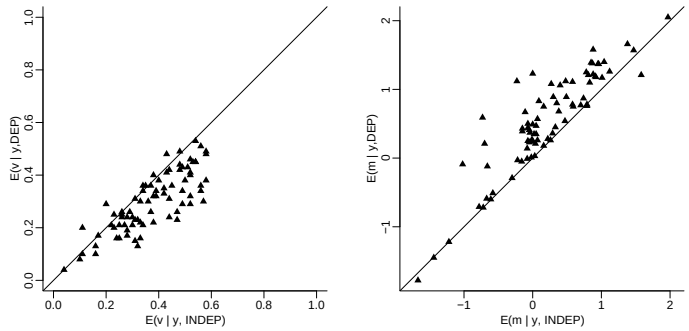
The loss function encourages the inclusion of few highly differentially expressed genes with low correlation. Correlation is formalized as the tetrachoric correlation of the trinary outcomes e_{it}. See, for example, Ronning and Kukuk (1996) for a discussion of polychoric correlations for ordinal outcomes.

Example 1 (ctd): Epithelial Ovarian Cancer (EOC)

Earlier we reported inference using the POE model and the loss functions L_N and L_m. We reanalyze the data with a variation of the POE model that includes dependent gene expression. In the implementation we specified (9) with $k_1 = 1$, $k_2 = 0.01$, $k_3 = k_4 = 0$, restricting to $D = 20$ (for comparability with the results under L_N and L_m), and setting $f_D(m_i) = m_i^2$. The inference summaries v_i and $\overline{m}_i$ change slightly when adjusting for dependence. The change in the estimates is shown in Figure 1. Although the changes are minimal for most genes, the impact in the final decision is visible. The first four rows of Figure 2 show the reported set of genes under L_N using the independent POE model (row 1) versus the dependent model (row 2), under L_m using the independent (row 3) and dependent (row 4) model. The last row shows inference under the loss function L_D.

7. A PREDICTIVE LOSS FUNCTION

Microarray experiments are often carried out as hypothesis generating experiments, to screen for promising genes to be followed up in later experiments. Abruzzo et al. (2005) describe a setup, using RT-PCR to validate a list of differentially expressed genes found in a microarray group comparison experiment. In particular,



(a) v_i (b) $\overline{m}_i$

Figure 1: *Inference with dependent prior (y-axis) vs. indep prior (x-axis). The changes are large enough to change the decisions.*

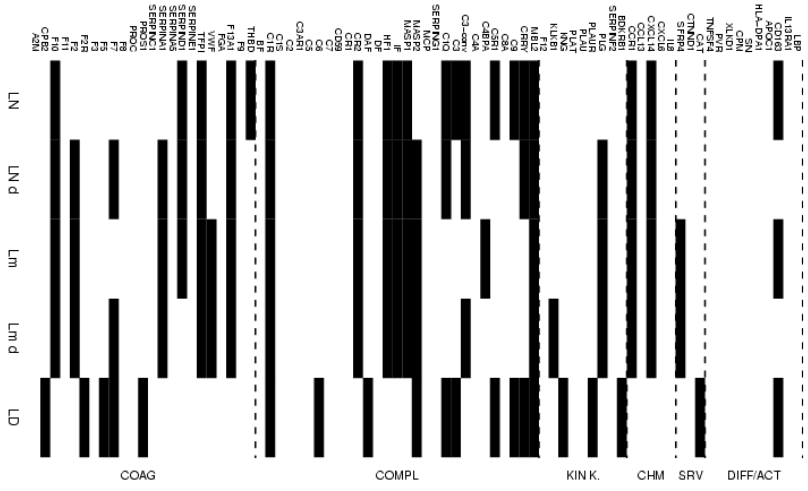


Figure 2: *Genes with $\delta_i = 1$, using L_N (top), L_N and dep model, L_m, L_m and dep model, and L_D (bottom).*

they use TaqMan Low-Density Arrays for real-time RT-PCR (Applied Biosystems, Foster City, USA). They consider inference for nine samples from patients with chronic lymphocytic leukemia (CLL), using the microarray experiment for a first step screening experiment, and the real-time RT-PCR to validate the identified genes. With a setup like this experiment in mind we define a utility function that is based on the success of a future followup study. To define the desired loss function we need to introduce some more detail of a joint probability model for the microarray and real time RT-PCR experiments. Eventually we will use a stylized description of inference in this model to define a loss function.

Let z_{it} be a suitably normalized measurement for gene i in sample t, with approximately unit marginal variance, $var(z_{it}) \approx 1$. For example, z_{it} could be the latent probit score in (7). Let y_{it} denote the recorded outcome of the RT-PCR for gene i in sample t. The data are copy numbers, interpreted as base two logarithm of the relative abundance of RNA in the sample. Abruzzo et al. (2005) use a normal linear mixed effects model to calibrate the raw data against a calibrator sample and an endogenous control, chosen to reduce the variance of corrected responses across samples. Let y_{it} denote the calibrated pre-processed data. An important conclusion of the discussion in Abruzzo et al. (2005) is inference about the correlation of the microarray and RT-PCR measurements. They find a bimodal distribution of correlations across genes. About half the genes show a cross-platform correlation $\rho_i \approx 0.8$, and half show essentially zero correlation, $\rho_i = 0$.

We introduce a simple hierarchical model to represent the critical features of the cross-platform dependence, and a realistic distribution of the RT-PCR outcomes. We build on the POE model described earlier, in (7) and (7), without necessarily including the dependent model extension. Let z_{it} denote the latent score in (7). For y_{it} we assume:

$$p(y_{it} \mid z_{it}, \rho_i) = \begin{cases} z_{it} & \text{with prob. } \rho_i \\ N(0,1) & \text{with prob. } (1-\rho_i) \end{cases} \tag{10}$$

with $Pr(\rho_i = \rho^\star) = p_\rho$ and $Pr(\rho_i = 0) = 1 - p_\rho$. We use $\rho^\star = 0.8$ and $p_\rho = 0.5$ to approximately match the reported inference in Abruzzo et al. (2005). Also, after standardization Abruzzo et al. (2005) found standard deviations in the RT-PCR outcomes for each gene across samples in the range of approximately 0.5 through 1.5 (with some outliers below and above). We chose the unit variance in the normal term above to match the order of magnitude of these reported standard deviations.

Consider now the problem of reporting a list of differentially expressed genes in a microarray group comparison experiment. We assume that the selected genes will be validated in a future followup real-time RT-PCR experiment, using, for example, the described TaqMan Low-Density Arrays. We build a loss function designed to help us to construct a rule that identifies genes that are most likely to achieve a significant result in the followup experiment. For a stylized description we assume that the followup experiment is successful for gene i if we can report a statistically significant difference of expression across the two biologic conditions.

In words, the construction of the proposed loss function proceeds as follows. For each identified gene i we first select an alternative hypothesis. We then carry out a sample size argument based on achieving a desired power for this alternative versus the null hypothesis of no differential expression, using a traditional notion of power. Next we find the posterior predictive probability of a statistically significant outcome (R_i) for the future experiment. Finally, we define a loss function with

terms related to the posterior predictive probability for R_i and the sampling cost for the future experiment. The stylized description is not a perfect reflection of the actual experimental process. It is not even a reasonable model for actual data analysis. But we believe it captures the critical features related to the desired decision of identifying differentially expressed genes. In particular, it includes a natural correction of statistical significance for the size of the effect.

Let $x_t \in \{-0.5, 0.5\}$ denote a (centered) indicator for the biologic condition of sample t. Recall from (7) that the level of differential expression for gene i, m_i, was defined as the probit regression coefficient for an indicator of the biologic condition. Let $(\overline{m}_i, s_i)$ denote the posterior mean and standard deviation of m_i. Let $\overline{\rho} = p_\rho \rho^\star$ denote the assumed average cross-platform correlation, averaged across genes. Let $\mu_{i1} = E(y_{it} \mid x_t > 0)$ and $\mu_{i0} = E(y_{it} \mid x_t < 0)$ denote the mean expression under the two conditions in the followup experiment. For the upcoming sample size argument we assume a test of the null hypothesis $H_0 : \mu_{i1} - \mu_{i0} = 0$ versus the alternative hypotheses $H_1 : \mu_{i1} - \mu_{i0} = m^*_{yi}$, with $m^*_{yi} = \overline{\rho}(\overline{m}_i - s_i)$, the mean difference under an assumed alternative $m_i = \overline{m}_i - s_i$. Let q_α denote the $(1-\alpha)$ quantile of a standard normal distribution. Assuming that upon conclusion of the followup experiment the investigators carry out a normal z-test, we find a required sample size

$$n_i(z) \geq 2\left[(q_\alpha + q_\beta)/m^*_{yi}\right]^2$$

for a given significance level α and power $(1-\beta)$. The sample size is a function of the data z, implicitly through the choice of the alternative m^*_{yi}. Here sample size refers to the number of samples under each biologic condition, i.e., the total number of samples is $2n$. Let $\overline{y}_{i0}$ and $\overline{y}_{i1}$ denote the sample average in the followup experiment, for gene i and the two conditions. Let $R_i = \{(\overline{y}_{i1} - \overline{y}_{i0})\sqrt{n/2} \geq q_\alpha\}$ denote the event of a statistically significant difference for gene i in the followup experiment. Let $\pi_i = Pr(R_i \mid Y)$ denote the posterior predictive probability of R_i. Let $\Phi(\cdot)$ denote the standard normal c.d.f.

$$\pi_i(z) = (1-p_\rho)\alpha + p_\rho \Phi\left[\frac{\rho^\star \overline{m}_{i1}\sqrt{n_i/2} - q_\alpha}{\sqrt{1 + \frac{n}{2}\rho^{*2}s_i^2}}\right].$$

Combining n_i and π_i to trade off the competing goals of small sampling cost and high success probability we define a loss function

$$L_F(\delta, z) = \sum_{\delta_i=1} [-c_1\pi_i(z) + n_i(z)] + c_2 D \tag{11}$$

Under the loss function L_F, the optimal rule is easily seen as

$$\delta^*_i = I(n_i + c_2 \leq c_1\pi_i).$$

If we replace classical power by Bayesian power (Spiegelhalter et al., 2004, chapter 6.5.3) then π_i remains constant by definition, leaving only the bound on the sample size n_i. Also, c_2 could be zero if the size of the reported short-list is not an issue. Figure 3 shows the optimal decision under L_F for the EOC example (squares). We used $c_1 = 3000$ and calibrated c_2 such that the optimal decision reports $D = 20$ genes, as before. The value for c_1 was chosen to have the reward c_1 match approximately 10 times the average sampling cost of a followup trial.

8. SUMMARY

Table 2 summarizes the proposed loss functions and rules. For all except L_D the optimal rule can be described as a threshold for an appropriate gene-specific summary statistic of the data. Storey (2005) describes such rules as significance thresholding. The summaries are the marginal posterior probability of differential expression v_i, posterior mean and standard deviation of the level of differential expression $(\overline{m}_i, s_i)$, the sample size for a followup experiment n_i, the posterior predictive probability of a significant outcome π_i, and the increment in posterior probability of non-differential expression $\Delta w_{B(i),i}$. The last three are functions of $(v_i, \overline{m}_i, s_i)$ only.

Table 2: Alternative loss functions and optimal rules

Loss function	Rule
$L_N = c\overline{\mathrm{FD}} + \overline{\mathrm{FN}}$	$\delta_i = I(v_i > t)$
$L_m = -\sum \delta_i m_i + k\sum(1-\delta_i)m_i + cD$	$\delta_i = I(\overline{m}_i > t)$
$L_D = -k_1 \log(\lvert\Sigma_1\rvert) - k_2 \sum \delta_i f_D(m_i) +$ $+k_3 \sum(1-\delta_i) f_N(m_i) + k_4 D$	no closed form
$L_F = \sum_{\delta_i=1}(-c_1\pi_i + n_i) + c_2 D$	$\delta_i = I(n_i + c_2 \le c_1 \pi_i)$
$L_U = \overline{\mathrm{FDR}}/\alpha - g(D)$	$\delta_i = I(v_i \ge v_j)$ with $j = \max_i \left\{ i : \Delta w_{i,B(i)} \le \alpha\, i/n \right\}$

In summary, all optimal rules are computed on the basis of only a few underlying summaries. This makes it possible to easily consider multiple rules in a data analysis. Critical comparison of the resulting rules leads to a finally reported set of comparisons. In some cases the application leads to a different loss function. Good examples are the loss functions considered in Lin et al. (2004). If a specific loss function arises from a specific case study, it should be used.

The loss function L_D requires the additional summary $\Sigma_1(\delta)$. Let S denote a covariance matrix of the relevant latent variables for *all* genes that are considered for reporting in L_D. The desired $\Sigma_1(\delta)$ for any subset of selected genes is then computed as the marginal correlation matrix for that subset. Let S_δ denote the submatrix defined by choosing rows and columns selected by δ. Let $\lambda_i = 1/\sqrt{S_{ii}}$, and λ_δ denote the vector of λ_i corresponding to the reported genes. We use $\Sigma_1(\delta) = \lambda_\delta[S_\delta]\lambda'_\delta$. This reduces δ_D to a function of $\overline{m}$ and S only.

Figure 3 compares the reported gene lists for the loss functions L_N, L_m and L_D, under the independent model and the dependent model. For many genes the decision remains unchanged across all loss functions. For some genes with high probability of differential expression, but small level of differential expression, and vice versa, the decision depends on whether or not the terms in the loss function are weighted by m_i.

9. DISCUSSION

We have reviewed alternative approaches to addressing problems related to massive multiple comparisons. Starting from traditional rules that control expected FDR, with the expectation over repeated sampling, we have discussed the limitation of the

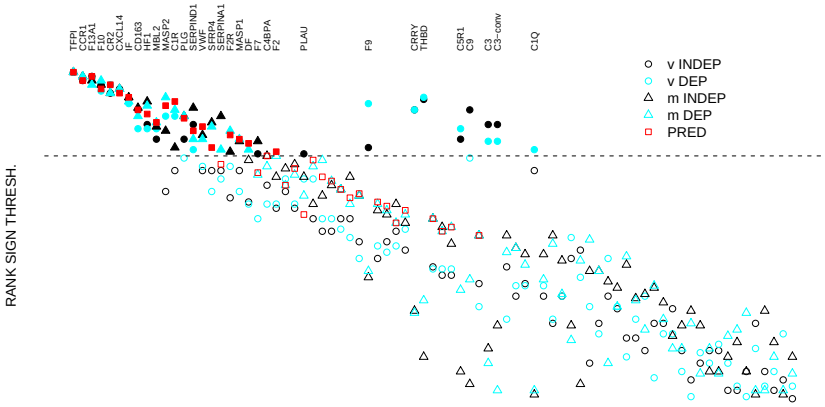


Figure 3: *Comparison of optimal rules for L_N (circles) under the independent model (black) and dependent model (light gray), L_m (triangle) under the independent (black) and dependent model (gray), and L_F (square). For each gene (horizontal axis) symbols are plotted against the rank of the corresponding significance threshold statistic (vertical axis). The reported genes under each rule are the top 20 ranked genes (above the dashed horizontal line). The symbols corresponding to selected genes are filled. The names of selected genes are shown on the vertical axis. Genes are sorted by average rank under the five criteria.*

interpretation of these decisions as Bayesian rules. We have argued for a solution of the massive multiple comparisons as a decision problem, and we have shown how this is implemented in structured probability models including dependence across genes. Most, but not all, loss functions lead to rules that can be defined in terms of a significance thresholding function, $S(data)$, as proposed in Storey (2005).

The proposed approaches are all based on casting the multiple comparison problem as a decision problem and thus inherit the limitations of any decision theoretic solution. In particular, we recognize that not all research is carried out to make a decision. The decision theoretic perspective might be inappropriate when an experiment is carried out for more heuristic learning, without being driven by specific decisions. Also, all arguments were strictly model-based. Even results that apply for any probability model still need a specific probability model to implement related approaches. Like any model-based inference, the implementation involves the often difficult tasks of prior elicitation, and the choice of appropriate parametric models. Additionally, our arguments require the choice of a utility function. A common feature of early stage hypothesis generating experiments is that they serve multiple needs. We might want to carry out a microarray experiment to suggest interesting genes and proteins for further investigation, to propose suitable candidates for a correlation study with clinical outcomes, and also to simply understand molecular mechanisms.

We caution against over-interpretation of results based on highly structured probability models and often arbitrary choices of utility functions. Data analysis for

high-throughput gene expression data is particularly prone to problems arising from data pre-processing. Often it is more important to understand the pre-processing of the raw data, and correct it if necessary, than to spend effort on sophisticated modeling. Specifically related to the dependent probability model, it is important to acknowledge limitation of pathway information that is used to select the set of possible parent nodes N_i in (8) when constructing the dependent probability model. Pathway information does not necessarily describe relations among transcript levels, although it carries some information about it.

We have focused on the inference problem of reporting lists of differentially expressed genes, and inference on massive multiple comparisons in general. A similar framework, using the same probability models and loss functions, can be used for other decision problems related to the same experiments. For example, one could consider choosing the sample size for a future experiment (sample size selection), ranking genes or selecting a fixed set of genes (for a panel of biomarkers).

Acknowledgments

Research was supported by NIH/NCI grant 1 R01 CA075981 and by NSF DMS034211. Kenneth Rice was supported by Career Development Funding from the Department of Biostatistics, University of Washington

REFERENCES

Baggerly, K. A., Morris, J. S., Wang, J., Gold, D., Xiao, L. C. and Coombes, K. R. (2003) A comprehensive approach to analysis of MALDI-TOF proteomics spectra from serum samples. *Proteomics*, **9**, 1667–1672.

Bayarri, M. J. and Berger, J. O. (2000) P values for composite null models. *J. Amer. Statist. Assoc.* **95**, 1127–1142.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* **57**, 289–300.

Berry, D. A. and Hochberg, Y. (1999) Bayesian perspectives on multiple comparisons. *J. Statist. Planning and Inference* **82**, 215–227.

Berry, S. and Berry, D. (2004) Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics* **60**, 418–426.

Casella, G. and Berger, R. L. (1987) Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82**, 106–111.

Cohen, A. and Sackrowitz, H. B. (2005) Decision theory results for one-sided multiple comparison procedures. *Ann. Statist.* **33**, 126–144.

— (2006) More on the inadmissibility of step-up. *Tech. rep.*, Rutgers University.

Do, K., Müller, P. and Tang, F. (2005) A bayesian mixture model for differential gene expression. *Appl. Statist.* **54**, 627–644.

Duncan, D.B. (1965) A Bayesian Approach to Multiple Comparisons, *Technometrics* **7**, 171-222.

Efron, B. and Tibshirani, R. (2002) Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, **23**, 70–86.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96**, 1151–1160.

Fortini, M., Liseo, B., Nuccitelli, A. and Scanu, M. (2001) On bayesian record linkage. *Research in Official Statistics*, **4**, 185–198.

— (2002) Modelling issues in record linkage: a bayesian perspective. In *Proceedings of the ASA meeting*. ASA, ASA.

Genovese, C., Lazar, N. and Nichols, T. (2002) Thresholding of statistical maps in neuroimaging using the false discovery rate. *NeuroImage*, **15**, 870–878.

Genovese, C. and Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. B* **64**, 499–518.

— (2003) Bayesian and Frequentist Multiple Testing. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press, pp. 145-162.

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

Keeney, R. L., Raiffa, H. A. and Meyer, R. F. C. (1976) *Decisions With Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley

Lin, R., Louis, T. A., Paddock, S. M. and Ridgeway, G. (2004) Loss function based ranking in two-stage hierarchical models. *Tech. rep.*, Johns Hopkins University, Dept. of Biostatistics.

Lindley, D. V. (1971) *Making decisions*. New York: Wiley

Müller, P. and Parmigiani, G. (2006) Modeling dependent gene expression. *Tech. rep.*, M.D. Anderson Cancer Center.

Müller, P., Parmigiani, G., Robert, C. and Rouseau, J. (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Amer. Statist. Assoc.* **99**, 990-1001.

Newton, M., Noueriry, A., Sarkar, D. and Ahlquist, P. (2004) Detecting differential gene expression with a semiparametric heirarchical mixture model. *Biostatistics*, **5**, 155–176.

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

Parmigiani, G., Garrett, E. S., Anbazhagan, R. and Gabrielson, E. (2002) A statistical framework for expression-based molecular classification in cancer. *J. Roy. Statist. Soc. B* **64**, 717–736.

Ronning, G. and Kukuk, M. (1996) Efficient estimation of ordered probit models. *J. Amer. Statist. Assoc.* **91**, 1120–1129.

Scott, J. and Berger, J. (2003) An exploration of aspects of bayesian multiple testing. *Tech. rep.*, Duke University, ISDS.

Sellke, T., Bayarri, M. J. and Berger, J. O. (2001) Calibration of p values for testing precise null hypotheses. *Amer. Statist.* **55**, 62–71.

Shen, W. and Louis, T. A. (1998) Triple-goal estimates in two-stage hierarchical models. *J. Roy. Statist. Soc. B* **60**, 455–471.

Shen, X., Huang, H.-C. and Cressie, N. (2002) Nonparametric hypothesis testing for a spatial signal. *J. Amer. Statist. Assoc.* **97**, 1122–1140.

Storey, J. (2002) A direct approach to false discovery rates. *J. Roy. Statist. Soc. B* **64**, 479–498.

— (2005) The optimal discovery procedure I: a new approach to simultaneous signficance testing. *Tech. Rep. 259*, University of Washington.

Storey, J. D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31**, 2013–2035.

Storey, J. D., Taylor, J. E. and Siegmund, D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. Roy. Statist. Soc. B* **66**, 187–205.