Biostatistics (2014), **15**, 1, *pp.* 60–73
doi:10.1093/biostatistics/kxt026
Advance Access publication on August 8, 2013

Prior robust empirical Bayes inference for large-scale data by conditioning on rank with application to microarray data

J. G. LIAO*
Division of Biostatistics and Bioinformatics, Penn State University, Hershey, PA 17033, USA
jliao@phs.psu.edu

TIMOTHY MCMURRY
Division of Biostatistics, University of Virginia, Charlottesville, VA 22908, USA

ARTHUR BERG
Division of Biostatistics and Bioinformatics, Penn State University, Hershey, PA 17033, USA

Summary

Empirical Bayes methods have been extensively used for microarray data analysis by modeling the large number of unknown parameters as random effects. Empirical Bayes allows borrowing information across genes and can automatically adjust for multiple testing and selection bias. However, the standard empirical Bayes model can perform poorly if the assumed working prior deviates from the true prior. This paper proposes a new rank-conditioned inference in which the shrinkage and confidence intervals are based on the distribution of the error conditioned on rank of the data. Our approach is in contrast to a Bayesian posterior, which conditions on the data themselves. The new method is almost as efficient as standard Bayesian methods when the working prior is close to the true prior, and it is much more robust when the working prior is not close. In addition, it allows a more accurate (but also more complex) non-parametric estimate of the prior to be easily incorporated, resulting in improved inference. The new method's prior robustness is demonstrated via simulation experiments. Application to a breast cancer gene expression microarray dataset is presented. Our R package *rank.Shrinkage* provides a ready-to-use implementation of the proposed methodology.

Keywords: Bayesian shrinkage; Confidence intervals; Ranking bias; Robust multiple estimation.

1. Introduction

Large-scale technologies, which measure many similar entities in parallel, have emerged as an important tool in biomedical research. For example, the expression level of thousands of genes is compared between cancer and normal tissues in microarray experiments. In genome-wide association studies, the log odds ratio of the association of disease status (disease vs. control) and single-nucleotide polymorphism

*To whom correspondence should be addressed.

(SNP) frequency is estimated for thousands or millions of SNPs in a single case–control study. There are two prominent features in large-scale data. First, different parameters (e.g. difference in expression levels between cancer and normal tissues for different genes) are often studied with the same set of subjects and using the same design. Second, a large majority of the underlying parameters are 0. Because of this, the unknown parameters can be profitably modeled as random effects in an empirical Bayes framework. A popular model is to treat the large number of parameters as draws from a spike-and-slab prior distribution that is a mixture of a large mass at 0 and a non-zero component. The parameters in the spike-and-slab prior can be estimated from the many parallel measurements, resulting in an empirical Bayes analysis that borrows information across different genes or SNPs. The empirical Bayes framework automatically adjusts for the large number of hypothesis tests or effect estimates. The application of empirical Bayes to large-scale testing naturally leads to the false discovery rate (FDR) and the local FDR (Benjamini and Yekutieli, 2005; Efron *and others*, 2001; Newton *and others*, 2004; Efron, 2008). The application to parallel estimation (e.g. of the log fold changes in expression level) in the microarray context includes Newton *and others* (2001), Kendziorski *and others* (2003), and Smyth (2004). There is substantial literature in this area and the reader is referred to Efron (2010) for a summary of the state of the art in the empirical Bayes approach to large-scale inference and for complete references.

This paper focuses on estimating the effect sizes, the log fold change in gene expression level in microarray data, for example. We show that a popular empirical Bayes random effects model can lead to poor performance if the form of the prior is mis-specified; this has important practical implications because in real applications we are never certain of the correct distributional form, especially in the tails of the distribution, which often produce the most interesting observations. Motivated by this sensitivity to the form of the random effects model, we propose a new rank-conditioned inference in which shrinkage and confidence intervals are based on the distribution of the measurement error conditioned on the rank of the data instead of on the data themselves as in a traditional Bayesian posterior. The primary advantage of the rank-conditioned method is that it is almost as efficient as standard Bayesian methods when the working prior is close to the true prior, and it is much more robust when the working prior is not close. In particular, the proposed method provides efficient and valid inference even when the working random effects model substantially deviates from the true model in location. The proposed method can, therefore, substantially improve empirical Bayes inference for microarray studies as well as other large-scale data such as for genome-wide association studies and flow cytometry.

To put the rank-conditioned method in the context of the broader literature, we note the following. First, we condition on rank of the observed data in order to obtain more robust estimation of effect size. This is different from Noma *and others* (2010), whose aim is to better rank the effect sizes. Second, the idea of combining a Bayes formulation with rank-based likelihood has previously been proposed and studied in other context; for example, Dunson and Taylor (2005) use the idea for estimating quantiles, Gu and Ghosal (2009) for estimating receiver operating characteristic curves, and Hoff (2007) in estimating semi-parametric copula. Large-scale data is an area where this idea can be more profitably used because rank of the observed data and the observed data themselves are closely correlated. We are, therefore, able to take advantage of the robust property of the rank method with little loss of efficiency compared with standard empirical Bayes. Third, improved robustness can also be achieved through a more diffuse prior for the random effects. For example, Do *and others* (2005) and Kim *and others* (2009) combine Dirichlet process with spike-and-slab prior in a non-parametric Bayes model for random effects. A more diffuse prior, however, necessarily weakens the effectiveness of shrinkage and information borrowing as seen in the simulation study in Section 4.2.

The paper is organized as follows. Section 2 describes a model for large-scale microarray data analysis. Section 3 presents our proposed ranked-conditioned inference. Section 4 consists of simulation studies assessing the performance of rank-conditioned inference. Section 5 applies the proposed method to a breast cancer microarray dataset. Section 6 is discussion.

2. EMPIRICAL BAYES MODEL FOR LARGE-SCALE DATA

For concreteness, throughout the remainder of the paper, we focus on estimating the standardized effect size in case–control microarray experiments; application of our method in other large-scale data, such as genome-wide association studies, is similar.

We begin by describing an empirical Bayes model for the log fold change in gene expression. Let $y_{ij}^{[1]}$ and $y_{ij}^{[2]}$ be the log expression level of the ith gene for the jth subject in the cancer and normal group, respectively. The total number of genes is n so that $i = 1, \ldots, n$. We start with the model

$$y_{ij}^{[1]} \sim \mathcal{N}(\alpha_i^{[1]}, \rho_i^2), \quad j = 1, \ldots, m_i^{[1]},$$
$$y_{ij}^{[2]} \sim \mathcal{N}(\alpha_i^{[2]}, \rho_i^2), \quad j = 1, \ldots, m_i^{[2]},$$

where ρ_i^2 is the variance of the ith gene expression common for the cancer and normal groups, and $m_i^{[1]}$ and $m_i^{[2]}$ are the respective sample sizes. The quantity $\alpha_i^{[1]} - \alpha_i^{[2]}$ is the average (log) fold change (Guo *and others*, 2006; Choe *and others*, 2005). Let $\bar{y}_i^{[1]}$ be the mean of $y_{ij}^{[1]}$ over j, and similarly let $\bar{y}_i^{[2]}$ be the mean of $y_{ij}^{[2]}$. It then follows that

$$z_i \equiv \rho_i^{-1}(\bar{y}_i^{[1]} - \bar{y}_i^{[2]}) \sim \mathcal{N}(\theta_i, \sigma_i^2),$$

where

$$\theta_i \equiv \rho_i^{-1}(\alpha_i^{[1]} - \alpha_i^{[2]})$$

is the standardized log fold change and $\sigma_i^2 = (m_i^{[1]})^{-1} + (m_i^{[2]})^{-1}$. Note that $m_i^{[1]}$ and $m_i^{[2]}$ typically do not vary much from gene to gene in a microarray experiment so that variance σ_i^2 should be relatively constant across i.

The first stage of our empirical Bayes model is

$$z_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ independently for $i = 1, \ldots, n$. In application, the ρ_i in the definition of z_i will be replaced by its pooled estimate using $y_{ij}^{[1]}$ and $y_{ij}^{[2]}$ and $z_i \sigma_i^{-1}$ will then follow a t-distribution. For simplicity, we shall use normal error model (2.1), since the t degrees of freedom, $m_i^{[1]} + m_i^{[2]} - 2 = 207$, is large for the breast cancer data in Section 6. For a smaller $m_i^{[1]} + m_i^{[2]} - 2$, a modified version of (2.1) based on a non-central t-distribution can be used instead. For a genome-wide association study, z_i can be the estimated log odds ratio from a logistic regression for the association between disease status and the ith SNP, θ_i be the unknown true log odds ratio and σ_i be the standard error of estimate z_i. Next, we will model θ_i as independent random draws from a prior π. Given prior π, the Bayesian inference for θ_i is based on the posterior distribution of θ_i given z_i with density

$$f(\theta_i | z_i, \pi) \propto \pi(\theta_i)\phi_i(z_i - \theta_i), \tag{2.2}$$

where ϕ_i is a $\mathcal{N}(0, \sigma_i^2)$ density. The posterior mean $\hat{\theta}_i^{\text{Bayes}} \equiv \mathrm{E}[\theta_i | z_i]$ is a standard Bayes estimator of θ_i and the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution provide the $1 - \alpha$ confidence limits ℓ_i^{Bayes} and u_i^{Bayes}.

The prior π, however, is unknown. Empirical Bayes analysis uses a working prior π_p in place of π with the parameters in π_p estimated from data $z_1, \ldots, z_n$ usually via maximum likelihood (Morris, 1983).

Our parametric working prior π_p is a three-component mixture

$$\pi_p = (1 - \eta_1 - \eta_2)\delta_0 + \eta_1 \mathcal{N}(\mu_1, \omega_1^2) + \eta_2 \mathcal{N}(\mu_2, \omega_2^2), \tag{2.3}$$

where δ_0 is the delta function (point mass) at 0, and $\mathcal{N}(\mu_1, \omega_1^2)$ and $\mathcal{N}(\mu_2, \omega_2^2)$, respectively, model the up- and down-regulated genes. This working prior is the same as that in Noma *and others* (2010) with one important difference: we use (2.3) to model the distribution of the standardized differences θ_i instead of the raw differences $\alpha_i^{[1]} - \alpha_i^{[2]}$. We show in Section 5 that modeling the standardized differences θ_i as draws from a common prior leads to a much better fit to a breast cancer microarray dataset.

An important practical advantage of working prior (2.3) is that the posterior distribution $\theta_i|z_i$ is also a mixture of the same form as (2.3) (see Noma *and others*, 2010; Muralidharan, 2010 for analytical formula), which makes programming much easier and computing time manageable for large-scale problems. Spike-and-slab priors such as (2.3) have been used in variable selection and shrinkage estimation (see, e.g. Ishwaran and Rao, 2005) and have played a prominent role in multiple testing (Efron *and others*, 2001).

3. Rank-conditioned inference

3.1 *Rank-conditioned shrinkage*

For our basic model (2.1), we have

$$\mathrm{E}[\theta_i|z_i] + \mathrm{E}[\varepsilon_i|z_i] = z_i.$$

The Bayesian estimate $\hat{\theta}_i^{\text{Bayes}} = \mathrm{E}[\theta_i|z_i]$ can also be written as

$$\hat{\theta}_i^{\text{Bayes}} = z_i - \mathrm{E}[\varepsilon_i|z_i],$$

which reflects the fact that the conditional mean of ε_i, given the observed z_i, is no longer 0.

For the dataset $z_1, \dots, z_n$, let $r(i)$ be the rank of z_i among $z_1, \dots, z_n$. Our rank-conditioned inference for θ_i is based on the conditional distribution

$$\varepsilon_i|r(i) = j, \tag{3.1}$$

where j is the realized value of rank $r(i)$. The rank-conditioned shrinkage estimator is defined as

$$\hat{\theta}_i^{\text{rank}} \equiv z_i - \mathrm{E}[\varepsilon_i|r(i) = j], \tag{3.2}$$

where $\mathrm{E}[\varepsilon_i|r(i) = j]$ is the conditional mean of the error ε_i, given that z_i has rank j among $z_1, \dots, z_n$. Given prior π, a draw from (3.1), $\varepsilon_{i,j}^*$, which is error ε_i^* conditional on $r(i) = j$, can be generated using the following three steps:

Step 1: Generate θ_i^* from density π independently for $i = 1, \dots, n$. Let $z_i^* = \theta_i^* + \varepsilon_i^*$, where $\varepsilon_i^* \sim \mathcal{N}(0, \sigma_i^2)$.

Step 2: Let $r^*(i)$ be the rank of z_i^* among $z_1^*, \dots, z_n^*$.

Step 3: Repeat Steps 1–2 until $r^*(i) = j$. Then output $\varepsilon_{i,j}^* = \varepsilon_i^*$.

Theorem 3.1 Let $\theta_i \sim \pi$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ independently for $i = 1, \dots, n$. Let z_i be defined as in model (2.1). Then $\hat{\theta}_i$ is unbiased in the sense that

$$\mathrm{E}[\hat{\theta}_i^{\text{rank}} - \theta_i|r(i) = j] = 0,$$

for any given i and j, when the expectation is evaluated with respect to the joint distribution of $\theta_1, \dots, \theta_n$ and $\varepsilon_1, \dots, \varepsilon_n$.

Proof. Theorem 3.1 follows directly from definition (3.2) and (2.1):

$$\mathrm{E}[\hat{\theta}_i^{\mathrm{rank}} | r(i) = j] = \mathrm{E}[z_i - \varepsilon_i | r(i) = j] = \mathrm{E}[\theta_i | r(i) = j].$$

Theorem 3.1 says that $\hat{\theta}_i^{\mathrm{rank}}$ corrects the ranking bias, a concept discussed as in Jeffries (2009).

In addition to point estimate $\hat{\theta}_i^{\mathrm{rank}}$, the proposed method provides a natural confidence interval for θ_i. Let $\ell_{i,j}$ and $u_{i,j}$ satisfy

$$\Pr\{\ell_{i,j} \leqslant \varepsilon_{i,j}^* \leqslant u_{i,j}\} \geqslant 1 - \alpha. \tag{3.3}$$

It follows that

$$\Pr\{z_i - u_{i,j} < \theta_i < z_i - \ell_{i,j}\} \geqslant 1 - \alpha.$$

We have, therefore, shown that the interval

$$(z_i - u_{i,j}, z_i - \ell_{i,j}) \tag{3.4}$$

contains the realized θ_i with $1 - \alpha$ probability, given $r(i) = j$ when $\theta_1, \ldots, \theta_n$ and $\varepsilon_1, \ldots, \varepsilon_n$ are drawn as in Theorem 3.1. □

Conditioning on $r(i) = j$ in the rank-conditioned shrinkage estimator (3.2) and confidence limits (3.4) is in fact closely related to conditioning on z_i itself in standard Bayes, as a larger z_i generally corresponds to a higher rank. More specifically, let G_n be the empirical distribution of $z_1, \ldots, z_n$. Suppose that σ_i, $i = 1, \ldots, n$, can be modeled as draws from some distribution F. It then follows from Glivenko–Cantelli theorem that G_n converges uniformly to G, the distribution of $\theta + \varepsilon$ with $\theta \sim \pi$, $\sigma \sim F$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. In such a case, conditioning on $r(i) = j$ is almost the same as conditioning on $z_i \approx G^{-1}(j/n)$ so long as j/n is not close to 0 or 1 (the difference can be more substantial for j/n close to 0 or 1). The proposed rank-conditioned inference, however, can be much more robust than standard empirical Bayes against mis-specification of π. For this, we have the following result.

THEOREM 3.2 Let $\theta_i \sim \pi$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ independently for $i = 1, \ldots, n$. Let z_i be defined as in model (2.1). In the case where the σ_i are equal, conditional distribution (3.1) and consequently rank-conditioned estimator (3.2) and confidence limits (3.4) remain the same (and valid) when the true prior density $\pi(\theta)$ is replaced by density $\pi(\theta - a)$ for any given constant a.

Proof. The proof is straightforward. Let $\theta_i^* \sim \pi$ and $z_i^* = \theta_i^* + \varepsilon_i^*$ as in the three steps above. When the σ_i are equal, the rank of z_i^* is not changed when θ_i^* are all translated by a constant a, so the distribution of $\varepsilon_{i,j}^*$ does not change. Theorem 3.2 then follows.

For unequal $\sigma_1, \ldots, \sigma_n$, Theorem 3.2 remains approximately valid so long as the variation in $\sigma_1, \ldots, \sigma_n$ is small. Section 5 demonstrates that the rank-based shrinkage is in general more robust, not just against location shift. This is a unique feature of rank-conditioned shrinkage: the ranking bias $\mathrm{E}[\varepsilon_i | r(i) = j]$ is negative for lower ranked j and positive for higher ranked j even when evaluated under a badly specified prior. In the three steps for generating $\varepsilon_{i,j}^*$ at the beginning of this section, the prior π determines which $\varepsilon_i^* \sim \mathcal{N}(0, \sigma_i^2)$ will be output as $\varepsilon_{i,j}^*$. As such, the effect of a grossly mis-specified π on $\hat{\theta}_i^{\mathrm{rank}}$ remains limited. A grossly mis-specified π can, however, have a much larger distorting impact on Bayes shrinkage estimator $\hat{\theta}_i^{\mathrm{Bayes}}$.

Finally, a confidence interval such as (3.4) that adjusts for ranking $r(i)$ can be crucial for valid inference; Benjamini and Yekutieli (2005) show that the unadjusted marginal confidence interval of θ_i can have coverage probability that differs substantially from the nominal coverage for top-ranked parameters selected based on the same data. They propose the false coverage rate controlled confidence interval as a solution to this problem. As shown in Efron (2010, pp. 230–233), however, this interval can differ markedly from the corresponding Bayes interval and can be frighteningly wide. Westfall (2005) suggests constructing empirical Bayes confidence intervals centered at the shrunken estimators; the same idea is used and further developed in Qiu and Gene Hwang (2007) and in Ghosh (2009). Our interval is similar, but is instead based on rank-conditioned shrinkage. It is generally very close to the corresponding Bayes interval when the working prior is close to the true prior. □

3.2 *Non-parametric update of the parametric prior*

In Section 2, a parametric working prior π_p is used in empirical Bayes to capture the primary structure of π. For the rank-conditioned method, we propose a non-parametric update of density π_p to density π_{pu} by formula

$$\pi_{pu}(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} f(\theta|z_i;\pi_p),$$

where the posterior density $f(\theta|z_i;\pi_p)$ is given by (2.2) with π replaced by π_p and with θ_i replaced by a generic θ. The parameters in π_p will take the values of their maximum likelihood estimates. π_{pu} can be interpreted as the average of the posterior densities for θ, given z_i with prior $\theta \sim \pi_p$. Vardi *and others* (1985) use a similar update to improve the estimated image density in positron emission tomography. They show that it is one expectation-maximization iteration and therefore always increases the (marginal) likelihood of $z_1, \ldots, z_n$. See also Eggermont and LaRiccia (1997). The use of π_{pu} in place of π_p does not significantly increase the computational burden for rank-conditioned inference. The density π_{pu} could potentially be further updated but the analytical complexity and computational cost will increase drastically.

3.3 *Algorithm and implementation*

The three-step algorithm for drawing $\varepsilon^*_{i,j}$ in Section 3.1 is greatly simplified for the special case of $\sigma_i = \sigma$ for all i because the distribution of $\varepsilon^*_{i,j}$ depends only on the rank j and not on i. Under this condition, Steps 2 and 3 become

> Let $z^*_{[1]} \leqslant \cdots \leqslant z^*_{[n]}$ be the order statistics of $z_1, \ldots, z_n$. Let $\varepsilon^*_{[j]}$ be the ε^*_i that corresponds to $z^*_{[j]}$. Output $\varepsilon^*_{[j]}$ for $j = 1, \ldots, n$.

In this way, one round of Steps 1–3 generates a complete and independent set of $\varepsilon^*_{[1]}, \ldots, \varepsilon^*_{[n]}$. The $\mathrm{E}[\varepsilon_i|r(i) = j]$ in (3.2) is now simply $\mathrm{E}[\varepsilon^*_{[j]}]$ and the $\ell_{i,j}$ and $u_{i,j}$ in (3.3) are defined by $\Pr\{\ell_j \leqslant \varepsilon^*_{[j]} \leqslant u_j\} \geqslant 1-\alpha$ irrespective of the value of i.

For most large-scale problems, the values of error standard deviation σ_i may not be constant but they are not far apart (say within a factor of 3 or 4) because of the inherent common design structure. For the microarray example in Section 2, σ_i for standardized difference θ_i depends only on sample size $m_i^{[1]}$ and $m_i^{[2]}$. Therefore, σ_i does not differ too much unless the number of missing data points varies dramatically between genes. Similarly, in genome-wide association study, each SNP is compared between the same set of cases and controls. For such dataset, we can partition the n observations, $z_1, \ldots, z_n$, into several sub-groups so that σ_i for observations within each sub-group varies within a factor of 1.5, for example. The simplified algorithm above can then be applied to each subgroup separately as an approximation. Monte Carlo Markov chain type of algorithm is under investigation to efficiently sample from the rank-conditioned distribution (3.1) without requiring σ_i to be constant.

Table 1. *Mean square error of the three methods under different model mis-specification*

Working prior	$\text{MSE}_{\text{Bayes}}$ with π_p	MSE_{rank} with π_p	$\text{MSE}_{\text{Bayes}}$ with π_{pu}
Same as true prior	0.675	0.677	0.677
$\mu_{1p} = \mu_1 + \mu_1/5$ $\mu_{2p} = \mu_2 + \mu_2/5$	0.763	0.710	0.681
$\mu_{1p} = \mu_1 - \mu_1/5$ $\mu_{2p} = \mu_2 + \mu_2/5$	0.762	0.685	0.683
$\omega_{1p} = 1.25\omega_1$ $\omega_{2p} = 1.25\omega_2$	0.702	0.686	0.683
$\omega_{1p} = \omega_1/1.25$ $\omega_{2p} = \omega_2/1.25$	0.699	0.681	0.679

4. ASSESSING PERFORMANCE OF RANK-CONDITIONED INFERENCE

4.1 *Simulation study* 1

This example is adapted from Efron (2010, pp. 230–233). Let $n = 10^4$ and $\sigma_i = 1$ for all i for model (2.1). The true π for random effects θ_i is (2.3) with $\eta_1 = \eta_2 = 0.1$, $\mu_1 = -3$, $\mu_2 = 3$, $\omega_1 = \omega_2 = 1$. These parameter values are chosen to have a moderate Bayes shrinkage effect. Monte Carlo simulation is conducted to compare the Bayes shrinkage estimates $\hat{\theta}_i^{\text{Bayes}}$ and rank-conditioned shrinkage estimates $\hat{\theta}_i^{\text{rank}}$ under five different specifications of working prior π_p. These working priors π_p have the same parametric form of (2.3) but with possibly different values of μ_{1p}, μ_{2p}, ω_{1p}, ω_{2p} as given in various rows of Table 1. Parameters not listed are the same as in true prior. For example, $\eta_1 = \eta_2 = 0.1$ for all the five working priors. Our simulation study is conducted as follows:

Step 1: Generate θ_i, $i = 1, \ldots, n$, from prior π. Let $z_i = \theta_i + \varepsilon_i$ as in model (2.1).

Step 2: Let $z_{[1]}, \ldots, z_{[n]}$ be the order statistics of $z_1, \ldots, z_n$. Let $\theta_{[j]}$ be the θ_i corresponding to $z_{[j]}$. The $\theta_{[j]}$ can, therefore, refer to different θ_i for different realizations of $z_1, \ldots, z_n$.

Step 3: Compute empirical Bayes estimate $\hat{\theta}_{[j]}^{\text{Bayes}}$ under working model π_p. Compute the rank-conditioned estimate $\hat{\theta}_{[j]}^{\text{rank}}$ under working model π_p and its non-parametric update π_{pu}, respectively.

Step 4: Let $S = \{j : j = 1, \ldots, 500,\ j = n - 501, \ldots, n\}$. Calculate the mean square loss

$$\frac{1}{1000} \sum_{j \in S} \left(\hat{\theta}_{[j]} - \theta_{[j]} \right)^2,$$

for estimator $\hat{\theta}_{[j]} = \hat{\theta}_{[j]}^{\text{Bayes}}$ and estimator $\hat{\theta}_{[j]} = \hat{\theta}_{[j]}^{\text{rank}}$ for both π_p and π_{pu}. We only include the 500 lowest and 500 highest j in S because these $\theta_{[j]}$ are most interesting in large-scale data analysis.

Steps 1–4 are replicated 1000 times and the mean square error $\text{MSE}_{\text{Bayes}}$ for Bayes method and mean square error MSE_{rank} for rank-conditioned inference are estimated by averaging the squared error loss over these replications. The estimated $\text{MSE}_{\text{Bayes}}$ and MSE_{rank} values are given in Table 1. In row 1, the working prior is the same as the true prior and the standard Bayes is therefore optimal. The rank-conditional inference under both π_p and π_{pu} shows little loss of efficiency with almost the same mean square error. In rows 2 and 3, μ_{1p} and μ_{2p} in the working prior are shifted away from μ_1 and μ_2 in the true prior. The $\text{MSE}_{\text{Bayes}}$ increases noticeably with this model mis-specification. The rank-conditioned inference, especially under π_{pu}, proves robust with a much smaller change in MSE_{rank} from row 1. In rows 4 and 5, ω_{1p} and ω_{2p} in the working prior are inflated or shrunk from ω_1 and ω_2 in the true prior. Again, rank-conditioned inference is more robust.

4.2 *Simulation study* 2

For simulation 2, we continue to let $n = 10^4$ and $\sigma_i = 1$ for all i in model (2.1). The true prior π of θ_i now has the form

$$\pi = 0.8\delta_0 + 0.1\{-\Gamma(k,\beta) - 5 + k\beta\} + 0.1\{\Gamma(k,\beta) + 5 - k\beta\}, \tag{4.1}$$

where $\Gamma(k,\beta)$ is the Gamma distribution with shape k and scale β. We take $\beta = k^{-1/2}$ so that $-\Gamma(k,\beta) - 5 + k\beta$ always has mean -5 and variance 1 and $\Gamma(k,\beta) + 5 - k\beta$ always has mean 5 and variance 1 for any k. For the simulation study, we generate $\theta_i \sim \pi$ and $z_i = \theta_i + \varepsilon_i$, $i = 1,\ldots,n$, as specified by (2.1) and (4.1). Let $z_{[j]}$ be the jth order statistic of $z_1,\ldots,z_n$ and let $\theta_{[j]}$ be the corresponding θ_i. For every dataset $z_1,\ldots,z_n$, we compute a point estimate and 90% confidence limits for $\theta_{[j]}$, $j = 1,\ldots,100$, using three different methods. Method 1 is Bayes estimate $\hat{\theta}_i^{\text{Bayes}}$ and confidence limits ℓ_i^{Bayes} and u_i^{Bayes} in Section 2 using π_p in (2.3) as the working prior. In accordance with empirical Bayes, parameters $\eta_1, \eta_2, \mu_1, \mu_2, \omega_1^2, \omega_2^2$ in π_p are substituted by their maximum likelihood estimates using data $z_1,\ldots,z_n$ under models (2.1) and (2.3). We shall call this parametric Bayes method. Method 2 is also based on Bayes posterior but with a more diffuse prior using Dirichlet process mixture. Let $\text{DP}(G_{j0})$ be the Dirichlet process with base distribution G_{j0} and scaling parameter 1 and let $f_j \sim \text{DP}(G_{j0})$ be a (random) distribution drawn from this Dirichlet process. Following Do *and others* (2005) and Dunson (2010), we take the more diffuse prior, π_{DP}, as

$$\pi_{\text{DP}} = (1 - \eta_1 - \eta_2)\delta_0 + \eta_1 f_1 + \eta_2 f_2,$$

where f_j, $j = 1, 2$ is generated as

$$f_j \sim \text{DP}(G_{j0}),$$
$$G_{j0}(\mu,\omega^2) = N(\mu|\mu_j,\omega^2)\text{Inv-Gamma}(\omega^2|a_j,b_j).$$

To be consistent with Method 1, η_1 and η_2 in π_{DP} and μ_j in $G_{j0}(\mu,\omega^2)$ are substituted by their maximum likelihood estimates under models (2.1) and (2.3) as in Method 1. For the inverse gamma distribution, we choose shape parameter $a_j = 2$ and scale parameter $b_j = \hat{\omega}_j^2$, where $\hat{\omega}_j^2$ is maximum likelihood estimate of variance ω_j^2 in (2.3), so that the mean of the inverse gamma equals $\hat{\omega}_j^2$. Note that the prior π_{DP} is considerably more diffuse and less informative than π_p due to the extra variation in $f_j \sim \text{DP}(G_{j0})$. We call Method 2 non-parametric DP Bayes. Method 3 is the proposed rank-conditioned inference under π_{pu}, the non-parametric update of π_p. Again, the parameters in π_p are substituted by their maximum likelihood estimates. Let $\hat{\theta}_{[j]}$ be the point estimate and $(\ell_{[j]}, u_{[j]})$ be the confidence limits of $\theta_{[j]}$ from one of the three methods. Let $1(\cdot)$ be the indicator function. The mean square error and actual coverage rate for parameter $\theta_{[j]}$, $j = 1,\ldots,100$, are estimated by averaging

$$(\hat{\theta}_{[j]} - \theta_{[j]})^2 \quad \text{and} \quad 1(\ell_{[j]} \leqslant \theta_{[j]} \leqslant u_{[j]})$$

over 1000 replications of $z_1,\ldots,z_n$.

The simulation study is conducted for $k = 1000$, $k = 8$, and $k = 2$ with prior π given by (4.1); the results are given in Figures 1–3, respectively. In each figure, the left panel shows the estimated root MSE for the lowest 100 ranked genes, $\theta_{[j]}$ for $j = 1,\ldots,100$, and the right panel shows the estimated actual coverage rate of nominal 90% confidence intervals for $\theta_{[i]}$. Figure 1 shows the case where $k = 1000$; when k is large, the normal distribution in the working prior π_p approximates $\Gamma(k,\beta)$ in true prior π extremely well. As expected, the parametric Bayes performs the best among the three methods with the smallest mean square errors and close (to nominal) actual coverage rates for all $\theta_{[j]}$, $j = 1,\ldots,100$. The non-parametric Bayes with Dirichlet process prior performs poorly as the mean square errors are large and the actual

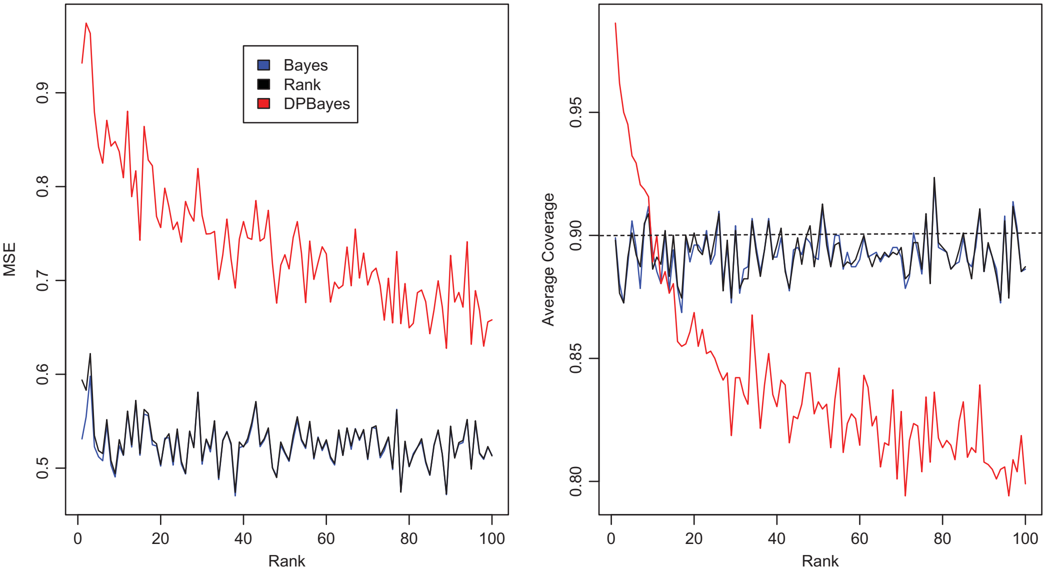


Fig. 1. Simulation study of Section 4.2 with the correct prior in (4.1) and parameter $k = 1000$. The left panel is the root mean square error for parameter estimate $\hat{\theta}_{[j]}$, $j = 1, \ldots, 100$, and the right panel is the actual coverage rate of confidence intervals for $\theta_{[j]}$ at the 90% nominal level. Parametric empirical Bayes and rank-conditional inference perform similarly in this case with smaller mean square error and correct actual coverage rates. The non-parametric Bayes model with Dirichlet prior performs considerably worse compared with the other two methods in both mean square error and actual coverage rate.

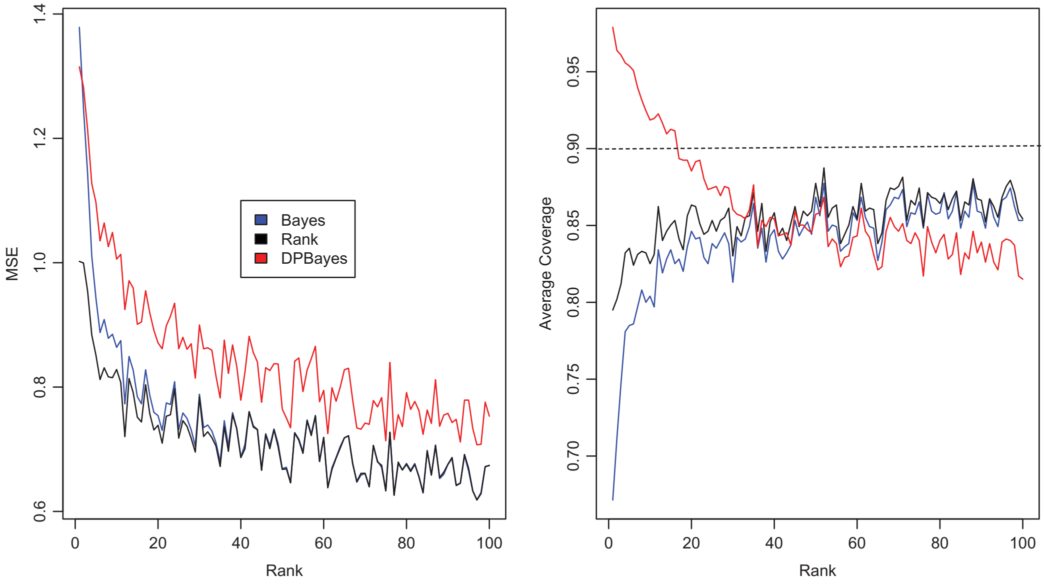


Fig. 2. Simulation study of Section 4.2 with the correct prior in (4.1) and parameter $k = 8$. The rank-conditioned method has the smallest root mean square errors and close actual confidence rates. The parametric Bayes and non-parametric Bayes with Dirichlet prior perform badly due to the large mean square error especially for $j < 10$.

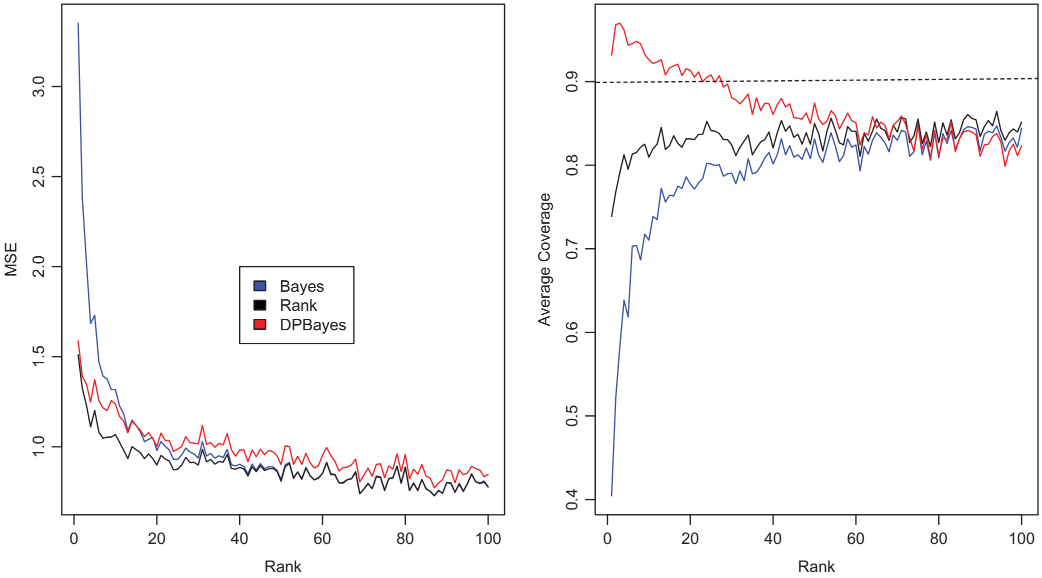


Fig. 3. Simulation study of Section 4.2 with the correct prior in (4.1) and parameter $k = 2$. The rank-conditional inference is the best in terms of MSE and non-parametric Bayes with Dirichlet prior is the close second best. The parametric Bayes is the distant third, with much larger MSE for $j < 10$. For the actual coverage rates, non-parametric Bayes method with Dirichlet prior is the best and the rank-conditioned inference is the close second. The parametric Bayes again is the distant third due to its much lower actual coverage rates.

coverage rates are far off. This is not surprising because an overly diffuse prior π_{DP} does not bring the needed shrinkage. Figure 3 shows the case for $k = 2$, in which the working prior deviates substantially from the true prior. The parametric Bayes performs poorly with huge MSE and far off actual coverage rates, while the non-parametric Bayes is much superior in both MSE and coverage rates. Figure 2 is for $k = 8$, an intermediate case between $k = 1000$ and $k = 2$. Neither of the two methods works well especially for $j < 15$. Our proposed rank-based inference, however, performs well for all the three cases. In particular, it is only slightly worse than the parametric Bayes for $k = 1000$ when the prior is correctly specified. When the prior is mis-specified as in $k = 2$ and $k = 8$, its mean square errors are the smallest among the three methods and its actual coverage rates are not too far off from the nominal 90%. The rank-conditioned inference therefore achieves robustness against mis-specified prior with minimal loss of efficiency under correctly specified prior. While not shown in the graphs, the superior performance of the rank-conditioned inference is similarly observed for the highest ranked θ_i such as $\theta_{[n]}, \theta_{[n-1]}, \ldots, \theta_{[n-99]}$. The difference between the methods is small for middle-ranked $\theta_{[i]}$ as their inference are primarily determined by the large mass at 0 which is present in both the true prior π and working prior π_p. Finally, our implementation of Method 2 is based on R package *DPpackage* (Jara *and others*, 2011). Two additional simulation studies are included in supplementary material available at *Biostatistics* online.

5. Application to breast cancer microarray dataset

We now apply our proposed method to the breast cancer data in Wang *and others* (2005). This was a large Affymetrix-based gene expression profiling study of $n = 22\,283$ genes on 286 untreated patients with lymph node-negative primary breast cancer. The data are available at http://www.ncbi.nlm.nih.gov/geo/

as dataset GSE2034. We will compare gene expression level between patients who developed distant metastasis (74 subjects) and patients who were relapse-free at 5 years (135 subjects) among the 209 estrogen receptor positive patients. These data were also analyzed in Noma *and others* (2010). We use the gene expression model described in Section 2 with z_i being the standardized sample mean difference in log gene expression level and θ_i being the true standardized mean difference for the ith gene. We have $m_i^{[1]} = 74$ and $m_i^{[2]} = 135$ for all i as there are no missing values for any gene. It then follows that σ_i^2 in (2.1) is $\frac{1}{74} + \frac{1}{135}$ for all i.

The maximum likelihood estimates of the parameters in the working prior π_p obtained by assuming models (2.1) and (2.3) for $z_1, \ldots, z_n$ are

η_1	μ_1	ω_1	η_2	μ_2	ω_2
0.0856	0.258	0.0426	0.315	−0.159	0.0470

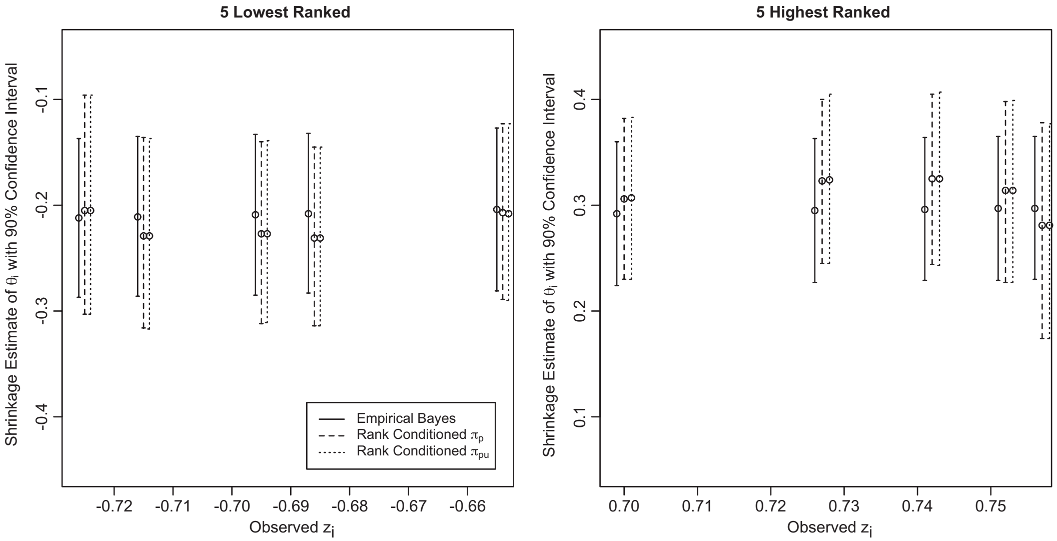
which suggests about 40% of non-zero θ_i among $n = 22\,283$ genes. In order to check the fit of the parametric prior π_p, we simulated new data from the fitted π_p and compared its distribution to that of the original data through the following algorithm. Let $z_k^* = \theta_k^* + \varepsilon_k^*$, where $\theta_k^* \sim \pi_p$ and $\varepsilon_k^* \sim \mathcal{N}(0, 1)$. The percentiles of $z_1, \ldots, z_n$ and $z_1^*, \ldots, z_n^*$ are given in the table below, which shows excellent fit of model π_p to data $z_1, \ldots, z_n$. As comparison, we also fit the model used in Noma *and others* (2010), which formulates in terms of the unstandardized log fold change $d_i \equiv \bar{y}_i^{[1]} - \bar{y}_i^{[2]}$. Using notation of this paper, their model is

$$d_i \sim N(\alpha_i^{[1]} - \alpha_i^{[2]}, \rho_i^2 \sigma_i^2), \quad \alpha_i^{[1]} - \alpha_i^{[2]} \sim \pi_p,$$

where π_p has form (2.3). The discrepancy between the percentiles of the original data $d_1, \ldots, d_n$ and the simulated new data $d_1^*, \ldots, d_n^*$ is much larger here. Modeling the standardized log fold change therefore provides much better fit to this dataset.

	Percentile	0	2.5	25	50	75	97.5	100
Modeling standardized	$z_1, \ldots, z_n$	−0.7248	−0.3848	−0.1512	−0.0325	0.0894	0.3578	0.7569
Log fold change	$z_1^*, \ldots, z_n^*$	−0.6847	−0.3831	−0.1528	−0.0334	0.0912	0.3639	0.8623
Modeling unstandardized	$d_1, \ldots, d_n$	−0.9333	−0.2688	−0.0796	−0.0159	0.0430	0.1959	0.7797
Log fold change	$d_1^*, \ldots, d_n^*$	−0.7248	−0.3848	−0.1512	−0.0325	0.0894	0.3578	0.7569

Coming back to the model in Section 2 and using the maximum likelihood estimates for η_1, μ_1, ω_1, η_2, μ_2, and ω_2 obtained above, the standard empirical Bayes estimates $\hat{\theta}_i^{\text{Bayes}}$ and the corresponding 90% confidence interval for all θ_i are then computed under π_p. Rank-conditioned estimates $\hat{\theta}_i$ and 90% intervals are also calculated under both π_p and the non-parametric update π_{pu}. Results for θ_i that correspond to the five lowest ranked z_i (−0.725, −0.715, −0.695, −0.686, −0.654) and to the five highest ranked z_i (0.700, 0.727, 0.742, 0.752, 0.757) are given in Figure 4 for the three methods. We make three observations. First, the three methods have a huge shrinkage effect on the raw estimate $z_{[j]}$ for these top genes. For example, $z_{[1]} = -0.725$ but $\hat{\theta}_{[1]}^{\text{Bayes}} = -0.212$ and $\hat{\theta}_{[1]} = -0.205$ (under both π_p and π_{pu}). Second, results from empirical Bayes and rank-conditioned inference under π_p and π_{pu} are very similar although the rank-conditioned confidence intervals are a little wider. The same is true for other θ_i not shown in Figure 4. This is not surprising, given the excellent fit of working prior π_p to $z_1, \ldots, z_n$ as discussed above. The agreement of the three methods and the robustness properties of the rank-conditioned inference should give us more


5 Lowest Ranked

5 Highest Ranked

Fig. 4. Parametric empirical Bayes and rank-conditional inference of θ_i for the five lowest ranked and the five highest ranked z_i for the breast cancer dataset in Section 5.

confidence in the result. Third, an oddity of the rank-conditioned inference is that $\hat{\theta}_{[1]}$ can be slightly larger than $\hat{\theta}_{[2]}$ even though $z_{[1]} \leqslant z_{[2]}$ by definition. This happens when the difference in rank-conditioned bias for $z_{[1]}$ and $z_{[2]}$ as random variables exceeds their observed difference in the observed $z_{[1]}$ and $z_{[2]}$. The same can happen to estimates of other ranks. This is generally a small peculiarity that is appropriately accounted for by the wide confidence intervals.

6. Discussion

We have proposed a rank-conditioned inference that can substantially improve the prior robustness of empirical Bayes inference with little loss of efficiency. More research is needed, however, to further develop and establish the proposed method. For example, in the simulations presented in Section 4, the actual coverage rates for the rank-conditioned intervals, in spite of being a substantial improvement over standard empirical Bayes, are still below the nominal 90% rate for $k = 2$ and $k = 8$. We expect that it is possible to further improve the actual coverage rate by drawing on similar research in the empirical Bayes literature, such as in Morris (1983), Laird and Louis (1987), He (1992), Qiu and Gene Hwang (2007), and Gene Hwang *and others* (2009). Second, model (2.1) assumes that errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent, which can be unrealistic in many applications. We are currently working to relax this requirement to accommodate a more general correlation structure. Preliminary results show that the method in this paper continues to work well if the correlation of $\varepsilon_1, \ldots, \varepsilon_n$ is mild. Details will be reported in a future manuscript. We hope this paper will stimulate more research in robust Bayes inference for large-scale data to meet the pressing analytical need in genomics and genetics.

7. Software

Our R package *rank.Shrinkage* provides a ready-to-use implementation of the proposed methodology. The R code for the simulation studies is available at https://sites.google.com/site/jiangangliao/.

Supplementary material

Supplementary Material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We are grateful for the extremely thorough review and suggestions by an anonymous reviewer and an associate editor, which led to a greatly improved manuscript. *Conflict of Interest*: None declared.

Funding

L.G.L.'s and T.M.'s work was supported by Penn State Cancer Institute internal grant. Berg's work was supported by Grant Number UL1 TR000127 from the National Center for Advancing Translational Sciences (NCATS). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association* **100**(469), 71–81.

Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M. and Halfon, M. S. (2005). Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biology* **6**(2), R16.

Do, K. A., Müller, P. and Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(3), 627–644.

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. *Bayesian Nonparametrics* **28**, 223.

Dunson, D. B. and Taylor, J. A. (2005). Approximate Bayesian inference for quantiles. *Nonparametric Statistics* **17**(3), 385–400.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, **23**(1), 1–22.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Volume 1. Cambridge: Cambridge University Press.

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**(456), 1151–1160.

Eggermont, P. P. B. and LaRiccia, V. N. (1997). Nonlinearly smoothed EM density estimation with automated smoothing parameter selection for nonparametric deconvolution problems. *Journal of the American Statistical Association* **92**(440), 1451–1458.

Gene Hwang, J. T., Qiu, J. and Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(1), 265–285.

Ghosh, D. (2009). Empirical Bayes methods for estimation and confidence intervals in high-dimensional problems. *Statistica Sinica* **19**(1), 125.

Gu, J. and Ghosal, S. (2009). Bayesian ROC curve estimation under binormality using a rank likelihood. *Journal of Statistical Planning and Inference* **139**(6), 2076–2083.

Guo, L., Lobenhofer, E. K., Wang, C., Shippy, R., Harris, S. C., Zhang, L., Mei, N., Chen, T., Herman, D. and Goodsaid, F. M. (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology* **24**(9), 1162–1169.

HE, K. (1992). Parametric empirical Bayes confidence intervals based on James-Stein estimator. *Statistics and Decisions* **10**(1–2), 121–132.

HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics* **1**(1), 265–283.

ISHWARAN, H. AND RAO, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics* **33**(2), 730–773.

JARA, A., HANSON, T. E., QUINTANA, F. A., MÜLLER, P. AND ROSNER, G. L. (2011). DPpackage: Bayesian non-and semi-parametric modelling in R. *Journal of Statistical Software* **40**(5), 1.

JEFFRIES, N. O. (2009). Ranking bias in association studies. *Human Heredity* **67**(4), 267–275.

KENDZIORSKI, C. M., NEWTON, M. A., LAN, H. AND GOULD, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22**(24), 3899–3914.

KIM, S., DAHL, D. B. AND VANNUCCI, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Analysis* **4**(4), 707–732.

LAIRD, N. M. AND LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association* **82**(399), 739–750.

MORRIS, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78**(381), 47–55.

MURALIDHARAN, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics* **4**(1), 422–438.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. AND TSUI, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**(1), 37–52.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. AND AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**(2), 155–176.

NOMA, H., MATSUI, S., OMORI, T. AND SATO, T. (2010). Bayesian ranking and selection methods using hierarchical mixture models in microarray studies. *Biostatistics* **11**(2), 281–289.

QIU, J. AND GENE HWANG, J. T. (2007). Sharp simultaneous confidence intervals for the means of selected populations with application to microarray data analysis. *Biometrics* **63**(3), 767–776.

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**(1), 3.

VARDI, Y., SHEPP, L. A. AND KAUFMAN, L. (1985). A statistical model for positron emission tomography. *Journal of the American Statistical Association* **80**(389), 8–20.

WANG, Y., KLIJN, J. G. M., ZHANG, Y., SIEUWERTS, A. M., LOOK, M. P., YANG, F., TALANTOV, D., TIMMERMANS, M., MEIJER-VAN GELDER, M. E., YU, J. *and others* (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* **365**(9460), 671–679.

WESTFALL, P. H. (2005). Comment. *Journal of the American Statistical Association* **100**(469), 85–89.

[*Received March 26, 2012; revised July 10, 2013; accepted for publication July 10, 2013*]