Improved Estimation of Correlation in Microarray Data Analysis

Marc Sobel[1], and Bud Mishra[2,3]*

[1] Statistics Department, Temple University, Philadelphia, PA, USA.

[2] Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY, USA 10012.

[3] Cold Spring Harbor Lab, 1 Bungtown Road, Cold Spring Harbor, NY, USA 11724.

May 28, 2004

ABSTRACT **In the original work on clustering due to Eisen et al., in which they performed one of the most highly-re-analyzed microarray dataset of gene expressions, the authors claimed to have "found in the budding yeast *Saccharomyces cerevisiae* that clustering gene expression data groups together efficiently genes of known similar function." However, they measured similarity between any pair of genes using a somewhat non-standard definition of correlation coefficient instead of Pearson's correlation coefficient, an unbiased estimator. Eisen et al.'s paper remains mysteriously silent about how drastically the clusters of genes would change if one changed the definition of similarity back to Pearson's correlation coefficient, or to any other from a larger family of estimators between Pearson's and Eisen et al.'s, obtained by a "shrinkage coefficient" taking a value between 0 and 1. Their approach raised several issues: what would be the best shrinkage coefficient; how can one compute it and whether it can be computed quickly in a closed-form. Mishra and his students answered these questions in a recent paper, but left it for future research to understand if there is even a wider family of similarity metrics, even though such metrics may not be computable in closed form. We take up this problem in this paper and suggest how to compute a somewhat better similarity metric using an MCMC algorithm; how to define an intuitively clear Bayesian risk assessment and finally, how to interpret the empirical results obtained through simulation.**

*To whom correspondence should be addressed. E-mail: mishra@nyu.edu

1 Problem Formulation

In the post-genomic biology, clustering genes by their similarity has now occupied many biologists and statisticians for almost half a decade. Although the relevance of such a succinct representation in understanding fundamental principles of biology is yet to be firmly established, there is much less disagreement that the resulting data reorganization may add clarity to the subsequent bioinformatic analysis and experiment design, e.g., interpreting ChIP-Chip experiments, looking for cis-regulatory elements, etc. There is now a rapidly mushrooming body of genomics literature devoted to clustering, co-clustering, bi-clustering, etc., with random or designed sets of conditions and different definitions of similarity, and yet, there is much less attention paid to derive a statistically robust definition for similarity of genes.

In the usual setting, starting with a series of expression microarray experimental data, one wishes to estimate similarity between the expression levels of a pair of genes because it is frequently indicative of functional relationships between them. Highly correlated transcriptomic behavior of a group of genes often suggests the presence of causal relationships, usually through common regulatory mechanisms. Identifying such potential relationships is of primary importance

1. In understanding and modeling microarray and other genetic data, and
2. In inferring functional relationships crucial to predictive and other kinds of inference.

These identifications frequently arise from partitioning genes into closely related groups, called clusters. Traditionally, algorithms for cluster analysis of expression data are based on statistical properties of gene expressions and result in organizing genes according to similarities between their patterns (see [1]).

One of the earliest and most popular clustering algorithms reported in the literature and due to Eisen et al. ([6] and [8])

1. Associates genes with expression vectors consisting of the logarithms of ratios of the amounts of transcribed mRNA under various conditions with respect to their values under normal conditions, and

2. Clusters these expression vectors using a hierarchical clustering algorithm with 'an appropriate' (see [6]) similarity metric.

In Eisen et al.'s paper [6], the authors claimed to have "found in the budding yeast *Saccharomyces cerevisiae* that clustering gene expression data groups together efficiently genes of known similar function," but did not address how best the similarity between a pair of genes is to be estimated. Eisen et al. used a somewhat non-standard (biased) definition of correlation coefficient instead of the usual unbiased estimator of Pearson. Mysteriously, this paper remained silent about how drastically the clusters would change if one changed the definition of similarity back to Pearson's estimator of correlation coefficient or to one of an infinite family of biased estimators between Pearson's and Eisen et al.'s that could be obtained by introducing a "shrinkage coefficient" (as discussed below) taking a value between 0 and 1.

Note that, with a large data samples, the best unbiased estimator for similarity is obtained through Pearson's Correlation Coefficient, referred to above. The similarity metric, proposed by Eisen et al. [6], is obtained by setting the mean expression value μ in Pearson's correlation coefficient to 0. The motivation for this modification was provided by their prior knowledge that the logarithm of mean expression ratios for genes was often known (apart from noise) to be 0. To avoid confusion between parameters below, we refer to the mean expression μ as the **baseline** value and employ a different parameterization for it. Cherepinsky et al. [3] argued that, in view of the Bayesian paradigm, prior knowledge regarding these baseline values should be combined with the observed gene expression levels for purposes of estimating the baseline. Intuitively, their definition implied modifying the baseline value to $\widehat{\beta} \cdot \mu$ ($0 \leq \widehat{\beta} \leq 1$), where the shrinkage parameter $\widehat{\beta}$ gives rise to a family of estimators with Eisen's and Pearson's being two extrema, for $\widehat{\beta} = 0$ and $\widehat{\beta} = 1$, respectively. In doing so, Cherepinsky et al. introduced flexible similarity metrics which depend on the data, but also raised several important questions:

1. What would the '*best*' shrinkage coefficient be in this context; and

2. How can such coefficients be computed.

While [3] addressed these questions in this limited setting, they left it for future research to characterize and solve these questions in general statistical settings, in which similarity metrics can depend on the (assumed) statistical model as well as the data.

In this paper, we take up this problem by,

1. Characterizing the '*true*' baseline (or offset) values as parameters whose estimation leads to the determination of intergene correlation and clustering (section 2),

2. Defining prior distributions over these intergene correlations using this characterization (also, section 2), and

3. Comparing Bayes risks for different intergene correlation estimators (section 3).

Different statistical models (see, section 3) arise depending on,

- Whether we assume all baseline parameter for the same gene are equivalent and/or

- Whether we assume baseline parameters for different genes are mutually independent.

We examine these models to compare the different correlation and estimators arising from them. We also assess the fitness of the statistical models devised in this paper and examine robustness properties which they enjoy (section 4). We also evaluate the global accuracy of our method by cataloging false positives and negatives in the clustering as similarity metrics and models are varied.

As mentioned earlier, algorithms for clustering expression data are designed to organize genes according to similarities in their pattern of expression. Coexpression of genes of known functionality with new genes leads to a discovery process designed to characterize the functionality of unknown or less-well-known genes. False negatives result when pairs of coexpressed genes are assigned to distinct clusters; false positives result when pairs of independent genes are assigned to the same cluster. The former indicates that the discovery process is ignoring useful information while the latter adds noise to the discovery process.

In this paper we employ self-organizing maps (SOM) [7] for clustering genes; they are specifically designed to produce clusters whose members have a high degree of similarity, according to the (parameter dependent) metric introduced below. We index the clustering procedures used below by specifying a smallest level of similarity (between an expression vector and its cluster center) necessary to belong to a cluster; we denote this by α ($0 < \alpha < 1$). Clustering procedures associated with large cutoff values are more robust

to the introduction of spurious expression features; those with smaller cutoff values are less robust to the introduction of these features. Procedures of the former variety tend to produce more clusters; these tend to exclude false positives. Procedures of the latter variety produce fewer clusters; these tend to exclude false negatives.

Our results demonstrate conclusively that, in such settings, the shrinkage estimators introduced by Cherepinsky et al. [3] more accurately estimate intergene similarity.

2 Model Formulation

We assume that the observed expression data consists of the matrix

$$\mathbf{X} = \{X_{g,c}\}_{g=1,\ldots,G;c=1,\ldots,C},$$

with $g = 1, \ldots, G$ indexing the genes and $c = 1, \ldots, C$ indexing the conditions (of the expression vectors). Our goal is to *estimate* the similarities between the observed expression vectors, $\mathbf{X}_g$

$$(\mathbf{X}_g = (X_{g,1}, \ldots, X_{g,C}), \quad g = 1, \ldots, G).$$

One class of quantitative measures of similarity take the form,

$$r_{\text{baseline}}(\mathbf{X}_g, \mathbf{X}_h) \tag{1}$$
$$= \frac{\sum_{c=1}^{C}(X_{g,c} - \text{baseline}_g)(X_{h,c} - \text{baseline}_h)}{\sqrt{\sum_{c=1}^{C}(X_{g,c} - \text{baseline}_g)^2 \sum_{c=1}^{C}(X_{h,c} - \text{baseline}_h)^2}}$$

where 'baseline_g' denotes the (central) baseline (or offset) value associated with expression vector ($g = 1, \ldots, G$). We note that all such quantities take values in the closed interval $[-1, 1]$; they differ only in regards to what is assumed regarding the baseline values for the given vectors. A more general class of quantitative measures of similarity can be constructed if the baseline values baseline_g are allowed to depend on both the gene 'g' and the condition 'c'.

Examples of such measures include:

1. *Pearson's Correlation*, defined by:

$$r_p(\mathbf{X_g}, \mathbf{X_h}) \tag{2}$$
$$= \frac{\sum_{c=1}^{C}(X_{g,c} - \overline{X}_g)(X_{h,c} - \overline{X}_h)}{\sqrt{\sum_{c=1}^{C}(X_{g,c} - \overline{X}_g)^2 \sum_{c=1}^{C}(X_{h,c} - \overline{X}_h)^2}}$$

   We note that the baseline values are given, in this setting, by the corresponding (vector) sample means.

2. Eisen's Correlation, defined [6] by:

$$r_e(\mathbf{X_g}, \mathbf{X_h}) \tag{3}$$
$$= \frac{\sum_{c=1}^{C}(X_{g,c})(X_{h,c})}{\sqrt{\sum_{c=1}^{C}(X_{g,c})^2 \sum_{c=1}^{C}(X_{h,c})^2}}$$

   We note that the baseline values are all taken to be 0 in this setting.

3. Linear Correlation estimators, [2] defined by:

$$r_{LC}(\mathbf{X_g}, \mathbf{X_h}) \tag{4}$$
$$= \frac{\sum_{c=1}^{C}(X_{g,c} - \widehat{\beta}\overline{X}_g)(X_{h,c} - \widehat{\beta}\overline{X}_h)}{\sqrt{\sum_{c=1}^{C}(X_{g,c} - \widehat{\beta}\overline{X}_g)^2 \sum_{c=1}^{C}(X_{h,c} - \widehat{\beta}\overline{X}_h)^2}}$$

   where $\widehat{\beta}$ is a real-valued coefficient estimated from the full data set as shown below. We note that this definition is motivated by using baseline values which shrink the sample mean vector toward 0.

   More generally, we can construct linear correlation estimators which shrink the sample mean vector toward an alternative fixed vector; this is done below.

4. Compound Linear Correlation estimators, defined by:

$$r_{CLC}(\mathbf{X}_g, \mathbf{X}_h) \tag{5}$$
$$= \frac{\sum_{c=1}^{C}(X_{g,c} - \widehat{B}_g\overline{X}_g)(X_{h,c} - \widehat{B}_h\overline{X}_h)}{\sqrt{\sum_{c=1}^{C}(X_{g,c} - \widehat{B}_g\overline{X}_g)^2 \sum_{c=1}^{C}(X_{h,c} - \widehat{B}_h\overline{X}_h)^2}}$$

   where $\widehat{B} = (B_1, \ldots, B_G)$ is a vector of real-valued coefficients estimated from the full data set as described below in more details. We note that this definition is motivated by using baseline values which shrink each component of the sample mean vector toward 0 by the (respective) factors $B_1, \ldots, B_G$

   Note that when $B_1 = \cdots = B_G = \widehat{\beta}$ then it becomes r_{LC}, and if in addition $\widehat{\beta} = 1$ (respectively, $= 0$) then it becomes r_p (respectively, r_e).

Each of these similarity measures (arguably) provides an appropriate measure of dependence between expression vectors in settings appropriate to its application. In order to facilitate comparisons between them, it is necessary to define what is, in effect, being estimated by these quantities. Pursuant to this purpose, we define a 'true' baseline value for each gene. Measures of similarity between each given pair of observed expression vectors are computed in terms of these true baseline values; the resulting values are referred

to below as their 'true' similarities (similar to 'clairvoyant' similarity metric of [3]) .

Procedures for simultaneously estimating the similarity between each given pair of genes are assessed by how closely they approximate the 'true' similarities between them. Assessing the efficiency of such procedures requires the assumption of a statistical model. Below, we describe three such models.

2.1 Statistical Models of Type I

Let $\{\theta_g\}$ be (real valued) parameters representing the 'true' baseline values for the (respective) observed expression vectors

$$\{\mathbf{X}_g\}_{g=1,\ldots,G}.$$

We assume that the baseline value θ_g is the common expected value $\{X_{g,c}\}$ of the components of the vector $\mathbf{X}_g$ ($g = 1, \ldots, G$) (i.e., $E_\theta X_{g,c} = \theta_g$; $g = 1, \ldots, G$; $c = 1, \ldots, C$). We also let $\{\sigma_g\}$ denote additional parameters which characterize the distribution of the observations. It is assumed that the parameters,

$$\{\theta_g, \sigma_g\}_{g=1,\ldots,G}$$

are independent and identically distributed (i.i.d.) with prior distribution $\pi(\bullet|\Gamma)$; Γ denotes the model's hyperparameter(s). The true similarity between vectors $\mathbf{X}_g$, and $\mathbf{X_h}$ is defined, in this setting, to be:

$$\begin{aligned} & r(\mathbf{X}_g, \mathbf{X}_h; \theta_g, \theta_h) \qquad (6) \\ = \; & \frac{\sum_{c=1}^{C}(X_{g,c}-\theta_g)(X_{h,c}-\theta_h)}{\sqrt{\sum_{c=1}^{C}(X_{g,c}-\theta_g)^2 \sum_{c=1}^{C}(X_{h,c}-\theta_h)^2}} \end{aligned}$$

Estimates $r(\mathbf{X}_g, \mathbf{X}_h; \widehat{\theta}_g, \widehat{\theta}_h)$ of the true similarity (between expression variables) $\mathbf{X}_g$ and $\mathbf{X}_h$) are formed by replacing the baseline parameters, θ_g, θ_h with the (respective) estimates, $\widehat{\theta}_g$, $\widehat{\theta}_h$ which depend on the observed data. We note that

- For Pearson similarity estimates, $\widehat{\theta}_g = \overline{X}_g$;
- For Eisen similarity estimates, $\widehat{\theta}_g = 0$;
- For linear correlation estimates, $\widehat{\theta}_g = \widehat{\beta} \cdot \overline{X}_g$,

($g = 1, \ldots, G$).

2.2 Statistical Model: Type II

Model II differs from Model I in dropping the assumption that the parameters $\theta_1, \ldots, \theta_G$ are mutually independent and identically distributed (i.i.d.). As an example, assume the θ parameters have a prior Gaussian distribution with common mean μ, common variance τ^2 and common correlation ρ.

2.3 Statistical Model: Type III

Let

$$\{\Theta_g\}_{g=1}^{G}$$

(with $\Theta_g = (\Theta_{g,1}, \ldots, \Theta_{g,C})$) be (vector valued) parameters representing the 'true' baseline component values for the (respective) observed expression vectors

$$\{\mathbf{X}_g\}_{g=1,\ldots,G}.$$

We assume that the baseline value $\Theta_{g,c}$ is the expected value of the component, $X_{g,c}$ of the vector $\mathbf{X}_g$ ($g = 1, \ldots, G$) — i.e.,

$$E_\theta X_{g,c} = \Theta_{g,c}; \quad g = 1, \ldots, G; \quad c = 1, \ldots, C.$$

We also let $\{\Sigma_g\}$ denote additional parameters which characterize the distribution of the observations. It is assumed that the parameters,

$$\{\Theta_g, \Sigma_g\}_{g=1,\ldots,G}$$

are independent and identically distributed (i.i.d.) with prior distribution $\pi(\bullet|\Gamma)$; Γ denotes the models hyperparameter(s). The true similarity between vectors $\mathbf{X}_g$, and $\mathbf{X}_h$ is defined, in this setting, to be:

$$\begin{aligned} & r(\mathbf{X}_g, \mathbf{X}_h; \Theta_g, \Theta_h) \qquad (7) \\ = \; & \frac{\sum_{c=1}^{C}(X_{g,c}-\Theta_{g,c})(X_{h,c}-\Theta_{h,c})}{\sqrt{\sum_{c=1}^{C}(X_{g,c}-\Theta_{g,c})^2 \sum_{c=1}^{C}(X_{h,c}-\Theta_{h,c})^2}} \end{aligned}$$

Estimates $r(\mathbf{X}_g, \mathbf{X}_h; \widehat{\Theta}_g, \widehat{\Theta}_h)$ of the true correlation (between variables $\mathbf{X}_g$ and $\mathbf{X}_h$) are formed by replacing the baseline parametric vectors, Θ_g, Θ_h with the (respective) estimates, $\widehat{\Theta}_g, \widehat{\Theta}_h$ which depend on the observed data. We note that for compound linear correlation estimates, $\widehat{\Theta}_g = \widehat{B}'\overline{X}_g$, ($g = 1, \ldots, G$). As an example of a model of this type let c_0 (between 1 and C) be a fixed cutpoint and assume the parametric vectors $\Theta_1, \ldots, \Theta_G$ are mutually independent and identically distributed with prior distribution described by:

$$\Theta_{g,c} = \begin{cases} \theta_{g,1}, & \text{if } 1 \le c \le c_0; \\ \theta_{g,2}, & \text{if } c_0 < c \le C. \end{cases}$$

and

$$\begin{aligned} \theta_{g,1} &= \mu + \epsilon_{g,1}; \\ \theta_{g,2} &= \mu + \gamma \cdot (\theta_{g,1} - \mu) + \epsilon_{g,2}; \end{aligned}$$

$\epsilon_{g,1}$, and $\epsilon_{g,2}$ are independent errors; and μ, and γ are independent parameters.

Our results concern the advantages of estimating similarity in model I and model II settings. We are currently exploring the advantages of Model III settings. Models of type III may be advantageous if there is a wide difference between conditions (i.e., if the array columns are non-homogeneous). *Henceforth, in this paper, we assume a model I and/or II setting only, unless otherwise specified.*

3 Model Assessment for the Observed Data

We initially assess the models described above for the data set described below. One mechanism for doing so involves the computation of cross-validated residuals [2]. Cross-validated residuals traditionally measure the difference, for each observation, between what is observed and what the model predicts (without benefit of the given observation). Since the similarities between pairs of expression vectors, as defined above, are not observed we need to alter this definition slightly. Keeping this in mind, for each (unordered) pair of expression vectors, we calculate residuals by comparing, for each unordered pair of observations, its marginal similarity with what we predict it to be (without benefit of the given pair of observations). In other words, if $\mathrm{CORR}_{(\theta)}(\bullet)$ denotes the similarity between observations for fixed baseline values θ, $E_{\widehat{\Gamma}}$ denotes expectation with respect to the prior distribution over θ when hyperparameters Γ are estimated using their empirical Bayes estimates $\widehat{\Gamma}$, $\mathbf{x}_{-\mathcal{A}}$ denotes the full set of observations with the index set $\mathcal{A}$ removed, and Θ_g denotes the vector all of whose components are θ_g, we define,

$$\begin{aligned}\text{Residual}(g,h) &= E_{\widehat{\Gamma}}\left\{\mathrm{CORR}_{\Theta}\left(\mathbf{X_g},\mathbf{X_h}\right)\right\} \\ &\quad -E_{\widehat{\Gamma}}\left\{\mathrm{CORR}_{\Theta}\left(\mathbf{X_g},\mathbf{X_h}\right)|\mathbf{X_{-g,-h}}\right\}; \\ &\quad \text{where } 1\le g<h\le G\end{aligned} \tag{8}$$

Large residual values suggest lack of fit between the data and the model for the correlation between the given pair of observations. We can measure residual size in the units of the problem or in standardized units; below, our results measure these in standardized units.

3.1 Clustering Assessment

We employ a self-organizing map (SOM) algorithm [7] for clustering, described in 4.4. This map yields a clustering of the gene expression data for each given baseline value θ and each critical value α (thus, distinguishing the measure of similarity needed for a gene to be placed in a given cluster). In such settings, the 'true' sensitivity (i.e., the fraction of true positives given the baseline value) and the 'true' specificity (i.e., the fraction of true negatives given the baseline value) must be estimated. Using the terminology e.g., $TP(\theta,\alpha)$ for the number of true positives given the baseline value θ and critical value α, the true sensitivity of a clustering procedure is given by the formula;

$$TSens(\alpha)=\left\{\frac{TP(\theta,\alpha)}{TP(\theta,\alpha)+FP(\theta,\alpha)}\right\} \tag{9}$$

The true specificity is given by the formula;

$$TSpec(\alpha)=\left\{\frac{TN(\theta,\alpha)}{TN(\theta,\alpha)+FN(\theta,\alpha)}\right\} \tag{10}$$

An empirical Bayes estimate of the True Sensitivity is given by the formula:

$$ESens(\alpha)=\frac{TP(\widehat{\theta},\alpha)}{TP(\widehat{\theta},\alpha)+FP(\widehat{\theta},\alpha)}, \tag{11}$$

and an empirical Bayes estimate of the true specificity is given by the formula:

$$ESpec(\alpha)=\frac{TN(\widehat{\theta},\alpha)}{TN(\widehat{\theta},\alpha)+FN(\widehat{\theta},\alpha)}. \tag{12}$$

We can then provide (estimated) ROC curves consisting of graphs of points with x-coordinate given by the estimated specificity and y-coordinate the estimated sensitivity. We assess clustering procedures by examining the relationship between confidence intervals for the true specificity/sensitivity and the corresponding values of specificity/sensitivity for Pearson and Eisen procedures (having the same critical value).

3.2 Bayes Risk Assessment

We assess estimates of the true similarity between expression vectors using a Bayes risk formulation. The Bayes risk, employed below, is the expectation of the sum of the squared differences between the estimated and true similarity measures averaged over the marginal distribution of the observations. It measures how well we can predict the similarity between gene pairs 'on average' under the given (likelihood/prior) model. The use of the Bayes risk, as a measure of estimator assessment avoids the bias implicit in using the observed vectors themselves to assess their own relationships to one another ([2] and [4]). Below, we argue that this

Bayes risk assessment is entirely general, applying as it does to all observations in which the same likelihood/priors are assumed. We begin by constructing the Bayes risk for estimating the similarity between expression vectors, $\mathbf{X}_g$, and $\mathbf{X}_h$. The Bayes risk, employed to estimate this using the baseline estimates $\widehat{\theta}_g$, $\widehat{\theta}_h$ is given by:

$$\begin{aligned} &\text{BAYESRISK}_{\widehat{\theta}}(g,h) \qquad (13)\\ &= E^{(\pi)}\, E^{(\bullet|\theta,\sigma)}\Big\{ \\ &\qquad r(X_g^*, X_h^*, \widehat{\theta}_g^*, \widehat{\theta}_h^*) - r(X_g^*, X_h^*, \theta_g, \theta_h)\Big\}^2, \qquad (14)\\ &\qquad \text{where } (1 \le g < h \le G). \end{aligned}$$

The inner expectation on the right hand side of equation 15 is over the assumed likelihood conditional on the parameters; the outer expectation is over the assumed prior. The notation '$\mathbf{X}_g^*$, and $\mathbf{X}_h^*$' (respectively, '$\widehat{\theta}_g^*$, and $\widehat{\theta}_h^*$') refers to simulated observed vectors (respectively, functions of observed random vectors) having the same conditional distribution as that assumed for the observed vectors '$\mathbf{X}_g$, and $\mathbf{X}_h$' (respectively, '$\widehat{\theta}_g$, and $\widehat{\theta}_h$'). The quantity, $r(X_g^*, X_h^*, \widehat{\theta}_g^*, \widehat{\theta}_h^*)$ is the estimate (obtained under simulation) of the similarity measure, $r(X_g, X_h; \theta_g, \theta_h)$.

The Bayes risk for simultaneously estimating all similarities is given, in the notation introduced in 15, by,

$$\begin{aligned} &\text{BAYESRISK}_{\widehat{\theta}}^{\text{MODEL I}} \qquad (15)\\ &= \sum_{1\le g<h\le G} \text{BAYESRISK}_{\widehat{\theta}}^{\text{MODEL I}}(g,h). \end{aligned}$$

3.2.1 Model I setting

In what follows we employ the notation, '$\mathcal{N}(\lambda,\eta)$' in reference to the normal distribution with mean λ and variance η. We assume that all gene expression levels are (conditionally) i.i.d. (independent and identically distributed) Gaussian with respective gene expression level means $\theta_1, \ldots, \theta_G$ — i.e.,

$$X_{g,c} \sim \mathcal{N}(\theta_g, \sigma_g^2); \quad g = 1,\ldots,G; \quad c = 1,\ldots,C.$$

Below, we use standard maximum likelihood estimates

$$\widehat{\sigma}^2 = \frac{1}{C}\sum_{c=1}^{C}(X_{g,c} - \overline{X}_g)^2$$

for σ_g^2 ($g = 1, \ldots, G$). We assume that the mean level parameters $\theta_1, \ldots, \theta_G$ are themselves i.i.d. with common Gaussian prior distribution $\mathcal{N}(\mu, \tau^2)$. Standard empirical Bayes theory [2] shows that, in this setting, the hyperparameters μ, τ should be estimated respectively by,

$$\widehat{\mu} = \frac{\sum\sum X_{g,c}}{G\cdot C} \quad \text{and} \quad \widehat{\tau^2} = \frac{\widehat{\sigma}^2\cdot(1-\widehat{B})}{C\cdot B}$$

with,

$$\widehat{B} = \frac{(C-2)\cdot G}{\sum_{g=1}^{G}\sum_{c=1}^{C}(X_{g,c}^2)} \qquad (16)$$

The empirical Bayes correlation estimator is constructed by estimating the baseline parameters by

$$\widehat{\theta}_g = (1-\widehat{B})\cdot\overline{X}_g.$$

The Bayes risks for such procedures are given in figures 1 and 2 below.

3.3 Robustness

Below, we employ the notation, $\phi(X,\lambda,\omega)$ for the normal Gaussian density with (component) mean(s) λ and (component) variance(s) ω. We examine the robustness of the Model I empirical Bayes similarity estimators for models in which the parameters $\theta_1, \ldots, \theta_G$ are assumed to have an ϵ-contamination prior [2] — i.e., a prior taking the form,

$$\pi(\theta) = \begin{cases} \phi(\mathbf{X},\mu,\tau^2), & \text{wprob } 1-\epsilon; \\ \phi(\mathbf{X},\mu,\text{LARGE}), & \text{wprob } \epsilon; \end{cases} \qquad (17)$$

with ϵ fixed at various small and LARGE at various large values. Below, we refer to those expression vectors $\mathbf{X}_g$ having *a priori* variance τ^2 as 'ordinary,' and those having *a priori* variance LARGE as 'outliers'. The hyperparameter τ whose estimate plays a role in the estimation of correlation is estimated (to maximize the contaminated marginal likelihood) using expectation-maximization (EM) via a gradient ascent strategy, i.e., using the notation $\mathcal{L}(\mathbf{X}|\tau)$ for the log of the marginal likelihood and $\pi(\mathbf{X}|\tau)$ for the probability that vector $\mathbf{X}$ is 'ordinary' (in the sense described above). The gradient ascent step then takes the form:

$$\begin{aligned} &\tau^{(new)} \qquad (18)\\ &= \tau^{(old)} - \frac{(1/C)\sum_g \frac{\partial\mathcal{L}(\mathbf{X_g}|\tau^{(\mathbf{old})})\cdot\pi(\mathbf{X_g}|\tau^{(\mathbf{old})}}{\partial\tau^{(old)}}}{\sum_g \frac{\partial^2\mathcal{L}(\mathbf{X_g}|\tau^{(old)})\cdot\pi(\mathbf{X_g}|\tau^{(\mathbf{old})}}{\partial[\tau^{old}]^2}} \end{aligned}$$

with

$$\begin{aligned} &\pi(\mathbf{X}|\tau) \qquad (19)\\ &= \frac{\phi\big(\mathbf{X},\widehat{\mu},(\sigma^{\mathbf{2}}/\mathbf{C})+\tau^{\mathbf{2}}\big)}{\phi\big(\mathbf{X},\widehat{\mu},(\sigma^{\mathbf{2}}/\mathbf{C})+\tau^{\mathbf{2}}\big)+\phi\big(\mathbf{X},\widehat{\mu},(\sigma^{\mathbf{2}}/\mathbf{C})+\text{LARGE}^{\mathbf{2}}\big)}. \end{aligned}$$

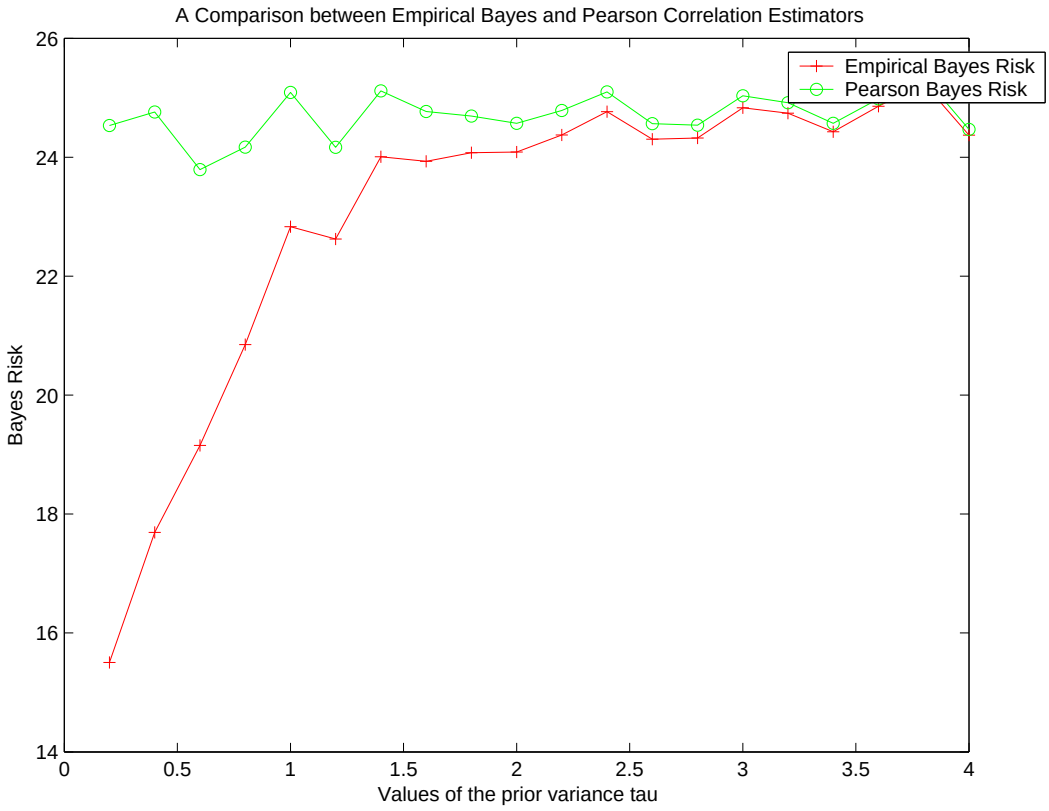

Figure 1: Comparison Between Bayes risks of Empirical Bayes versus Pearson Correlation estimates for prior mean= 0.0 and prior standard deviation between 0.2 and 4.0. in a Model I setting (Example 1)

Our results demonstrate that, in settings where the assumed prior is contaminated by noise, the correlation estimates given by the empirical Bayes estimator are no less accurate than Pearson's estimator. To demonstrate this, we have (separately) explored comparisons between the aforementioned Bayes risks when the amount of contamination varies between 0.01 and 0.1 and when the outlier standard deviations vary between 2.0 and 22.0 (see section 4).

3.3.1 Assessing similarity in Model II setting

We assume, as above, that all gene expression levels are (conditionally) (independent and identically distributed) i.i.d. Gaussian with respective mean expression levels, θ_1, ..., θ_G — i.e.,

$$X_{g,c} \sim \mathcal{N}(\theta_g, \sigma_g^2); \quad g = 1, \ldots, G; \quad c = 1, \ldots, C.$$

We drop our assumptions (cf, 3.2.1) regarding the mutual independence of the mean gene expression level parameters (θ_1, ..., θ_G) and assume instead that they have a common correlation ρ, called the prior correlation (see e.g., [5] for a discussion of settings like this one). This parameter, together with the prior variance τ, is estimated using a Model II Maximum likelihood procedure; these estimates are employed in the calculation of the Bayes risk. The resulting Bayes risks are given in figures 4 and 5 below.

4 Biological Examples of 'Simultaneous' Estimators

Below, we consider two examples of microarray data.

Microarray Data Example 1: This data was selected from the *Arabidopsis* Data generated by the Coruzzi group. Ninety genes observed under 20 conditions having the smallest noise to signal ratio were selected. The ratio of their expression values with normal ones were log transformed to form the complete data set.

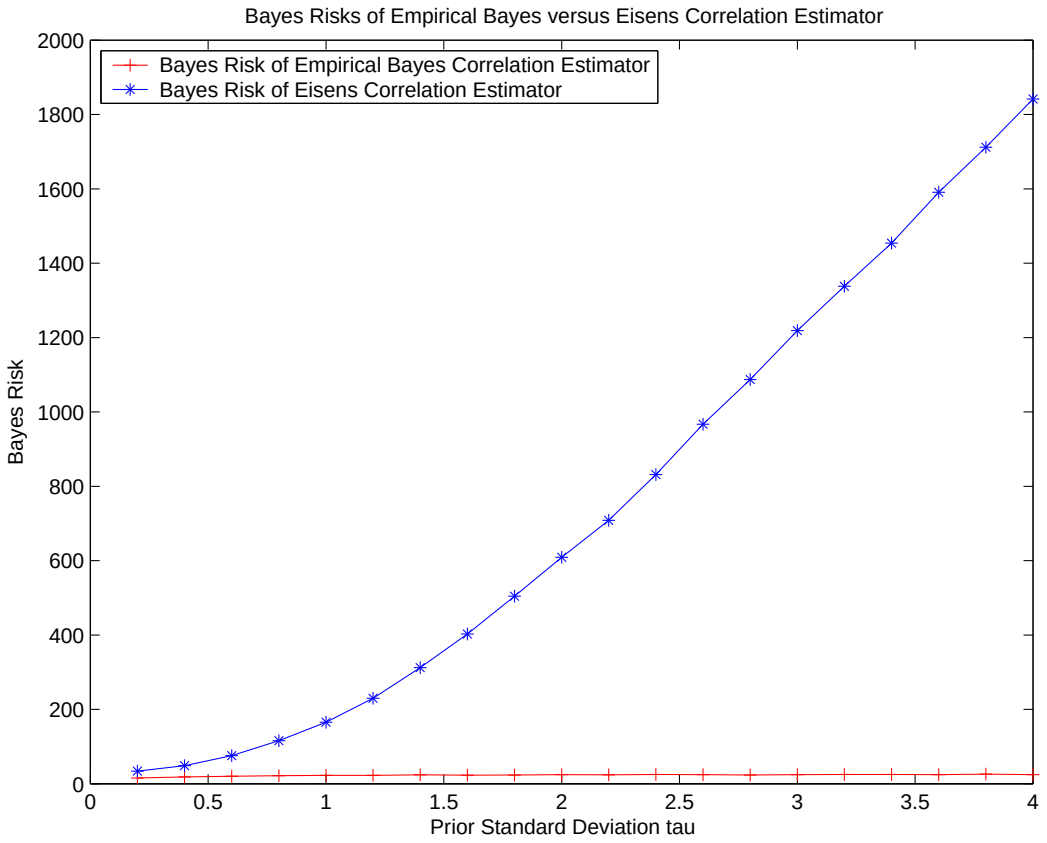

Figure 2: Comparison Between Bayes risks of Empirical Bayes versus Eisen's Correlation estimates when the outlier prior standard deviation varies between 0.2 and 4.0. in a Model I setting (Example 1)

Microarray Data Example 2: Expression data was selected from that available at `http://genome-www.stanford.edu/clustering/`. Forty four Genes were grouped by transcriptional activators and cell cycle functions (see [3] for a full table of these). Similarity measures were evaluated against this "ground truth" clustering.

1. Empirical Bayes estimator have smaller Bayes risk than Pearson and Eisen estimators;

2. Improvements of the Bayes risk for empirical Bayes estimators over those for their Pearson counterparts are clearly greater for smaller values of the prior variance.

.

4.1 Bayes Risk: Example 1

In the example analyzed below, $G = 90$ and $C = 19$. We compute the Bayes risk associated with using the empirical Bayes versus Pearson correlation estimators when the true prior variance τ^2 takes on values between 0.2 and 4.0 (see 3.2). Results comparing empirical Bayes with Pearson correlation estimators are given in figure 1; those comparing empirical Bayes with Eisen correlation estimators are given in figure 2. Our results demonstrate that,

4.2 Residuals: Example 2

Using the microarray data from example 2, we calculated residuals for each pair of genes (as discussed in 3). In figure 3 we construct a histogram for the set of paired standardized residuals. We note that the overwhelming bulk (99.9% of the residual data) are smaller in absolute value than 2, lending credence to the good fit of the model. This suggests that the model adequately describes correlation for most pairs of genes.

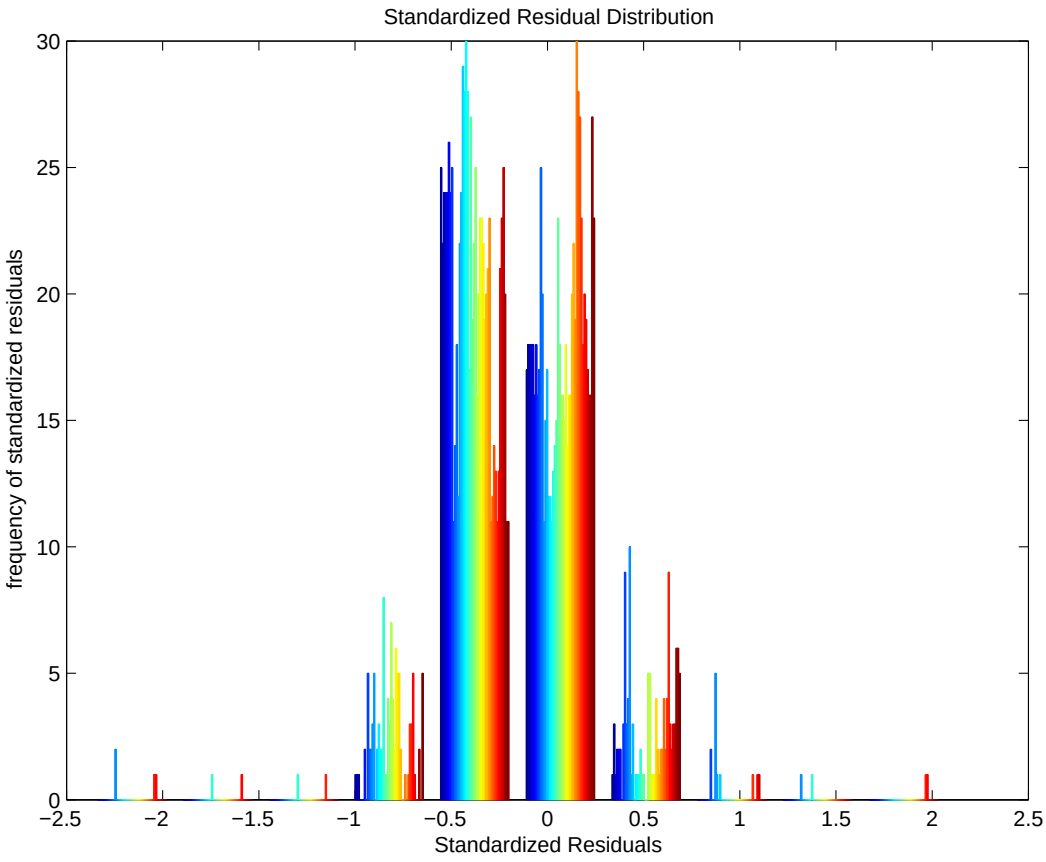

Figure 3: Histogram of the (8100) Bayesian residuals in a Model I setting. (Example 2)

4.3 Clustering: Example 2

Using the microarray data from example 2, we calculated confidence intervals for the true sensitivity/specificity for our procedures and compared this with sensitivity/specificity estimates for Pearson's estimator and Eisen's estimator. The results demonstrated that empirical Bayes estimator performs better for values of $\alpha < 0.6$, in the sense that the true sensitivity/specificity was superior. (Data not shown.)

4.4 Robustness: Example 1

For analyzing the robustness of empirical Bayes estimates, we calculate similarity measures in settings where the prior distribution is contaminated by a Gaussian distribution with high variance (see section 3.3). We calculate correlation in this setting when the amount ϵ of contamination is 5% and the *a priori* variance is between 0.1 and 4.0. Our results demonstrate that empirical Bayes estimators do as well as Pearson's estimator in such settings. (Data not shown.)

Acknowledgments

We would like to thank Vera Cherepinsky, Jiawu Feng and Marc Rejali for some insightful discussions and thank Gloria Coruzzi for the opportunity to test our software on microarray data generated from our collaborative N2010 project. We also thank Mike Wigler of CSHL for discussions of similar topics that stimulated our research. The work reported in this paper was supported by grants from NSF's Qubic program, NSF's ITR program, NSF's N2010 program, Defense Advanced Research Projects Agency (DARPA), Howard Hughes Medical Institute (HHMI) biomedical support research grant, the US Department of Energy (DOE), the US air force (AFRL), National Institutes of Health (NIH) and New York State Office of Science, Technology & Academic Research (NYSTAR).

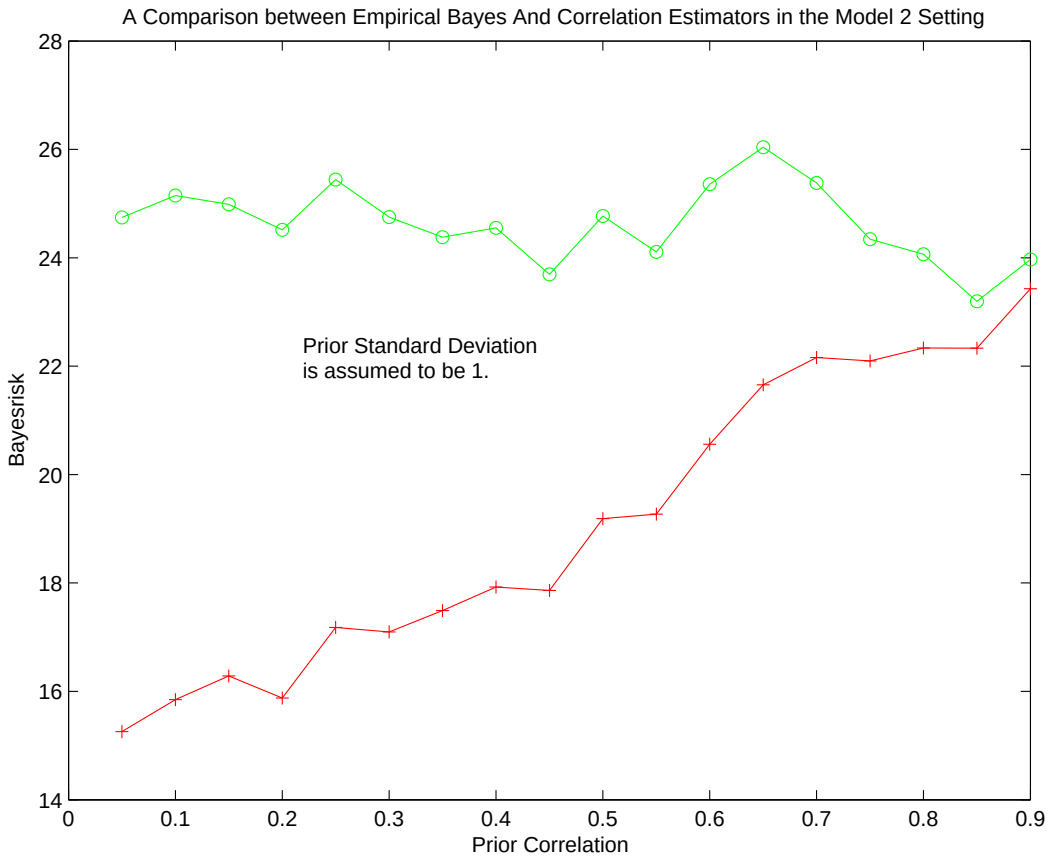

Figure 4: Comparison Between Bayes risks of Empirical Bayes versus Pearson Correlation estimates for prior mean= 0.0, prior intergene correlation $\rho = 0.3$ and prior standard deviation varying between 0.1 and 0.7 in a Model II setting (Example 1)

A Self-Organizing Map Clustering with Similarity at θ

We base our discussion on the Self-Organizing Map algorithm of Kohonen [7]. For given baseline vector θ, and reference vectors $\mu_1, \ldots, \mu_t$ we use the notation

$$\rho(X(j,:),\mu_l) = \frac{\sum_c (X(j,c)-\theta(j))(\mu_l(c)-\theta(j))}{\sqrt{\sum_c (X(j,c)-\theta(j))^2}\sqrt{\sum_c (\mu_l(c)-\theta(j)^2}} \quad (20)$$

for the similarity between the expression vector $X(j,:)$ and the reference vector μ_l. We employ the notation,

$$\mu_{c[j]} = \text{argmax}\{\rho(X(j,:),\mu_t)\} \quad (21)$$

and use ϵ as a smoothness factor. At stage j, we update the reference vector $\mu_{c[j]}^{(old)}$ by

$$\mu_{c[j]} = \mu_{c[j]}^{(old)} + \epsilon \cdot \Big\{ \big(X(j,:)-\theta(j)\big) - \rho(X(j,:), \quad (22)$$

$$\theta(j)) \cdot \big(\mu_{c[j]}^{(old)} - \theta(j)\big) \frac{\|X(j,:)-\theta(j)\|}{\|\mu_{c[j]}^{(old)} - \theta(j)\|} \Big\} \quad (23)$$

$$(24)$$

We continue to update the reference vectors in this fashion until convergence. In order to adjust this procedure to have critical value α, we only transform $\mu_{c[j]}^{(old)}$ if the similarity between $X(j,:)$ and $\mu_{c[j]}^{(old)}$ is greater than that of the critical value.

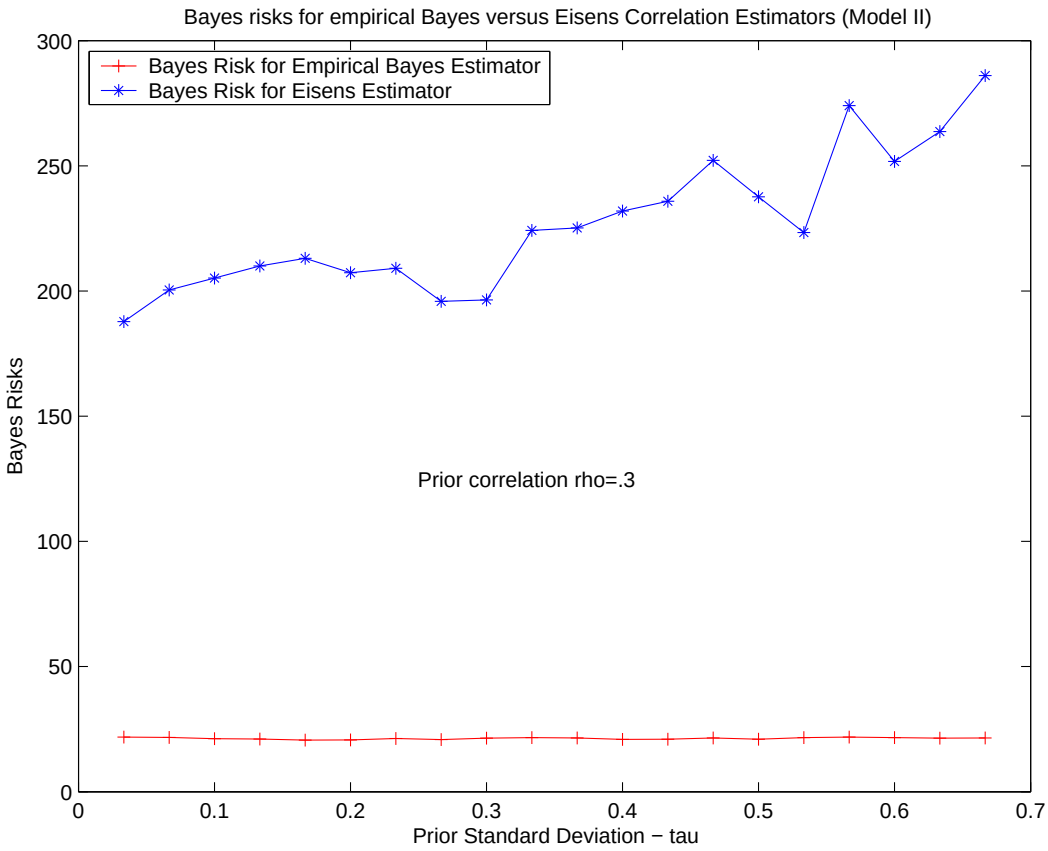

Figure 5: Comparison Between Bayes risks of Empirical Bayes versus Eisen's Correlation estimates for prior mean= 0.0, prior intergene correlation $\rho = 0.3$ and prior standard deviation varying between 0.1 and 0.7 in a Model II setting (Example 1)

References

[1] Baldi, P., and Hatfield, G.W., *DNA Microarrays and Gene Expression*, Cambridge, 2002.

[2] Carlin, B., and Louis, T., *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall, 1996.

[3] Cherepinsky, V., Feng, J., Rejali, M., and Mishra, B., "Shrinkage-Based Similarity Metric for Cluster Analysis of Microarray Data," *Proceedings of the National Academy of Sciences*, **100**(17): 9668–9673, 2003.

[4] Efron, B., "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, **1**(1): 55–75, 1986.

[5] Efron, B., "Empirical Bayes Methods for Combining Likelihoods," *Journal of the American Statistical Association*, **91**(434): 538–550, 1996.

[6] Eisen, M.B., Spellman, P., Brown, P.O., and Botstein, D., "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences*, **95**(25): 14863–14868, 1998.

[7] Kohonen, T., *Self Organizing Maps*, Springer, N.Y., 2001.

[8] Parmigiani, G., Garrett, E.S., Irizarry, R.A., and Zeger, S., *The Analysis of Gene Expression Data: Methods and Software*, Springer 2003.