The International Journal of Biostatistics

Volume 5, *Issue* 1 2009 *Article* 18

Likelihood Estimation of Conjugacy Relationships in Linear Models with Applications to High-Throughput Genomics

Brian S. Caffo, *Johns Hopkins University*
Dongmei Liu, *London School of Hygiene & Tropical Medicine*
Robert B. Scharpf, *Johns Hopkins University*
Giovanni Parmigiani, *Johns Hopkins University*

Recommended Citation:
Caffo, Brian S.; Liu, Dongmei; Scharpf, Robert B.; and Parmigiani, Giovanni (2009) "Likelihood Estimation of Conjugacy Relationships in Linear Models with Applications to High-Throughput Genomics," *The International Journal of Biostatistics*: Vol. 5: Iss. 1, Article 18.
DOI: 10.2202/1557-4679.1129

Likelihood Estimation of Conjugacy Relationships in Linear Models with Applications to High-Throughput Genomics

Brian S. Caffo, Dongmei Liu, Robert B. Scharpf, and Giovanni Parmigiani

Abstract

In the simultaneous estimation of a large number of related quantities, multilevel models provide a formal mechanism for efficiently making use of the ensemble of information for deriving individual estimates. In this article we investigate the ability of the likelihood to identify the relationship between signal and noise in multilevel linear mixed models. Specifically, we consider the ability of the likelihood to diagnose conjugacy or independence between the signals and noises. Our work was motivated by the analysis of data from high-throughput experiments in genomics. The proposed model leads to a more flexible family. However, we further demonstrate that adequately capitalizing on the benefits of a well fitting fully-specified likelihood in the terms of gene ranking is difficult.

KEYWORDS: multilevel models, hierarchical models, EM, microarray, gene-expression

Author Notes: The work on this grant was supported by NIH grants K25EB003491 and 5T32HL007024 and NSF grant DMS034211.

1 Introduction

Multilevel models (Carlin and Louis, 2000) are widely used for the simultaneous estimation of an ensemble of related quantities. In recent years, they have been applied successfully in situations where the number of quantities to be estimated simultaneously is very high. For example, estimands could be associated to each of the zip codes in a nation, or each of the genes in an organism. In these applications, it can be useful to rely on parametric models that allow for fast computation. From this standpoint a very convenient default approach has been to choose a so called conjugate multilevel model (see Ando and Kaufman, 1965; Lindley and Smith, 1972). However, this approach can be too restrictive, as they impose specific constraints on the relationship between sources of variations at different levels of the model. Moreover, it is typically applied without validating these assumptions. In this paper we explore the ability of the likelihood to detect specific departures from conjugacy. Furthermore, we propose a general class of multilevel models that highlights a particularly important continuous scale of departures from conjugacy. We investigate the potential robustness offered by estimating the conjugacy relationship as a component of model building.

Our development is motivated by the analysis of data from gene expression microarrays, and is illustrated in that context, although the techniques can be applied much more broadly. These experiments study the expression levels of thousands of genes across experimental conditions. Because the number of replicates is typically small, the borrowing of strength afforded by multilevel models can be critical and multilevel models are widely applied (see Cui et al., 2005; Baldi and Long, 2001; Newton et al., 2001; Lönnstedt and Speed, 2002; Ibrahim et al., 2002; Parmigiani et al., 2002; Su et al., 2007). While commonly used, the assumption of conjugacy is often violated by the data, as shown in Liu et al. (2004). On the other hand, the large number of genes potentially gives one the opportunity to check this assumption and reliably fit more complex models.

To provide an idea of the nature of the generalization investigated in this article, consider experiments that compare two groups across several units of interest, such as genes. Following Liu et al. (2004), we focus on three parameters of interest. The first is a mean difference in the response between two groups for each unit, referred to as the unit-specific signal; the second is the overall expression for a gene across the groups, called the unit-specific abundance and the third is the standard deviation of the expression measure for a unit, called the unit-specific noise. Throughout we assume the noise is common across the two groups being compared.

Generally, conjugacy refers to the property that both the prior and the posterior distribution of these parameters, conditional on hyperparameters, belong to the same family, in this case a normal-gamma. In this context, to achieve conjugacy one needs to assume a) independence of the signal to noise ratio and the noise, and b) independence of the abundance to noise ratio and the noise. Liu et al. (2004) consider the impact of relaxing this assumption in favor of independence between the signals and the noise and the abundances and the noises. In this manuscript, we present a class of distributions that formally estimates the relationship between the signal and noise. In addition, our formulation allows for separate variances on the distribution of signals and abundances and the possibility of a positive correlation between the signals and abundances. All analyses and simulations were performed in the statistical language R (Ihaka and Gentleman, 1996). Software fore fitting can be downloaded from

`http:\\www.biostat.jhsph.edu\~bcaffo\downloads.html`

The paper is laid out as follows. In Section 2 we present the new model while Section 3 gives an easily implemented expectation maximization (EM) fitting algorithm. In Section 4 we evaluate the performance of the power conjugate model in a simulation study. In Section 5 we analyze data from three lung cancer studies using the model. Finally, Section 6 presents a summary discussion.

2 A class of two-stage multilevel models

We develop the model in general though focus our examples and discussion on two group comparisons. Let y_g be a vector of length J comprised of outcome measures for unit $g = 1, \ldots, G$. In the microarray setting, typically the y_g are log expression ratios from a two channel array, or background adjusted and normalized log expressions from an oligonucleotide array. In the settings we have in mind, the number of units, G, is large. Let X be a full rank $J \times p$ matrix of independent variables of interest. A common model of interest for a single unit, g, would specify

$$y_g \mid \beta_g, \lambda_g \sim \mathrm{N}(X\beta_g, \lambda_g I) \tag{1}$$

where β_g is a p-vector of unit-specific effects and λ_g is univariate, unit-specific, variance.

While such a model is generally appropriate for one or a few units, applications such as gene expression arrays require hundreds or thousands of models

DOI: 10.2202/1557-4679.1129

such as (1). In this setting gene-specific maximum likelihood is wastefully ignorant of the aggregate information contained across units. Therefore, we specify a hierarchical model by assuming

$$\beta_g \mid \lambda_g, \delta, \mu, F \sim \mathrm{N}(\mu, F\lambda_g^{\delta}) \quad \text{and} \quad \lambda_g \mid \nu, \tau \sim \mathrm{IG}(\nu, \tau), \tag{2}$$

where IG is short hand for the inverted gamma density,

$$\frac{\tau^{\nu}}{\Gamma(\nu)} \lambda_g^{-\nu-1} \exp(-\tau/\lambda_g).$$

Here μ is the $(p \times 1)$ inter-unit mean of the β_g and F is an $(p \times p)$ unstructured variance matrix. The parameter δ embodies the development discussed in the paper. Notice that δ controls the a-priori dependence between β_g and λ_g; if $\delta = 0$, then there is a-priori independence. Also, notice that if $\delta = 1$, then a standard conjugate model is obtained.

In this manuscript, we explore the ability of the marginal likelihood (obtained by integrating over β_g and λ_g) to identify δ, as well as the benefits and limitations for its likelihood estimation. When δ is estimated to be either 0 or 1, the problem is simply model selection between the conjugate and independence models. However, intermediate values of δ could also be of interest (see Scharpf et al., 2008).

To illustrate this parameter, consider Figure 1 which displays an example of the conditional density (in gray scale) of β_g given λ_g for various values of δ. Negative values of δ suggest that the variation in the signals decreases with the magnitude of the noise, while positive values suggest an increase. The latter situation is more common in practice, and so negative values of δ are usually not of interest and will not be considered.

The rate of increase of the conditional variability of β_g for given noise gets larger with δ (see Figure 1). Hence, a large value of δ assumes that large variation in the signals implies large residual error. In contrast, a δ value of zero suggests independence, in which case the variation in the signals offer no evidence regarding the noise. In addition to illustrating the impact of δ, Figure 1 suggests how δ can be identified given observed data. For example, one could plot histograms or boxplots of the sample signals by binned values of the estimated noises (Liu et al., 2004).

Recall, that the conditional variation in the β_g increases with λ_g when $\delta > 0$. Therefore, when averaged over the distribution of λ_g, β_g will have heavier tails than the independence case. This is illustrated in Figure 2, which displays an example marginal distribution of β_g for different values of δ.

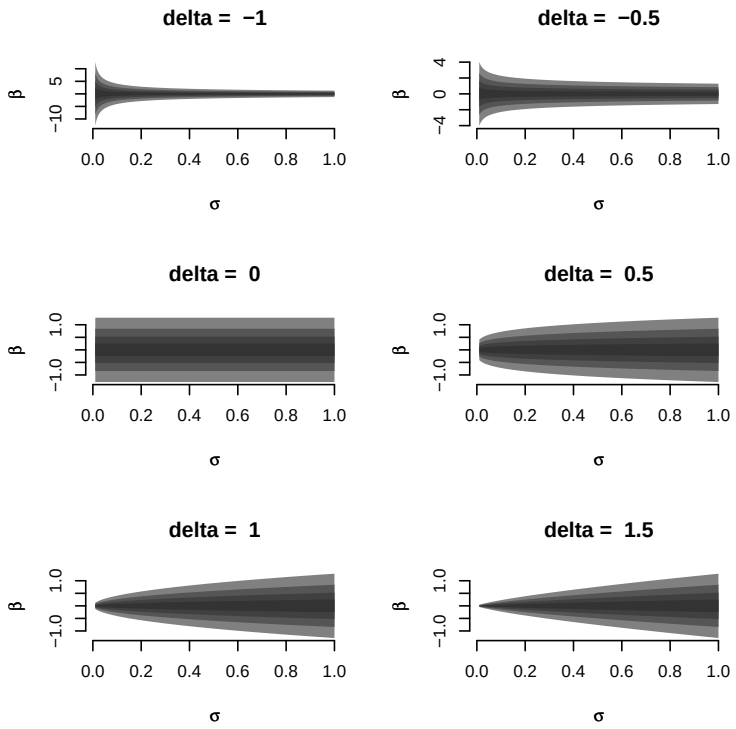

Figure 1: Illustration of the impact of δ. Plotted is the conditional distribution of a scalar random effect slope parameter, β, with mean value 0, conditional on varying values of λ. Gray scales represent the height of the density (darker being higher). The independence model is given when $\delta = 0$ while the conjugate model occurs when $\delta = 1$.

In this article we focus on two group comparisons. For example,

$$X = \begin{pmatrix} 1 & .5 \\ \vdots & \vdots \\ 1 & .5 \\ 1 & -.5 \\ \vdots & \vdots \\ 1 & -.5 \end{pmatrix} \qquad \beta_g = \begin{pmatrix} \beta_{0g} \\ \beta_{1g} \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} \qquad F = \begin{pmatrix} F_0 & F_{01} \\ F_{01} & F_1 \end{pmatrix}.$$

Here we refer to β_{0g} as the gene-specific abundance, β_{1g} as the gene-specific signal and $\lambda_g^{1/2}$ as the gene-specific noise. For two group comparisons, if an intercept is fit, it is important to code the columns of X corresponding to group status as .5 for one experimental group and $-.5$ for the other, rather

DOI: 10.2202/1557-4679.1129

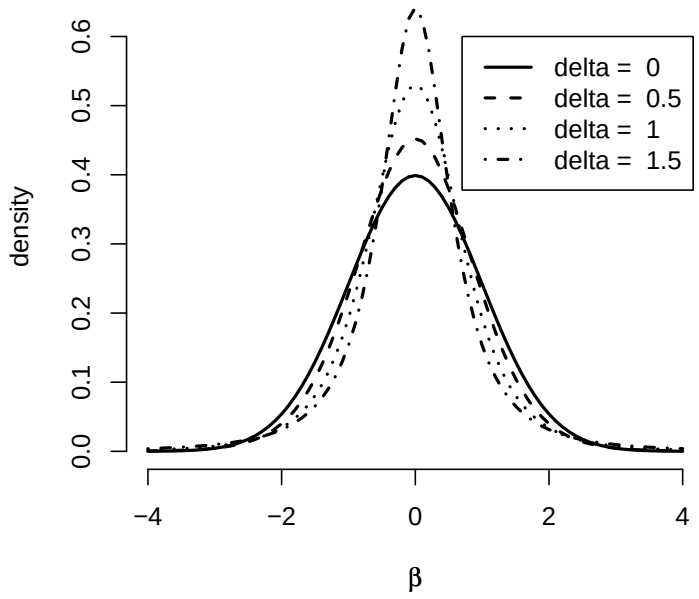

Figure 2: Illustration of the impact of δ. Plots of the marginal density of β for differing assumptions regarding δ.

than the 0 and 1 coding commonly used for fixed effects. The latter coding can have the perhaps unintended effect of assuming a larger marginal variance for the second group.

A closely related modeling approach is presented in Liu et al. (2004). In particular, their model specified that the conditional variance of β_g given λ_g is

$$\begin{pmatrix} F_0\lambda_g^{\delta_0} & 0 \\ 0 & F_1\lambda_g^{\delta_1} \end{pmatrix}. \tag{3}$$

They subsequently characterized the model with $(\delta_0, \delta_1) = (0, 0)$ as complete independence, the model with $(\delta_0, \delta_1) = (1, 1)$ as complete conjugacy. These are special cases of model (2) with $\delta = 0$ and 1 respectively and $F_{01} = 0$. However, they also considered cases $(\delta_0, \delta_1) = (0, 1)$ and $(\delta_0, \delta_1) = (1, 0)$, which they referred to as independence-conjugate and conjugate-independence models respectively. Thus, their model explicitly allows for separate conjugacy relationships for the signal and noise.

Our approach differs markedly from that study by: (i) allowing for the correlation, F_{01}, to be non-zero, (ii) focusing on maximum likelihood estimation and fully specified models rather than real-time moment-based statistics, (iii) assuming that $\delta_0 = \delta_1$, (iv) considering values of δ other than 0 and 1.

3 Maximum likelihood using the EM algorithm

We discuss computational fitting with regard to model (2). Maximum likelihood estimates of F, μ, ν and τ are obtained by maximizing the marginal likelihood

$$\int f(y_g|\beta_g, \lambda_g) f(\beta_g|\lambda_g; \delta, \mu, F) f(\lambda_g; \nu, \tau) d\beta_g d\lambda_g \tag{4}$$

where f denotes a generic density. We label these maxima, $\hat{F}$, $\hat{\mu}$, $\hat{\nu}$ and $\hat{\tau}$. We suggest calculating these maxima using the EM algorithm. It will be seen that this choice results in considerable simplification in calculations. In practice, we have seen that the convergence of the algorithm is quite acceptable, usually requiring only a few minutes. While not quite real-time in the sense of Liu et al. (2004), we find that the algorithm is generally stable and requires virtually no user input.

When $\delta = 1$, $\hat{F}$, $\hat{\mu}$, $\hat{\nu}$ and $\hat{\tau}$ are easily obtained using EM, since all of the relevant integrals are tractable. For arbitrary δ, the relevant integrals for applying EM or marginal maximization are not tractable and hence we apply numerical integration techniques based on Gaussian quadrature. Below, we outline the fitting. One important note is that it is much easier to calculate a profile likelihood for δ rather than incorporating maximization with respect to δ into the M step of the EM algorithm. In general, rather than maximizing the profile likelihood numerically, we suggest using a grid search. This is largely because we have found that a fine grid for δ is not necessary.

Before we discuss the details of the fitting algorithm, we briefly discuss EM. A more complete introduction can be found in Lange (1999). To put this problem into the framework of EM, we consider the y_g as the "observed data", the (β_g, λ_g) as the "missing data" and their combination as the "complete data". Given starting values, the EM algorithm maximizes (4) by iteratively maximizing the expected value of the complete data log-like likelihood given the observed data and the previous parameter estimates, see (6) in the Appendix. In what follows we discuss the details of a single EM update for fixed values of δ. Minor modifications are necessary to adapt the algorithm for model (3). We note that starting values can be obtained from the complete conjugacy

DOI: 10.2202/1557-4679.1129

case ($\delta = 1$), where no numerical integration is required, or by the method of moments approximations (see Liu et al., 2004).

The $(t+1)$st parameter estimates, say, are obtained by separately maximizing the marginal log likelihood (see Appendix equation (7)) with respect to μ and F and (8) with respect to ν and τ. Benefits of applying EM are that (7) and (8) can be maximized independently and that (7) has the closed form solutions

$$\mu^{(t+1)} = \sum_{g=1}^{G} E_t^*[\beta_g]/G$$

and

$$F^{(t+1)} = \sum_{g=1}^{G} E_t^*[(\beta_g - \mu^{(t+1)})(\beta_g - \mu^{(t+1)})'\lambda_g^{-\delta}]$$

provided F is unstructured. The notation, E_t^*, is defined in Appendix A. Closed form solutions for μ and F for fixed (δ_0, δ_1) also exist if model (3) is employed.

Maximizing (8) with respect to ν and τ is equivalent to maximizing a likelihood of independent, identically distributed gamma variates for which we employ an algorithm from Johnson et al. (1995). The previous parameter estimates, $\nu^{(t)}$ and $\tau^{(t)}$, provide good starting values.

It is possible to obtain REML-type estimates of ν and τ to avoid using this iterative algorithm within the M step. Notice that $z_g \equiv Hy_g$ is normally distributed given λ_g with mean 0 and variance $HH'\lambda_g$ where H is any $J-p \times J$ matrix whose rows span the null column space of X. For example H could be comprised of any of the $J - P$ row vectors of $I - X(X'X)^{-1}X'$. The inverted gamma marginal distribution for λ_g is indeed conjugate; therefore, estimates of ν and τ can be obtained from the distribution of the z_g via marginal maximization independently of μ and F. We obtain these REML-type estimates for ν and τ, which only need to be calculated once, before starting the EM algorithm to obtain estimates for μ and F.

The following facts are useful for application of the algorithm:

1. Since the distribution of $\beta_g|\lambda_g, y_g$ is normal, each intractable integral can be reduced (by iterating expectations) to a univariate integral involving only λ_g.

2. Only one pass through the gene index g is required per EM iteration. In particular, only the five quantities $\sum_{g=1}^{G} E_t^*[\log \lambda_g]/G$, $\sum_{g=1}^{G} E_t^*[\lambda_g^{-1}]/G$, $\sum_{g=1}^{G} E_t^*[\beta_g]/G$, $\sum_{g=1}^{G} E_t^*[\beta_g \lambda_g^{\delta}]/G$ and $\sum_{g=1}^{G} E_t^*[\beta_g \beta_g' \lambda_g^{\delta}]/G$ need to be calculated and stored, where E^* is defined in the Appendix.

3. Without accurate starting values, cycling through the entire gene index is unnecessary early on in the algorithm. Instead one can cycle through a random subset of genes increasing the size of the subset as the algorithm progresses (Caffo et al., 2002).

Remaining is a discussion of how to approximate the intractable expectation. Because the distribution of $\beta_g|\lambda_g, y_g$ is tractable, the E-step can be reduced to finding a method for approximating integrals with respect to the distribution of $\lambda_g|y_g$ (see 9). Since this approximation must be applied for every gene once for every EM step, numerical integration is perhaps preferable to Monte Carlo (though see Caffo et al., 2002). We implement Gauss/Laguerre integration (see Press et al., 1992) described in the Appendix. In application, we have found 50 quadrature points are often sufficient to obtain stable parameter estimates. Gauss/Laguerre integration is only one of many techniques for numerical integration and other approximations may be worthy of further study these applications.

4 A simulation study

In this section we evaluate model performance when estimating δ via an extensive simulation study. We focus on three aspects of modeling: i) estimation of δ, ii) the importance of modeling the covariance between β_{1g} and β_{2g} and iii) robustness of rankings offered by estimating δ. For the third goal we assume, as is the case in many microarray experiments, that the short-term objective is to generate a list of genes that are likely to be different across the two conditions. Differences of interest depend on the experimental goal, and so do the statistics used to select genes. Here we focus our attention on detecting reliably measured differences via signal-to-noise statistics. Notationally, the experimental goal is to select the genes with the largest absolute value of $\beta_{1g}/\lambda_g^{1/2}$. A natural statistic from the multilevel model is the posterior mean

$$E[\beta_{1g}/\lambda_g^{1/2} \mid y_g, \hat{\mu}, \hat{F}, \hat{\nu}, \hat{\tau}], \tag{5}$$

where the estimated parameters are plugged-in after taking the expectation. In our simulation study we compare rankings based on this statistic assuming conjugacy ($\delta = 1$) and independence ($\delta = 0$) to those with δ estimated.

We assumed $G = 1,000$ or $10,000$ genes with 4 or 32 replicates per group, $J = 8$ or 64. We considered a variety of parameter settings and simulated 100 complete data sets per setting. The following transformed parameters are most useful for summarizing the results: (i) the overall ratio of the standard

DOI: 10.2202/1557-4679.1129

deviations of the abundances to the signals, $\phi = \sqrt{F_0/F_1}$, (ii) the correlation between the signals and abundances, $\rho = F_{01}/F_0^{1/2}F_1^{1/2}$, and (iii) the coefficient of variation, CV, of the inverted gamma distribution and (iv) the mean, M, of the inverted gamma distribution on the λ_g. Finally we also considered four values of δ. To summarize, the specific parameter values are all of the combinations of

$$\phi \in \{.5, 1, 2\} \quad \rho \in \{0, .5\} \quad CV \in \{.5, 1, 2\} \quad M \in \{.5, 1, 1.5\} \quad \delta \in \{0, .5, 1, 1.5\}.$$

The simulations were conducted using the R open source programming language (Ihaka and Gentleman, 1996). Because of the extensiveness of the simulations, we highlight the most salient results that were consistent across simulation settings.

Estimating δ

To obtain maximum likelihood estimates for δ, we performed both coarse, $\delta \in \{0, 1\}$, and fine, $\delta \in \{0, .1, .2, \ldots, 1.5\}$, grid searches for each simulation setting. Table 1 shows the modal estimates of δ from the coarse grid search across the 100 simulations by the true value of δ for $G = 1,000$ and $G = 10,000$ and $J = 8$ and 64; results are collapsed over the remaining simulation parameters and combined in a single contingency table. This table highlights a consistent finding; that is, low estimated values of δ tend to be accurate. For example, Table 1 depicts counts across simulation settings of the modal estimate of δ within each simulation setting. The table suggests that when considering only conjugacy or independence, an estimated δ of 0 strongly suggests independence, while an estimated δ of 1 apparently offers less insight into the the true conjugacy relationship across the simulation parameters considered. However, this conclusion is the result of the fitted model tending to prefer the conjugacy model over independence for the parameter settings chosen. Such a result may be due to the fact parameter settings are not entirely exchangeable; for example, fixing the remaining parameters and switching from independence to conjugacy adds marginal variability in the signals and abundances.

The pattern was similar for the finer grid search for δ. Figure 3 shows the average estimated value of δ by the true value across the remaining parameters when $J = 8$ and $G = 1,000$. Again this figure suggests greater accuracy for smaller estimated values of δ and, in general, the preference of the model for choosing larger values of δ.

	Modal estimated δ							
	$J = 8$				$J = 64$			
	$G = 1,000$		$G = 10,000$		$G = 1,000$		$G = 10,000$	
True value	0	1	0	1	0	1	0	1
0	31	23	34	20	29	24	37	11
.5	1	53	6	48	0	52	0	48
1	0	54	0	54	0	54	0	47
1.5	0	54	0	54	0	54	0	48

Table 1: Simulation results for likelihood-based estimation of δ. For each parameter setting, the modal estimate of δ (either 0 or 1) across simulations was obtained. Shown above are the counts of each mode for various values of J, G and true values of δ, collapsing over the remaining simulation parameters.

Impact of the signal/abundance correlation

We compared Model (2) that estimates the correlation between signals and abundances, F_{01}, to with the model where $F_{01} = 0$. When it is assumed to be zero, we considered both estimating δ and fixing it at either complete conjugacy or complete independence. Unlike (3), however, we do not consider instances where the conjugacy relationship differs for the signals and abundances. We compare areas under the receiver operating characteristic curve (AUC) for detecting the largest 1%, 5% and 10% signal to noise ratios across the two groups. We compare the median AUC across the 100 simulations for each of the parameter settings.

Incorrectly assuming that the correlation between the signals and abundances is zero can have a very negative impact on results, especially for very small sample sizes. For example, consider comparing the model that performs a coarse grid search for δ and estimates the correlation between signals and abundances and to the model that assumes the correlation is zero. When the number of observations and genes was smaller ($J = 8$ and $G = 1,000$) there were numerous instances where fixing the covariance to be zero ($F_{12} = 0$) resulted in very poor operating characteristics. Figure 4 displays the ratio of the median AUCs for the best performing model presuming a correlation versus the best performing model presuming independence. Points to the right of the figure have a higher assumed covariance between the signals and abundances. The figure displays that the penalty for incorrectly assuming a correlation is generally less than incorrectly assuming independence. This figure omits some 30% of the simulation settings under independence where the models assuming independence had very poor performance, sometimes with AUCs less than

DOI: 10.2202/1557-4679.1129

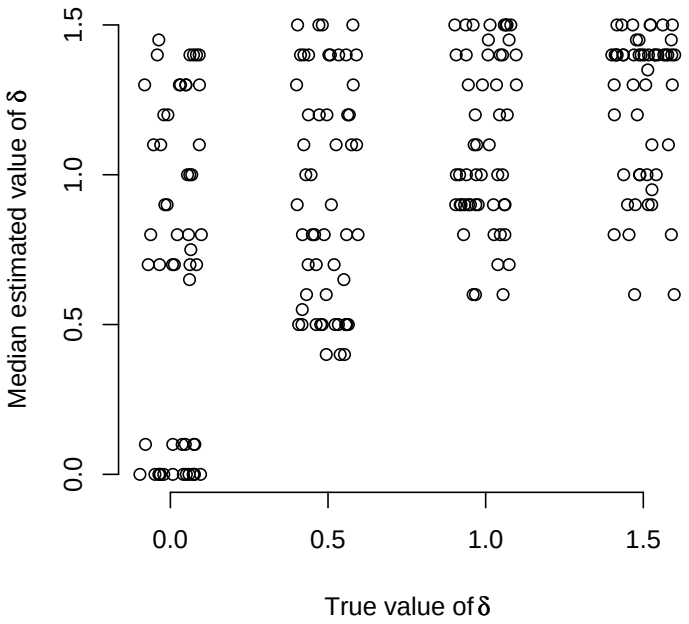

Figure 3: Median estimated values of δ for the simulation study across the remaining simulation parameters, with $J = 8$ and $G = 1,000$ by the the true value used for simulation.

0.50. These are not shown, as we stipulate that those extremely poor operating characteristics may be due to numerical instabilities, though no evidence suggests that this is the case.

There were instances where the best model allowing for possible correlation outperformed the assumption of independence, even when even when the actual correlation used for simulation was zero. We believe this is a result of the additional joint variability in the signals and abundances implied by the conjugacy relationship. However, the general pattern suggests that one is safer allowing for the possibility of a correlation rather than assuming independence, even if some higher order aspects of the model are incorrectly specified. Hence these results mirror conventional wisdom regarding random effect slopes and intercepts.

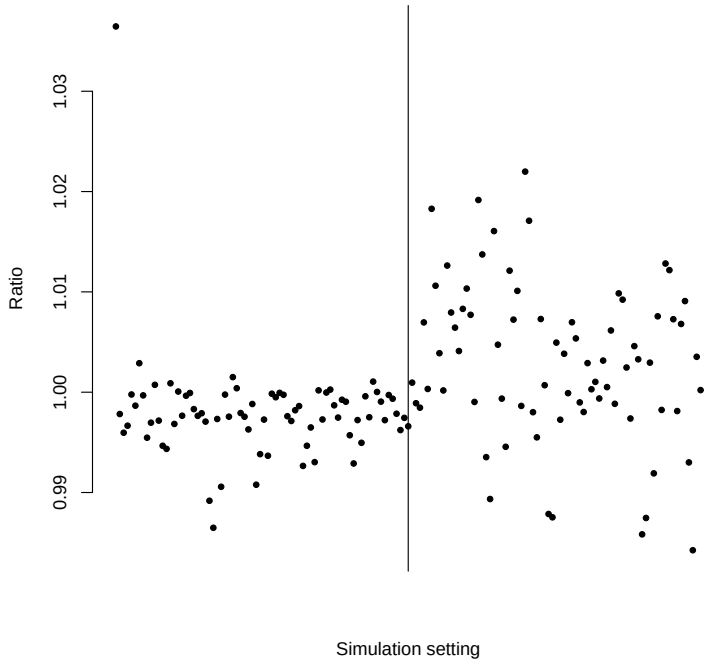

Figure 4: Ratios of median areas (taken across simulation iterations) under the receiver operating characteristic curve comparing the best performing model allowing for a non-zero signal/abundance correlation to the best fitting model assuming independence for each simulation setting for $J = 8$ and $G = 1,000$. Shown are the results for detecting the top 5% of signal-to-noise ratios. A horizontal line divides the points where the correlation used for simulation was 0 (left) .5 (right).

Impact of misspecification of δ

In this section we consider the potential robustness that estimating δ offers to misspecification. To focus the discussion, consider only the coarse search where $J = 8$ and $G = 1,000$; the remaining cases were similar, though more favorable toward estimating δ as the number of genes and samples increased and less favorable with the finer grid search.

In terms of the median AUC for the coarse search, the model with the estimated δ outperformed one of either the conjugate or independence model 23% of the time. Overwhelmingly, the instances where estimating the conjugacy relationship performed well in the terms of AUC was when $F_1 > F_0$ or $\rho = .5$.

DOI: 10.2202/1557-4679.1129

Moreover, the instances where estimating δ showed an improvement, was when the true model was conjugacy and the model selected conjugacy. Moreover, the improvements tended to be modest, with an average 1% improvement over the worst model, in the cases where estimating δ represented the best model.

In contrast, the model assuming $\delta = 0$ was the best on 50% of the simulation settings. The conjugate model, $\delta = 1$, was the best on 50%, though tied with model with an estimated δ on 23% of those cases. Surprisingly, the model performance was largely unaffected by the true value of delta. For example $\delta = 1$ for only 50% of the simulation settings where assuming conjugacy resulted in the best model. This suggests a complex interplay between: the conjugacy relationship, the other model parameters, the statistic used for ranking, and ranking performance.

Because of concerns over the role that fitting played a part in the simulation conclusions we conducted a smaller scale simulation study using Gibbs sampling with diffuse priors for fitting. Again the program was coded in the R statistical programming language. Since the results from this simulation confirmed the ML results, they are not presented.

Practical recommendations based on simulation evidence

The results of the simulation study suggest that we cannot give an unqualified recommendation for using the likelihood to estimate the conjugacy relationship. The higher order components of the model defining the conjugacy relationship are only weakly identified, even in genomic settings. The simulation results were more clear on the benefit and robustness for modeling a non-zero correlation between signals and abundances, even if one specifies an incorrect conjugacy model.

With regard to the coarse grid search for δ, i.e. likelihood based model selection between the conjugate and independence model, the simulation results suggest that an estimated independence model appears more trustworthy than an estimated conjugate model. We believe this is in part due to the added marginal variability imposed by the conjugate model. That is, the simulation settings for the conjugate model had greater variability in the signals and abundances than those of the independence model. As a consequence, it appears that confirming conjugacy via the likelihood requires a larger sample. With regard to finer estimation of δ, because of the difficulty in estimation, a very fine grid, or numerical maximization algorithm does not seem warranted. Again, the results of the finer grid search suggested that lower values appear to be more trustworthy than larger.

Further complicating the discussion was that there was no clear winner

among the strategies "always choose conjugacy" and "always choose independence" for ranking signal-to-noise ratios. The strategy of estimating δ is not preferable to the strategy of "always choosing conjugacy." However, such a statement only considers ranking and selection of the topmost signal-to-noise ratios. It appears that likelihood-based estimation of the δ parameter is most accurate and offers no penalty in performance when there is a large degree of correlation between the signals and abundances and when there is potential imbalance in the variance of the signals and abundances. We believe that this largely depends on the assumption of the same conjugacy relationship across the signals and abundances. For example, having low variation in the abundances allowed for a better estimate of δ while having high variation in the signals increased the impact that the model assumptions had on the shrinkage of the test statistics.

5 Differential expression between lung cancer subtypes

The data for this example were collected from three separate studies from researchers at Harvard University (using Affymetrix Hu95a), University of Michigan (using Affymetrix HG6800) and Stanford University (using cDNA) (Bhattacharjee et al., 2001; Beer et al., 2002; Garber et al., 2001). We demonstrate that the Stanford data set has properties that would make estimating the conjugacy relationship worthwhile. Here we focus on measures of expression taken for 307 genes using cDNA microarrays. These 307 genes were selected as the most variable subset of a larger group of several thousand in an analysis that combined the three studies (Parmigiani et al., 2004) to compare survival rates over gene expression profiles. In the Stanford data set, there were 68 samples with 31 cases and 37 controls. The outcome is the log-relative expression from the two-channel array.

Figure 6 in Appendix B shows boxplots of the separately estimated signals minus their overall means divided by the noises to the power δ for $\delta = 0, 1$ and estimated using a fine grid search. Figure 7 in Appendix B replicates this plot for the separately fit abundances. Here "separately estimated" refers to taking the simple gene-specific overall means, difference in means and pooled variances. None of the models presented in this paper appear to hold for the Harvard or Michigan studies, because of a marked increase of the noises with the abundances. For the Stanford study, the assumption of model (2) appears to hold, with a similar dependence structure for both the signals and variance and abundances and variances. It is difficult, however, to determine

DOI: 10.2202/1557-4679.1129

a reasonable value for δ from the plot alone. Figure 5, shows the profile likelihood for δ (with reference lines drawn at 1/32 and 1/8, see Royall, 1997) for the Stanford data set. The model fits suggest a value between conjugacy and independence.

Interestingly, there is significant correlation between the signals and abundances in the Stanford data set. That is, $F_{01}/F_0^{1/2}F_1^{1/2}$, was estimated to be $-.64$, which is validated by the estimated correlations of the separately fit signals and abundances of $-.61$. Also, the estimates of the between-gene abundances and signal variances (F_0 and F_1 respectively) are quite different, with F_0 being 56% larger than F_1. That is, this data set has some of the hallmarks where the simulation study suggested that estimation of the conjugacy relationship offered no penalty in ranking performance for signal to noise ratios.

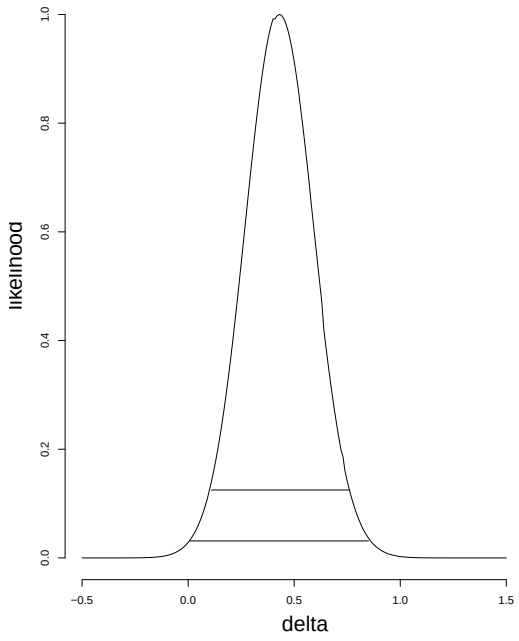

Figure 5: Profile likelihood for δ for the lung cancer data from Section 5.

6 Discussion

In this article we investigated the ability of the likelihood to discern specific departures from conjugacy in multilevel models. We gave a real-data example illustrating this estimation that has some of the hallmarks where the approach may work well. The proposed class of models estimates the relationship between the variation of group-specific means and the noise with which those are measured, while traditional approaches impose restrictions on such relationship. Therefore, this class allows for a wider variety of data patterns than the most common multilevel models while still retaining a completely specified likelihood. We presented an efficient EM algorithm for fitting data using this family of models.

We explored the potential of using this class of multilevel models for determining differential expression in two group comparisons of gene expression arrays. We emphasized two important modeling considerations: diagnosing the correct form of the conjugacy relationship between gene specific mean parameters and the variance parameter, and addressing correlation between signals and abundances. A potentially more robust approach for future research would employ non-parametric methods to fit more flexible models in these settings (see Newton et al., 2004, for example.).

Our simulation study suggests tempered enthusiasm for likelihood-based estimation of conjugacy relationships for the purposes of gene ranking by signal-to-noise ratios. Identifying higher-order components of the model is difficult, even with the ensemble of information contained in thousands of genes. We found that estimates closer to independence tend to be more reliable than those suggesting conjugacy. We found that when there is a high correlation between signals and abundances, while a large discrepancy in their variances, estimating the conjugacy relationship resulted in good performance. However, in these settings, the model tended to select conjugacy. Hence, there is evidence to suggest that the likelihood can select between conjugate and independence models and potentially reap the benefits of a better fitting model. However, the simulation results suggest that adequately capitalizing on a better fitting model is a difficult task. In addition, the simulation study suggests a complicated relationship between the degree of conjugacy and the remaining parameters, statistic used, ranking objective and performance.

One must consider the following caveats when interpreting the simulation evidence. First, though extensive and requiring several months of computer time, the simulation settings studied represented only a small subset of possible real-world settings. Secondly, our focus on estimating reliably measured genes via signal-to-noise statistics is only a small subset of possible methods to rank

DOI: 10.2202/1557-4679.1129

genes. For example, one could similarly rank genes via posterior medians, Bayes factors, posterior tail probabilities, best linear predictors of the signal, the joint posterior distributions of the ranks and so on. In fact one could consider the plethora of methods in which a fully specified model could be used as both a strength and weakness of such an approach.

Our investigations showed that other signal-to-noise, such as the moderated T statistic of Smyth (2004), statistics performed well across simulations. Their performance notwithstanding, it is worth considering the strengths of a completely specified model-based approach over robust signal-to-noise statistics, which includes a more complete description and summary of the data, as well the ability to rank genes by other properties and a mechanism for extending results to a population. In addition, there may be benefits to a complete Bayesian solution that are overlooked in the current study. These include the diversity of statistics available from Bayesian machinery. Moreover, Bayesian methods for model averaging and model search might improve over strict likelihood-based methods for selecting conjugacy relationships.

The use of multilevel models to rank gene expression differences is a much more subtle process than using moment-based signal-to-noise statistics. Of particular importance is the fact that the large volume of genes under investigation can magnify the negative effects of unmet assumptions in models. Therefore, more flexible models need to be considered for statistics based on multilevel models to be on par with robust statistics. However, our investigation also illustrates that, while addressing higher order features in the data is necessary, constructing robust models to perform the estimation typically can force a compromise between a better fitting model and the added complexity, additional parameters, and ranking performance.

A Notes on the EM algorithm

Let $\nu^{(t)}$, $\tau^{(t)}$, $\mu^{(t)}$ and $F^{(t)}$ be the current parameter estimates (for fixed δ), EM maximizes

$$\sum_{g=1}^{G} E\left[\log\left\{f(y_g|\beta_g,\lambda_g)f(\beta_g|\lambda_g;\delta,\mu,F)f(\lambda_g;\nu,\tau)\right\}|\, y_g,\nu^{(t)},\tau^{(t)},\mu^{(t)},F^{(t)}\right]. \quad (6)$$

To simplify notation, we denote expectations with respect to this distribution as E^*.

The contribution of the terms involving μ and F to the expected complete

data log-likelihood is

$$-\frac{1}{2}\log|F| - \sum_{g=1}^{G} E_t^*[(\beta_g - \mu)' F^{-1}(\beta_g - \mu)\lambda_g^{-\delta}]/2G. \tag{7}$$

The contribution of the terms involving ν and τ to the expected complete data log-likelihood is

$$\nu \log \tau - \log \Gamma(\nu) - \nu \sum_{g=1}^{G} E_t^*[\log \lambda_g]/G - \tau \sum_{g=1}^{G} E_t^*[\lambda_g^{-1}]/G. \tag{8}$$

Let $\theta_g = 1/\lambda_g$; then

$$\begin{aligned} & f(\theta_g|y_g; F, \mu, \tau, \nu; \delta) \\ \propto\ & \theta_g^{J/2+p\delta/2+\nu-1}|X^tX\theta_g + F^{-1}\theta_g^{\delta}|^{-1/2} \\ \times\ & \exp\{-y_g^t y_g \theta_g/2 - \theta_g\tau + y_g^t X(X^tX\theta_g + F^{-1}\theta_g^{\delta})^{-1}X^t y_g \theta_g^2/2\} \\ =\ & |X^tX\theta_g + F^{-1}\theta_g^{\delta}|^{-1/2} \exp\{y_g^t X(X^tX\theta_g + F^{-1}\theta_g^{\delta})^{-1}X^t y_g \theta_g^2/2\} K(\theta_g) \end{aligned} \tag{9}$$

where K is a gamma kernel.

In Gaussian/Laguerre integrations, integrals are approximated via a trapezoidal rule,

$$\int_0^{\infty} h(\theta_g) dK(\theta_g) \approx \sum_{i=1}^{n} h(n_i) w_i,$$

where the nodes, n_i, are the zeros of the n^{th} order Laguerre polynomial and the weights are constructed so that the resulting approximation is exact when h is a polynomial of degree n or lower.

DOI: 10.2202/1557-4679.1129

B Figures

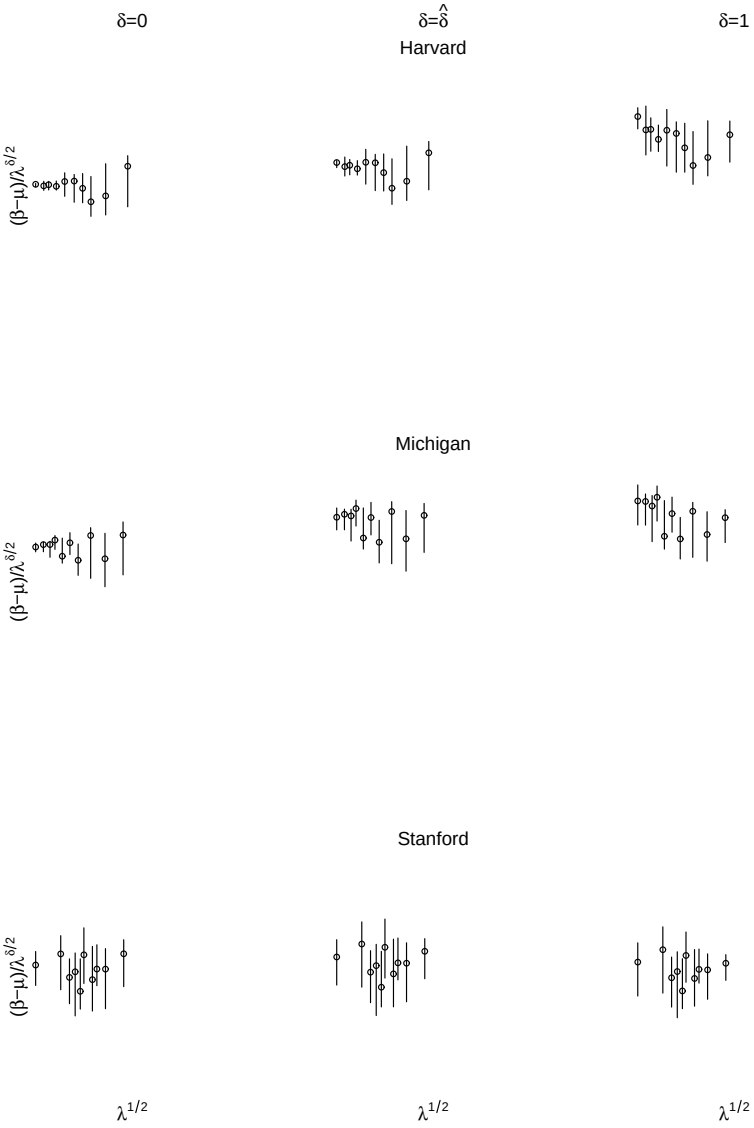

Figure 6: Boxplots of $(\hat{\beta}_{1g} - \hat{\mu}_1)/\hat{\lambda}_g^{\delta/2}$ by deciles of $\hat{\lambda}_g^{1/2}$ for $\hat{\beta}_{1g}$ and $\hat{\lambda}_g$ estimated by the gene-specific sample means and variances for three values of δ for the Harvard, Michigan and Stanford lung cancer data sets. Note that $\delta = 0$ implies the independence model, $\delta = 1$ implies the conjugate model and $\delta = \hat{\delta}$ uses the estimated δ from a fine grid search. The vertical axis scale is fixed both within and across rows.

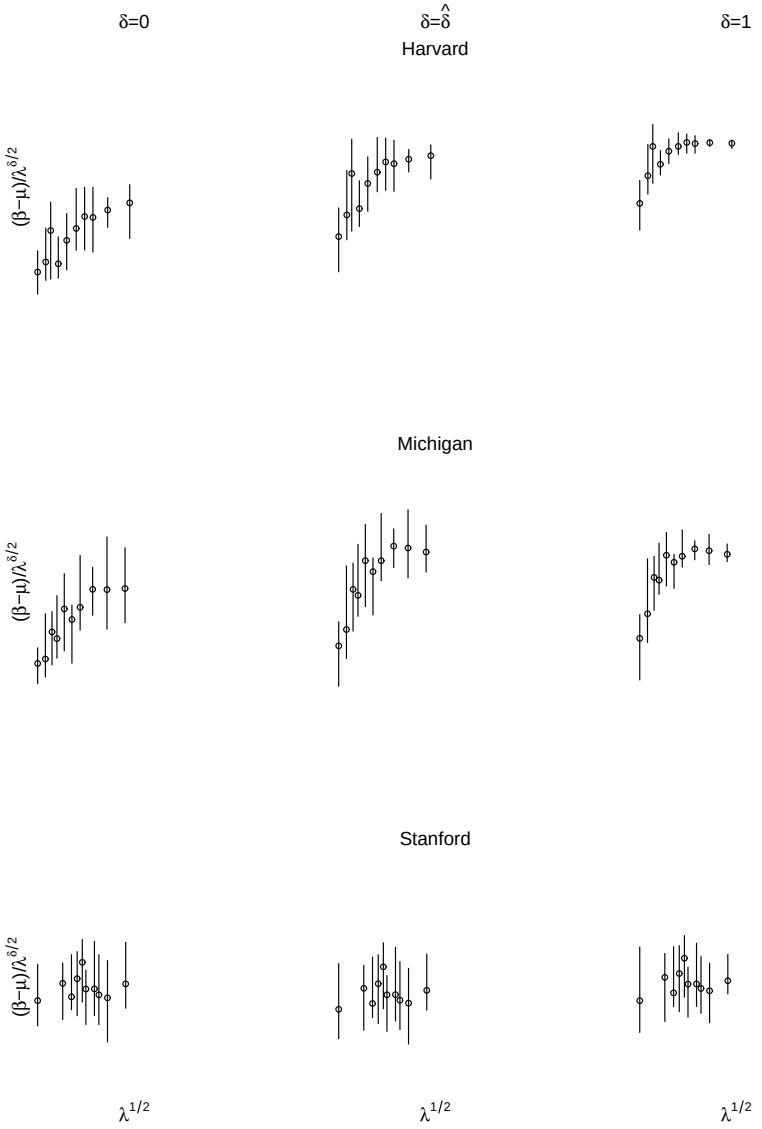

Figure 7: Boxplots of $(\hat{\beta}_{0g} - \hat{\mu}_0)/\hat{\lambda}_g^{\delta/2}$ by deciles of $\hat{\lambda}_g^{1/2}$ for $\hat{\beta}_{2g}$ and $\hat{\lambda}_g$ estimated by the gene-specific sample means and variances for three values of δ for the Harvard, Michigan and Stanford lung cancer data sets. Note that $\delta = 0$ implies the independence model, $\delta = 1$ implies the conjugate model and $\delta = \hat{\delta}$ uses the estimated δ from a fine grid search. The vertical axis scale is fixed both within and across rows.

DOI: 10.2202/1557-4679.1129

References

Ando, A. and Kaufman, G. M. (1965). Bayesian analysis of the independent multinormal process – Neither mean nor precision known. Journal of the American Statistical Association, 60:347–358.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t–test and statistical inferences of gene changes. Bioinformatics, 17(6):509–519.

Beer, D., Kardia, S., Huang, C., and et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine, 8:816–824.

Bhattacharjee, A., Richards, W., Staunton, J., and et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct subclasses. The Proceedings of the National Academy of Sciences of the United States of America, 98:13790–13795.

Caffo, B., Jank, W., and Jones, G. (2002). Ascent based Monte Carlo EM. Technical report, Johns Hopkins University Department of Biostatistics.

Carlin, B. P. and Louis, T. A. (2000). Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall, Boca Raton, FL.

Cui, X., Hwang, J., Qiu, J., Blades, N., and Churchill, G. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. Biostatistics, 6(1):59–75.

Garber, M., Troyanskay, O., Schluens, K., Petersen, S., Z, T., Pacyna-Gengelbach, van de Rijn, M., Rosen, G., Perou, C., Whyte, R., Altman, R., Brown, P., Botstein, D., and Petersen, I. (2001). Diversity of gene expression in adenocarcinoma of the lung. The Proceedings of the National Academy of Sciences of the United States of America, 98(24):13784–13789.

Ibrahim, J. G., Chen, M. H., and Gray, R. J. (2002). Bayesian models for gene expression with DNA microarray data. Journal of the American Statistical Association, 97:88–99.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5(3):299–314.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). Continuous Univariate Distributions. Volume 2. Wiley, New York, second edition.

Lange, K. (1999). Numerical Analysis for Statisticians. Springer-Verlag.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). Journal of the Royal Statistical Society, Series B, 34:1–41.

Liu, D., Parmigiani, G., and Caffo, B. (2004). Screening for differentially expressed genes: Are multilevel models helpful? Technical report, Johns Hopkins University.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. Statistica Sinica, 12(1):31–46.

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

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, 8:37–52.

Parmigiani, G., Garrett, E. S., Anbazhagan, R., and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. Journal of the Royal Statistical Society, Series B, Methodological, 64:717–736.

Parmigiani, G., Garrett-Mayer, E., Ramaswamy, A., and Gabrielson, E. (2004). A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clinical cancer research, 10:2922–2927.

Press, W., Teukolsky, S., Vetterling, W., and Flanner, B. (1992). Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge, second edition.

Royall, R. (1997). Statistical Evidence: A Likelihood Paradigm. Chapman and Hall.

Scharpf, R. B., Tjelmeland, H., Parmigiani, G., and Nobel, A. (2008). A Bayesian model for cross-study differential gene expression. JASA. To appear.

DOI: 10.2202/1557-4679.1129

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

Su, S., Caffo, B., Garrett-Mayer, E., and Bassett, S. (2007). modified test statistics by inter-voxel variance shrinkage with an application to fMRI. Johns Hopkins University, Dept. of Biostatistics Working Papers, page 138.