Funct Integr Genomics (2005) 5: 32–39
DOI 10.1007/s10142-004-0123-0

ORIGINAL PAPER

Jode W. Edwards · Grier P. Page · Gary Gadbury · Moonseong Heo · Tsuyoshi Kayo · Richard Weindruch · David B. Allison

Empirical Bayes estimation of gene-specific effects in micro-array research

Received: 14 October 2003 / Revised: 11 February 2004 / Accepted: 30 August 2004 / Published online: 29 September 2004

Abstract Micro-array technology allows investigators the opportunity to measure expression levels of thousands of genes simultaneously. However, investigators are also faced with the challenge of simultaneous estimation of gene expression differences for thousands of genes with very small sample sizes. Traditional estimators of differences between treatment means (ordinary least squares estimators or OLS) are not the best estimators if interest is in estimation of gene expression differences for an ensemble of genes. In the case that gene expression differences are regarded as exchangeable samples from a common population, estimators are available that result in much smaller average mean-square error across the population of gene expression difference estimates. We have simulated the application of such an estimator, namely an empirical Bayes (EB) estimator of random effects in a hierarchical linear model (normal-normal). Simulation results revealed mean-square error as low as 0.05 times the mean-square error of OLS estimators (i.e., the difference between treatment means). We applied the analysis to an example dataset as a demonstration of the shrinkage of EB estimators and of the reduction in mean-square error, i.e., increase in precision, associated with EB estimators in this analysis. The method described here is available in software that is available at http://www.soph.uab.edu/ssg.asp?id=1087.

Keywords Micro-array · Shrinkage · Estimation · Empirical Bayes

Introduction

The advent of micro-array technology has allowed investigators to measure the expression of thousands of genes simultaneously. Such technology can be used for

J. W. Edwards
United States Department of Agriculture, Agricultural Research Service (USDA-ARS), Department of Agronomy, Iowa State University,
Ames, IA, 50014, USA

G. P. Page · D. B. Allison (✉)
Section on Statistical Genetics, Department of Biostatistics,
RPHB 327, 1530 3rd Ave. S.,
Birmingham, AL, 35294-0022, USA
e-mail: dballison@ms.soph.uab.edu
Tel.: +1-205-9759167
Fax: +1-205-9752540

G. Gadbury
Department of Mathematics and Statistics, University of Missouri-Rolla,
202 Rolla Building,
Rolla, MO, 65409, USA

M. Heo
Department of Psychiatry/Westchester, Cornell Institute of Geriatric Psychiatry, Weill Medical College of Cornell University,
21 Bloomingdale Road,
White Plains, NY, 10605025, USA

R. Weindruch
Department of Medicine and the Wisconsin Primate Research Center, University of Wisconsin,
Madison, WI, USA

R. Weindruch
The Geriatric Research, Education, and Clinical Center,
William S. Middleton VA Hospital,
Madison, WI, 53705, USA

D. B. Allison
Clinical Nutrition Research Center,
WEBB Building, Room 402, 1530 3rd Ave. S.,
Birmingham, AL, 35294-3360, USA

T. Kayo
Life Gen Technologies, LLC c/o Mirus Corporation,
505 S. Rosa Rd.,
Madison, WI, 53719, USA

many purposes including evaluating whether and to what extent certain conditions affect gene expression. These conditions or factors can include genotype or environmental manipulation. An example of a study of this type was published by Lee et al. (1999) in which they examined the effects of age and caloric restriction on gene expression levels in over 6,000 genes in mice. Such a rich data set offers many possibilities, but also poses many challenges to summarize the volume of available data.

A primary objective of many micro-array experiments is to identify which genes are most likely to be differentially expressed. Secondarily, investigators may be interested in examining the amount of change in gene expression, i.e., estimation of the degree of change in expression. Results in statistics show that when *estimates* (e.g., estimated differences in gene expression) of many *parameters* (e.g., true gene expression differences) are examined simultaneously, the estimators have a much larger variance, i.e., range in values, than the true parameters. This is because of the additive property of variances; it could be assumed that the parameters (true unobservable values) are random samples from some distribution, such as the normal. The distribution of parameters thus has a mean and a variance and each parameter is assumed to be a random sample from that distribution in which the variance of the parameters describes the variability among parameter values. Likewise, the estimators of the parameters have errors, i.e., differences between the true parameter value and the estimator, which also can be assumed are random samples from a normal distribution. Hence, each estimator may be thought of as the sum of two random variables, a parameter value and an estimation error. An important result from statistics dealing with normally distributed random variables states that the variance of a sum equals the sum of the variances. In this context, the variance of the estimators equals the variance of the parameters (e.g., true differences in gene expression) *plus* the variance of the estimation errors. As a result, the estimators have a much larger variance than the true parameters so that estimators of positive differences in gene expression tend to be larger than the true values (the parameters) and estimators of large negative differences in gene expression tend to be more negative than corresponding true differences (assuming differences in expression were centered about zero).

In practice, only the estimators are known; true values are unobservable. Thus, we need to think about the problem in reverse, i.e., to infer from the data to unknown parameters rather than from parameters and errors to the data as in the foregoing discussion. The foregoing discussion emphasized inflation of estimators away from zero relative to the true values; in practice, only the estimators are observed with the knowledge that they are inflated and thus need to be *shrunk* towards zero (or some grand mean value) so they are closer, on average, to the true parameter values. Such estimators, i.e., shrinkage estimators, have been developed in a Bayesian context (James and Stein 1961) and in a frequentist context (Henderson 1984; Robinson 1991).

A fundamental requirement for application of shrinkage estimators is the assumption that the true parameter values can be modeled as realizations of random variables from some probability distribution. In the context of Bayesian statistical estimation, it is assumed that parameters are *exchangeable* samples from some distribution, commonly a normal distribution. A set of random variables, say θ_1, $\theta_2,\cdots$, θ_k, is said to be *exchangeable* if the parameters are samples from a common joint distribution, $p(\theta_1, \theta_2,\cdots, \theta_k)$, that is invariant to permutation of the indexes, i.e., the order, of the random variables (Gelman et al. 2003; Everitt 1998; Good 1994). In practical terms, exchangeability implies a symmetry in the prior information available on the parameters. If a parameter, θ_i, is chosen and another parameter, θ_j, is chosen at random the indices i and j convey no information that can distinguish θ_i and θ_j so that a priori we have the same expectations about their values. Gelman et al. (2003) describe exchangeability as implying “ignorance”, i.e., to claim parameters are exchangeable is to include no prior information about the parameters in the prior distribution. This is not to say that the parameters would be exchangeable if we knew everything about them (in which case we would not need an experiment), but it is to say that we do not assume any prior knowledge, such as what pathway a gene is in or what its regulatory elements are. Other information could be used in the model, if it could be formalized into a statistical model, which is perhaps the biggest challenge in model-driven approaches. While the assumption that gene expression differences are exchangeable samples from a normal distribution appears to be a restrictive assumption, it is in fact just the opposite, it is an assumption of ignorance, i.e., no prior information.

In this paper, we have applied a straightforward (EB) analysis, described in Morris (1983), in which differences between means of log base-2 and normalized intensity measures were regarded as being exchangeable samples from a normal distribution with normally distributed observational errors. We illustrate the approach with both simulated data and real data and we quantify the reduction in mean-square error of the estimated difference in gene expression via simulation.

Materials and methods

Micro-arrays offer an excellent example of simultaneous estimation of many parameters, where parameters in this case refers to true differences in gene expression. The objective is to obtain an estimator of the ensemble of gene expression differences that on average have the lowest mean-square error. Assuming that errors in gene expression measurement are exchangeable samples from a normal distribution, the classical estimator of the difference between treatment means, the ordinary least squares (OLS) estimator, is known to have the smallest variance (i.e., to be the most precise estimator) among unbiased estimators of the difference in gene expression for a particular gene (Searle 1971; Searle et al. 1992). However, if we consider the ensemble of parameters (differences in gene expression) being estimated to be an *exchangeable* set of parameters, we have additional information about them, which can provide better (more precise) estimators. In such a case, we have additional information, namely that true gene expression differences are random samples from a common distribution, such as a normal distribution, with known (or estimable) mean and variance. Morris (1983) has given a thorough review and description of how empirical Bayes (EB) estimation can be used to obtain estimators of parameters that are considered the result of a known stochastic sampling process. A review of frequentist methods, namely best linear unbiased prediction (BLUP), for obtaining estimates of random effects under a very similar model is given by Robinson (1991).

Experimental conditions

We model an experiment in which gene expression measurements are made on n_1 subjects in group 1 and n_2 subjects in group 2. The treatments can be assumed to be any factor separating the two groups such as age, sex, disease, or drug treatment. For each subject, the tissue(s) of interest is collected, extraction is performed, and a single gene expression measurement is generated for each of k genes. The objective is to estimate the parameters, θ_i (i $=1\cdots k$), which are the true differences in gene expression between treatment groups 1 and 2 for genes indexed $1\cdots k$. The parameters θ_i are defined as

$$\theta_i \equiv \mu_{1i} - \mu_{2i},$$

where μ_{1i} is the true mean gene expression for gene i in treatment group 1 and μ_{2i} is the true mean gene expression for gene i in treatment group 2.

The traditional OLS estimator of θ_i is given by

$$D_i = \bar{X}_{1i} - \bar{X}_{2i},$$

where

$$\bar{X}_{1i} = 1n_1 \sum_{j=1}^{n_1} X_{1ij},$$

$$\bar{X}_{2i} = 1n_2 \sum_{j=1}^{n_2} X_{2ij},$$

and X_{hij} is the gene expression measurement for the group h, gene i, and subject j.

The estimated variance of D_i is

$$\begin{aligned} V_i & \\ &= 1n_1 \left(1n - 1 \sum_{j=1}^{n} \left(X_{1ij} - \bar{X}_{1i} \right)^2 \right) \\ &\quad + 1n_2 \left(1n - 1 \sum_{j=1}^{n} \left(X_{2ij} - \bar{X}_{2i} \right)^2 \right). \end{aligned}$$

The OLS estimator is well known to have the smallest variance among all unbiased estimators (Searle 1971). Conditional on parameters θ_i and V_i of the D_i, the distribution of the D_i is

$$D_i | \theta_i, V_i \sim N(\theta_i, V_i) \tag{1}$$

The conditional distribution of differences in gene expression, D_i, in Eq. 1 is conditioned on fixed values of the true difference in gene expression, θ_i, and the observation error variance for gene i, V_i, which in reality are not known. If the ensemble of parameters of interest, namely, θ_i for i $=1\cdots k$, were considered an *exchangeable* set of random variables, estimators with lower mean-square errors are obtainable, which has been shown in a Bayesian context (James and Stein 1961) and in a frequentist context (Henderson 1984; Robinson 1991). In the hierarchical model reviewed by Morris (1983), parameters of interest were modeled as exchangeable samples from the normal distribution, meaning that the set of random variables represents a set of similar quantities. In the case of gene expression micro-array data, each θ_i represents the true (unobservable) difference in gene expression between two treatment groups for one gene. The θ_i are all similar in that they all represent differences in mRNA levels corresponding to a specific gene. A set of θ_i that might not be considered exchangeable would be three parameters representing differences in levels of preprocessed RNA, mRNA, and protein (i.e., gene product), respectively, corresponding to a particular gene. While the three parameters are associated with a single gene, they are unknown quantities representing three fundamentally different biological processes, so we might not think of them as *exchangeable* quantities. Assuming the parameters θ_i are exchangeable, we can model them as independent and identically distributed (iid) samples from some distribution, such as a normal distribution. In the classical model reviewed by Morris (1983), they are assumed to be iid samples from the normal distribution, i.e.,

$$\theta_i | \mu, A \sim N(\mu, A), \quad i = 1, \ldots, k. \tag{2}$$

An ensemble of estimators is obtained by application of Bayes law to obtain estimators of the θ_i conditional on the data (Morris 1983). The data consist of the OLS estimators

D_i, and known parameters V_i, μ, and A, where

$$\theta_i | D_i, V_i, \mu, A \sim N\left(\theta_i^*, V_i(1 - B_i)\right), \tag{3}$$

$$\theta_i^* = (1 - B_i)D_i + B_i\mu, \tag{4}$$

and

$$\hat{B}_i = k - 3k - 1V_i(V_i + A). \tag{5}$$

Equations 3, 4 and 5 are simplified forms of Eqs. 1.7, 1.8, and 5.5, respectively, in Morris (1983), where we have given the estimated form of the Bayesian shrinkage coefficient, B_i. EB estimates of the θ_i were obtained by using point estimates of D_i, V_i, μ, and A estimated directly from the data. Estimation of model parameters followed by substitution of the estimates as known values makes the analysis "empirical" Bayes as opposed to "Bayesian." The parameters μ and A were estimated using the iterative equations outlined in Morris (1983).

$$\hat{\mu} = \sum_{i=1}^{k} W_i D_i \sum_{i=1}^{k} W_i \tag{6}$$

$$W_i = 1W_i + \hat{A} \tag{7}$$

and

$$\hat{A} = \sum_i W_i \left\{ (kk - 1)(D_i - \hat{\mu})^2 - V_i \right\} \sum_i W_i \tag{8}$$

The EB estimators, θ_i^* for θ_i are obtained by substitution of $\hat{\mu}$ and $\hat{A}$ from Eqs. 6 and 7 for μ and A in Eqs. 4 and 5. The EB estimator θ_i^* in Eq. 4 is a compromise estimator between the prior estimate of θ_i, $\hat{\mu}$ (estimated from the data), and the data, represented by the OLS estimator, D_i. Equation 4 can be rearranged to accentuate the shrinkage property of this estimator as $\theta_i^* = D_i - \hat{B}_i (D_i - \hat{\mu})$, where the term $-\hat{B}_i(D_i - \hat{\mu})$ is the "shrinkage", i.e., the amount by which the OLS estimator is "shrunk" or "regressed" towards the mean, $\hat{\mu}$.

We found the algorithm straightforward to implement in common packages such as SAS and Splus, as well as programming languages such as JAVA and Fortran. Equations 6, 7 and 8 converged quickly for data sets that we analyzed. The analysis is implemented in a publicly available software package, HDBStat!, available at http://www.soph.uab.edu/ssg_content.asp?id=1164.

Simulation methods

To demonstrate the reduction in mean-square error of EB estimators compared to OLS estimators, gene expression differences were simulated from known distributions and estimated by OLS and by EB. Micro-array datasets with 10,000 genes were simulated with three chips in a control group and three chips in a treatment group. It was assumed that 30% of genes were differentially expressed between the treatment and control groups. We chose 30% based on experience in our group with several micro-array data sets. Differentially expressed genes were assumed to follow a standard Laplace distribution with a mean of zero and scale parameter 1 (Evans et al. 1993). The Laplace distribution was chosen because it is highly kurtotic and therefore a test of non-normality of gene expression differences. Furthermore, it was chosen because it was thought to be a biologically reasonable departure from normality in that, under this distribution, there are many more values close to zero than expected under normality. Error variances of observed values were sampled from a uniform distribution with lower bound 1 and upper bounds of 1, 2, 10, and 100 to simulate both different ratios of error to treatment variance and heteroscedasticity of error variances. The observational error for gene i with subject j was generated with a correlation to gene i −1 within the same subject. Correlation coefficients between successively simulated genes within subjects of −0.99, −0.5, 0, 0.5, or 0.99 were used to simulate different levels of correlation of gene expression within subjects. The ordered way in which we simulated errors induced an autoregressive correlation structure such that errors for subjects i and i +1 had a correlation of r, i, and i +2 had a correlation of r^2, and so on (subjects i and i + n and a correlation of r^n). Two error distributions were examined, the standard normal and chi-squared distribution with a single degree of freedom, to test EB estimators under assumed normality of errors and with non-normal errors. The chi-square-shaped distribution was generated by squaring the standard normal errors and dividing by the square root of 2. Finally, randomly generated errors were multiplied by the square root of the randomly generated error-variance to simulate heteroscedastic error variances. The chi-square distribution was chosen to simulate errors with a long tail of the distribution, as commonly observed in micro-array data. Final simulated data were obtained by adding the random gene expression difference to the errors in the treatment group (for those genes that were truly differentially expressed). The process was repeated to produce 10,000 data sets for each set of simulation conditions, where simulation conditions included two error distributions (normal and chi-square), four upper bounds for the random error-variance and five levels of correlation between genes for 40 sets of simulation conditions. These simulation conditions cannot possibly represent all biological possibilities. The conditions chosen were thought to reasonably represent biological situations and were chosen to test departures from assumptions.

OLS and EB estimates of differences in gene expression between the control and treatment group were computed for all simulated datasets. The estimation error for each gene, $\hat{\theta}_i - \theta_i$, was computed for the OLS estimator and for the EB estimator for each gene. Average bias was estimated as the average of all estimation errors across all genes for a dataset and mean-square error was estimated as the sum of squared estimation errors for the dataset divided by number of genes. Bias and mean-square error were then averaged across the 10,000 simulation runs for each set of conditions.

Example data

The example data were taken from a study of mouse liver (T. Kayo and R. Weindruch, unpublished data, University of Wisconsin-Madison). A comparison was made between calorically restricted mice and young control mice. Four mice in each group were sacrificed and mRNA was extracted from each mouse. The mRNA from each mouse was analyzed on a single Affymetrix GeneChip (Affymetrix, Santa Clara, Calif.) micro-array chip. Further details on experimental conditions have been given in a related experiment by Hagopian et al. (2003). Individual intensity measurements were log base-2 transformed and the mean of each chip was subtracted from each observation on the chip to remove confounding of chip-to-chip variation. The log base-2 transformation was chosen to make the effects being estimated interpretable; by transforming with a base-2 logarithm, a "difference" of 1 unit corresponds to a doubling of expression, a difference of 2 units corresponds

Table 1 Average bias and mean-square error of ordinary least squares (OLS) and empirical Bayes (EB) estimators of gene expression differences

Error distribution[a]	Correlation[b]	Error ratio[c]	Bias OLS[d]	Bias EB	MSE OLS[e]	MSE EB[f]	MSE ratio[g]
1	−0.99	1	0.0000	0.0000	0.67	0.37	0.55
1	−0.5	1	0.0001	0.0000	0.67	0.36	0.54
1	0	1	−0.0001	−0.0002	0.67	0.36	0.54
1	0.5	1	0.0000	0.0000	0.67	0.36	0.54
1	0.99	1	−0.0005	−0.0008	0.67	0.37	0.56
1	−0.99	2	0.0000	0.0000	1.00	0.46	0.46
1	−0.5	2	0.0000	−0.0001	1.00	0.45	0.45
1	0	2	0.0002	0.0003	1.00	0.45	0.45
1	0.5	2	0.0002	0.0002	1.00	0.45	0.45
1	0.99	2	0.0006	0.0011	1.00	0.47	0.47
1	−0.99	10	0.0001	0.0002	3.67	0.89	0.24
1	−0.5	10	0.0000	0.0000	3.67	0.87	0.24
1	0	10	0.0001	0.0000	3.67	0.87	0.24
1	0.5	10	−0.0007	−0.0007	3.67	0.87	0.24
1	0.99	10	−0.0006	−0.0012	3.66	0.93	0.25
1	−0.99	100	0.0003	0.0003	33.73	4.76	0.14
1	−0.5	100	0.0000	−0.0005	33.67	4.55	0.13
1	0	100	−0.0001	0.0004	33.67	4.54	0.13
1	0.5	100	−0.0010	−0.0011	33.66	4.54	0.13
1	0.99	100	0.0083	0.0100	33.66	5.00	0.15
2	−0.99	1	0.0010	0.0003	0.67	0.26	0.39
2	−0.5	1	0.0000	0.0000	0.67	0.26	0.39
2	0	1	−0.0001	−0.0001	0.67	0.26	0.39
2	0.5	1	0.0000	−0.0001	0.67	0.26	0.39
2	0.99	1	−0.0002	−0.0003	0.67	0.26	0.39
2	−0.99	2	0.0004	0.0001	1.01	0.32	0.32
2	−0.5	2	0.0001	0.0002	1.01	0.32	0.31
2	0	2	0.0000	0.0001	1.01	0.32	0.31
2	0.5	2	−0.0002	0.0000	1.01	0.32	0.31
2	0.99	2	0.0004	−0.0001	1.01	0.32	0.32
2	−0.99	10	0.0020	0.0012	4.02	0.56	0.14
2	−0.5	10	0.0001	0.0001	4.02	0.54	0.14
2	0	10	−0.0002	−0.0001	4.02	0.54	0.14
2	0.5	10	−0.0005	−0.0004	4.02	0.54	0.14
2	0.99	10	0.0012	0.0006	4.02	0.56	0.14
2	−0.99	100	−0.0053	0.0004	38.90	2.12	0.05
2	−0.5	100	−0.0007	−0.0004	38.91	2.03	0.05
2	0	100	−0.0007	−0.0001	38.89	2.03	0.05
2	0.5	100	0.0013	0.0006	38.91	2.03	0.05
2	0.99	100	0.0005	0.0007	38.92	2.11	0.05

[a] Distribution of errors, 1 for standard normal and 2 for chi-squared

[b] Correlation between adjacent genes (adjacent meaning only adjacent in the covariance matrix of observations within subjects)

[c] Error variances were randomly simulated from a uniform distribution with lower bound one and upper bound given in the "Error ratio" column

[d] Bias columns give the average bias (across 10,000 simulation replicates) of ordinary least squares estimators (OLS) and empirical Bayes estimators (EB)

[e] Mean-square error of OLS estimators

[f] Mean-square error of EB estimators

[g] Ratio mean-square error of EB estimators divided by mean-square error of OLS estimators

to a change in expression of four times (2 to the second power) and so forth (Baldi and Long 2001; Wolfinger et al. 2001; Yang et al. 2002). For each gene, the mean expression and variance of transformed measurements for each of the two groups were calculated and used to obtain EB estimators. The example data presented here do not validate nor invalidate use of this model, but rather they provide a demonstration of the shrinkage property of the model. Because in any real data set we do not know the truth, real data sets cannot be used to test the use of the model.

Results

Simulation

Simulations clearly demonstrated that shrinkage estimators have smaller mean-square error than OLS estimators under a wide range of conditions. For the simulation conditions chosen in this particular study, the ratio of mean-square error of EB estimators to mean-square error of OLS estimators ranged from 0.05 to 0.56 (Table 1; Fig. 1). Correlation between genes had almost no impact on the mean-square error (Table 1). The assumed observational error distribution (normal vs chi-square) and the variance ratio however had a much greater impact on the reduction in mean-square error of the EB estimators. Across all conditions, a greater reduction in mean-square error was observed for EB estimators with the chi-square distribution as the error distribution. The reduction in mean-square error also increased with increasing range of error variance. The larger the upper bound of the randomly simulated error variance (with lower bound always set equal to 1), the larger the reduction in mean-square error. No appreciable average bias was detected in either the OLS estimators or the EB estimators.

Example data

The example dataset provided a demonstration of the shrinkage produced by the EB estimators used here (Fig. 2). OLS estimators $(\bar{x}_{1i} - \bar{x}_{2i})$ ranged from −5.28 to 4.33, whereas EB (shrunken) estimators ranged from −2.06 to 1.69. Figure 2 also shows that not all gene expression differences are equally shrunken. For a given OLS estimate of the difference in gene expression, some EB estimates had a wide range in value, i.e., the degree to which they were shrunk towards the mean. The variability in shrinkage coefficients among genes is a function of the variability in estimated variances of gene expression differences among genes. The higher the estimated variance of the difference in gene expression, the less information there is for a gene, and hence, the larger the degree of shrinkage for that gene.

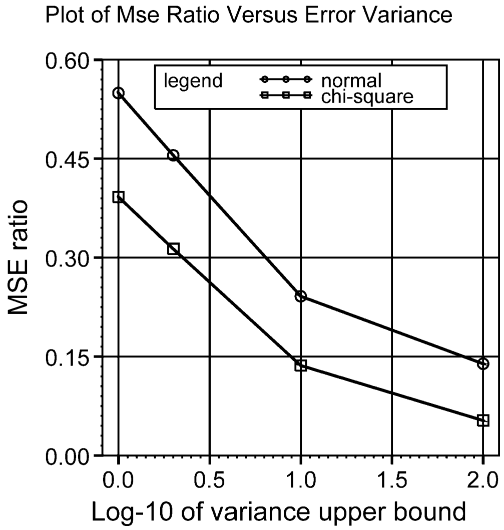

Fig. 1 Plot of ratio of mean-square error of empirical Bayes (EB) estimators divided by mean-square error of ordinary least squares (OLS) estimators (*y*-axis) versus the base-10 logarithm of the upper bound of observational error variances (*x*-axis) for simulated data. Each plotted observation is the ratio of mean-square errors averaged across five values of correlation between successively simulated genes (see Materials and methods for description of correlation between genes)

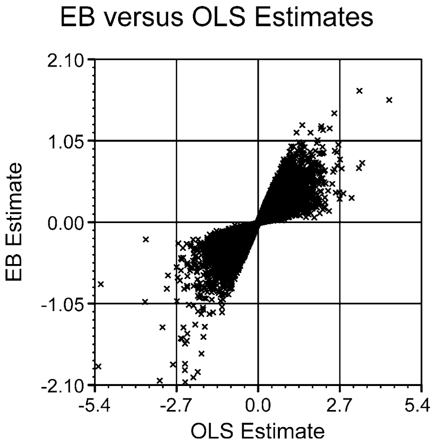

Fig. 2 Plot of empirical Bayes (EB) estimates versus ordinary least squares (OLS) estimators

Discussion

Much attention has been given to hypothesis testing in the micro-array literature, but less emphasis has been placed on estimation of effect sizes. If investigators are interested in examining empirical estimates of differences in gene expression as a means of identifying genes with relatively large differences in expression, estimation of effect sizes becomes an important issue. Micro-arrays are a clear example of a situation in which many parameters are estimated, each parameter with very little relative information due to very small sample size.

Morris (1983) outlined an EB solution for a hierarchical linear model in which the underlying parameters (in this case true differences in gene expression) are modeled as exchangeable samples from a normal distribution. By application of this algorithm to micro-array data, we have shown that estimators of gene expression differences for a large sample of genes are readily obtained with greatly reduced mean-square errors compared to ordinary estimators. Simulation results demonstrated that reduction in mean-square error was robust to departures from certain assumptions, including normality of both gene expression differences and normality of measurement errors. The reduction in mean-square error was particularly robust to correlation between genes. The autoregression correlation structure we introduced in which pairs of genes had correlations ranging from essentially zero to as high as 0.99 produced a distribution of errors that appears to be quite non-exchangeable if one knew the correlations and how they are distributed among genes. However, as we have shown, if ignorance is assumed, a large degree of variable correlation between genes did not impact the mean-square error of shrinkage estimators. Heteroscedasticity and magnitude of error variances had a large impact on the relative reduction in mean-square error in that the greater the magnitude of error variances, the greater the reduction in mean-square error of the EB estimators.

In conducting our investigations, we found that this analysis was easily implemented in standard statistical packages and converges quickly. Hence, the EB estimator from a hierarchical normal model reviewed by Morris (1983) not only has a desirable statistical property, but also has relative computational simplicity compared to more fully Bayesian techniques that require Markov Chain Monte Carlo approaches.

There are clear benefits to obtaining shrunken estimators of gene expression differences in order to reduce the mean-square error of the estimate. Several other authors have proposed EB and Bayesian methods for micro-array data that provide shrinkage estimators (Ibrahim et al. 2002; Kendziorski et al. 2003; Newton et al. 2001, 2003). Alternatively, several authors have proposed methods (both frequentist and Bayesian) that model gene expression differences from a common distribution (or mixture) but that do not necessarily provide shrinkage estimators (Broet et al. 2002; Efron et al. 2001; Lee et al. 2000, 2002; Pan et al. 2003). All of the methods we have cited here employed some type of mixture distribution for gene expression differences to model differentially and non-differentially expressed genes (or expressed and unexpressed) as being from different distributions. The exception was that of Lee et al. (2002) who, like our approach, did not employ a mixture distribution.

The EB approach we have described does have limitations. First, the analysis assumed known error variances, which we have estimated for each gene and treatment group from a very small number of observations. Second, the analysis assumes that gene expression differences are exchangeable samples from a normal distribution. Assuming that genes are exchangeable samples is not wholly unreasonable. However, assuming that they are normally distributed may be a difficult assumption (Figs. 3,4). Despite the assumption that gene expression differences were exchangeable samples from a normal distribution, EB estimators had much lower mean-square errors than traditional OLS estimators across all conditions. In particular, the correlation structure we assumed generates a correlation structure that is definitely not exchangeable because some genes have very high correlations while others have very low correlations. Based on the results of our study, the amount of correlation simulated had almost no impact on reduction in mean-square error for EB estimators, and therefore, it would appear that the exchangeability assumption is of little consequence. In reality, it is unknown what correlation structure to expect in micro-array data. It is expected that biological relationships among genes should induce correlations, but the extent to which such expected correlations are realized in micro-array data is very difficult to assess given the large numbers of genes and small numbers of observations per gene. As Gelman et al. (2003) state, "the less we know about a problem, the more confidently we can make claims of exchangeability." In other words, exchangeability is a claim of ignorance more than it could be considered a restriction on the model.

The liver data example revealed evidence that the true distribution of gene effects had larger tails than expected under the normality assumption (Fig. 4). The large tails in the estimates of gene expression differences suggests some departure from the normal-normal model that was assumed. If the normal-normal model is an inadequate description of the underlying sampling process, the estimates of some θ_k may have very large mean-square errors, even though the MSE on average is less (Louis and Shen 1999). We have not investigated in detail the extent to which departure of the model from the assumed normal-normal structure affects the MSE of specific θ_k. One might expect that if the tails of the distribution of the θ_k appear larger than predicted by the model, that the largest (and smallest) gene expression differences are truly larger (or smaller) than predicted, and hence the model predictions, i.e., their estimators, would likely be less than their true values.

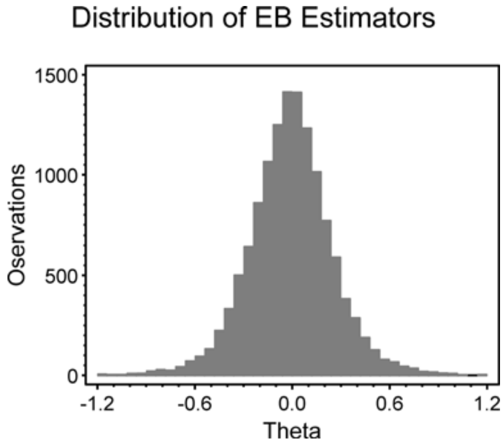

Fig. 3 Histogram of empirical Bayes (EB) estimates of gene expression differences for the example data. Nineteen observations were less than −1.2 and seven observations were greater than 1.2, so these are not represented in the histogram

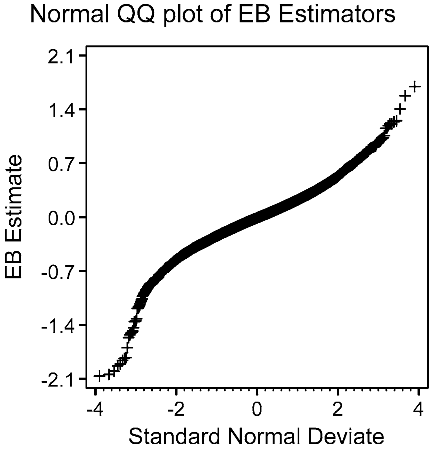

Fig. 4 Quantile-quantile probability plot of empirical Bayes (EB) estimators showing heavy tails of the distribution of EB estimates from the example data

Acknowledgements We wish to acknowledge Dr Eva Gropp for her contributions to earlier versions of this manuscript and Dr Alfred Bartolucci for reading two versions of this manuscript and providing many helpful comments. This research was supported in part by NSF grants 0217651 and 0090286, NIH grants T32AR007450, P01AG11915, and R01AG18922, an intramural award from the University of Alabama Health Services Foundation, and the Frederick Gardner Cottrell Foundation. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

References

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

Broet P, Richardson S, Radvanyi F (2002) Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. J Comput Biol 9:671–683

Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160

Evans M, Hastings N, Peacock B (1993) Statistical distributions, 2nd edn. Wiley, New York

Everitt BS (1998) Cambridge dictionary of statistics, 2nd edn. Cambridge University Press, Cambridge

Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis. Chapman and Hall, New York

Good P (1994) Permutation tests. Springer, New York Berlin Heidelberg

Hagopian K, Ramsey JJ, Weindruch R (2003) Influence of age and caloric restriction on liver glycolytic enzyme activities and metabolite concentrations in mice. Exp Gerontol 38:253–266

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph

Ibrahim JG, Chen MH, Gray RJ (2002) Bayesian models for gene expression with DNA microarray data. J Am Stat Assoc 97:88–99

James W, Stein C (1961) Estimation with quadratic loss. In: Neyman J (ed) Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, pp 361–379

Kendziorski CM, Newton MA, Lan H, Gould MN (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Technical report no. 166. Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wis.

Lee CK, Klopp RG, Weindruch R, Prolla TA (1999) Gene expression profile of aging and its retardation by caloric restriction. Science 285:1390–1393

Lee ML, Kuo FC, Whitmore GA, Sklar J (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 97:9834–9839

Lee ML, Lu W, Whitmore GA, Beier D (2002) Models for microarray gene expression data. J Biopharm Stat 12:1–19

Louis TA, Shen W (1999) Innovations in Bayes and empirical Bayes methods: estimating parameters, populations and ranks. Stat Med 18:2493–2505

Morris CN (1983) Parametric empirical Bayes inference: theory and applications. J Am Stat Assoc 78:47–55

Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 8:37–52

Newton MA, Noueiry A, Sarkar D, Ahlquist P (2003) Detecting differential gene expression with a semiparametric hierarchical mixture method. Technical report no. 1074. Department of Statistics, University of Wisconsin, Madison, Wis.

Pan W, Lin J, Le CTA (2003) mixture model approach to detecting differentially expressed genes with microarray data. Funct Integr Genomics 3:117–124

Robinson GK (1991) That BLUP is a good thing: the estimation of random effects. Stat Sci 6:15–51

Searle SR (1971) Linear models. Wiley, New York

Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol 8:625–637

Yang C, Bakshi BR, Rathman JF, Blower PE Jr (2002) Multiscale and Bayesian approaches to data analysis in genomics high-throughput screening. Curr Opin Drug Discov Dev 5:428–438