Empirical Bayes in the presence of exceptional cases, with application to microarray data

Belinda Phipson¹,³, Stanley Lee²,⁴, Ian J. Majewski²,⁴, Warren S. Alexander²,⁴, Gordon K. Smyth¹,³,⁵

22 May 2013

(1) Bioinformatics Division and (2) Cancer and Haematology Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia. (3) Department of Mathematics and Statistics and (4) Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia. (5) Corresponding author.

Abstract

Empirical Bayes is a statistical approach for estimating a series of unknown parameters from a series of associated data observations. It provides an effective means to “borrow strength” from the ensemble of cases when making inference about each individual case. Such methods are ideally suited to genomic applications where data is collected for tens of thousands of genes simultaneously. Empirical Bayes methods can however be less effective when highly exceptional cases are present. This article proposes a practical solution whereby exceptional cases are identified and the amount of “learning” from the ensemble appropriate for each case is assessed on a case-specific basis. The approach is developed in detail for the problem of estimating genewise variances from microarray data. In this context, the proposed robust empirical Bayes procedure recognizes and protects against hyper-variable genes. The new procedure improves statistical power for most genes in many microarray data sets. Simulations show that the robust estimation procedure correctly controls the type I error rate and does not increase the number of false discoveries when no hyper-variable genes are present. In the presence of hyper-variable genes, the robust method improves power to detect differential expression for the majority of genes that are not outliers. The proposed robust method is applied to an example microarray data set, on which it correctly identifies and downweights genes associated with a hidden covariate and detects more genes likely to be scientifically relevant to the experimental conditions. The new procedure is implemented in the limma software package which is freely available from the Bioconductor repository.

Keywords

Empirical Bayes, outliers, robustness, gene expression, microarrays

1 Introduction

Empirical Bayes is a statistical approach for estimating a series of unknown parameters from a series of associated data observations (Robbins, 1956). The approach assumes a Bayesian hierarchical model but, instead of basing the prior distribution on prior knowledge, the prior distribution is estimated from the marginal distribution of the observed data. This article focuses on parametric empirical Bayes methods, wherein the prior distribution is specified up to a finite
number of unknown parameters (Efron and Morris, 1977; Morris, 1983; Efron, 2010). Efron (2003) credits Fisher and others (1943) with the first use of parametric empirical Bayes, but the method was made popular by a series of seminal papers by Efron and Morris (1972, 1973, 1975) showing that empirical Bayes is a flexible alternative to the James-Stein estimator for estimating high dimensional parameters.

Empirical Bayes is well suited to scientific contexts in which data is collected as a series of cases or experiments and analogous statistical models are to be fitted to the data from each case. Empirical Bayes then provides a compromise between fitting the model separately to the data for each case, yielding case-specific parameter estimates, and fitting the model to all the cases simultaneously under the assumption that the same parameter estimates are suitable for all cases. By squeezing the case-wise parameter estimates towards to the global parameter estimates, empirical Bayes provides a practical means to “borrow strength across experiments” (Efron and Morris, 1973).

Empirical Bayes is ideally suited to genomic applications in which data is collected simultaneously for tens of thousands of genes or genomic locations. For the past 16 years, microarrays have been a popular genomic technology for measuring gene activity levels, also known as gene expression. A number of empirical Bayes statistical approaches have been developed for detecting changes in gene expression levels between treatment conditions (Newton and others, 2001; Lonnstedt and Speed, 2002; Broberg, 2003; Wright and Simon, 2003; Smyth, 2004; McCarthy and Smyth, 2009). The most common use of empirical Bayes in genomics has been to moderate genewise variance estimators. For example, the empirical Bayes moderated t-statistic of Smyth (2004), which replaces the sample variance in the denominator of the t-statistic with the posterior variance, has proved to offer much improved statistical power and false discovery rate relative to the ordinary genewise t-statistic (Kooperberg and others, 2005; Murie and others, 2009; Ji and Liu, 2010; Jeannoufin and others, 2010). These and related methods have been very successful and have been used in tens of thousands of genomic publications over the past decade.

Parametric empirical Bayes procedures assume that the true casewise parameter values can be viewed as a sample from a member of a specified parametric family of distributions. Although, empirical Bayes procedures are insensitive to the exact form of the prior distribution family (Berger, 1982; Berger and Berliner, 1986), the issue nevertheless can arise that there are a small number of exceptional cases that do not seem to fit in with the distribution of the bulk of the cases. Efron and Morris (1971, 1972) noted that some exceptional cases appeared to “learn too much” from the ensemble of cases as they appeared to come from a different subpopulation. Efron and Morris (1975, 1977) and Efron (2010) exposited the case for special treatment for exceptional cases in the context of baseball data. The batting averages of 18 major league baseball players was recorded for their first 45 times at bat during the 1970 season. The aim was to predict the batting average for each player for the remainder of the baseball season. Although empirical Bayes performs well overall, the cohort included an exceptionally good hitter, Clemente, for whom the empirical Bayes estimate was a worse prediction than his individual record. For an exceptional player like Clemente, the ensemble performance of the group was not as relevant as the Bayesian model implies.

Efron and Morris (1971, 1972) proposed limited translation rules to combat the “Clemente problem” when estimating a normal mean. This article proposes an alternative solution that can be adapted in a practical manner to most empirical Bayes contexts. We view exceptional cases as coming from an alternative prior distribution that is more diffuse than that for the bulk of cases. The prior distribution for the non-outlier cases is estimated robustly, so that exceptional cases have limited influence, and each case is then tested for concordance with this
prior distribution. Finally the relevance of the prior is evaluated for each case in terms of the posterior probability that the case is exceptional.

We develop our robust empirical Bayes strategy in detail in the context of estimating genewise variances. Our procedure estimates, on a case-specific basis, the effective degrees of freedom that should be associated with the prior global variance when estimating each individual variance. The procedure integrates into the limma software pipeline for differential expression analysis of microarray data (Smyth, 2005). In this context, the procedure recognises hyper-variable genes (Dozmorov and others, 2004), improves statistical power for other genes, and protects against hyper-variable genes being falsely called as differentially expressed.

Our approach to robustness is at the gene level, viewing hypervariable genes as outliers rather than viewing individual expression values as outliers. The latter approach was taken by Gottardo and others (2006) using a t-distribution rather than normal for the log-expression values.

The plan for the remainder of this article is as follows. Section 2 reviews the standard parametric empirical Bayes method. Section 3 outlines our robust parametric empirical Bayes approach. Section 4 reviews the conjugate empirical Bayes model for variances. Robust hyper-parameter estimation for variances is detailed in Sections 5.1 and 6. Sections 7 and 8 describe the problem in terms of microarray data. Simulation results are shown in Section 9 and an application of the robust method to a real microarray dataset involving pro-B cells is shown in Section 10. Concluding remarks are offered in Section 11.

2 The parametric empirical Bayes method

The parametric empirical Bayes method can be summarized as follows. A series of observations \( y_g, g = 1, \ldots, G \), are observed. Each \( y_g \) represents a case and each comes from a different sampling distribution. Specifically, the distribution of \( y_g \) depends on an unknown parameter \( \theta_g \) and, given \( \theta_g \), \( y_g \) follows a known probability density function that we will write as \( f(y; \theta_g) \).

The problem is to estimate all the \( \theta_g, g = 1, \ldots, g \), so that reliable inference can be conducted about each case \( g \). The difficulty that empirical Bayes addresses is that the number of cases may be large and each \( y_g \) may provide only limited information for estimating \( \theta_g \). The idea of empirical Bayes is to improve on individual casewise estimation of \( \theta_g \) by “borrowing strength” from the whole ensemble of cases. The cases are assumed to be connected through a prior distribution. Specifically the \( \theta_g \) are assumed to be sampled from a distribution with density \( g(\theta_g; \theta_0, \tau_0) \), where \( \theta_0 \) parametrizes the location of the distribution and \( \tau_0 \) parametrizes its precision. We will assume that \( \tau_0 = 0 \) corresponds to an uninformative diffuse prior while \( \tau_0 = \infty \) corresponds to a point distribution with all mass at \( \theta_0 \). The hyperparameters \( \theta_0 \) and \( \tau_0 \) are unknown, and it is this fact that distinguishes empirical Bayes from regular Bayesian inference.

The empirical Bayes method is to estimate the hyperparameters \( \theta_0 \) and \( \tau_0 \) by fitting the marginal distribution of the \( y_g \),

\[
h(y; \theta_0, \tau_0) = \int f(y; \theta)g(\theta; \theta_0, \tau_0)d\theta
\]
to the observed empirical distribution of the \( y_g \). Casewise estimators of the \( \theta_g \) are then obtained from the posterior distribution of \( \theta_g \) given \( y_g \), plugging in the estimates \( \hat{\theta}_0 \) and \( \hat{\tau}_0 \) as if they had been prior specified. The weight given to the prior is determined by \( \tau_0 \). If the \( \tau_0 \) is zero, the posterior estimates \( \hat{\theta}_g \) will be equal to the individual casewise maximum likelihood estimates of
If \( \tau_0 \) is large, then \( \tilde{\theta}_g \) will be squeezed towards \( \theta_0 \). If \( \tau_0 = \infty \) then the \( y_g \) are treated as an identically distributed sample.

### 3 Robust empirical Bayes with exceptional cases

This section outlines in general terms our proposed approach to robust empirical Bayes. Our robust empirical Bayes approach envisages that there may be a minority of cases for which it is not reasonable to treat the \( \theta_g \) as coming from the same prior distribution \( g(\theta_g; \theta_0, \tau_0) \) as the bulk of the cases. We therefore consider the possibility that there are a minority of cases for which \( \theta_g \sim g(\theta_g; \theta_0, \tau_1) \) with \( \tau_1 < \tau_0 \).

The first step of our approach is to estimate \( \hat{\theta}_0 \) and \( \hat{\tau}_0 \) robustly from the marginal distribution of the \( y_g \). Any suitably robust estimators could be used. Note that if outlier cases are present, then the robust \( \hat{\tau}_0 \) will usually be larger than a non-robust estimator would have been.

An estimate is also required for the outlier precision \( \tau_1 \). Our methodology requires only a rough estimate for this parameter, which we can obtain by fitting the marginal distribution \( h(y; \theta_0, \tau_1) \) to the most extreme value of \( y_g \).

We then assess whether each case \( \theta_g \) can be reasonably viewed as coming from the prior \( g(\theta_g; \hat{\theta}_0, \hat{\tau}_0) \). For each \( g \), we conduct a hypothesis test of the null hypothesis that \( \theta_g = \hat{\theta}_0 \). Let \( p_g \) be the \( p \)-value from this test. The \( p \)-values can be converted into posterior probabilities in the following way. Let \( q \) be the prior probability that case \( g \) is not an outlier and let \( r_g \) be the marginal probability of observing an observation more extreme than \( y_g \). Using Bayes theorem, the posterior probability that case \( g \) is not an outlier given \( y_g \) is \( \pi_g = p_g q / r_g \). We conservatively put \( q = 1 \) and estimate \( r_g \) empirically from the rank of \( y_g \) amongst all the observed values of \( y \).

The crux of our approach is to allow a case-specific value for the prior precision \( \tau_1 \). The case-wise posterior mean estimator for \( \tau \) is

\[
\tilde{\tau}_g = \pi_g \hat{\tau}_0 + (1 - \pi_g) \hat{\tau}_1.
\]

Finally the posterior estimator of \( \theta_g \) is obtained from the posterior distribution of \( \theta_g \) given \( y_g \) with \( \theta_0 = \hat{\theta}_0 \) and \( \tau_0 = \hat{\tau}_0 \). This has the effect that any case judged to be a possible outlier will have smaller \( \tilde{\tau}_g \), so the prior will receive less weight when estimating \( \theta_g \), and hence \( \tilde{\theta}_g \) will be squeezed less strongly towards the consensus value.

The robust approach produces estimates for outlying observations that are allowed to be more case-specific. Meanwhile, the estimates for the bulk of cases that are not outliers are squeezed more heavily to the global average than would be done by a non-robust approach, allowing more information to be borrowed from the ensemble. Potentially this improves the accuracy of estimation for both outlier and non-outlier cases.

### 4 Conjugate empirical Bayes for variances

The specific empirical Bayes application that we consider in this article is the problem of estimating a series of true variances from a corresponding series of sample variances. This section reviews the conjugate Bayesian model for variances and the hyperparameter estimation strategy proposed by Smyth (2004). The notation used here follows Smyth (2004). Suppose that for each case \( g, g = 1, \ldots, G \), a sample variance \( s_g^2 \) on \( d_g \) degrees of freedom is available to estimate the true variance \( \sigma_g^2 \). Given \( \sigma_g^2, s_g^2 \) is assumed to follow a scaled chi-square distribution with \( d_g \)
degrees of freedom and mean $\sigma^2_g$. Abusing notation slightly, we will write this as
\[ s^2_g \sim \sigma^2_g \lambda_g / d_g. \]
It is convenient to assume a conjugate prior for the true variances. The $\sigma^2_g$ are assumed to be sampled from a scaled inverse chi-square prior distribution with degrees of freedom $d_0$ and location $s^2_0$:
\[ \sigma^2_g \sim s^2_0 d_0 / \chi^2_d. \]
Here $d_0$ determines the precision of the prior distribution and $s^2_0$ its location. Once the prior distribution is specified, the posterior distribution for the variances is
\[ \sigma^2_g | s^2_g \sim s^2_0 d_0 / \chi^2_d. \]
so the posterior expectation of $\sigma^{-2}_g$ given $s^{-2}_g$ is $\tilde{s}^{-2}_g$ with
\[ \tilde{s}^{-2}_g = \frac{d_0 s^2_0 + d_g s^2_g}{d_0 + d_g}. \]

Empirical estimates for $s^2_0$ and $d_0$ are obtained from the marginal distribution of the $s^2_g$. Under the hierarchical model, the marginal distribution of $s^2_g$ is $s^2_0 F_{d_g, d_0}$, where $F_{d_g, d_0}$ represents the $F$-distribution on $d_g$ and $d_0$ degrees of freedom. Accurate moment estimators for $s^2_0$ and $d_0$ have been described by Smyth (2004). Moment estimation is in terms of the log-variances, $z_g = \log s^2_g$. The $z_g$ follow a Fisher’s z-distribution which, unlike the $F$-distribution, is roughly symmetric and has finite moments of all orders. The hyperparameters $s^2_0$ and $d_0$ are estimated by matching the theoretical mean and variance of the z-distribution to the observed sample mean and variance of the $z_g$.

The empirical estimates $s^2_0$ and $d_0$ are then plugged into the above formulas to obtain $\tilde{s}^2_g$. The prior variance $s^2_0$ is an appropriate average of the sample variances $s^2_g$, so the posterior variances $\tilde{s}^2_g$ are equal to the casewise variances squeezed towards the average variance. The strength of the squeezing is determined by the degrees of freedom attributed to the prior relative to the degrees of freedom associated with the sample variances.

5 Robust empirical Bayes for variances

5.1 Robust hyperparameter estimation

We now develop a robust empirical Bayes approach for estimating variances. The robust approach uses the same hierarchical model described above for the majority of cases, but in addition it allows for the possibility that a minority of the variances $\sigma^2_g$ might be sampled from an alternative more diffuse prior, $\sigma^2_g \sim s^2_1 d_1 / \chi^2_{d_1}$ with $0 \leq d_1 < d_0$.

The first step is to modify the moment estimation scheme of Smyth (2004) to estimate the hyperparameters $s^2_0$ and $d_0$ robustly for the bulk of the cases. Our approach is to apply moment estimation to the Winsorized sample variances. The idea of Winsorizing is to reset a specified proportion of the most extreme sample variances to less extreme values (Tukey, 1962).

Let $p_l$ be the maximum proportion of outliers allowed in the lower tail of the $s^2_g$, and let $p_u$ be the maximum proportion of outliers in the upper tail. Let $q_l$ and $q_u$ be the corresponding quantiles of the empirical distribution of $s^2_g$, so that $p_l$ of the variances are less than or equal to
\( q_1 \) and \( p_u \) are greater than or equal to \( q_u \). The empirical Winsorizing transformation resets the extreme values of the sample variances to the lower and upper quantiles:

\[
\text{win}(s_g^2) = \begin{cases} 
q_l & \text{if } s_g^2 \leq q_l \\
\frac{s_g^2}{\bar{g}} & \text{otherwise} \\
q_u & \text{if } s_g^2 \geq q_u.
\end{cases}
\]

Write \( z_g = \log \text{win}(s_g^2) \) for the Winsorized variances on the log-scale. Let \( \bar{z} \) and \( s^2_{\bar{z}} \) be the mean and variance of the observed values of \( z_g \).

Similarly define the Winsorized \( F \)-distribution as follows. If \( f \sim F_{d_1,d_0} \) then the Winsorized random variable is

\[
\text{win} (f) = \begin{cases} 
q_l & \text{if } f \leq q_l \\
f & \text{otherwise} \\
q_u & \text{if } f \geq q_u.
\end{cases}
\]

where now \( q_l \) and \( q_u \) are the lower tail \( p_l \) and upper tail \( p_u \) quantiles of the \( F_{d_1,d_0} \) distribution.

Write \( \nu(d_g, d_0) \) and \( \phi(d_g, d_0) \) for the expected value and variance of \( \log \text{win}(f) \). These distributional quantities can be computed by numerical integration. An efficient computation using Gaussian quadrature is described in the next section.

Assuming that the \( d_g \) are all equal, the hyperparameter \( d_0 \) is estimated by equating \( s_g^2 = \phi(d_g, d_0) \) and solving for \( d_0 \) using a modified Newton algorithm (Brent, 1973). Having estimated \( d_0 \), the logarithm of the parameter \( s_g^2 \) is estimated by \( \bar{z} - \nu(d_g, d_0) \).

In practice, residual degrees of freedom usually are equal for all genes, but occasionally some genes may have reduced \( d_g \) because of missing expression values for some samples. If this is the case, then the \( s_g^2 \) with reduced \( d_g \) are transformed to equivalent random variables with the same \( d_g \) as the other genes before applying the above algorithm. Let \( d \) be the maximum \( d_g \). First the hyperparameters \( d_0 \) and \( s_0^2 \) are estimated by the non-robust algorithm. Then the \( s_g^2 \) are transformed to \( s_0^2 F_{d,d_0}^{-1} F_{d_g,d_0}(s_g^2/s_0^2) \) where \( F_{k_1,k_2} \) denotes here the cumulative distribution function of the \( F \)-distribution on \( k_1 \) and \( k_2 \) degrees of freedom. This yields transformed \( s_g^2 \) that can be treated as all on \( d \) degrees of freedom.

### 5.2 Case-specific prior degrees of freedom

The alternative prior degrees of freedom \( d_1 \) is estimated by maximum likelihood from the maximum value of \( s_g^2 \). We then obtain a case-specific estimate of the prior degrees of freedom by combining \( d_0 \) and \( d_1 \) according to the probability that each case is an outlier.

Let \( p_g \) be the \( p \)-value for testing the hypothesis that case \( g \) is an outlier, defined by \( p_g = P(f > s_g^2/s_0^2) \) where \( f \sim F_{d_g,d_0} \). Write \( r_g \) for the proportion of all observed \( s^2 \) values greater than a particular \( s_g^2 \). Specifically \( r_g \) is \((r - 0.5)/G\) where \( r \) is the rank of \( s_g^2 \). Using Bayes theorem, and assuming that most cases are not outliers, we estimate the probability that case \( g \) is not an outlier by \( \pi_g = p_g/r_g \).

This expression for \( \pi_g \) is not necessarily monotonic in \( s_g^2 \) or \( p_g \). We ensure that \( \pi_g \) is a non-decreasing function of \( p_g \) in the following manner. First the cases are ordered in increasing order of \( p_g \). Then the cumulative mean \( \bar{\pi}_g = \sum_{i=1}^{g} \pi_i \) is computed for each \( g \). Let \( g_m \) be the first value of \( g \) for which \( \bar{\pi}_g \) achieves its minimum. All \( \pi_g \) for \( g = 1, \ldots, g_m \) are set to the minimum value of \( \bar{\pi}_g \). This is to allow for the possibility that \( \pi_g \) might be small for a group of cases but not for the most extreme case. Finally, a cumulative maximum filter is applied to the \( \pi_g \), after which the \( \pi_g \) are non-decreasing.
Having computed the $\pi_g$, the casewise prior degrees of freedom are estimated by

$$\hat{d}_0 g = \pi_g \hat{d}_0 + (1 - \pi_g) \hat{d}_1.$$ 

6 Numerical implementation

6.1 Computing Winsorized moments

Here we describe the computation of the mean $\nu(d_g, d_0)$ and variance $\phi(d_g, d_0)$ of the log Winsorized $F$-distribution. Let $Z = \log \text{win}(f)$ denote the log-Winsorized $F$ random variable defined in Section 5.1. The expected value is

$$\nu(d_g, d_0) = \mathbb{E}(Z) = p_l \log q_l + p_{lu} \mathbb{E}(Z | q_l < Z < q_u) + p_u \log q_u$$

with $p_{lu} = (1 - p_l - p_u)$. After transforming $Z$ to the unit interval, the conditional expectation can be re-interpretted as

$$p_{lu} \mathbb{E}(Z|q_l < Z < q_u) = (b - a) \mathbb{E}(h(U))$$

where $U$ is uniformly distributed on the interval from $a = q_l/(1 + q_l)$ to $b = q_u/(1 + q_u)$ and

$$h(u) = \log \left( \frac{u}{1 - u} \right) \frac{1}{(1 - u)^2} \text{pdf} \left( \frac{u}{1 - u} \right)$$

where pdf is the probability density function of the $F$-distribution on $d_g$ and $d_0$ degrees of freedom. The expectation in terms of the uniform random variable can be evaluated efficiently by Gauss-Legendre quadrature as described below.

Having computed the mean, the variance of $Z$ can be obtained as

$$\phi(d_g, d_0) = \mathbb{E}\{(Z - \nu)^2\} = p_l (\log q_l - \nu)^2 + p_{lu} \mathbb{E}\{(Z - \nu)^2 | q_l < Z < q_u\} + p_u (\log q_u - \nu)^2$$

where $\nu = \nu(d_g, d_0)$. The second term containing the conditional expectation can be re-interpretted as $(b - a) \mathbb{E}(h(U))$ with

$$h(u) = \left( \log \left( \frac{u}{1 - u} \right) - \nu \right)^2 \frac{1}{(1 - u)^2} \text{pdf} \left( \frac{u}{1 - u} \right)$$

to permit evaluation by Gauss-Legendre quadrature.

6.2 Evaluating an integral using Gaussian quadrature

The gauss.quad.prob function of the statmod package (Smyth, 2006) implements Gaussian quadrature strategies for evaluating the expected values of random variables from a selection of distributions. Consider the random variable $h(U)$ where $U$ follows a uniform distribution on the interval $[a, b]$. For any desired order $k$, gauss.quad.prob computes nodes $u_i$ and Gauss-Legendre weights $w_i$ such that

$$\mathbb{E}\{h(U)\} \approx \sum_{i=1}^k w_i h(u_i)$$

The approximation is exact if $h(u)$ can be expressed as a polynomial of order $2k - 1$ or less on the interval $[a, b]$. The accuracy of Gauss-Legendre quadrature is excellent if $h(u)$ is a reasonably smooth function taking finite values on the interval. The weights and nodes are computed using an adaption of an algorithm and Fortran code by Golub and Welsch (1969).

All results reported in this article use $k = 128$ nodes. This was sufficient for close to double-precision accuracy provided $a$ is bounded above zero and $b$ is bounded below one.
7 Application to microarray data analysis

Accurate estimation of a series of variances is of great interest in microarray data analysis, because of the need to conduct \( t \)-tests or \( F \)-tests, with the genewise variance appearing as the denominator of the test statistic, for tens of thousands of genes simultaneously.

Microarray technology is applied to a set of independent RNA samples, and yields a measure of gene activity, known as gene expression, for each known gene in each RNA sample. In the simplest case, the RNA samples may be divided into two treatment groups, so that a two-sample \( t \)-test is conducted for differential expression for each gene. More generally, a linear model is fitted to the log-expression values for each gene, leading to genewise coefficient estimates and genewise residual variances \( s_g^2 \) (Smyth, 2004).

Typically the residual degrees of freedom \( d_g \) are small, so that the sample variance is an imprecise estimator of the true genewise variance. Smyth (2004) has shown further that, in the context of a linear regression model for each case \( g \), the posterior variance \( s_g^2 \) may be inserted in place of the ordinary sample variance in \( t \)-test statistics to obtain empirical Bayes \( t \)-statistics. The empirical Bayes \( t \)-statistics follow a \( t \)-distribution on \( d_g + d_0 \) degrees of freedom under the null hypothesis that the regression coefficient is zero. This shows that \( d_0 \) represents the information that is “borrowed” from the ensemble of cases to assist with inference about case \( g \).

Accurate estimation of the variances is vital for correctly controlling the false discovery rate. In many microarray data applications, \( d_g \) is small and the number of genes \( G \) is large, so the information borrowed is of vital importance. Many studies have shown that the empirical Bayes \( t \)-tests have markedly better power to detect true differential expression than ordinary genewise \( t \)-tests in this context.

Unusually large genewise variances can arise from a variety of technical or biological causes, including unidentified batch effects or genetically heterogeneous samples. Very small variances can also arise from technical causes. Outlier variances cause two potential problems in microarray analysis. The first and most pervasive problem is that outlier variances cause the prior degrees of freedom \( d_0 \) to be under-estimated for the majority of genes that are not outliers. This means that less strength is borrowed between genes. The result is that the empirical Bayes \( t \)-tests for the bulk of genes have less power than they would if the outliers had been absent, being based on fewer posterior degrees of freedom. The second problem is that a hypervariable gene might be incorrectly identified as differentially expressed because its variance has been squeezed towards to global consensus variance, resulting in a \( t \)-statistic that is unrealistically large. This potential problem is partly mitigated by under-estimation of \( d_0 \).

When we apply robust empirical Bayes to genewise variances from microarray data, we consider outliers in both tails, that is either unusually large or small variances, when estimating \( s_0^2 \) and \( d_0 \). However we consider only outliers in the right tail, that is large variances, when computing \( \pi_g \) and \( d_{0g} \). This ensures that very small variances are still squeezed strongly towards the global estimate when computing empirical Bayes \( t \)-tests, to avoid a large \( t \)-statistic arising from a very small fold change but an even smaller variance.

Our robust empirical Bayes strategy addresses both of the problems described above. It protects against false discoveries from hypervariable genes, while at the same time providing more degrees of freedom and more statistical power for the empirical Bayes \( t \)-tests for the bulk of genes.
8 Covariate dependent priors

It is common in microarray experiments for the variance of the log-expression values to depend partly on the magnitude of the expression level. Although background correction and pre-processing algorithms and transformations are successful in ameliorating the mean-variance trend in the processed expression values (Huber and others, 2002; Ritchie and others, 2007; Shi and others, 2010), some trend usually remains (Sartor and others, 2006). Figure 3a shows a trend between the residual standard deviations and the average log-expression level of each gene for the example dataset.

It is therefore helpful to extend the empirical Bayes principle to permit the prior variance $s_0^2$ to depend on the average log-expression $A_g$ of each gene (Sartor and others, 2006). This generalizes the prior distribution for $\sigma^2_g$ to be gene-specific:

$$\sigma^2_g \sim s_0^2 g \chi^2 /\nu_0$$

where the $s_0^2 g$ vary smoothly with $A_g$. In other words, the prior distribution depends on the covariate $A_g$. Such a strategy is implemented in the eBayes function of the limma package.

Our strategy for robust empirical Bayes with a variance trend is as follows. First we fit a robust lowess trend (Cleveland, 1979) to log $s^2_g$ as a function of $A_g$. We detrend the log $s^2_g$ by subtracting this trend, then apply the robust empirical Bayes algorithm described above to the detrended variances. The final genewise prior values $s^2_{0g}$ are the product of the unlogged lowess trend and the $s^2_0$ estimated from the detrended variances.

9 Simulation study

9.1 Simulation strategy

Simulations were used to benchmark the performance of the proposed method. The simulations were designed as follows. One thousand datasets each with 10 000 genes were generated. A two group scenario with a sample size of three for each group was assumed. The hyperparameters chosen were $\nu_0 = (2, 4, 10)$ and $\sigma_0 = 0.2$. Since the residual degrees of freedom is four, $\nu_0 = 2$ reflects the situation where there is not much shrinkage to $s_0^2$, $\nu_0 = 4$ reflects a balanced design and $\nu_0 = 10$ shrinks quite heavily to the prior variance. The true variance $\sigma^2_g$ was sampled from $\nu_0 s_0^2 /\chi^2_{\nu_0}$ and the noise signal generated from $N(0, \sigma^2_g)$. For simulations with differential expression, the log fold changes for 5% of the genes were generated from a normal distribution with mean zero and variance four. For simulations with outliers, 250 sample variances were generated from a scaled inverse chi-square distribution with degrees of freedom equal to a half. These replaced a random subset of the sample variances for the generated data. No differentially expressed genes were also outliers. The data was analysed using procedures in the limma package from Bioconductor. The performance of the non-robust and the robust hyperparameter estimation was compared. We denote the non-robust method “ebayes” and the robust method “robust”.

9.2 The robust method does not compromise the accuracy of hyperparameter estimates when no outliers are present

First datasets were generated as described without any outliers. Figure 1 shows the estimate of the prior degrees of freedom in panel (a) and the estimate of the prior variance in (b). The empirical Bayes and robust empirical Bayes methods accurately estimate the hyperparameters.
Figure 1: Hyperparameter estimation with no differentially expressed genes and no outliers present in the data. (a) Estimate of the prior degrees of freedom. True values are 2, 4 and 10. (b) Estimate of the prior variance. True value is 0.04.

Table 1: Comparing the type I error rate of the empirical Bayes and robust empirical Bayes estimation procedures. Mean type I error rates are reported for 1000 simulations with no differentially expressed genes and no outliers present in the data. The standard deviation with which the error rate is estimated ranges from approximately 0.0003 for rates near 0.001 to 0.003 for rates near 0.1 with no sizeable difference between the robust and ebayes method.

<table>
<thead>
<tr>
<th>$d_0$</th>
<th>Test method</th>
<th>Nominal P-Value</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ebayes</td>
<td>0.0009961</td>
<td>0.0099820</td>
<td>0.0500065</td>
<td>0.0999357</td>
<td></td>
</tr>
<tr>
<td></td>
<td>robust</td>
<td>0.0009964</td>
<td>0.0099810</td>
<td>0.0499645</td>
<td>0.0998315</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ebayes</td>
<td>0.0010082</td>
<td>0.0100249</td>
<td>0.0500806</td>
<td>0.1001226</td>
<td></td>
</tr>
<tr>
<td></td>
<td>robust</td>
<td>0.0010131</td>
<td>0.0100381</td>
<td>0.0500599</td>
<td>0.1000470</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>ebayes</td>
<td>0.0010167</td>
<td>0.0100541</td>
<td>0.0501832</td>
<td>0.1000803</td>
<td></td>
</tr>
<tr>
<td></td>
<td>robust</td>
<td>0.0010326</td>
<td>0.0100986</td>
<td>0.0502130</td>
<td>0.1000533</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2: Estimation in the presence of outliers. (a) Estimate of $d_0$. The ebayes method consistently underestimates the true prior degrees of freedom in the presence of outliers. (b) Estimate of $s_0^2$. True value is 0.04. The ebayes method consistently underestimates the true prior variance in the presence of outliers. (c) Log squared error of the squeezed variance estimate. The robust method results in lower median logged squared error of the variance.

When $d_0 = 10$, there is slightly more variability in the estimate of $d_0$ for both methods and the robust method very slightly over-estimates $d_0$. The estimate of the prior variance is slightly more variable for smaller $d_0$ with both methods and becomes more consistent as $d_0$ increases. When $d_0 = 2$ the robust method very slightly over-estimates $s_0^2$. Overall the accuracy of the robust and non-robust methods are almost identical.

Accurate hyperparameter estimation translates into good error rate control. Table 1 shows the actual type I error rates for the robust and non-robust method for nominal p-values of 0.001, 0.01, 0.05 and 0.1. Both robust and non-robust methods control the type I error rate correctly. In summary, there is no penalty in using the robust method to estimate the hyperparameters when no outliers are present.

### 9.3 Robust method more accurately estimates the hyperparameters in the presence of outliers

The median estimate of the prior degrees of freedom and prior variance was recorded for each of the 1000 datasets. Figure 2a shows that the ebayes method always underestimates the hyperparameters in the presence of outliers. This is more severe as the prior degrees of freedom becomes larger. When $d_0 = 10$, the median ebayes estimate is approximately three. The robust method slightly under-estimates $d_0$ (median $d_0 = 8.5$), however it is a great improvement over the ebayes estimate. For the prior variance the robust estimate very slightly over-estimates the true value for all $d_0$. Using ebayes, $s_0^2$ is always under-estimated. As $d_0$ increases, the estimate of $s_0^2$ becomes progressively worse (Figure 2b). The robust method has more accurate estimates for both of the hyperparameters in the presence of the 250 outliers.

Figure 2c shows the log squared error for the posterior variance estimates for the 1000 simulations in the presence of outliers. It is clear that the robust method has the minimum log squared error for all values of $d_0$. Using the robust method for calculating the posterior variances results in better estimates of the true gene-wise variances.
Figure 3: Detection of differential expression. (a) False discovery rate curves for the robust and eBayes method for each value of $d_0$ with 5% differential expression and no outliers. The eBayes and robust curves overlap each other. (b) False discovery rate curves for the robust and eBayes method for each value of $d_0$ with 5% differential expression and 250 outliers. Modest improvement of the false discovery rate using the robust method as $d_0$ increases. (c) The Benjamini and Hochberg adjusted P-values for each method for every value of $d_0$. The dotted horizontal line represents a 5% threshold. Robust estimation increases number of genes called differentially expressed at 5% when 250 outliers are spiked into the data.

9.4 Robust estimation does not increase the false discovery rate when no outliers are present

The simulations were modified to include 5% or 500 differentially expressed genes. No outliers were simulated. Figure 3a shows the false discovery rate curves for the two methods for each value of $d_0$. The eBayes and robust false discovery rate curves overlap each other. As $d_0$ increases, the number of false discoveries decreases in the top 500 genes for both methods. With no outliers present in the data, the number of false discoveries between the two methods is identical.

9.5 Robust estimation modestly improves the false discovery rate in the presence of outliers

In this scenario, the simulations were modified to incorporate 500 differentially expressed genes and 250 outliers generated from a scaled inverse chi-square distribution on half a degree of freedom. Robust estimation shows a modest improvement on the number of false discoveries, which becomes more apparent as $d_0$ increases (Figure 3b).

9.6 Robust estimation improves power to detect differential expression in the presence of outliers

When there are 500 differentially expressed genes and 250 outliers in the simulated data, Figure 3c shows that the robust estimation method results in more genes called differentially expressed at a 5% Benjamini and Hochberg adjusted p-value cut-off. When $d_0 = 2$, eBayes resulted in an average of 294 genes over the 1000 simulations called differentially expressed. The robust method had a mean of 334 genes identified as differentially expressed. At $d_0 = 4$, the eBayes method had a mean of 334 genes called differentially expressed and with robust estimation a mean of 350 genes were identified as differentially expressed. When $d_0 = 10$, eBayes had an average of 355 genes differentially expressed and robust had an average of 386 genes identified as differentially expressed. The difference in the power between eBayes and robust is more marked at higher
prior degrees of freedom, although at every level of $d_0$ the robust estimation procedure resulted in more genes being identified at significantly differentially expressed without an increase in the number of false discoveries (Figure 3b).

10 Case study: loss of polycomb repressor complex 2 function in pro B cells

Polycomb group proteins are transcriptional repressors that play a central role in the establishment and maintenance of gene expression patterns during development. Suz12 is a core component of Polycomb Repressive Complex 2 (PRC2). Majewski and others (2008) and Majewski and others (2010) studied mice with a mutation in the Suz12 gene that results in loss of function of the Suz12 protein and hence PRC2. They profiled gene expression in hematopoietic stem cells from these mice. Here we describe a gene expression study of a different hematopoietic cell type from the same Suz12 mutant mice strain. This study profiles gene expression in pro-B cells, an early progenitor immune cell intermediate in a series of development stages between hematopoietic stem cells and mature B-cells.

Our interest is to study development, so cells were isolated from 16-day embryonic mice. For this study RNA, was extracted from foetal pro B cells that were isolated from the liver of four wild-type mice and four Suz12 mutant mice. RNA was hybridized at the Australian Genome Research Facility to Illumina Mouse Whole-Genome-6 version 2 BeadChips, a microarray platform containing about 48,000 60-mer DNA sequences probing most genes in the genome. Summary probe intensity profiles were exported from GenomeStudio and analysed using the limma software package version 3.17.13 (Smyth, 2005) in R. Intensities were background corrected, quantile normalised and transformed to the log$_2$-scale using the neqc function (Shi and others, 2010). One of the Suz12 mutant samples was discarded because it clustered with the wildtype instead of the Suz12 samples, leaving four wildtype and three Suz12 mutant samples. Probes were filtered from further analysis if they failed to achieve a detection p-value of less than 0.01 in at least two of the remaining samples. This left 14084 probes for analysis.

Linear modelling was applied to normalized log-expression values, resulting in a residual sample variance on 5 residual degrees of freedom for each probe. Figure 4 shows the square-root sample standard deviations plotted against the average log intensity for each probe. The grey curve shows the estimated trend for the prior variance $s_1^{1/2}$. The non-robust estimate of the prior degrees of freedom was 11.9. The robust algorithm identified a number of outlier variances marked on Figure 4a. The robust algorithm estimated prior degrees of freedom 14.1 for most genes, but with prior degrees of freedom as low as 0.5 for the outlier variances (Figure 4b).

Further examination showed that many of the probes identified as outliers corresponded to genes known to have sex-linked expression, including many on the X or Y chromosomes (Figure 4a). The most outlying variances corresponded to Y chromosome genes Erdr1 and Eif2s3y up-regulated in males, and X chromosome gene Xist, known to be up-regulated in females. Other outliers gene were ribosomal genes Rn18s and Rpl7a, suggesting ribosomal RNA contamination in one or more samples, and hemoglobin genes Hbb-y and Hbb-b1 suggesting red blood or bone marrow content in some tissue samples. None of these genes should be related to the Suz12 mutation.

Differential expression between the Suz12 mutants and the wildtype mice was assessed using empirical Bayes moderated t-statistics. The p-values were adjusted to control the false discovery rate at less than 5% (Benjamini and Hochberg, 1995). The non-robust and robust procedures found 251 down-regulated and 35 up-regulated probes in common (Figure 5). However 22 and
Figure 4: The quarter root sample variance is plotted against the average log intensity for each gene. (a) The grey line shows the estimate of the prior variance. In this case there is a trend on $s^2_0$. The chromosomes of the larger observations are shown. Many of the genes with larger variability come from the X and Y chromosome, indicating there may be a sex factor in this data that we are not aware of. (b) The coloured points show the binned estimate of the prior degrees of freedom. The points with larger sample variances have smaller $d_{log}$.

Figure 5: Venn diagram showing overlap of significant genes for Suz12 versus wildtype found using ebayes and robust methods. Genes found significant using ebayes but not robust tend to be sex linked genes. Additional genes found by robust are biologically relevant and include Bcl2l1, Ccne2 and Myst2.
16 down-regulated genes were found only by the robust or non-robust procedures respectively. The non-robust unique genes tended to be sex-linked or hemoglobin related (Xist, Apoa2, Hbb-b1 etc) whereas the robust unique genes were related to programmed cell death (Bcl2l1), cell cycle (Ccne2) or chromatin remodelling (Myst2). For up-regulated genes, 8 and 10 unique probes were found by the robust and non-robust procedures respectively. The non-robust unique genes tended to be Y chromosome sex-linked genes (Ddx2y, Erdr1 etc) whereas the robust unique genes appeared related to the PRC2 process of interest.

Further investigation confirmed that two of the Suz12 mutant embryos were in fact female, whereas all the other mice were male. This was an unwanted complication in the experiment, difficult to avoid without sex-typing of the embryo mice at the time of tissue collection. The results show that the robust empirical Bayes method was successful in identifying and down-weighting genes that are associated with the hidden covariate. The robust procedure results in more statistical power to detect other genes that are more likely to be of scientific significance.

11 Discussion

We have developed a method that robustly estimates the hyperparameters of the conjugate empirical Bayes model for variances and negates the effects of outliers. The simulation studies showed that when no outliers and no differentially expressed genes were present the robust method controlled the type I error rate correctly. The robust method also accurately estimated the hyperparameters, showing no negative effects when no outliers were present. When 500 differentially expressed genes were introduced into the simulations, the false discovery rate between the two methods was identical.

Introducing 250 outlier sample variances as well as the 500 differentially expressed genes showed that the non-robust method lost power and could not estimate the hyperparameters correctly. Using robust estimation showed a small improvement in terms of numbers of false discoveries. The main improvement was to the Benjamini and Hochberg adjusted p-values, particularly as \( d_0 \) became larger. More genes were identified as significantly differentially expressed at a 5% cut-off using the robust method compared to the non-robust method. This was due to the robust method more accurately estimating the hyperparameters and hence performing the correct amount of shrinkage. The posterior variances evaluated using the robust estimates more accurately reflected the true variances.

We applied the robust hyperparameter estimation to a polycomb repressor complexes microarray dataset using pro-B cell samples. We showed that the robust method correctly identified hypervariable genes associated with an unwanted covariate and with other technical variations between the RNA samples.

12 Software

The robust method for estimating the hyperparameters is freely available in the limma software package (Smyth, 2005; Smyth and others, 2013) available from the Bioconductor repository. The method can be used as part of standard limma analysis pipeline by using the option \texttt{robust=TRUE} when calling the \texttt{eBayes} function.
Acknowledgments

BP was supported by a PhD scholarship from Science Faculty of The University of Melbourne. GKS was supported by a Research Fellowship from the National Health and Medical Research Council.

References


