Differential variation and expression analysis

Haim Bar* and Elizabeth D. Schifano
Department of Statistics, University of Connecticut

March 2018

Abstract

We propose an empirical Bayes approach using a three-component mixture model, the L_2N model, that may be applied to detect both differential expression (mean) and variation. It consists of two log-normal components (L_2) for the differentially expressed (dispersed) features (one component for under-expressed [dispersed] and the other for over-expressed [dispersed] features), and a single normal component (N) for the null features (i.e., non-differentially expressed [dispersed] features). Simulation results show that L_2N can capture asymmetries in the numbers of over- and under- expressed (dispersed) features (e.g., genes) when they exist, can provide a better fit to data in which the distributions of the null and non-null features are not well-separated, but can also perform well under symmetry and separation. Thus the L_2N model is particularly appealing when no a priori biological knowledge about the mixture density is available. The L_2N model is implemented in an R package called DVX, for **D**ifferential **V**ariation and e**X**pression analysis. The package also includes an implementation of differential expression analysis via the `limma` package, and a differential variation and expression analysis using a three-way normal mixture model. `DVX` is a user-friendly, graphical interface implemented via the 'Shiny' package [6], so that a user is not required to have R programming knowledge. It offers a set of diagnostics plots, data transformation tools, and report generation in Microsoft Excel- and Word-compatible formats. The package is available on the web, at https://haim-bar.uconn.edu/software/DVX/ .

KEY WORDS: Differential dispersion, empirical Bayes, mixture model

1 Introduction

High-dimensional data arise frequently in health sciences and biomedical studies, and has emerged in recent years as a consequence of the rapid advance of "-omic" research. For example, array-based technologies allow scientists to simultaneously collect measurements on hundreds of thousands of genetic markers from an experimental sample. Due to high

*Correspondence: haim.bar@uconn.edu

cost, it is common that these markers are measured for a relatively small number of independent samples in a given study, and as a consequence, one faces the 'large G, small n' problem, where G is the total number of markers or features and n is the number of samples. The resulting datasets consist of observed values quantifying relative abundance levels of platform-dependent biological material at multiple sites across the genome. Particularly for gene expression and DNA methylation array experiments, the goal is often to identify genetic markers that are differentially expressed (methylated) across clinically relevant subgroups. Similar settings and goals are also found in next-generation sequencing platforms (e.g., RNA-seq), as well as fields beyond genomics (e.g., metabolomics, brain imaging), however, the terminology of gene expression will be used below for simplicity of exposition. As such, the objective is to conduct simultaneous tests across the G genetic markers for differential mean expression due to a particular predictor variable of interest (e.g., clinical subgroup, age group, etc).

The existing methods for detecting differential mean expression between two populations are numerous, as classical parametric (t-/F-) statistics evaluated at each marker do not provide a reliable methodology for determining differential mean expression across the genome. The large number of markers with relatively few samples not only induces a severe multiple testing problem, but also yields marker-wise variance estimates that are often inaccurate [e.g. 20, 8]. More powerful tests have since been proposed that combine information across genetic markers for stabilization. Indeed, the most widely-used methods currently for detecting differential gene expression 'borrow strength' across genetic markers by treating marker-specific effects as random variates [e.g., 18, 7, 10, 1, 3].

In this work, we take a broader interpretation of 'differential expression' and allow for identification of genes which are differential across populations not only in terms of their mean expression levels, but also possibly their variances. Differential variation is important, for example, in the analysis of heritability of complex diseases [16], epigenetic analysis [9], and gene network regulation [17]. Recently, Bar et al. [2] proposed an extension of the 'lemma' model in [1, 3] by introducing a bivariate modeling strategy which accounts for both differential mean expression and differential dispersion across two populations, and yields a substantial gain in power to detect differential mean expression when differential dispersion is present. As with the limma [18] and lemma [1] models, Bar et al. [2] uses a mixture model, but in contrast to limma and lemma, it is based on a mixture of three normal distributions with one component designed to capture the non-differentially expressed (dispersed) genes, and the remaining two components designed to capture the underexpressed (underdispersed) and overexpressed (overdispersed) genes relative to a reference group. The two-component mixture models of limma [18] and lemma [1] implicitly assume that the differentially expressed genes are symmetrically up- and down-regulated, which may or may not be reasonable depending on the data at hand (see, e.g., [11]). In the extreme case where only up- or down-regulated genes are expected or relevant, Ivanek et al. [11] proposed first using limma [18] to estimate certain hyperparameters and rank the genes, and then fitting an extreme value distribution to the tail of interest. While the three-component Normal mixture of Bar et al. [2] can adequately capture asymmetries in the numbers of over- and under- expressed (dispersed) genes when they exist, the performance of the three-component Normal mixture model can be suboptimal when there is a large degree of overlap between the three mixture components. In this situation, for example, a large portion of

the density near zero may be allocated counterintuitively to the nonnull components, the mixture component corresponding to the over-expressed genes may attribute an overly large probability to negative values of test statistic (e.g., difference in means), and vice versa.

To overcome these challenges, we propose an empirical Bayes approach using a different three-component mixture model, the L_2N model, that consists of two log-normal components (L_2) for the differentially expressed (dispersed) genes (one component for under-expressed (dispersed) and the other for over-expressed (dispersed) genes), and a single normal component (N) for the null genes (i.e., non-differentially expressed (dispersed) genes). Our simulation results show that this approach can still capture asymmetries in the numbers of over- and under- expressed (dispersed) genes when they exist, but can also provide a better fit to data exhibiting a large degree of overlap in the three-component Normal mixture while still providing a good fit when the numbers of over- and underexpressed (dispersed) genes are similar. Thus the L_2N model is particularly appealing when no a priori biological knowledge about the mixture density is available. This type of mixture, in which the non-null components have zero mass or density at the center of the null distribution and a negligible mass in a sufficiently small neighborhood around it, is similar to what was defined in [12] as 'non-local alternative priors' in the Bayesian hypothesis testing framework. In that paper, Johnson and Rossell use point-mass null and a different form of alternative which, unlike the L_2N model, is symmetric and assigns equal probabilities to over- and underexpressed genes. They further show (under additional constraints) the gain that their model yields in terms of rate of convergence of the Bayes factor.

The L_2N model is implemented in an R package called `DVX` [4], for **D**ifferential **V**ariation and e**X**pression analysis. The package also includes an implementation of a three-way normal mixture model, which we refer to as N_3, similar to the one proposed in [2]. The software also allows for differential expression analysis via the `limma` package. `DVX` is a user-friendly, graphical interface implemented via the 'Shiny' package [6], so that a user is not required to have R programming knowledge. It offers a set of diagnostics plots, data transformation tools, and report generation in Microsoft Excel- and Word-compatible formats. All three models implemented in `DVX` (namely, L_2N, N_3, and limma) allow for adjustment for covariates. Furthermore, `DVX` allows users to define more general contrasts than the simple two-group comparison. An extensive documentation of the software is provided in the Supplementary Materials.

The paper is organized as follows. We introduce the L_2N model in Section 2 and further describe the relationship between the three models implemented in DVX. In Section 3 we present results from the three model implementations under various simulation experiments as well as a case-study in which we identify change in gene expression levels as a result of aging (across four age groups) in normal brains [15]. We conclude with a brief discussion in Section 4.

2 Methods

We are interested in identifying which genes have different distributional properties of expression levels between two populations. With normalized expression data this is usually interpreted as testing for differences between the means of two distributions. However, we

use a more general interpretation and ask which genes are differentially expressed (different means) and/or differentially dispersed (different variances) between the two populations.

In the following, the normalized expression levels for gene g in group $i = 1, 2$ have mean μ_{ig} and variance σ^2_{ig}. The sample versions are m_{ig} and s^2_{ig}, respectively. A gene is differentially expressed (DE) if $\beta_g = \mu_{1g} - \mu_{2g} \neq 0$. With normalized data, testing the null hypotheses $H_{0g} : \beta_g = \theta_0$ relies on a statistic of the form

$$z_g = \frac{d_g - \theta_0}{sd(d_g)}\,, \tag{2.1}$$

where $d_g = m_{1g} - m_{2g}$. For non-DE genes, if $var(d_g)$ are assumed to be known, z_g has a standard normal distribution for non-DE genes. Otherwise, z_g follows a t distribution. A common approach to detecting DE genes is to use a mixture model in which non-DE genes are assumed to follow a 'null distribution' and belong to one component, and DE genes follow another distribution and belong to a different mixture component. For example, in the limma model [18], the variances are assumed equal across groups, i.e., $\sigma^2_{ig} \equiv \sigma^2_g$, $i = 1, 2$, and $d_g \sim N(\beta_g, v_g\sigma^2_g)$ such that for non-DE genes $\beta_g = 0$ and for DE genes $\beta_g \sim N(0, v_0\sigma^2_g)$. This results in a two-component mixture model for $d_g|\sigma^2_g$, in which both components have mean zero, but expression levels of genes in the DE component have greater variability. The model in [2] is a three component mixture in which for non-DE genes $d_g \sim N(\theta_0, \kappa^2_g)$, where $\kappa^2_g = \sigma^2_{1g}/n_{1g} + \sigma^2_{2g}/n_{2g}$, and for DE genes $d_g \sim N(\theta_0 + \theta_g, \kappa^2_g)$ where $\theta_g \sim N(\pm\theta_D, \kappa^2)$. This model consists of three normal components, with means θ_0, $\theta_0 - \theta_D$, and $\theta_0 + \theta_D$, where $\theta_D > 0$. Here, we use a more general model in which the means of the non-null components are $\theta_0 + \theta_{D1}$ and $\theta_0 - \theta_{D2}$ and where θ_{D1} may be different than θ_{D2}. We denote this model by N_3. Table 1 summarizes the differences between three approaches for detecting DE: the standard t-test (one gene at a time), limma, and N_3. A key difference between the three models is in how $var(d_g)$ is estimated. In the standard t-test approach the variances are estimated separately for each gene and each group, while in the mixture models (limma and N_3) the variances are estimated by borrowing information across all genes. This is achieved by assuming a prior distribution for the error variances of all genes. Typical gene expression experiments involve small sample sizes, and thus lack power to detect differentially expressed genes. Power is further reduced when one accounts for multiple testing and adjusts the significance level for the large number of hypotheses (one per gene). When sample sizes are small (or even moderate) and the number of genes is large, the mixture models yield a substantial increase in power for detection of DE genes, due to the so-called 'shrinkage estimation' [e.g., 18, 10].

The L_2N Model: In the model presented in this paper we also assume that $\{z_g\}$ come from a mixture distribution, and that non-DE genes follow a normal distribution, $z_g \sim N(\theta_0, \kappa^2_g)$, where κ^2_g are also assumed to follow a prior distribution. For the DE genes, we assume that

$$z_g - \theta_0 | [z_g > \theta_0,\, g \in DE] \;\sim LogNormal(\theta_{D_1}, \kappa^2_{D_1})\,, \tag{2.2}$$

$$-(z_g - \theta_0) | [z_g < \theta_0,\, g \in DE] \;\sim LogNormal(\theta_{D_2}, \kappa^2_{D_2})\,. \tag{2.3}$$

The parameter θ_0 represents an overall difference between non-DE genes in the two groups. While it may be close to 0 in many applications, this is not always the case, and models

Table 1: Models for differential expression. $\beta_g = \mu_{1g} - \mu_{2g}$, $d_g = m_{1g} - m_{2g}$. In all three models inference is based on z_g from (2.1).

Model	$\beta_g \vert H_0$	$\beta_g \vert H_1$	$var(d_g)$
t-test	$\theta_0 = 0$	$\neq 0$	$\frac{\sigma_{1g}^2}{n_{1g}} + \frac{\sigma_{2g}^2}{n_{2g}}$
limma	$\theta_0 = 0$	$\sim N(0, v_0\sigma_g^2)$	$\sigma_g^2\left(\frac{1}{n_{1g}} + \frac{1}{n_{2g}}\right)$
N_3	θ_0	$\sim N(\theta_0 + \theta_{D1}, \kappa^2)$ if $\beta_g > \theta_0$ $\sim N(\theta_0 - \theta_{D2}, \kappa^2)$ if $\beta_g < \theta_0$	$\frac{\sigma_{1g}^2}{n_{1g}} + \frac{\sigma_{2g}^2}{n_{2g}}$

which assume that $\theta_0 = 0$ when it is not so, yield many false discoveries as we discuss in Section 3.3.

Denote the non-DE probability density function (p.d.f.) of z_g by f_0 and the p.d.f.s of the two DE components by f_1 and f_2. Denote the three components of the mixture by C_0, C_1, or C_2, and let the corresponding proportions of genes belonging to each component $j = 0, 1, 2$, be p_j such that $\sum_{j=0}^{2} p_j = 1$. Classifying genes into one of the three components is then achieved by computing their posterior probabilities

$$Pr(g \in C_j) = \frac{p_j f_j(z_g)}{p_0 f_0(z_g) + p_1 f_1(z_g) + p_2 f_2(z_g)}, \ j = 0, 1, 2. \tag{2.4}$$

This mixture model, which we call L_2N, allows for different proportions of overexpressed and underexpressed genes. Under this model, the probability that a gene with a positive (negative) z_g statistic is misclassified as underexpressed (overexpressed) is zero. In contrast, the limma model assumes that the distribution of DE genes is symmetric, and while the N_3 model allows for different proportions of over and underexpressed genes, the unbounded support of the DE components implies that an overexpressed gene has a non-zero probability of being classified as underexpressed, and vice versa. Figure 1 demonstrates the three types of mixtures mentioned here, with the limma model on the left, the N_3 model in the middle, and the L_2N model on the right. The dotted blue lines represent the distributions of the non-DE genes, which are distributed normally, and in all three models we set $Pr(g \in C_0) = 0.8$. The dashed red lines represent the distributions of DE genes, and the thick gray lines represent the mixture distributions. In the limma model, the assumptions imply symmetry, and thus, that the proportions of overexpressed and underexpressed genes are the same. This may be a biologically reasonable assumption in some situations, but not in others. In both the limma and N_3 models, the probability of a type II error is greater than in the L_2N model - for a DE gene, i.e. $g \notin C_0$, and for some small $\epsilon > 0$, $P(|z_g| < \epsilon \mid g \notin C_0)$ is smaller for L_2N than for the limma and N_3 models.

Both the N_3 and L_2N models use a hierarchical mixed model in which σ_{ig}^2, $i = 1, 2$, are assumed to have a common prior distribution for all g, within group i. This allows 'borrowing information' across genes and shrinks larger variances toward the overall mean of variances. To obtain 'shrinkage estimates' for the variances, σ_{ig}^2, we follow [2] and define

$$u_{ig} = \log(s_{ig}^2) - \log(X^2/f_{ig}), \ i = 1, 2, \tag{2.5}$$

where X^2 is a Chi-squared random variate with $f_{ig} = n_{ig} - 1$ degrees of freedom. Given

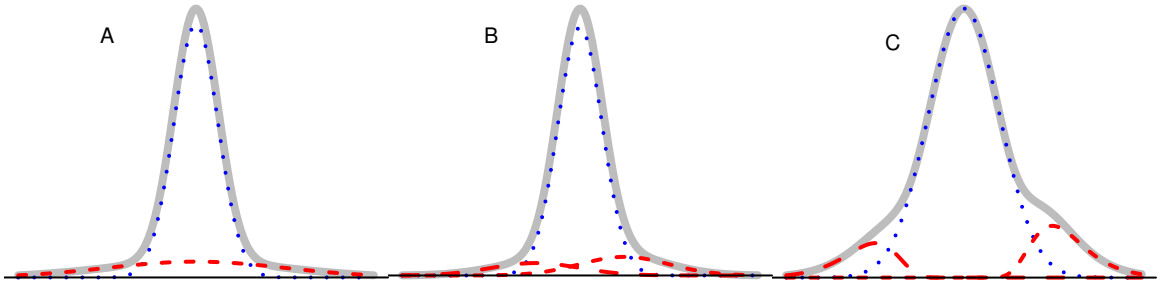

Figure 1: Three mixture models for differential expression. A. The limma model [18] B. the N_3 model [2] C. the L_2N model. In all cases, $Pr(g \in C_0) = 0.8$.

$\log(\sigma_{ig}^2)$,

$$u_{ig} \approx N(\log(\sigma_{ig}^2),\ 2\psi'(f_{ig}/2)),\ i = 1, 2, \tag{2.6}$$

where ψ' is the trigamma function. The normal approximation in (2.6) allows us to test for differential dispersion, which may be of interest in its own right. To do that, we place a three-component L_2N prior on $\log(\sigma_{2g}^2/\sigma_{1g}^2)$. Then, the differential dispersion statistics

$$u_g = u_{2g} - u_{1g}\,, \tag{2.7}$$

follow an L_2N mixture distribution having the same form as (2.2) and (2.3): using the superscript v to indicate that the model parameters for the dispersion statistics are different from those for the expression model, for differentially dispersed (DD) genes we assume

$$u_g - \theta_0^v | [u_g > \theta_0^v,\ g \in DD] \sim LogNormal(\theta_{D_1^v}, \kappa_{D_1^v}^2)\,, \tag{2.8}$$

$$-(u_g - \theta_0^v) | [u_g < \theta_0^v,\ g \in DD] \sim LogNormal(\theta_{D_2^v}, \kappa_{D_2^v}^2)\,. \tag{2.9}$$

Similarly to the expression model, we denote the mixture components for u_g, by C_j^v, $j = 0, 1, 2$ (which may be different from the DE mixture components, C_j), and the posterior probabilities $Pr(g \in C_j^v)$, have the same form as (2.4).

We compute the posterior mean of σ_{ig}^2, $i = 1, 2$, denote it by $\tilde{\sigma}_{ig}^2$ and fit the L_2N model to $\{z_g\}$, with $\sqrt{\tilde{\sigma}_{ig}^2}$ replacing the $sd(d_g)$ in (2.1). To fit the L_2N model to $\{z_g\}$ (or $\{u_g\}$), we use the EM algorithm, where the 'missing data' are indicator variables such that $b_{gj} = 1$ for $g \in C_j$ (or $b_{gj}^v = 1$ for C_j^v), $j = 0, 1, 2$, and $b_{gj} = 0$ if $g \notin C_j$ (or $b_{gj}^v = 0$ if $g \notin C_j^v$). We estimate b_{gj} (b_{gj}^v) by taking its expectation (i.e., (2.4) for expression; similar for dispersion), given the current estimates of the mixture component parameters. Maximum likelihood estimates for θ_0, θ_{D_j}, and $\kappa_{D_j}^2$ (or θ_0^v, $\theta_{D_j^v}$, and $\kappa_{D_j^v}^2$) are obtained in each iteration, while holding the values b_{gj} fixed at their current estimates. Additional details regarding the estimation procedure are provided in the Supplementary Materials. If $Pr(g \in C_1 \cup C_2)$ is sufficiently large, we conclude that gene g is DE. Similarly, if the posterior probability $Pr(g \in C_1^v \cup C_2^v)$ is sufficiently large, we conclude that gene g is DD. The normality of the null component under the L_2N model also allows us to use the frequentist approach so that gene g is DE (DD) if $h[2(1 - \Phi(|z_g|))] < q$ (or, $h[2(1 - \Phi(|u_g|))]$, when testing for differential dispersion), where h is a function which adjusts the p-values to account for multiple testing.

3 Results

Experiments which aim to identify differentially expressed genes are usually performed under the assumption that the proportion of DE genes is relatively small. When the total number of genes is large and the sample sizes are small or moderate, as is often the case, the challenge from the statistical point of view is to maximize the power (i.e., find the largest number of true DE genes), while limiting the false discovery rate (FDR). The goal of the simulations presented here was to evaluate and compare the performance of the three models implemented in `DVX` (limma, N_3, and L_2N) in terms of power and FDR. To clarify the terminology used in this section, when we refer to the performance of limma, N_3, or L_2N we mean the software implementation of the corresponding mixture model.

In each of almost 50 different configurations we simulated 5,000 genes and $n = 5$ subjects in each group. Each configuration was simulated 30 times. We varied the number of DE genes, the mean of the difference between expression for DE genes between the two groups, which we denote by θ, and the variance of the random error. The error variances were simulated from an inverse gamma distribution: $\sigma_{ig}^{-2} \sim Gamma(\alpha, \beta)$. The shape and scale parameters, α and β, were set so that $E[Var(z_g)] = 1$, so that we can control the mean signal to noise ratio (SNR) in the simulation only in terms of the signal, θ. With this setting, $\beta = n(\alpha-1)/2$, and the variance of $\{Var(z_g)\}$ is determined by α. We let $\alpha \in \{4.5, 6, 7.5, 9\}$, corresponding to $Var[Var(z_g)] \in \{1, 0.63, 0.45, 0.36\}$. Here, we show results with $\alpha = 9$. We note that as α increases, the variances across genes become more homogenous, and the power to detect DE genes increases for all three methods (for a fixed mean difference between the two groups, θ; results not shown).

With each of the three methods, the null distribution is obtained, and a gene is declared as DE if its Benjamini-Hochberg [5] adjusted p-value is less than 0.05. Note that all three methods allow for Bayesian inference, as well, in terms of posterior probabilities or Bayes' factors.

3.1 Simulation Study – Differential Expression, no Differential Dispersion

First, we consider scenarios in which the data are generated according to the N_3 model, when there are no differentially dispersed genes. We compare the three methods under consideration for different signal strengths, and different number of DE genes. Table 2 shows the median power of the three methods for $\theta \in \{2.5, 3\}$. The median for each configuration was taken over 30 iterations. Note that the observed FDR for all three methods in the configurations described here, was indeed, on average, controlled at the desired level. In the table, D_1 and D_2 denote the number of genes that are over-expressed in groups 1 and 2, respectively. When θ and $D = D_1 + D_2$ are large, the three methods have similar power (defined as the fraction of the D differentially expressed genes correctly classified.) However, as the signal strength or the total number of DE genes decreases, N_3 and L_2N tend to have higher power than limma. Of course, the power of all methods increases with θ. What is perhaps less obvious is that the power decreases with D. The difference in power between the methods in these situations is particularly important because in many cases, it is believed that the proportion of DE genes is small. For example, when $G = 20,000$ and $D = 400$

Table 2: Simulation results – median true positive DE genes. G =5,000 genes, $n_1 = n_2 = 5$.

		$\theta = 2.5$			$\theta = 3$		
D_1	D_2	limma	N_3	L_2N	limma	N_3	L_2N
500	500	0.31	0.3	0.31	0.59	0.57	0.57
225	25	0.1	0.13	0.13	0.35	0.36	0.35
50	50	0.05	0.08	0.08	0.2	0.22	0.24
35	15	0.03	0.06	0.06	0.14	0.19	0.19

(proportional to the third row in Table 2) , and $\theta = 3$, the 4% difference in power between L_2N and limma translates into 16 additional true DE discoveries. Arguably, from a practical standpoint, finding these additional 16 genes could have important consequences.

We also simulated data according to the limma model, and in this scenario the three method have identical power. For example, with a total of $D = 250$ DE genes, the powers obtained by all three methods for $v_0 = 6, 9, 12$ are 0.18, 0.28, and 0.35, respectively. When we hold v_0 fixed and decrease D we again observe that the power decreases. For example, with $v_0 = 9$, the powers of all three methods for $D = 1000, 500, 200, 100$ are 0.34, 0.3, 0.27, 0.24, respectively.

3.2 Simulation Study – Differential Expression and Differential Dispersion

We simulated datasets in which some genes were differentially dispersed across the two groups, as well as differentially expressed. The goal was to test not only whether the L_2N method can detect those differentially dispersed genes, but also to check if and how the presence of DD genes affects the power to detect DE genes. The results presented here were obtained with $\alpha = 9$ (relatively low variability of the error variance), and $\theta \in \{2.5, 3\}$. We simulated 50 over-dispersed genes in group 1 and 50 over-dispersed genes in group 2, by dividing the standard deviation for the DD genes in one of the groups by 6.

Table 3 shows the results for $D = 500$ and $D = 250$, for $\theta = 2.5, 3$. It is clear that the approaches of N_3 and L_2N, where the variance estimates are obtained from a three-way mixture model for differential dispersion, greatly increase the power to detect DE genes when DD is present as compared with limma. We also see that L_2N is slightly more powerful than N_3. When the variance of the null distribution increases (i.e., when α decreases) the difference in power between L_2N and N_3 increases. For example, with $\alpha = 6$, $\theta = 2.5$, and 250 DE genes, L_2N has a power of 0.46, vs. 0.44 for N_3 (and limma has a power of 0.15).

Table 3: Simulation results – median true positive DE genes when 100 genes are also differentially dispersed. G =5,000 genes, $n_1 = n_2 = 5$.

		$\theta = 2.5$			$\theta = 3$		
D_1	D_2	limma	N_3	L_2N	limma	N_3	L_2N
250	250	0.16	0.35	0.36	0.38	0.55	0.55
200	50	0.18	0.45	0.46	0.35	0.60	0.61

As was the case in the previous subsection, the observed FDR for all three methods was controlled at the desired level.

3.3 Simulation Study – Mean Shift

The three-way mixture models, N_3 and L_2N, include a term for an overall difference between the mean expression in the two groups, whereas the limma model assumes there is none. This difference was denoted earlier by θ_0. To understand why it is important to account for this difference, we also simulated data sets in which $\theta_0 = 0.5$ and $\theta_0 = 1$. In these simulations no genes were differentially dispersed.

The overall difference is estimated correctly by the N_3 and L_2N methods automatically, and accounting for $\theta_0 \neq 0$ yields results which are practically identical to the simulations in which $\theta_0 = 0$ in terms of power and FDR. In contrast, when we use limma the FDR is no longer controlled at the desired level (0.05, in these simulations.) For example, with $\alpha = 9$, $\theta \in \{2.5, 3\}$, $D = 250$, and there is a small overall mean shift, $\theta_0 = 0.5$, limma's actual FDR is 0.11. When the overall mean shift is greater, $\theta_0 = 1$, limma's actual FDR is 0.39.

To ensure that the FDR is controlled at the desired level when using the limma modeling approach in `DVX`, one has to center the gene expression data around the group means (or medians) for each gene. That is, if y_{ijg} if the expression level of gene g for subject j in group i, transforming it to $y_{ijg} - \bar{y}_{\cdot\cdot g}$ resolves the problem of high false positive rate. The `DVX` software provides an easy way to transform gene expression data, and in particular it allows the user to perform median-centering.

3.4 Case Study

Understanding the mechanisms that preserve normal neuronal functionality is very important for treating Alzheimer's disease (AD) patients. REST/NRSF (repressor element 1-silencing transcription/neuron-restrictive silencer factor) is known to regulate neuronal genes during embryonic development, and Lu et al. [15] showed that it is "induced in the aging human brain and regulates a network of genes that mediate cell death, stress resistance and AD pathology." Lu et al. [15] observed that REST is lost from the nucleus of cells among AD and mild cognitive impairment (MCI) patients, which leads to dysregulation of this gene network.

In the experiment, gene expression levels for 54,675 genes were obtained from 41 people, in four groups: extremely aged (95-106yr) (n=4), normal aged (70-94yr) (n=16), middle aged (40-69yr) (n=9), and young (<40yr) (n=12). There are 21 females and 20 males in this sample. The data has been deposited in the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository with accession number GSE53890.

The deposited data has been normalized, but we observed some skewness which can be explained by a large number of low-abundance and low-variance genes. We eliminate these genes, as they may be indistinguishable from 'background noise'. We filter out genes which have an overall median log-expression ≤ 5.5, across all subjects. We also equalized the medians across all samples, in order to ameliorate any subject-specific effects. The resulting dataset contained 17,833 genes. Since the change in neuronal condition is known to deteriorate gradually over time for adults, we tested three contrasts: young vs. middle

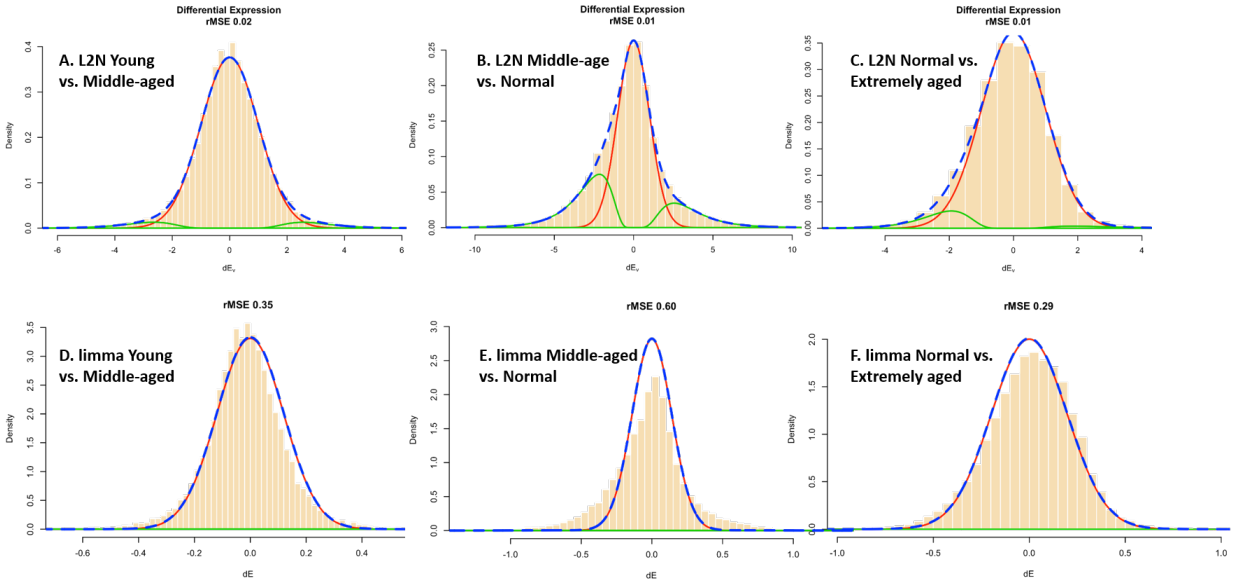

Figure 2: Fitted distributions for the differential expression statistics. The sub-figures in the top row (A-C) were obtained by fitting the L_2N model. Sub-figures D-F were obtained by using limma. Panels A and D depict the fitted distributions for the comparion between middle aged and young (the baseline); B and E correspond to the comparison between Middle-aged (baseline) and Normal-aged, and C and F correspond to the contrast Normal-aged (baseline) vs. Extremely-aged.

aged, middle aged vs. normal aged, and normal aged vs. extremely aged. We performed the differential analysis using L_2N, N_3 and limma, and in all cases we controlled for gender.

Figure 2 shows the fitted distributions for each of the three contrasts, namely, Young as the baseline group vs. Middle-aged as treatment (left), Middle-aged vs. Normal-aged (center), and Normal-aged vs. Extremely aged (right). Sub-figures A-C were obtained by fitting the L_2N model, and D-F were obtained by using limma. The red curve represents the distribution of the 'null' (non differentially expressed) genes, and the green curves show the distribution of the non-null genes, per the selected model. Note that limma, by default, uses P(non-null) = 0.01. The dashed blue line is the fitted mixture distribution. These plots also show the estimated goodness of fit, in terms of the root mean squared error (rMSE). For all three contrasts, the L_2N model yields a better fit than limma. For example, rMSE=0.02 vs. 0.35 for the Young vs. Middle-aged comparison. Note that the scales on the x-axis are different for limma and L_2N (and N_3, which is not shown here.) For limma, dE is the estimated contrast between the two groups, accounting for predictors. If no predictors are included in the model, dE is just the difference between the mean expression level in the treatment group and that in the control group. When using N_3 or L_2N to fit the data, the x-axis is labeled dE_v which is the estimated standardized contrast between the two groups, accounting for predictors. The standardized contrasts are obtained by dividing dE by the estimated gene-specific standard deviations.

It is clear from the plots that in the comparisons young vs. middle aged and normal aged vs. extremely aged, the vast majority of genes are not differentially expressed, whereas

in the comparison between middle aged vs. normal aged many genes are estimated to be differentially expressed. Furthermore, plots B and C show that according to the L_2N model, there are more genes which are overexpressed in the younger cohort (the baseline group in each comparison), since the mixture component in the $(-\infty, 0]$ range has a higher weight than the component in the $[0, \infty)$ range.

Both L_2N and N_3 also test for differential variation. With this dataset and with a q-value [19] threshold of 0.01, no genes are differentially dispersed in any of the three comparisons. The number of differentially expressed genes obtained from each method for each of the three contrasts (using q$\leq$0.01) is summarized in Table 4. The columns labeled "B>T" ("T>B") contain the total number of genes found to be overexpressed in the baseline (treatment) group. The numbers in parentheses are the root mean squared errors for the fitted model. In all three comparisons, L_2N gives the best fit, in terms of the rMSE. In the comparison between Young and Middle Aged, no genes are found to be DE when using limma, while N_3 and L_2N find 157 DE genes with q-value ≤ 0.01. Similarly, in the comparison between "Normal Aged" and "Extremely Aged" limma detects no DE genes, and N_3 and L_2N find 64 DE genes. All three methods detect hundreds of DE genes in the comparison between "Middle Aged" and "Normal Aged" (limma: 522, N_3: 2,682, L_2N: 2,725), suggesting that a significant change, as pertains to cognition, takes place after 70yrs.

A detailed version of this case study which demonstrates several features of the `DVX` package (e.g., use of diagnostic plots and data transformation tools, generation of report and result files in Microsoft Word- and Excel-compatible formats, respectively) is available in the supplementary material.

4 Discussion

We introduced the L_2N three-component mixture model and corresponding empirical Bayes implementation as a flexible approach for assessing differential variation and expression. The proposed model is particularly well-suited for situations in which there is no a priori knowledge of the mixture distribution of the data: it can capture asymmetries in the numbers of over-expressed [dispersed] and under-expressed [dispersed] genes when they exist and can provide a better fit to null and non-null components that are not well-separated, while still performing well under symmetry and little overlap. In our power analysis, we compared

Table 4: Case study - the REST dataset (GSE53890) - total number of differentially expressed genes obtained from limma, N_3, and L_2N.

Test		limma		N_3		L_2N	
Baseline	Treatment	B<T	B>T	B<T	B>T	B<T	B>T
< 40	$[40-70]$	0	0	66	91	66	91
	(rMSE)	(0.35)		(0.02)		(0.02)	
$[40-70]$	$[71-94]$	200	322	1130	1552	1087	1638
	(rMSE)	(0.6)		(0.05)		(0.01)	
$[71-94]$	$[95-106]$	0	0	24	40	24	40
	(rMSE)	(0.29)		(0.02)		(0.01)	

L_2N with two other hierarchical mixture models, namely, limma [18], and N_3 which consists of three normal distributions [2]. A key feature common to all three models is that the differential expression d_g of non-DE genes are assumed to come from a normal distribution. Each model assumes that the DE genes arise from a common distribution, but the choice of nonnull distribution differs across the three models. This hierarchical modeling approach results in 'borrowed information' across genes, which leads to greater power, when compared to naive (one gene at a time) approaches.

Both L_2N and N_3 have two advantages over limma. First, the DE genes are allowed to have a non-symmetrical prior distribution, which makes the three-component mixture models less restrictive, and more realistic in many cases, as demostrated by the case study. Second, the same mixture model can be used to account for differential variation when evaluating differential (mean) expression. We see in our simulations that this leads to improved power when differential variation truly exists.

L_2N was shown to be a bit more powerful than N_3 when differential dispersion exists, and more so when the variability of the random errors increases. Furthermore, in our simulations and analysis of real datasets, L_2N seems to provide the best fit to the differential expression statistics in terms of rMSE. Also, of the three mixture models, L_2N is the only one which uses non-local priors for the DE genes. Conceptually, this is advantageous because it implies that there is a negligible probability of declaring a gene as DE when the actual differential expression statistic is close to 0.

To make the method easily accessible to biologists, we created a user-friendly R Shiny interface called `DVX` [4], which also includes the implementations of N_3 and `limma`. With all three models, it is possible to control for the effects of other factors and covariates, and to set up linear contrasts, beyond the simple two-group comparison. `DVX` offers a set of diagnostic plots and transformation tools, and options for exporting Word-compatible reports and Excel-compatible result tables. `DVX` may also be used with count data (e.g., RNA-seq read counts) with proper data transformation, using tools such as 'voom' (see, e.g., [14] and [13].)

Acknowledgements

We wish to thank Ved Deshpande and M. Henry Linder for their help with earlier versions of the code and documentation.

Funding

This work was partially supported by the University of Connecticut Research Foundation [EDS].

References

[1] Haim Bar, James Booth, Elizabeth Schifano, and Martin Wells. Laplace approximated EM microarray analysis: an empirical Bayes approach for comparative microarray experiments. *Statistical Science*, 25(3):388–407, 2010.

[2] Haim Bar, James Booth, and Martin Wells. A bivariate model for simultaneous testing in bioinformatics data. *JASA*, 2014.

[3] Haim Bar and Elizabeth Schifano. Empirical and fully Bayesian approaches for random effects models in microarray analysis. *Statistical Modelling*, 11(1):71–88, 2011.

[4] Haim Bar and Elizabeth D. Schifano. *DVX: an R package for Differential Variation and eXpression analysis*, 2018.

[5] Yoav Benjamini and Yosef Hochberg. Controlling the false dicovery rate-a practical and powerful approach to multiple testing. *Journal Of The Royal Statistical Society Series B*, 57(3):499–517, 1995.

[6] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2017. R package version 1.0.5.

[7] Bradley Efron. Microarrays, empirical bayes and the two groups model. *Statistical Science*, 23(1):1–22, 2008.

[8] Bradley Efron, Robert Tibshirani, J. Storey, and V. Tusher. Empirical bayes analysis of microarray experiment. *JASA*, 96:1151–1160, 2001.

[9] Andrew P. Feinberg and Rafael A. Irizarry. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*, 107(suppl 1):1757–1764, January 2010.

[10] J. T. Gene Hwang and Peng Liu. Optimal Tests Shrinking Both Means and Variances Applicable to Microarray Data Analysis. *Statistical Applications in Genetics & Molecular Biology*, 9(1):1–33, 2010.

[11] R. Ivanek, Y. T. Grohn, M. T. Wells, S. Raengpradub, M. J. Kazmierczak, and M. Wiedmann. Extreme value theory in analysis of differential expression in microarrays where either only up- or down-regulated genes are relevant or expected. *Genet Res (Camb)*, 90(4):347–361, Aug 2008.

[12] Valen E Johnson and David Rossell. On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170, 2010.

[13] C. W. Law, M. Alhamdoosh, S. Su, G. K. Smyth, and M.E. Ritchie. RNA-seq analysisis easy as 1-2-3 with limma,Glimma and edgeR. *F1000Research*, 5(1408), 2016.

[14] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, 15(2):R29, Feb 2014.

[15] Tao Lu, Liviu Aron, Joseph Zullo, Ying Pan, Haeyoung Kim, Yiwen Chen, Tun-Hsiang Yang, Hyun-Min Kim, Derek Drake, X. Shirley Liu, David A. Bennett, Monica P. Colaicovo, and Bruce A. Yankner. Rest and stress resistance in ageing and alzheimer?s disease. *Nature*, 507(7493):448–454, 2014.

[16] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, October 2009.

[17] Jessica C. Mar, Nicholas A. Matigian, Alan Mackay-Sim, George D. Mellick, Carolyn M. Sue, Peter A. Silburn, John J. McGrath, John Quackenbush, and Christine A. Wells. Variance of Gene Expression Identifies Altered Network Constraints in Neurological Disease. *PLoS Genetics*, 7(8), August 2011.

[18] Gordon K. Smyth. Linear models for empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004. Article 2.

[19] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[20] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.*, 98:5116–5121, Apr 2001.