Hypothesis Testing in Normal Admixture Models to Detect Heterogeneous Genetic Signals

Qian Fan1, Richard Charnigo*, Zohreh Talebizadeh1 and Hongying Dai3,4
1Department of Statistics, 725 Rose Street, University of Kentucky, Lexington KY, USA
2Division of Experimental and Translational Genetics, 2401 Gillham Road, Children’s Mercy Hospital, Kansas City MO, USA
3Research Development and Clinical Investigation, 2401 Gillham Road, Children’s Mercy Hospital, Kansas City MO, USA
4Department of Biomedical & Health Informatics, 5100 RockHill Road, University of Missouri-Kansas City, Kansas City MO, USA

Abstract

In this work we consider a three-component normal mixture model in which one component is known to have mean zero and the other two contaminating components have nonnegative and nonpositive mean respectively, while all three components share a common unknown variance parameter. One potential application of this model may be in prioritizing statistical scores obtained in biological experiments, including genetics data. Such a mixture model may be useful in describing the distribution of numerous Z test statistics corresponding to different genes or SNPs, such that a “significant” Z test statistic for a particular gene suggests its connection to a medical condition. More specifically, the inferences drawn from such a mixture model may be useful in a filtration algorithm to remove large subsets of genes or SNPs from consideration, thereby reducing the need for stringent and power-depleting multiplicity adjustments for controlling type I error rates on the remaining genes. We show how to test whether there is contamination in at least one direction (i.e., the mixture model truly requires at least two components) and, if so, how to test whether there is contamination in both directions (i.e., the mixture model truly requires all three components). We assess our testing procedures in simulation studies and illustrate them through application to LOD scores in a genome-wide linkage analysis from an autism study.

Keywords: Autism; Bilaterally contaminated normal model; Gene expression; Genome-wide linkage; LOD; Microarray; Nuisance parameter; Omnibus null hypothesis; Unilateral null hypothesis

Introduction

Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) with probability density function (pdf)

$$f(x) = \sum_{i=1}^{3} \pi_i f_i(x),$$

where $f_i(x)$ are the density functions of the three components.

For a three-component normal mixture model, the pdf can be written as

$$f(x) = \sum_{i=1}^{3} \pi_i N(x; \mu_i, \sigma^2),$$

where $N(x; \mu, \sigma^2)$ is the normal distribution with mean $\mu$ and variance $\sigma^2$.

If $\sigma^2$ is known while the other parameters are unknown, then (1) is called a bilaterally contaminated normal (BCN) model. Charnigo et al. [1] observed that, in this case, one may test the omnibus null hypothesis that $\gamma_1 \mu_1 = 0$ and $\gamma_2 \mu_2 = 0$ by comparing the second sample moment to a chi-square quantile. Dai and Charnigo [7,8] subsequently adapted the modified likelihood ratio test to accommodate two-component mixtures in which some or all of the parameters for one mixture component were known a priori. They referred to such a mixture model as a contaminated density model or, for short, contaminated model. Modified likelihood ratio testing for mixture models with more than two components does not appear to have a similarly tractable asymptotic theory, which has inspired the development of the EM-test [9-11]. While we have some optimism that the EM-test may be useful for inference in the BCN+NP model, the present paper will develop tests based primarily on moments.

The practical motivation for the BCN+NP model is largely the same as for the BCN model, as described by Charnigo et al. [1]. To briefly recap, suppose that $X_i$ is a test statistic for comparing cases to controls on the expression level of gene $i$ in a microarray [12,13], such that $X_i - N(0, \sigma^2)$ if patients and controls have the same mean for either the BCN or BCN+NP model [2-5].

Due to difficulties with likelihood ratio testing in mixture modeling, Chen et al. [6] developed a modified likelihood ratio test to address whether a two-component mixture could be reduced to a homogeneous distribution. Applicable under fairly general circumstances, their test was supported by an asymptotic theory and simulation results showing that chi-square quantiles could be used as critical values. Dai and Charnigo [7,8] subsequently adapted the modified likelihood ratio test to accommodate two-component mixtures in which some or all of the parameters for one mixture component were known a priori. They referred to such a mixture model as a contaminated density model or, for short, contaminated model. Modified likelihood ratio testing for mixture models with more than two components does not appear to have a similarly tractable asymptotic theory, which has inspired the development of the EM-test [9-11]. While we have some optimism that the EM-test may be useful for inference in the BCN+NP model, the present paper will develop tests based primarily on moments.

The practical motivation for the BCN+NP model is largely the same as for the BCN model, as described by Charnigo et al. [1]. To briefly recap, suppose that $X_i$ is a test statistic for comparing cases to controls on the expression level of gene $i$ in a microarray [12,13], such that $X_i - N(0, \sigma^2)$ if patients and controls have the same mean
expression level, \( X_i \sim N(\mu, \sigma^2) \) if patients have greater expression, and \( X_i \sim N(\mu, \sigma^2) \) if patients have lesser expression. Testing the omnibus null hypothesis asks whether some genes are differentially expressed in patients versus controls, and testing the unilateral null hypothesis asks whether there exist simultaneously some genes that are over expressed and other genes that are under expressed.

As suggested by Dai and Charngio [14] in a somewhat different context, omnibus testing may help a researcher to avoid overly stringent adjustments for controlling Type I error rates in multiple testing. For example, if an omnibus test suggests no differential expression within a subset of genes, then that subset of genes can be “filtered out” and not contribute to the adjustments for controlling Type I error rates in other gene-specific tests across the genome. Thus, gene-specific tests in subsets of genes that have not been filtered out may become more powerful, improving the researcher’s ability to detect differential expression within such subsets of genes. As an aside, we mention that gene expression data are ordinarily normalized prior to performing expression within such subsets of genes. As an aside, we mention that gene expression data are ordinarily normalized prior to performing

\[
\frac{2\pi n^{-1/2} \| X \|_1}{\sqrt{n}}
\]

has a distribution that does not depend on \( \sigma^2 \). In fact, the asymptotic null distribution of \( \frac{2\pi n^{-1/2} \| X \|_1}{\sqrt{n}} \) is standard normal by Cramer’s Theorem [15]. This seems to suggest that we reject the null hypothesis when \( \frac{2\pi n^{-1/2} \| X \|_1}{\sqrt{n}} > z_{\alpha} \), where \( \alpha \) is the desired significance level.

Unfortunately, such a procedure is not consistent against all alternative hypotheses. Indeed, \( R \) may converge in probability to a number less than 1 under some alternative hypotheses (e.g., when \( \sigma^2 = 1, \mu_1 = 1, \mu_2 = 2, \gamma = y = 0.05 \)), to 1 under others (e.g., when \( \sigma^2 = 1, \mu_1 = 2, \mu_2 = 2, \gamma = y = 1/6 \)), or even to a number greater than 1 (e.g., when \( \sigma^2 = 1, \mu_1 = 1, \mu_2 = 1, \gamma = y = 0.3 \)).

The above example does not prove the non-existence of a simple test based on moments, but other obvious candidates for a test statistic involving low-order moments suffer similar difficulties. On the other hand, as Lemma 4.1 shows, \( H_2 \) can be equivalently expressed as

\[
H_2 : \mu_1 \neq \mu_2, \gamma \neq y.
\]

Lemma 4.1 If \( H_2 \) is false, then at least one of the following conditions must hold:

(a) \( m_1 / (15m_0) > 1 \), (b) \( m_2 / (15m_0) > 1 \),

(c) \( m_1 / (15m_0) < 1 \), (d) \( m_2 / (15m_0) < 1 \).

Proof. First suppose that \( H_2 \) is false with \( \mu_1 = \mu_2 > 0 \) and \( y = y > 0 \). If \( m_1 / (15m_0) > 1 \), then \( y = 1/6 \) and thus \( m_2 / (15m_0) < 1 \).

Next suppose that \( H_2 \) is false with \( \mu_1 = \mu_2 < 0 \). If \( m_2 / (15m_0) < 1 \), then \( y = 1/6 \) and thus \( m_1 / (15m_0) > 1 \).

This motivates the development of a union-intersection test [16]. Let \( \hat{m}_1 = n^{-1} \sum_{i=1}^{n} X_i \) denote the \( k \)th sample moment and put

\[
T := \sqrt{4m_1 / m_0}, \quad U := \sqrt{4n / 3(4m_1 / (3m_0) - 1)},
\]

\[
V := \sqrt{4n / 136(4m_1 / (150m_0) - 1)}, \quad W := \sqrt{4m_1 / 135m_0}.
\]

Theorem 4.1 establishes a test based on \( T, U, V, \) and \( W \).

Theorem 4.1 Let \( \alpha_1, \alpha_2, \alpha_3, \alpha_4 \in (0, 1) \) be given such that \( \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = \alpha \). Consider a test that rejects \( H_2 \) if and only if at least one of the following conditions holds: (a) \( T > z_{1-\alpha_1} \), (b) \( U > z_{1-\alpha_2} \),

(c) \( V > z_{1-\alpha_3} \), (d) \( W > z_{1-\alpha_4} \). Then the probability of correctly rejecting the null hypothesis under \( H_2 \) is asymptotically less than or equal to \( \alpha \), while the probability of rejecting the null hypothesis under \( H_2 \) converges to 1.

Proof. First suppose that \( H_2 \) is true. In this case, \( T, U, V, \) and \( W \) are all asymptotically standard normal by Cramer’s Theorem. The probability of incorrectly rejecting the null hypothesis is bounded above by \( P(T > z_{1-\alpha_1}) + P(U > z_{1-\alpha_2}) + P(V > z_{1-\alpha_3}) + P(W > z_{1-\alpha_4}) \), which converges to \( \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = \alpha \).

Next suppose that \( H_2 \) is false. In this case, at least one of \( n^{1/2}T \), \( n^{1/2}U \), \( n^{1/2}V \), and \( n^{1/2}W \) converges in probability to a nonzero quantity. Thus, at least one of \( T, U, V, \) and \( W \) diverges to infinity in probability, so that the corresponding probability in the preceding paragraph tends to 1, and this probability is a lower bound for the probability of correctly rejecting the null hypothesis.

We note that \( \hat{m}_2 \) in the described procedures is available upon request to the corresponding author.
The question of practical interest is how to choose $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. A simple choice is to set each of them equal to $\alpha/4$, but this may not optimize the power to correctly reject $H_0$. For example, if there is bilateral contamination and the contamination is symmetric, then neither $T$ nor $W$ will be useful. Rather, the contamination will be detectable via $U$ or $V$, suggesting that $\alpha_1$ and $\alpha_2$ be chosen close to $\alpha/2$ after examining the data, as this may result in a hidden inflation of Type I error rate.

**Testing the Unilateral Null Hypothesis**

Now we are concerned with testing the unilateral null hypothesis against its corresponding alternative,

$$H_i: y \mu_i = 0 \text{ or } y \mu_i \neq 0 \text{ against } H_i: y \mu_i = 0 \text{ and } y \mu_i 
eq 0.$$  

The alternative hypothesis indicates that there is contamination in both directions. Because this test is developed assuming that the omnibus null hypothesis is false, in practice we recommend sequential testing: first perform the test described in Section 4, and then, if only the omnibus null hypothesis is rejected, proceed to the test that we describe presently.

If $\sigma^2$ were known, a test of the unilateral null hypothesis could be obtained [1] based on

$$h(m, \sigma^2) := (m_i - \sigma^2)^2 + 3m_i^2\sigma^2 - m_i,$$

Where $m_i := (m_i, m_i, m_i, m_i)^T$. More specifically, since $h(m, \sigma^2) \geq 0$ with equality if and only if $H_i$ is true, one could reject $H_i$ if $h(m, \sigma^2)$ exceeded a critical value suggested by Cramer's Theorem, where $m_i := (m_i, m_i, m_i, m_i)^T$. However, since $\sigma^2$ is unknown, $h(m, \sigma^2)$ cannot be calculated and must be replaced by $h(\hat{m}, \hat{\sigma^2})$ for some $\hat{\sigma^2}$. The question then becomes, what is an appropriate critical value?

Before addressing the question, some comment on the choice of $\hat{\sigma^2}$ is warranted. In the simulation studies of Section 6, we take $\hat{\sigma^2}$ to be the maximum likelihood estimator of $\sigma^2$. If $H_i$ is indeed true (and a compact parameter space is imposed), then $\hat{\sigma^2}$ is anticipated to converge to $\sigma^2$ at the rate of $n^{-1/2}$. If $H_i$ is true, then $\hat{\sigma^2}$ is anticipated to converge to the slower rate of $n^{-1/4}[17]$. However, maximum likelihood estimation of $\sigma^2$ is not essential. Rather, we assume that $\hat{\sigma^2}$ is chosen such that there exists known $\delta_i$ with $\mathcal{P}(\sigma^2 - \hat{\sigma^2} \leq \delta_i) \rightarrow 1$ and $\delta_i \rightarrow 0$. For instance, if $\sigma^2$ converges at a rate of $n^{-1/4}$ or better, one may take $\delta_i = n^{1/4+e}$ for some $e \in (0, 1/4)$.

The following lemmas are preparatory to identifying a critical value for $h(\hat{m}, \hat{\sigma^2})$. In what follows, we define $A$ to be the $3 \times 3$ matrix with $A_{ii} = m_i - m_i^2$ and $b$ to be the vector of partial derivatives of $h(\hat{m}, \hat{\sigma^2})$ with respect to $m$. We also define $\hat{A}$ and $\hat{b}$ to be the corresponding estimators.

**Lemma 5.1** Suppose $H_i$ is true. Then, for any $\alpha \in (0, 1)$,

$$\mathcal{P}(h(\hat{m}, \hat{\sigma^2}) \leq z_{1-\alpha}/\sqrt{\hat{b}^T \hat{A} \hat{b}/n}) \rightarrow 1 - \alpha.$$

**Proof.** By Cramer's Theorem, $h(\hat{m}, \hat{\sigma^2})/\sqrt{\hat{b}^T \hat{A} \hat{b}/n}$ converges in law to standard normal. Since $\hat{A}$ and $\hat{b}$ are continuous functions of $m$ and $\sigma^2$, and since $\sqrt{\hat{b}^T \hat{A} \hat{b}} > 0$ when the omnibus null hypothesis is false, the Continuous Mapping Theorem implies that $\sqrt{\hat{b}^T \hat{A} \hat{b}} \rightarrow 1$ in probability. By Slutsky's Theorem, $h(\hat{m}, \hat{\sigma^2})/\sqrt{\hat{b}^T \hat{A} \hat{b}/n}$ converges in law to standard normal. The desired result is an immediate consequence.

**Lemma 5.2** Suppose $H_i$ is true. Then,

$$\mathcal{P}(h(\hat{m}, \hat{\sigma^2}) \leq h(\hat{m}, \hat{\sigma^2}) + \delta_i^2 + 2\delta_i \hat{m}_i + 3\hat{m}_i \delta_i) \rightarrow 1.$$

**Proof.** Suppose that $|\sigma^2 - \hat{\sigma^2}| \leq \delta_i$ and $\hat{m}_i > \sigma^2$, which occur with probability approaching 1.

Then

$$(m_i - \sigma^2)^2 \leq (m_i - \sigma^2)^2 + \delta_i^2 + 2\delta_i \hat{m}_i + 3\hat{m}_i \delta_i,$$

Adding these two inequalities yields the desired result.

We are now in a position to describe our testing procedure.

**Theorem 5.1** Let $\alpha \in (0, 1)$ be given. Consider a test that rejects $H_i$ if and only if $h(\hat{m}, \hat{\sigma^2}) > z_{1-\alpha}/\sqrt{\hat{b}^T \hat{A} \hat{b}/n} + \delta_i^2 + 2\delta_i \hat{m}_i + 3\hat{m}_i \delta_i$. Then the probability of incorrectly rejecting the null hypothesis under $H_i$ is asymptotically less than or equal to $\alpha$, while the probability of correctly rejecting the null hypothesis under $H_i$ converges to 1.

**Proof.** The first part of the conclusion follows from Lemma 5.1 and Lemma 5.2. The second part of the conclusion follows from the facts that $h(\hat{m}, \hat{\sigma^2})$ converges in probability to $h(\hat{m}, \hat{\sigma^2})$, the latter quantity is positive under $H_i$, and the critical value converges in probability to 0.

Our guidelines for $\delta$ are asymptotic. However, to carry out the test in practice, one must specify $\delta_i$ for a finite sample. If $\delta_i$ is too small, then the Type I error rate of the test will be too large. If $\delta_i$ is too large, then the power of the test will be unnecessarily low. The choice of $\delta_i$ is explored empirically in the simulation studies of Section 6.

**Simulation Results**

First we present results from simulation studies to assess the Type I and Type II error rates of our omnibus testing procedure in finite samples from the BCN+NP model. Table 1 pertains to omnibus testing based on sample sizes of $n = 100$ and $n = 1000$, with $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.0125$ and parameter values as shown in the left column. (We put $\sigma = 1$ to generate data but subsequently treated $\sigma$ as unknown. The simulation size was 1000.) To illustrate use of Table 1, consider two examples: First, when $n = 100$, there is approximately a 3.8% Type I error rate. Second, when $n = 1000$, there is approximately 78.0% power against the specific alternative $(\mu_1, \mu_2, \gamma_1^2, \gamma_2^2) = (0, -1, 0.2, 0.1)$.

Tables 2 and 3 also pertain to omnibus testing but with different choices of $\alpha_1$ to $\alpha_4$. In Table 2 more consideration for rejecting the omnibus null hypothesis is given to $T$ and $W$, whereas in Table 3 more consideration is given to $U$ and $V$.

As shown in Tables 1-3, the omnibus test appears conservative at both sample sizes and all three combinations of $\alpha_i$ through $\alpha_4$, in that the observed Type I error rate is less than 5%. As anticipated, power tends to be greater with a larger sample size. The power
Shown above are results from testing the omnibus null hypothesis based on a simulation of size 1000 and using the procedure from Section 4.

Table 1: Type I Error and Power for BCN+NP Data: $\alpha_1=\alpha_2=\alpha_3=\alpha_4=0.0125$.

<table>
<thead>
<tr>
<th>$(\mu_1, \mu_2, \gamma_1, \gamma_2)$</th>
<th>n=100</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0,0)</td>
<td>0.037</td>
<td>0.046</td>
</tr>
<tr>
<td>(1, -1, 0, 0, 0, 0, 0, 0)</td>
<td>0.038</td>
<td>0.004</td>
</tr>
<tr>
<td>(0, -1, 0, 0, 0, 0, 0, 0)</td>
<td>0.011</td>
<td>0.046</td>
</tr>
<tr>
<td>(0, 0, -1, 0, 0, 0, 0, 0)</td>
<td>0.000</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Shown above are results from testing the omnibus null hypothesis based on a simulation of size 100 and using the procedure from Section 4.

Table 2: Type I Error and Power for BCN+NP Data: $\alpha_1=\alpha_2=\alpha_3=\alpha_4=0.005$.

<table>
<thead>
<tr>
<th>$(\mu_1, \mu_2, \gamma_1, \gamma_2)$</th>
<th>n=100</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0,0)</td>
<td>0.032</td>
<td>0.038</td>
</tr>
<tr>
<td>(1, -1, 0, 0, 0, 0, 0, 0)</td>
<td>0.033</td>
<td>0.056</td>
</tr>
<tr>
<td>(0, -1, 0, 0, 0, 0, 0, 0)</td>
<td>0.024</td>
<td>0.024</td>
</tr>
<tr>
<td>(0, 0, -1, 0, 0, 0, 0, 0)</td>
<td>0.080</td>
<td>0.050</td>
</tr>
<tr>
<td>(0, 0, 0, -1, 0, 0, 0, 0)</td>
<td>0.002</td>
<td>0.005</td>
</tr>
</tbody>
</table>

shown above are results from testing the omnibus null hypothesis based on a simulation of size 1000 and using the procedure from Section 4.

Table 3: Type I Error and Power for BCN+NP Data: $\alpha_1=\alpha_2=\alpha_3=\alpha_4=0.005$, $\alpha_1=\alpha_2=\alpha_3=\alpha_4=0.02$.

under unilateral contamination is better when the contaminating mean is larger ($\pm 2$) rather than smaller ($\pm 1$) and when the weight of contamination is larger (0.2) rather than smaller (0.1). The power under asymmetric bilateral normal contamination tends to be better when the contaminating means are different (in absolute value) than when the weights are different. The power under symmetric bilateral contamination is relatively low except when the sample size is larger and $(\mu_1, \mu_2, \gamma_1, \gamma_2)=(-2, -2, 0, 1, 0, 1)$.

For added realism, we also consider scenarios in which the BCN+NP model is mispecified, in that the data originate from the “bilaterally contaminated and scaled T model with nuisance parameter” (BCT+NP model) but then are subjected to forward and reverse cumulative distribution function (cdf) transformations.

More specifically, suppose that $T_1, \ldots, T_n$ are i id with pdf

$$(1 - \gamma_j, \gamma_j)\sigma^2 f((u - \mu_j)/\sigma) + (1 - \mu_j, \mu_j)\sigma^2 f((u - \mu_j)/\sigma).$$

Where $f_j$ denotes the T pdf on $v$ degrees of freedom and $F_j$ denotes the corresponding cdf. We assume that $v$ is known; in the context of microarray data analysis, $v$ will relate to the number of persons (or experimental units) on which gene expression data have been obtained. Data arising from this model can be transformed by $X_j = \Phi^{-1}(F_j(T_j))$, where $\Phi$ is the standard normal cdf. The transformed data are then analyzed as if they had arisen from the BCN+NP model. If $\sigma = 1$ and $\gamma_j = \mu_j = 0$, then the transformed data will actually be standard normal (and hence from the BCN+NP model), but otherwise the transformed data may not truly be from the BCN+NP model.

Table 4 is organized analogously to Table 1, except that there are additional columns corresponding to various choices of $v$ from 5 to 100. (We have also obtained results analogous to those in Tables 2 and 3 but have omitted including them in tabular form to streamline this manuscript.) The omnibus test appears conservative in most scenarios and is not markedly anticonservative in any. As anticipated, power tends to be greater with a larger sample size. The power under unilateral contamination tends to be greater if either the contaminating mean or the contaminating weight is larger, and the power seems to increase with the degrees of freedom. The power under asymmetric bilateral contamination often increases with the degrees of freedom and tends

Table 4: Type I Error and Power for Transformed BCT+NP Data: $\alpha_1=\alpha_2=\alpha_3=\alpha_4=0.0125$.

<table>
<thead>
<tr>
<th>$(\mu_1, \mu_2, \gamma_1, \gamma_2)$</th>
<th>df=5</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>df=5</th>
<th>10</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,0,0,0)</td>
<td>0.027</td>
<td>0.03</td>
<td>0.026</td>
<td>0.039</td>
<td>0.038</td>
<td>0.037</td>
<td>0.045</td>
<td>0.051</td>
</tr>
<tr>
<td>(1, -1, 0, 0, 0, 0, 0, 0)</td>
<td>0.010</td>
<td>0.01</td>
<td>0.009</td>
<td>0.025</td>
<td>0.020</td>
<td>0.019</td>
<td>0.024</td>
<td>0.029</td>
</tr>
<tr>
<td>(0, -1, 0, 0, 0, 0, 0, 0)</td>
<td>0.010</td>
<td>0.01</td>
<td>0.009</td>
<td>0.025</td>
<td>0.020</td>
<td>0.019</td>
<td>0.024</td>
<td>0.029</td>
</tr>
<tr>
<td>(0, 0, -1, 0, 0, 0, 0, 0)</td>
<td>0.010</td>
<td>0.01</td>
<td>0.009</td>
<td>0.025</td>
<td>0.020</td>
<td>0.019</td>
<td>0.024</td>
<td>0.029</td>
</tr>
<tr>
<td>(0, 0, 0, -1, 0, 0, 0, 0)</td>
<td>0.010</td>
<td>0.01</td>
<td>0.009</td>
<td>0.025</td>
<td>0.020</td>
<td>0.019</td>
<td>0.024</td>
<td>0.029</td>
</tr>
<tr>
<td>(0, 0, 0, 0, -1, 0, 0, 0)</td>
<td>0.010</td>
<td>0.01</td>
<td>0.009</td>
<td>0.025</td>
<td>0.020</td>
<td>0.019</td>
<td>0.024</td>
<td>0.029</td>
</tr>
<tr>
<td>(0, 0, 0, 0, 0, -1, 0, 0)</td>
<td>0.010</td>
<td>0.01</td>
<td>0.009</td>
<td>0.025</td>
<td>0.020</td>
<td>0.019</td>
<td>0.024</td>
<td>0.029</td>
</tr>
<tr>
<td>(0, 0, 0, 0, 0, 0, -1, 0)</td>
<td>0.010</td>
<td>0.01</td>
<td>0.009</td>
<td>0.025</td>
<td>0.020</td>
<td>0.019</td>
<td>0.024</td>
<td>0.029</td>
</tr>
<tr>
<td>(0, 0, 0, 0, 0, 0, 0, -1)</td>
<td>0.010</td>
<td>0.01</td>
<td>0.009</td>
<td>0.025</td>
<td>0.020</td>
<td>0.019</td>
<td>0.024</td>
<td>0.029</td>
</tr>
</tbody>
</table>

shown above are results from testing the unilateral null hypothesis based on a simulation of size 1000 and using the procedure from Section 4.

Table 5: Type I Error and Power for BCN+NP Data: Small $\delta$. 

<table>
<thead>
<tr>
<th>$(\mu_1, \mu_2, \gamma_1, \gamma_2)$</th>
<th>n=100</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, -1, 0, 0, 0, 0, 0, 0)</td>
<td>0.100</td>
<td>0.090</td>
</tr>
<tr>
<td>(0, -1, 0, 0, 0, 0, 0, 0)</td>
<td>0.080</td>
<td>0.070</td>
</tr>
<tr>
<td>(0, 0, -1, 0, 0, 0, 0, 0)</td>
<td>0.060</td>
<td>0.050</td>
</tr>
<tr>
<td>(0, 0, 0, -1, 0, 0, 0, 0)</td>
<td>0.040</td>
<td>0.030</td>
</tr>
<tr>
<td>(0, 0, 0, 0, -1, 0, 0, 0)</td>
<td>0.020</td>
<td>0.010</td>
</tr>
<tr>
<td>(0, 0, 0, 0, 0, -1, 0, 0)</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(0, 0, 0, 0, 0, 0, -1, 0)</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(0, 0, 0, 0, 0, 0, 0, -1)</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

shown above are results from testing the unilateral null hypothesis based on a simulation of size 1000 and using the procedure from Section 5 with $\delta=0.05$ for $n=100$ and $\delta=0.025$ for $n=1000$.
To illustrate our new testing procedures, we analyzed LOD scores obtained in a whole genome linkage analysis from an autism study [18]. Autism [19] is a complex neuro developmental condition that might be affected by multiple genetic and non-genetic factors. Furthermore, there is a high degree of phenotypic heterogeneity both within and among families. To address the heterogeneity in disease phenotypes, Talebizadeh et al. [18] proposed a novel multi-step stratification method that divides subjects with autism into subgroups using previously developed cluster analyses of severity scores from an autism diagnostic test [20]. The objective of the applied stratification method was to identify subgroups representing more homogeneous autism subjects by reducing both inter and intra-family heterogeneity. Linkage analysis [21] was then performed to identify genetic markers linked with autism within each subgroup. Linkage analysis is a method to find the approximate chromosomal position of disease genes by testing for co-segregation of a trait of interest relative to known genetic markers. The likelihood of co-segregation (linkage) is estimated by calculating LOD scores [21].

After data quality control and filtration, 16973 SNPs (autosomal and X-linked) from a total of 392 multiplex families were included for the linkage analysis. Subjects were stratified into a total of 16 subgroups considering the following: affected individual’s disease severity [20], intra-family heterogeneity, and affected individual’s gender [i.e., male only (M) and female-containing (Fc) pedigrees]. The LOD score from the linkage analysis is a measure of the strength of association between a genetic marker and disease in familial data. A LOD score that is less than or equal to 0 suggests no genetic linkage. To characterize the distribution of LOD scores, we applied the BCN+NP model to LOD scores within the “G4Fc” subgroup. The distributions of LOD scores in most other subgroups were not deemed suitable for BCN+NP modeling; they might have been amenable to a normal mixture model in which different components could have different variances, but such

<table>
<thead>
<tr>
<th>(μ1, μ2, γ1, γ2)</th>
<th>n=100</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 1, 0, 0)</td>
<td>0.003</td>
<td>0.007</td>
</tr>
<tr>
<td>(2, 2, 0, 0)</td>
<td>0.028</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 6: Type I Error and Power for BCN+NP Data: Medium δ.

<table>
<thead>
<tr>
<th>(μ1, μ2, γ1, γ2)</th>
<th>n=100</th>
<th>n=1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1, 0, 0, 0)</td>
<td>0.010</td>
<td>0.007</td>
</tr>
<tr>
<td>(0, 1, 0, 0)</td>
<td>0.023</td>
<td>0.030</td>
</tr>
</tbody>
</table>

Table 7: Type I Error and Power for BCN+NP Data: Large δ.

The fitted BCN+NP and UCN+NP models, virtually indistinguishable, are displayed.

Figure 1: Fitted models from LOD case study.
Moving to unilateral testing for the BCN+NP model, we obtain $h(m, \sigma^2) = 7.4 \times 10^{-4}$. With $\alpha=0.05$ and $\delta=0$, the critical value for rejection of the unilateral null hypothesis is $9.5 \times 10^{-4}$. Since the critical value is an increasing function of $\delta$ (for example, the critical value with $\delta=0.05$ is $3.1 \times 10^{-4}$), there is no choice of $\delta$ for which the unilateral null hypothesis will be rejected at $\alpha=0.05$.

To understand why the unilateral null hypothesis is not rejected, we can juxtapose the fitted BCN+NP model against the fitted “unilaterally contaminated normal model with nuisance parameter” (UCN+NP model), a special case of the BCN+NP model in which either $\gamma_1 \mu_1 = 0$ or $\gamma_2 \mu_2 = 0$ but not(necessarily) both. Both models are fitted by maximum likelihood and are displayed in Figure 1. They are numerically indistinguishable to three decimal places on the parameters, as $0.545N(0, 0.119) + 0.135N(0.980, 0.119) + 0.320N(0, 0.119)$ and $0.865N(0, 0.119) + 0.135N(0.980, 0.119)$ respectively.

One might have anticipated that the estimate of $\mu_1$ equaling zero in the BCN+NP model should have forced $h(m, \sigma^2)$ to equal zero as well. However, the former estimate is based on maximum likelihood, whereas the latter test statistic is based primarily on moments. Even so, neither the fitted BCN+NP model nor the test statistic argues against the unilateral null hypothesis.

In light of the simulation results in Section 6, one may also be concerned about the possibility of inadequate power for the unilateral testing procedure. However, because the sample size in this case study was more than 16 times the larger sample size from the simulation results, and because the estimate of $\mu_1$ was zero, we do not believe that there was an undetected deviation of any importance from the unilateral null hypothesis in this case study, at least to the extent that the BCN+NP model approximation was valid.

The final fitted model for the G4Fc subgroup, one of the female-containing subgroups, is $0.865N(0, 0.119) + 0.135N(0.980, 0.119)$. This suggests that about 13.5% of genetic variants belong to a mixture component with $\mu_2>0$. We further calculated the posterior probabilities for genetic variants belonging to this mixture component. Such a posterior probability is a monotone function of the LOD score but may provide some insight that a LOD score does not, namely the probabilistic interpretation of how likely the genetic variant is to belong to the mixture component with $\mu_2>0$.

There are 253 SNPs with posterior probability greater than 99% (corresponding LOD score, 1.27) and 669 SNPs with posterior probability above 98% (corresponding LOD score, 1.19). Using 50% posterior probability as a threshold, the cutoff point for LOD scores is 0.765. In other words, if a LOD score is less than 0.765, then the genetic variant will be assigned to the mixture component with mean zero. If a LOD score is greater than 0.765, then the genetic variant will be assigned to the mixture component with mean $\mu_2>0$. Alternatively, if one wishes to assign 13.5% of genetic variants to the mixture component with mean $\mu_2>0$, then one may use a cutoff of 0.427. On the other hand, a LOD score of 0.427, 0.765, or even 0.980 may not be sufficiently large to argue for a clear connection of the genetic variant with autism. Thus, caution is required in interpreting the results of the fitted model.

Conclusions

We have presented and theoretically justified new procedures...
for testing omnibus and unilateral null hypotheses in a bilaterally contaminated normal model with nuisance parameter representing the unknown within-component variability. As our case study makes clear, there will arise situations in which assuming the within-component variability to be known and equal to unity is not a viable modeling strategy, and thus the procedures in the earlier work by Charnigo et al. [1] will not be applicable.

Our case study also illustrates that having a unilateral testing procedure is worthwhile. One may be inclined to assume that, if contamination is present in one direction, contamination should also be present in the other direction. Such an assumption may be true in many instances, but being able to declare that contamination is exclusively (or, at least, primarily) in one direction may be of scientific importance. Thus, even if one has an adequate sample size to estimate parameters for a model with two contaminating components, adopting such a model may be neither necessary nor desirable.

The primary limitation of the omnibus testing procedure proposed herein is that a union-intersection test with non-exclusive mechanisms to reject the null hypothesis (i.e., more than one of $T$ through $W$ could call for rejection simultaneously) will tend to be conservative. Even so, the simulation results suggest that the omnibus testing procedure may exhibit good power in many situations with unilateral contamination or asymmetric bilateral contamination. Symmetric bilateral contamination appears considerably more difficult to detect, presumably because such contamination is not easily distinguished from a larger value of the nuisance parameter under the omnibus null hypothesis. A secondary limitation is that the data analyst must specify $a_i$ through $a_m$. However, a “default” choice of $a_i=a_2=a_m=1/a_4$ may work reasonably well, if not optimally, in many situations.

The primary weakness of the unilateral testing procedure is its sensitivity to model misspecification. If the data originated from a bilaterally contained and scaled $T$ model with nuisance parameter on low degrees of freedom and were transformed so that the bilaterally contaminated normal model with nuisance parameter could be applied, the Type I error rates may be surprisingly high. Of course, this can be corrected by adjusting $\delta$, but we have not discovered a mechanism for adjusting $\delta$ under model misspecification, other than by trial and error. Indeed, a secondary weakness of the unilateral testing procedure is that the data analyst must specify $\delta$. However, if the model has been correctly specified, $\delta$ is interpretable as a high-probability bound between the true and estimated values of the nuisance parameter, and so choosing $\delta$ may not pose undue difficulty. The simulation results herein may also provide some guidance.

Future research should attempt to address the above issues, and one possibility may be a likelihood-based inferential framework. While ordinary likelihood ratio testing may not be helpful, because tractable asymptotic null distributions are not anticipated, an extension of the EM-test to the bilaterally contaminated normal (or scaled $T$) model with nuisance parameter may be viable, since the EM-test has previously been helpful in addressing null hypotheses that posit more than one component. Furthermore, methodology is needed that allows for differences in within-component variability, in effect changing the nuisance parameter into the second part of a component-specific vector characterizing that component probability distribution. Based on the work of Dai and Charnigo [7] as well as that of Chen et al. [5], we conjecture that a modified likelihood ratio test might have an asymptotic chi-square distribution under the omnibus null hypothesis in a bilaterally contaminated normal model with component-specific variances. An extension of the EM-test might be helpful to address the unilateral null hypothesis in such a scenario.

References