P1, 2, 3, 8, 10, 12, 16, 17, 19, 26, 29, 38
22

Asymptotic distribution theory for contamination models

Francisco Vera,*† David Dickey* and James Lynch*‡

April 19, 2010

Abstract

In many situations one is interested in identifying observations that come from sources of variation other than the normal background or baseline source. A simple model for such situations is a two point mixture model where one component in the mixture corresponds to the baseline model and the second to the other sources (the contamination component). Here the goal is two-fold: (i) detect the overall presence of Contamination c and (ii) identify observations that may be contaminated. A locally most powerful test is presented which gives some insights on how to accomplish this. Surprisingly, the test statistic can have an asymptotic distribution that is based on a stable law that is not the normal distribution. Examples and simulations are given to illustrate the approach.

Lower Case 'c'

Keywords: multiple testing, anomaly detection, stable law, false discovery rate

AMS 2000 Subject Classification: Primary 62F03, 62-07

Secondary 62E17, 62F30

*Research partially supported by NSF Grants DMS-0112069

†Research partially supported by NSF Grants EIA-0131884

‡Research partially supported by NSF Grants DMS-0805809 and DMS-0213594

1 Introduction

The point of this paper is to investigate the asymptotic distribution of the maximum likelihood estimator (MLE) and the locally most powerful (LMP) test of the parameter p in the two point mixture model

$$f_p(x) = \bar{p} f_0(x) + p f_1(x), \text{ where } \bar{p} = 1 - p, \text{ and } 0 \leqslant p \leqslant 1 \tag{1.1}$$

The distribution, f_p, is the so called *contaminated distribution model* which is sometimes used to model outliers or anomalies from the baseline model f_0.

~~anomalies~~

In this simple setting we shall see that, when $p = 0$, the MLE and the LMP test have asymptotic distributions that are nonstandard. They exhibit the Chernoff phenomena (Chernoff, 1954) of being two point mixtures. These two points mixtures each have point masses at zero where the second component in the mixture is based on an α-stable law depending on the tail behavior of the likelihood ratio $f_1(X)/f_0(X)$ under f_0.

In low contaminated situations(space p small), these asymptotics suggest using the LMP test to detect the presence of contamination. If the LMP test rejects $p = 0$, then ~~we use~~ the empirical posterior $\frac{p^* f_1(x)}{f_{p^*}(x)}$, where p^* is the mle, can be used to investigate what observations may be contaminated (from f_1). Confidence bounds for this posterior can also be constructed using confidence intervals for p^*.

The asymptotics indicate that the determination of contamination when p is small can be problematic using classical frequentist approaches, especially if parameters need to be estimated. In addition, this has similar implications for multiple testing problems. E.g., in the analysis of microarrays, a mixture model f_0 is the model for the expression levels of the nonexpressed genes and f_1 for the differentially expressed genes. In particular, there can be a justification for the use of a central t distribution where the degrees of freedom is determined by the amount of replication in the experiment or a central normal if the degrees of freedom is large. A similar justification can be ~~used to use~~ given for the use of noncentral t's or noncentral normals to model the differentially expressed genes. Here p is the proportion of differentianlly expressed genes.

The layout of the paper where we address the above ideas is as follows. In Section 2 we motivate a two point mixture data analytic model to analyze ~~data~~ pooled data. In Section 3 we discuss the LMP test for the mixture model where the asymptotic...

2 Pooling and mixtures

In many data analytic problems the observations $X_1, \ldots, X_n$ arise from pooling data from various sources of variation. In many cases, the pooling model has the following formulation for two sources of variation. In this formulation, a configuration C which is a subset of $\{1, 2, \ldots, n\}$ indicates which observations come from one source and C^c from the other. For example, such a pooling model might occur in a binary network where the network is modeled by a Markov random field. In the spread of an infectious disease over the network, the nodes are partitioned into two groups, C and C^c, where C^c is the collection of sites that have elevated levels of infections and C is the collection of sites which are normal. In the normal case the number of infections is governed by f_0 while for the elevated level by f_1. Then,

Commás →

$$p((C, C^c), X_1, \ldots, X_n) = K \exp\{E((C, C^c))\} \prod_{i \in C} f_0(X_i) \prod_{i \in C^c} f_1(X_i)$$

where $E((C, C^c))$ is related to the energy of the partition (C, C^c) (Huang, 1963) and where we have suppressed parameters in $E((C, C^c))$ and the normalizing constant K. Here we have assumed the positivity condition that all partitions have positive probabilities.

In general, the *pooling model* is given as follows.

- Generate a configuration C with probability p(C)
- Given C, for $i \in C$, X_i are iid $\sim f_0$ and, for $i \in C^c$, X_i are iid $\sim f_1$
  - C and C^c model a spatial or temporal (e.g., a change-point) pattern
  - You are "pooling" observations based on the configuration C where the configuration $C$is a hidden variable
  - The likelihood is then

Apend

$$\sum_{C} p(C) \prod_{i \in C} f_0(X_i) \prod_{i \in C^c} f_1(X_i)$$

Throughout we assume that all densities f are absolutely continuous with respect to a common measure m and absolutely continuous with respect to one another.

The *basic data analytic method* is as follows:

- Envision that the data are the effects of pooling observations from f_0 and f_1 where f_0 is the background distribution and f_1 is the distribution of the contaminated observations.
- Treat the data as if it is from a mixture model and use a mixture model to estimate the mixing proportions for f_0 and f_1, that is, the proportions in C and C^c. Use the estimates to test the null hypothesis that one of the mixing proportions is equal to zero. If this hypothesis is rejected, see if the fitted mixture model can give insights into which observations came from f_0, that is, into the configuration C

Formally, the basic data analytic model is the *simple contaminated model*

- $X_1, \ldots, X_n$ iid $\sim f_p = (1 - p)f_0 + pf_1$
  - f_0 is the density of the background mode
  - f_1 models the contamination
  - The likelihood is then

$$\prod_{i=1}^{n} \{(1-p) f_0(X_i) + p f_1(X_i)\}$$
$$= \sum_{j=0}^{n} \sum_{C_j} (1-p)^j p^{n-j} \prod_{i \in C_j} f_0(X_i) \prod_{i \in C_j^c} f_1(X_i)$$

where C_j denotes a subset of size j from $\{1, \ldots, n\}$

For low contaminated models one approach is to calculate the mle, p^*, of p. Use p^* to test $H_0 : p = 0$ versus $H_1 : p > 0$. If H_0 is rejected see if the mixture model can give insights into the configuration C_j. For example, calculate the empirical Bayes posterior with prior $p(C_j) = (1 - p^*)^j p^{*n-j}$. Then

$$p(C_j | X_1, \ldots, X_n) \propto (1 - p^*)^j p^{*n-j} \prod_{i \in C_j} f_0(X_i) \prod_{i \in C_j^c} f_1(X_i) \qquad (2.1)$$

Another approach is the following two stage multiple testing type of method for $p \approx 0$. This suggests using the *locally most powerful (LMP)* test statistic family with *mixing distribution* Q by

$$f_Q = \int f_\theta dQ(\theta)$$

For $X_1, \ldots, X_n$ being iid from f_Q, the likelihood and log likelihood are given by

$$L(Q) = \prod_i f_Q(X_i) \text{ and } \phi(f_Q) = \log \prod_i f_Q(X_i)$$

where $f_Q = (f_Q(X_1), \ldots, f_Q(X_n))$. The directional derivative of ϕ at f_{Q_0} towards f_{Q_1} is

$$\begin{aligned}\Phi(f_{Q_1}; f_{Q_0}) &= \lim_{\varepsilon \to 0}\Big(\varphi((1-\varepsilon)f_{Q_0} + \varepsilon f_{Q_1}) - \phi(f_{Q_0})\Big)/\varepsilon \\ &= \sum_{i=1}^{n} \frac{f_{Q_1}(X_i) - f_{Q_0}(X_i)}{f_{Q_0}(X_i)} = \sum_{i=1}^{n}\left\{\frac{f_{Q_1}(X_i)}{f_{Q_0}(X_i)} - 1\right\} \\ &= \int D(\theta; Q) dQ_1(\theta).\end{aligned}$$

The directional derivative D is used to identify when a *k-point MLE*, Q_k^*, for $L(Q)$ is the global mle Q^* (a k-point mle maximizes the likelihood function restricted to mixtures with k components). The basic idea is that $D(\theta; Q) = 0$ at the support points of the k-point MLE Q^* and $D(\theta; Q) \leqslant 0$ if and only if Q^* is the global MLE (Lindsay, 1983a,b).

4 Asymptotic considerations

In this section, we determine the asymptotic distributions of the MLE p^* of p and the LMP test statistic for testing $H_0 : p = 0$. When testing $H_0 : p = p_0$ and p_0 is in the interior of the parameter space, *i.e.*, $0 < p_0 < 1$, the usual asymptotics go through, since they are based on sums of bounded random variables. Therefore, we focus only in the case when testing $H_0 : p = 0$

Section 1.1 considers the case when the true value of the parameter $p = 0$. Since $p = 0$ is on the boundary, this leads to asymptotics under nonstandard conditions. In particular, the asymptotic distribution of the MLE p^* is a mixed distribution, where one of the components is degenerate at 0, and the other is

either half normal when the Fisher information $I_0 = E_0([(f_1 - f_0)/f_0]^2) < \infty$ or is a stable law when $I_0 = \infty$.

Section 4.2 considers the distribution of the LMP test statistic for testing $H_0 : p = 0$ when the true value of the parameter $0 < p < 1$. The results therein can be used for power calculations.

Section 1.3 gives the distributional properties of the ratio of two densities for the cases used in the examples and simulations.

one word

throughout

Need an "r".

Thoughout this section, let $X_1, \ldots, X_n$ be iid with density $f_p(x) = (1 - p)f_0(x) + pf_1(x)$ where all the random variables are assumed to be defined on the same probability space. Also let $Z_i = f_1(X_i)/f_0(X_i)$ and $L_i = Z_i - 1 = \frac{f_1(X_i) - f_0(X_i)}{f_0(X_i)}$. The LMP test statistic from Section 3 corresponding to the null hypothesis $H_0 : p = 0$ is denoted by $T_n = \sum_{i=1}^{n} L_i$. Let $I_0 = E_0(L_i^2)$ and $W_i = E_0(|L_i|^3), i = 0, 1$, where the expectations are taken under H_0. Note that I_0 is the Fisher information under H_0.

Also, though out this section, G_α represents the cumulative distribution function of a stable law with parameter $\alpha \in (0, 2]$, *i.e.*, its characteristic function is (A.1). Define $\bar{G}_\alpha = 1 - G$.

The next proposition is used in some parts of this section and is the basis for the claim that when p_0 is in the interior of the parameter space the terms in the LMP test statistic are all bounded.

Proposition 4.1.

$$\frac{f_1(x) - f_0(x)}{(1-p)f_0(x) + pf_1(x)}$$

is bounded for $0 < p < 1$ *(hence all its moments are finite).*

Proof. Notice that

$$\frac{f_1(x) - f_0(x)}{(1-p)f_0(x) + pf_1(x)} = \frac{1}{p}\left(\frac{f_1(x)}{\frac{1-p}{p}f_0(x) + f_1(x)}\right) - \frac{1}{1-p}\left(\frac{f_0(x)}{f_0(x) + \frac{p}{1-p}f_1(x)}\right)$$

Therefore,

$$\left|\frac{f_1(x) - f_0(x)}{(1-p)f_0(x) + pf_1(x)}\right| \leqslant \frac{1}{p} + \frac{1}{1-p}$$

□

4.1 First case: $p = 0$

The next few lemmas show the distribution of the MLE p^* when $p = 0$ under different conditions.

Lemma 4.2. *Under $H_0 : p = 0$, p^* converges to 0 almost surely.*

Proof. Let

$$l(p) \equiv \phi(f_p)$$

$$l'(p) = \frac{\partial}{\partial p}\phi(f_p) = \frac{\partial}{\partial p}\log\prod_{i=1}^{n} f_p(X_i) = \sum_{i=1}^{n}\frac{f_1(X_i) - f_0(X_i)}{f_p(X_i)}$$

and note that

$$l''(p) = -\sum_{i=1}^{n}\frac{[f_1(X_i) - f_0(X_i)]^2}{f_p(X_i)^2} \leqslant 0.$$

So $l(p)$ is concave and attains its maximum, p^*, either at 0 or 1 or on $(0, 1)$.

Let $U_n(p) = l'(p)$ where $U_n = U_n(0)$ and note that $U_n(p)$ is the sum of n iid random variables with

$$E_0\left(\frac{f_1(X_i) - f_0(X_i)}{f_p(X_i)}\right) = 0 \text{ when } p = 0 \text{ and } < 0 \text{ for } p > 0 \qquad (4.1)$$

When $U_n \leqslant 0$, $U_n(p) \leqslant U_n$ since $l(p)$ is concave. Thus, $l(p)$ attains it maximum at 0 on $\{U_n \leqslant 0\}$.

When $U_n > 0$, $l(p)$ attains its maximum on $(0, 1]$. Since $U_n(p)$ has mean less than 0 for $0 < p < 1$, $U_n(p)/n$ converges almost surely to a negative number (because of Proposition 1.1). When $U_n(p) < 0$ and $U_n > 0$, $0 < p^* < p$ with $U_n(p^*) = 0$ since $U_n(0) > 0$ and $U_n(p) < 0$. Thus, $\lim p* < p$ almost surely on the set where $U_n(p)/n$ converges to its mean. Since $p > 0$ is arbitrary, this with the previous paragraph implies that p^* converges to zero almost surely. □

Lemma 4.3. *If $I_0 < \infty$ and $W_i < \infty, i = 0, 1$, then, under H_0, $n^{-.5}p^*$ converges in distribution to X where $X = 0$ with probability .5 and $= |N(0, I_0^{-1})|$ with probability .5.*

Proof. If $p^* \in (0, 1)$, then $l'(p^*) = 0$ and

$$l'(0) = l'(0) - l'(p^*) = -l''(0)(p^*) - \frac{l'''(p')(p^*)^2}{2} \qquad (4.2)$$

where p' is between 0 and p^* and

$$l'''(p) = 2\sum_{i=1}^{n}\frac{[f_1(X_i) - f_0(X_i)]^3}{f_p(X_i)^3}.$$

Note that since the derivative of $l'''(p)$ is nonpositive, $l'''(p)$ is decreasing and $l'''(1) \leqslant l'''(p) \leqslant l'''(0)$. Thus, since $W_i < \infty$ for $i = 0, 1$, the sequence $l'''(p)/n, n = 1, 2, \ldots$ is bounded almost surely. It follows from (4.2) that when $U_n > 0$ and $U_n(1) < 0$, $p^* \in (0, 1)$ and

$$\begin{aligned}\frac{l'(0)}{\sqrt{n}} &= \frac{-l''(0)}{n}(\sqrt{n}p^*) - \frac{l'''(p')\sqrt{n}(p^*)^2}{2n} \\ &= \frac{-l''(0)}{n}(\sqrt{n}p^*)(1 + R_n).\end{aligned} \qquad (4.3)$$

Note that ~~since~~ R_n goes to zero almost surely since p^* converges to zero almost surely, the sequence $l'''(p)/n, n = 1, 2, \ldots$ is bounded almost surely and $-l''(0)/n$ converges almost surely to I_0.

When $U_n \leqslant 0$, $p^* = 0$. Since $U_n/\sqrt{n}$ is asymptotically $N(0, I_0)$ and $U_n(1)/n$ converges almost surely to a negative number by (4.2), $P(U_n \leqslant 0)$ and $P(U_n > 0$ and $U_n(p) < 0)$ both converge to $1/2$. The second part of this lemma follows from this and from (4.3) since $-l''(0)/n$ converges almost surely to $E_0([(f_1 - f_0)/f_0]^2) = I_0$ and $-l'(0)/\sqrt{n}$ converges in distribution to $N(0, I_0)$. □

For the next lemma, let a_n denote a sequence of real numbers and let

$$V_n(p) = \frac{1}{a_n^2}\sum\frac{(f_1(X_i) - f_0(X_i))^2}{f_0(X_i)f_p(X_i)}.$$

Lemma 4.4. *If Z_1 satisfies (A.2) for some $1 < \alpha \leqslant 2$ (i.e., is in the domain of attractation of an α-stable law) and a_n satisfies (A.3), then, under H_0, $a_n p^* V_n(p^*)$ converges in distribution to X where $X = 0$ with probability $G_\alpha(0)$ and $P(X > d) = \bar{G}_\alpha(d)$ for $d > 0$.*

Proof. For $p^* \in (0,1)$, $l'(p^*) = 0$ and

$$\begin{aligned} l'(0) &= l'(0) - l'(p^*) \\ &= \sum_{i=1}^{n} (f_1(X_i) - f_0(X_i)) \left(\frac{1}{f_0(X_i)} - \frac{1}{f_{p^*}(X_i)} \right) \\ &= p^* \sum \frac{(f_1(X_i) - f_0(X_i))^2}{f_0(X_i) f_{p^*}(X_i)} \\ &= p^* a_n^2 V_n(p^*). \end{aligned} \tag{4.4}$$

Note that

$$\frac{l'(0)}{a_n} = a_n p^* V_n(p^*)$$

and $l'(0)/a_n$ converges in distribution to G_α. Since $p^* = 0$ when $l'(0) < 0$, the results follows. □

From the proof of Lemma 4.4, by setting $a_n = \sqrt{n}$, we can get the asymptotic distribution of p^* without the third moment assumption given in Lemma 4.3.

Lemma 4.5. *If $I_0 < \infty$, then, under H_0, $\alpha = 2$ and $p^* \sqrt{n} V_n(p^*)$ converges in distribution to X where $X = 0$ with probability .5 and $= |N(0, I_0^{-1})|$ with probability .5. Moreover, $V_n(0)$ converges almost surely to I_0*

Remark 4.6. When $1 < \alpha < 2$, $V_n(0)$ converges in distribution to a stable law with parameter $\alpha/2$ (by Corollary A.3). So, one could replace $V_n(p^*)$ with $V_n(0)$ in (4.4), except that one could not justify this replacement without putting some condition on $l'''(p)$

The next few lemmas show the distribution of the LMP test statistic T_n for various cases.

Lemma 4.7. *If $I_0 < \infty$, then, under H_0, $T_n/\sqrt{n}$ converges in distribution to $N(0, I_0)$.*

Proof. Direct application of the central limit theorem. □

Lemma 4.8. *If Z_1 and a_n satisfy conditions (A.2) and (A.3), respectively, for some $1 < \alpha \leqslant 2$, then, under H_0, T_n/a_n converges in distribution to G_α (a stable law with parameter α).*

what the asymptotic behavior of T_n is.

Proof. Direct application of Lemma A.1 □

If f_1 has an unknown parameter and $p = 0$, an identifiability issue surfaces that makes it impossible to estimate that parameter. In this case, if the parameter is estimated from the data and used to calculate T_n, it is not clear what ~~is~~ the ~~limiting distribution~~ asymptotic behavior of T_n is,

4.2 Case 2: $p > 0$

The asymptotic distribution of T_n given in Lemmas 4.7 and 4.8 is for $p = 0$. The next two lemmas give the asymptotic distribution of T_n when $p > 0$. For this, assume that $X_i \sim f_p = (1 - p)f_0 + pf_1$ and let $W_0' = E_0(L_1^3)$

Lemma 4.9. *If $I_0 < \infty$ and $W_0' < \infty$, then $(T_n - npI_0)/\sqrt{n}$ converges in distribution to $N(0, I_0 + pW_0' - p^2 I_0^2)$.*

Proof. It is easy to prove that $E_p(L_1) = pI_0$ and $E_p(L_1^2) = I_0 + pW_0'$. The result then follows from a direct application of the central limit theorem. □

Lemma 4.10. *Suppose Z_1 and a_n satisfy conditions (A.2) and (A.3), respectively, for some $1 < \alpha \leqslant 2$. If $1 < \alpha \leqslant 2$, then $I_0 < \infty$ and $(T_n - npI_0)/a_n$ converges in distribution to G_α, while if $0 < \alpha < 1$, T_n/a_n converges in distribution to a stable law with parameter α. If $\alpha = 1$, then $(T_n - \mu_n)/a_n$ converges in distribution to a stable law with parameter 1, where μ_n is defined as in (A.4).*

Proof. If $1 < \alpha \leqslant 2$ then $E_p(L_1) = pI_0$ is finite and the result follows from the results in Appendix A. □

4.3 Distributional properties of density ratios

In this section, we consider the properties of $Z = f_1(X)/f_0(X)$ for some frequently used distributions. ~~This~~ These properties are required to use the lemmas in Sections 4.1 and 4.2. Section 4.3.1 considers the case when both f_0 and f_1 are exponential distributions, and Section 4.3.2 considers the case of the normal distribution.

5 can be used to get an appropriate normalizing constant a_n which satisfies (A.3).

The mean of the density ratio is

$$E_{\mu,\sigma^2}(Z) = \begin{cases} e^{\left(\frac{1}{2}\left[\sigma^2\delta_1^2 + 2\mu\delta_1 - \delta_2\right]\right)} & \text{when } \sigma_1^2 = \sigma_0^2 \\ \sqrt{\frac{\alpha^2}{\alpha-1}}\, e^{\left(\frac{1}{2\sigma^2}\left[\frac{\alpha}{\alpha-1}(\sigma^2\delta_1+\mu)^2 - \mu^2 - \sigma^2\delta_2\right]\right)} & \text{when } \sigma_1^2 < \sigma_0^2 \text{ or} \\ & (\sigma_1^2 > \sigma_0^2 \text{ and } \alpha > 1) \\ \infty & \text{when } \sigma_1^2 > \sigma_0^2 \text{ and } \alpha \leqslant 1 \end{cases}$$

where $\delta_j = \frac{\mu_1^j}{\sigma_1^2} - \frac{\mu_0^j}{\sigma_0^2}$ for $j = 1, 2$ (for $j = 0$ let $\delta_0 = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}$).

The variance is given by

$$\mathrm{var}_{\mu,\sigma^2}(Z) = \begin{cases} \left(e^{\sigma^2\delta_1^2} - 1\right) e^{\left(\sigma^2\delta_1^2 + 2\mu\delta_1 - \delta_2\right)} & \text{when } \sigma_1^2 = \sigma_0^2 \\ \sqrt{\frac{\alpha\sigma^2}{\alpha-2}}\, e^{\left(\frac{1}{2\sigma^2}\left[\frac{1}{\alpha-2}(2\sigma^2\delta_1+\mu)^2 - \mu^2\right] - \delta_2\right)} & \text{when } \sigma_1^2 < \sigma_0^2 \text{ or} \\ -\frac{\alpha\sigma^2}{\alpha-1}\, e^{\left(\frac{1}{\sigma^2}\left[\frac{\alpha}{\alpha-1}(\sigma^2\delta_1+\mu)^2 - \mu^2\right] - \delta_2\right)} & (\sigma_1^2 > \sigma_0^2 \text{ and } \alpha > 2) \\ \infty & \text{when } \sigma_1^2 > \sigma_0^2 \text{ and } 1 < \alpha \leqslant 2 \\ \text{undefined} & \text{when } \sigma_1^2 > \sigma_0^2 \text{ and } \alpha \leqslant 1 \end{cases}$$

When $f_{\mu,\sigma^2} = f_0$ the mean reduces to $E_0(Z) = 1$ and the variance, which is I_0, reduces to

$$I_0 = \mathrm{var}_0(Z) = \begin{cases} e^{\left(|\mu_1-\mu_0|^2/\sigma_0^2\right)} - 1 & \text{when } \sigma_1^2 = \sigma_0^2 \\ \frac{\sigma_0^2}{\sigma_1\sqrt{2\sigma_0^2 - \sigma_1^2}}\, e^{\left(\frac{|\mu_1-\mu_0|^2}{2\sigma_0^2-\sigma_1^2}\right)} - 1 & \text{when } \sigma_1^2 < 2\sigma_0^2 \\ \infty & \text{when } \sigma_1^2 \geqslant 2\sigma_0^2 \end{cases}$$

For $f_{\mu,\sigma^2} = f_1$ the mean becomes $E_1(Z) = \mathrm{var}_0(Z) + 1$ and the variance becomes

$$\mathrm{var}_1(Z) = \begin{cases} \left(e^{\left(|\mu_1-\mu_0|^2/\sigma_1^2\right)} - 1\right) e^{\left(2|\mu_1-\mu_0|^2/\sigma_1^2\right)} & \text{when } \sigma_1^2 = \sigma_0^2 \\ \frac{\sigma_0^3}{\sigma_1^2\sqrt{2(1.5\sigma_0^2-\sigma_1^2)}}\, e^{\left(1.5\frac{|\mu_1-\mu_0|^2}{1.5\sigma_0^2-\sigma_1^2}\right)} & \\ -\frac{\sigma_0^4}{\sigma_1^2(2\sigma_0^2-\sigma_1^2)}\, e^{\left(2\frac{|\mu_1-\mu_0|^2}{2\sigma_0^2-\sigma_1^2}\right)} & \text{when } \sigma_1^2 < 1.5\sigma_0^2 \\ \infty & \text{when } 1.5\sigma_0^2 \leqslant \sigma_1^2 < 2\sigma_0^2 \\ \text{undefined} & \text{when } \sigma_1^2 \geqslant 2\sigma_0^2 \end{cases}$$

Intro paragraph describing what you are doing.

5 Simulations

For the first set of simulations, background and contamination data are generated from exponential distributions with means $\theta_0 = 166.206$ and $\theta_1 = 592.922$, respectively, which are the estimated means for the 2-point rule for the Mining Data in the next section. Samples of sizes $n = 100, 500, 1000$ are generated, with 0, 1 and 5 percent of contamination ($p = 0, 1.01, .05$). With each sample we calculate $T_n = \sum_{i=1}^{n} \frac{f_1(X_i) - f_0(X_i)}{f_0(X_i)}$ and $S_n = \sum_{i=1}^{n} \frac{\hat{f}_1(X_i) - \hat{f}_0(X_i)}{\hat{f}_0(X_i)}$, where $\hat{f}_0$ and $\hat{f}_1$ are the densities based on the maximum likelihood estimators of θ_0 and θ_1. These estimates are used as if they were the true parameters and a normalizing constant and critical value are calculated based on these estimates. The process is repeated $N = 10000$ times and the number of rejections of the null hypothesis $H_0 : p = 0$ at the .05 level are recorded.

Following the results from Section 4.3.1, the variance of the terms in T_n corresponding to f_0 is infinite, but the tail behavior of the density ratio, under H_0, follows that of the first line of Table 5 with $\alpha = 592.922/(592.922 - 166.206) = 1.3895$ and $c = (166.206/592.922)^{1.3895} = 0.1708081$. Hence, the normalizing constant is $a_n = (0.1708081 s_{1.3895} n)^{1/1.3895} = .4881128n^{0.7196832}$. From Lemma A.1, the rejection region defined by $T_n/(.4881128n^{0.7196832}) > 4.40186$ would reject the null hypothesis with probability 0.05 if there are no anomalies. A similar process is done to calculate the rejection region $S_n/a_n > d_{.05}$ in each sample, where the normalizing constant a_n and the critical value $d_{.05}$ change from sample to sample, and are calculated based on either the normal or the stable distribution[1]. The results of these simulations are shown in Table 1.

~~was~~ Commas

The simulations show that, when using the true parameters to calculate T_n, the proportion of samples that rejected the null hypothesis under the null about 0.05, as expected. Notice that with the background and contamination means fixed at $\theta_0 = 166.206$ and $\theta_1 = 592.922$, the power increases as the

[1] if $\theta_1 < 2\theta_0$, and these are assumed to be the actual parameters, the variance of the terms in S_n is finite

	Proportion of anomalies					
	Based on T_n			Based on S_n		
Sample sizes	0	0.01	0.05	0	0.01	0.05
100	0.0492	0.1646	0.5316	0.2853	0.3898	0.6340
500	0.0483	0.3836	0.9361	0.3248	0.5725	0.9252
1000	0.0544	0.5454	0.9935	0.3387	0.6898	0.9786

Table 1: Proportion of rejections of H_0, no anomalies, out of 10000 simulations with background and contamination generated from exponential distributions with means 166.206 and 592.922, respectively

proportion of anomalies, p, increases and as the sample size increases.

The exact asymptotic power for the LMP test can be calculated using Lemma 4.10 and the results from Section 4.3.1 and Table 5, with $\alpha = \theta_0/(\theta_1 - \theta_0) = 0.3895003$ and $c = p(\theta_0/\theta_1)^\alpha = 0.6093393p$, which gives $(cs_\alpha)^{1/\alpha} = 0.1173926p^{2.567392}$. Rejection occurs when $T_n/(.4881128n^{0.7196832}) > 4.40186$, which is equivalent to rejecting if

doesn it agree with α in previous paragraph

$$T_n(n,p) \equiv T_n/(0.4473926p^{2.567392}n^{2.567392}) > 4.802503/(p^{2.567392}n^{1.847705}) \equiv C(p,n)$$

The left hand side of this inequality converges to a stable law with parameter $\alpha = 0.3895003$. The power can be obtained for each value of n and p. ~~For instance, if $n = 100$ and $p = 0.05$, the power is the probability that a value from a stable law is larger than 2.119844, that is, 0.510141. In the case of $n = 500$ and $p = 0.01$, the tail starts at 6.750576, which gives a power of~~ 0.3517792. For $n = 1000$ and $p = 0.05$, the power is 0.996379. The simulations confirm these values.

$C(p,n)$ ~~= 2.119844~~

Confusing

Keep

If estimates are used as if they were the true parameters, rejection occurs 28.5% of the time when there are no anomalies present ($p = 0$) and $n = 100$. This is troubling and indicates that false discovery is a serious problem in this case. This is not the case when $p > 0$ and Appendix B indicates the necessary adjustments that need to be made when estimating parameters.

~~For instance, for~~ the

~~The~~ ~~for~~ With this notation the power is $P(T(n,p) > C(n,p)) \equiv P(n,p)$.

For the second set of simulations data are generated from normal distributions, where the background consists of standard normal variables ($\mu_0 = 0$ and $\sigma_0^2 = 1$) and the anomalies consist of a normal with mean $\mu_1 = 0$ and variance $\sigma_1^2 = 3$. Samples of sizes $n = 100, 500, 1000$ are generated, with 0, 1 and 5 percent of the observations being anomalies.

The results from Section 4.3.2 and Table 5 are used to determine the rejection region for each sample. Since the variance of f_1/f_0 is infinite under f_0, the tail behavior of the distribution of the ratio needs to be taken into consideration to see what stable law applies when $\mu = \mu_0 = 0$ and $\sigma^2 = \sigma_0^2 = 1$. This is described in Section 4.3.2 with $\alpha = 3/(3-1) = 1.5$, $\beta = 1/3$, $b = 0$, $c_1 = 2\beta^{\alpha/2}/\sqrt{2\pi} = 2\left(\frac{1}{3}\right)^{1.5/2}/\sqrt{2\pi} = 0.350025$. The rejection region is now found using the results from Table 5: reject the null hypothesis $H_0 : p = 0$ if

$$\frac{T_n}{0.7274158(n/\sqrt{\log n})^{2/3}} > 3.824235.$$

The results of these simulations are found in Table 2. The LMP test seems somewhat conservative for $n = 100$, possibly because this sample size is too small to observe convergence to the stable law. This resulted in a not too powerful test. For $n = 500$ and $n = 1000$, the test seems to perform better.

	Proportion of anomalies		
Sample sizes	0	0.01	0.05
100	0.0371	0.0788	0.2331
500	0.0447	0.1533	0.5649
1000	0.0408	0.2295	0.7885

Table 2: Proportion of rejections of H_0, no anomalies, out of 10000 simulations with background and anomalies generated from normal distributions with mean 0 and variances 1 and 3, respectively

The exact asymptotic power can be calculated for these tests by using Lemma 4.10, Table 5 and Section 4.3.2. Suppose the true proportion of anomalies p is positive ($p > 0$) and let $\alpha = 1/(3-1) = 0.5$ and $c = p2\left(\frac{1}{3}\right)^{0.5/2}/\sqrt{2\pi} =$

Below is $P(n,p)$ tabulated for $n=100$, 500, 1000 for $p = .01$ and $p = .05$

Table.

Put this in a table.

#1 (n,p) Exact Simulation

(100, .05)

$0.6062612p$. Thus, to get a stable law we need to normalize T_n by $0.2886751p^2(n/\sqrt{\log n})^2$. A rearrangement of the rejection region (which was normalized originally by $0.7274158(n/\sqrt{\log n})^{2/3}$) would result in rejections when

$T_n(n,p)$

$$\frac{T_n}{0.2886751p^2(n/\sqrt{\log n})^2} > 3.824235\frac{0.7274158(n/\sqrt{\log n})^{2/3}}{0.2886751p^2(n/\sqrt{\log n})^2} = \frac{9.636468(\log n)^{2/3}}{p^2n^{4/3}} \geq C(n,p)$$

Rewrite like paragraph on page 17

The exact asymptotic power when $n = 100$ and $p = .05$ is the probability that a stable law variable with parameter 0.5 is greater than $\frac{9.636468(\log 100)^{2/3}}{.05^2 100^{4/3}} = 22.98661$. This probability is 0.1652201. Similarly, for $n = 500$ and proportions $p = .01$ and $p = .05$, the probabilities of rejecting the null hypothesis are 0.08789053 and 0.4189768, respectively, whereas the simulations estimated these numbers as 0.1533 and 0.5619, respectively.

6 Data Examples

Intro

In this section we analyze two data sets ~~Therefore~~ using the ideas developed here. The first is mining accident data from Maguire, ... ← describe the data set. while the second is an analysis of microarray data.

Following the analysis from Grego et al. (1990) of the mining accident data, Figure 1 has the gradient functions for the 2 and 3-point mixture mle's where the mixing is over the mean of an exponential distribution and Figure 2 has the assignment function for the second component in the 3-point mle. The estimates of the means and mixing proportions are given in Table 3. The gradient plot indicates that the 2-point mle is not the global mle but the 3-point is. The assignment function indicates a distinct difference in the first 53 times and rest of the times. Further analysis by Grego et al indicates that the first 53 are well fit by a single exponential and the rest by a 3- point mixture.

	$\mu_1(p_1)$	$\mu_2(p_2)$	$\mu_3(p_3)$
2-point mle	592.922 (.175149)	166.206 (.824851)	
3-point mle	595.495 (.171379)	171.587 (.805528)	29.0972 (.023093)

Table 3: Maximum likelihood estimates for the mining data

For the mining data we will use an exponential with mean 171.587 as f_0 and a 2-point mixed exponential with means 595.495 and 29.0972 and mixing

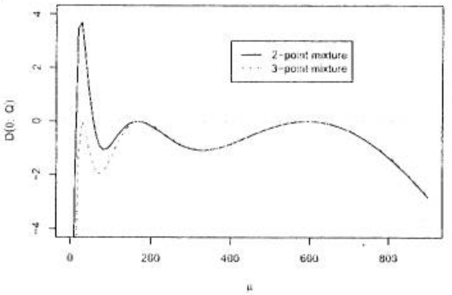

Figure 1: Gradient plots of a 2- and 3-point mixtures (mle) of exponentials for the Mining Data

proportions proportional to .171379 and .023093, respectively, as f_1. That is, $f_1 = f_{Q_1}$ where Q_1 has point masses at 595.495 and 29.0972 with mixing proportions .881253 and .118747 and the family, $\{f_\mu\}$, being mixed over is the exponential with its mean parameterization. These are assumed as the true parameters and Lemma 1.8 along with Table 5 can be used to calculate critical values for the LMP test statistic.

The LMP test statistic, T_n, assuming all the parameters are known, is then given by

$$T_n = 0.8812528\sum_{i=1}^{n}\left(\frac{\frac{1}{595.495}e^{-X_i/595.495}}{\frac{1}{171.587}e^{-X_i/171.587}} - 1\right)+0.1187472\sum_{i=1}^{n}\left(\frac{\frac{1}{29.0972}e^{-X_i/29.0972}}{\frac{1}{171.587}e^{-X_i/171.587}} - 1\right).$$

Under the null hypothesis, i.e., $X_i \sim f_0$, the terms in the first sum have infinite variance whereas the terms in the second sum have finite variance (see Appendix 1.3.1 for details). Using the notation of Appendix 1.3.1 with $\theta_0 = 171.587$ and $\theta_1 = 595.495$, let $\alpha = 595.495/(595.495 - 171.587) = 1.401774$, $c = (171.587/595.498)^{1.401774} = 0.1741281$ and $a_n = 0.5052435n^{0.7118582}$. If T_n is normalized by a_n, the second sum will quickly converge to zero as $n \to \infty$. The

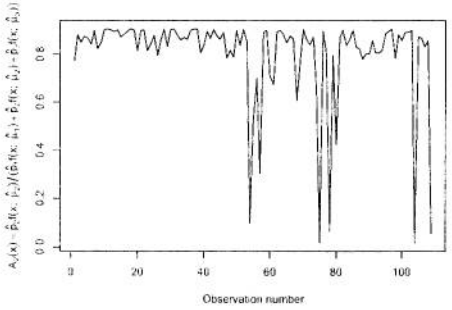

Figure 2: Assignment function for the second component of the 3-point mixture (mle) of exponentials for the Mining Data

first sum converges in distribution to a stable law with parameter $\alpha = 1.404774$. Therefore $T_n/(.8812528a_n)$ converges in distribution to the same stable law.

For the mining data, $T_n = 574.871$ and $T_n/(.8812528a_n) = 45.77311$. Using Table 1 for $\alpha = 1.4$ we can see that the p-value is between .005 and .001. The actual p-value is 0.002102145 (calculated using a computer and $\alpha = 1.404771$). This indicates that there is strong evidence that some observations come from f_1.

Note that parameters in both f_0 and f_1 are being estimated based on the 3-pt global mle. These estimates have to be taken into consideration in using the LMP test statistic to determine if spurious observations are present. As pointed out in the appendix, it would be impossible to estimate f_1 if $p = 0$, and hence the distribution of the LMP is not quite clear in this case.

If $p > 0$, then we only need to check the regularity conditions discussed in Lemma B.1 and Remark B.2. For the exponential distribution these conditions reduce to the finiteness of the first three moments.

We now illustrate some of these ideas using gene expression data. The

approach here will be first use the assignment function to identify possible anomalies (expressed genes) to get a pooled model. After that, we do the LMP test.

Efron (2007) compared prostate data of $m_1 = 50$ non-tumor subjects with $m_2 = 52$ tumor patients for each of $n = 6033$ genes (see Singh et al., 2002). For each gene they perform a two-sample t-test to compare the mean gene-expression between cancer and noncancer subjects.

Let t_i for $1 = 1, \ldots, n$ denote the test statistics used for each gene. For genes that have the same mean expression values for both groups t_i will follow a central t-distribution with $m_1 + m_2 - 2 = 100$ degrees of freedom.

Efron defines $z_i = \varphi^{-1}(F_{100}(t_i))$, where F_ν denotes the cumulative distribution function (cdf) of a t-distribution with ν degrees of freedom and φ denotes the cdf of a standard normal distribution. Then the distribution of z_i is standard normal for those genes that have the same mean expression for both groups of subjects.

Efron then fits the mixture $f = (1 - p)f_0 + pf_1$ as follows. Suppose f is a 7-parameter exponential family and estimate this density from the z-values, obtaining $\hat{f}$. Suppose f_0 is the standard normal density and estimate p by using $\log((1 - p)f_0)$ as a "quadratic approximation" of $\hat{f}$. From this he estimates the assignment function A_0 (false discovery rate). These calculations can be done using the R package `locfdr`. He discovered 51 genes using false discovery rate, declaring an anomaly when $A_0 < 0.2$.

the

Here we will first use the ratio $(f_1 - f_0)/f_0$ rather than A_0 to discover anomalies. To estimate f_1 output of `locfdr` is used. Then we plot $(f_1 - f_0)/f_0$ versus the z-values (Figure 3). It can be seen that the maximum of the ratio $(f_1 - f_0)/f_0$ is 17649.25. An observation is declared as an anomaly if $(f_1 - f_0)/f_0 > 23$, where the ad-hoc cutoff point 23 makes the average of $(f_1 - f_0)/f_0$ close to 0 for $(f_1 - f_0)/f_0 < 23$. This is reasonable since the mean of $(f_1 - f_0)/f_0$ is 0 under f_0. This same plot is shown in Figure 4 but with vertical axis scaled to show only those with $(f_1 - f_0)/f_0$ below 50. A total of 109 genes have

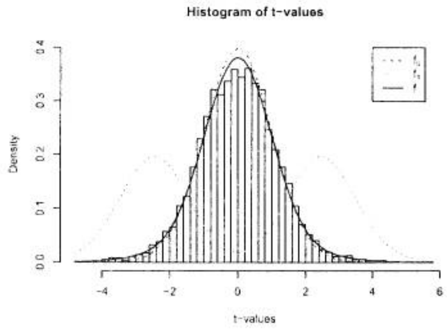

Figure 5: Histogram of t *values* corresponding to the prostate gene-expression data

Figure 5 shows a histogram of the t-values with the estimated densities superimposed. It can be seen that the central t-distribution f_0 (dashed line) does not fit the t-values very well since the t-values have a heavier tail. The mixture of the two non-central t-distributions with parameters δ and δ, f_1 (dotted line), help to explain the tails. When these two distributions f_0 and f_1 are mixed with p as the proportion for f_1, then the fitted distribution f (solid line) fits the histogram quite nicely using only two parameters (compared to fitting 8 parameters).

To do the LMP test, we need to explore the distribution of the density ratio f_1/f_0, and this is quite hard to do with the non-central t-distribution. To work around this, we suppose that the variance of the ratio is finite and just use the regular central limit theorem. A random sample of one million values from a central t-distribution with 100 degrees of freedom was generated and the ratio f_1/f_0 was calculated for each value, where the sample variance was 102.2010. This estimate of the variance of f_1/f_0 is assumed to be the true variance.

For the prostate data we get $T_n = \sum_i \frac{f_1(t_i)-f_0(t_i)}{f_0(t_i)} = 27186.5$. We assume the variance of f_1/f_0 is 102.2010. Then, if all observations are from f_0 and the the

observations were independent then $T_n/\sqrt{n\sigma^2} = 34.62255$. When compared to the quantiles of a standard normal distribution, this value indicates very strong evidence that some of the genes have different mean expression values for tumor and non-tumor patients.

A t-value is declared as coming from one of the non-central t-distributions if $(f_1 - f_0)/f_0 > 43$, where the ad-hoc cutoff point makes the mean of $(f_1 - f_0)/f_0$ close to zero for all genes with $(f_1 - f_0)/f_0 < 43$. With this cutoff point, 88 genes are declared as having different mean expression.

Since Efron (2007) found 51 anomalies, the LMP test ~~is~~ will be used to verify the hypothesis $H_0 : p = p_0 = 51/6033 = 0.00845$. In this case the test statistic is

$$T_n = \sum_{i=1}^{n} \frac{f_1 - f_0}{(1-p_0)f_0 + p_0 f_1} = 5081.583.$$

Since T_n is the sum of bounded r.v.'s (Proposition 4.1), the regular central limit theorem can be used to decide on a rejection region. The sample variance of $(f_1 - f_0)/((1-p_0)f_0 + p_0 f_1)$ is 10.67491 from the same sample of one million t-values as above. This gives us $T_n/\sqrt{n\sigma^2} = 20.02397$, which is quite significant when compared to quantiles of a standard normal distribution. Therefore, we conclude that the proportion of anomalies is greater than 0.00845.

Since the use of t-values gave 88 anomalies, we now repeat the exercise from the previous paragraph to test the hypothesis $H_0 : p = p_0 = 88/6033 = 0.01459$. From the same sample of one million t-values, the sample variance of $(f_1 - f_0)/((1-p_0)f_0 + p_0 f_1)$ is 7.16. The test statistic is $T_n = 3172.833$, which normalized gives $T_n/\sqrt{n\sigma^2} = 15.2649$. This indicates that the proportion of anomalies is greater than 0.01459.

A similar test to check if the proportion of anomalies is greater than 109/6033 = 0.01807 gives that $T_n/\sqrt{n\sigma^2} = 13.04648$. This suggests that a better cutoff for $(f_1 - f_0)/f_0$ point is needed to identify anomalies. This will be the subject of future research. “δ”

It is worthwhile to mention that the independence assumption between genes may not be realistic. With regard to the choice of $f_1 = .5g_\delta + .5g_{\ \delta}$,

?

	Right tail probabilities				
α	0.1	0.05	0.01	0.005	0.001
1	6.7612	13.6373	65.653	129.7645	NA
1.05	-6.3069	-0.5517	40.1627	87.9868	415.1734
1.1	-0.5213	4.355	36.8106	73.2907	330.9228
1.15	1.1151	5.2897	31.5672	59.932	250.189
1.2	1.7651	5.3706	26.929	49.3496	192.8381
1.25	2.0504	5.1871	23.0733	41.0467	151.2418
1.3	2.1718	4.9181	19.8967	34.4791	120.4426
1.35	2.2118	4.628	17.2687	29.2202	97.1814
1.4	2.2089	4.3431	15.0761	24.9537	79.2942
1.45	2.1832	4.071	13.2282	21.4463	65.3058
1.5	2.1457	3.8242	11.6541	18.5251	54.1916
1.55	2.1028	3.5946	10.2983	16.0605	45.2258
1.6	2.0584	3.3845	9.1171	13.9535	37.8839
1.65	2.0147	3.1931	8.0759	12.1273	31.7795
1.7	1.9733	3.0195	7.1466	10.5209	26.6205
1.75	1.9352	2.8631	6.3068	9.0842	22.179
1.8	1.9011	2.7231	5.5394	7.774	18.2669
1.85	1.8716	2.6002	4.8357	6.5525	14.7128
1.9	1.8468	2.4931	4.2054	5.3942	11.3245
1.95	1.827	2.402	3.6825	4.361	7.787
2	1.8124	2.3262	3.29	3.6428	4.3702

Table 4: Quantiles of the stable distribution

For the case when $P(Z_1 > z)$ is regularly varying with index −1, we define a_n as a sequence of real numbers such that $n\pi P(Z_1 > a_n)/2 \to 1$ as $n \to \infty$.

We also define

$$\mu_n = \int_{\min(a,0)}^{0} P(Z_1 \leqslant t)dt + \int_{\max(0,a)}^{a_n} P(Z_1 > t)dt. \qquad \text{(A.4)}$$

Lemma A.4. *For $\alpha = 1$,*

$$\frac{1}{a_n}\sum_{i=1}^{n}(Z_i - \mu_n)$$

converges in distribution to a stable law with parameter 1.

For the normal case, let $h(t) = \int_{\min(a,0)}^{0} zP(Z_1 \leqslant z)dz + \int_{\max(0,a)}^{t} zP(Z_1 > z)dz$ and let a_n be a sequence of real numbers such that $nh(a_n)/a_n^2 \to 1$ as $n \to \infty$. Note that if the variance of Z_1 is finite then $h(t) \to 2\,\mathrm{var}(Z_1)$ as $t \to \infty$, so $a_n = \sqrt{2n\,\mathrm{var}(Z_1)}$.

Lemma A.5. *If $h(t)$ is slowly varying, i.e., regularly varying of order 0, then $\mu = E(Z_1) < \infty$ and*

$$\frac{1}{a_n}\sum_{i=1}^{n}(Z_i - \mu)$$

converges in law to a normal distribution with mean 0 and variance 2.

If $h(t)$ is slowly varying, we say that Z_1 is in the domain of attraction of a 2-stable law (normal distribution with variance 2).

These lemmas above can be specialized to more specific situations. In particula, Table 5 shows several possible tail behaviors of the distribution of Z_1 which will satisty (A.2), along with their corresponding sequence a_n which satisfies (A.3).

B Asymptotics for the MLE of a k-point mixture

The next lemma states the asymptotic distribution for the k-point MLE of the mixing distribution Q when Q is a discrete probability measure on Θ with k distinct mass points, $\theta_1, \ldots, \theta_k$, and respective masses, $p_1, \ldots, p_k$. Here we assume that $\theta_1, \ldots, \theta_k$ are in the interior of Θ and that all the masses are positive.

where $\hat{Q}_n$ is the k-point mle. Since ρ is arbitrary, $\hat{Q}_n$ converges almost surely to Q_0.

Acknowledgements

This work was initiated when all three authors were involved in the Anomaly Detection Working Group at the Statistical and Applied Mathematical Sciences Institute (SAMSI). The first author was a postdoctoral fellow at SAMSI and the National Institute of Statistical Science (NISS). The authors would like to thank SAMSI and NISS for a stimulating environment and support while there. In particular, support from SAMSI for all three authors was funded by NSF under grant DMS-0112069 and support from NISS for the first author was funded under grant EIA0131884. In addition, the third author was partially supported by NSF Grants DMS 0243594 and DMS 0805809.

References

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics 25*(3), 573–578.

Efron, B. (2007). Size, power and false discovery rates. *Annals of Statistics 35*(4), 1351–1377.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretical Approach*. Academic Press, NY.

Geluk, J. L. and L. de Haan (2000). Stable probability distributions and their domains of attraction: a direct approach. *Probability and Mathematical Statistics 20*(1), 169–188.

Grego, J., H.-L. Hsi, and J. D. Lynch (1990). A strategy for analyzing mixed and pooled exponentials. *Applied Stochastic Models and Data Analysis 6*(1), 59–70.

Huang, K. (1963). *Statistical mechanics*. Wiley, NY.

Lehmann, E. L. (1983). *Theory of Point Estimation*. Probability and mathematical statistics. Wiley, NY.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: A general theory. *Annals of Statistics 11*(1), 86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, part ii: The exponential family. *Annals of Statistics 11*(3), 783–792.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell 1*, 203–209.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics 20*(4), 595–601.

Maguire Reference