



Tests and Diagnostics for Heterogeneity in the Species Problem

By CHANGXUAN MAO and BRUCE G. LINDSAY

Technical Report #01-09-10

2001

Center for Likelihood Studies

DEPARTMENT OF STATISTICS

THE PENNSYLVANIA STATE UNIVERSITY

UNIVERSITY PARK, PA 16802

Tests and diagnostics for heterogeneity in the species problem

Chang Xuan Mao and Bruce G. Lindsay

September 10, 2001

Chang Xuan Mao (Correspondence author)
Interdepartmental Group of Biostatistics
University of California, Berkeley
367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.
Email: cmao@stat.berkeley.edu

and

Bruce G. Lindsay
Department of Statistics
Pennsylvania State University
326 Thomas Building
University Park, PA 16802-2111, U.S.A.

Abstract

Suppose a random sample of individuals is drawn from a population with N disjoint classes. The population is said to be homogeneous when all the classes have the same abundance. Otherwise, it is said to be heterogeneous. The number of individuals from each class in the sample is assumed to be Poisson distributed. There is a vast literature towards the estimation of N . Although the performance of estimators is related to the homogeneity assumption, testing homogeneity has received little investigation. In this paper, we first discuss the χ^2 goodness-of-fit test. Next, a dispersion score test is presented and two graphic diagnostics are developed to detect the existence of heterogeneity. Two datasets from epidemiological and genomic studies are used to illustration of these tests and diagnostics.

Key words Number of species; Poisson mixture; Heterogeneity; Graphical diagnostics.

1 Introduction

Suppose there are N distinct classes in a population, and these classes are indexed by $1, 2, \dots, N$, where N is unknown. Let x_i be the number of individuals from the i th class present in the sample. If $x_i = 0$, then the i th class is undetected. Thus determining N is akin to determining the number of x_i 's that are zeros. The x_i 's are called *frequencies*. A common assumption is that x_i is from a Poisson distribution with mean parameter λ_i . The population is said to be *homogeneous* if all the λ_i 's are identical. Otherwise, the population is said to be *heterogeneous*. In the later case, the λ_i 's are often assumed to arise as a random sample from a latent distribution, which results in a mixture model, see Norris and Pollock (1998), Ord and Whitmore (1986) and Mao and Lindsay (2001b).

This is the famous species problem, which has applications in many different fields, such as astronomy, ecology, epidemiology, genomics, linguistics and numismatics; see Bunge and Fitzpatrick (1993) for a review. The present authors have been studying gene expression from the sequence data of cDNA libraries. Gene expression is highly differentiated and has been an important topic in biological research for a long time. We were motivated to address the species problem because it arises in our genomic studies.

Many estimators have been proposed for N , such as Chao (1984 and 1992), Darroch and Ratcliff (1980), Mao and Lindsay (2001b) and Zelterman (1988). Although homogeneity or heterogeneity is assumed explicitly or implicitly for all estimators, there has been little emphasis on methods for assessing the model fit.

Diagnosing homogeneity based on the complete data x_1, x_2, \dots, x_N , can be represented as a hypothesis testing problem: a simple Poisson model versus a Poisson mixture. However, the observed data exclude classes with zero frequency, so that the number of classes with zero frequency and the sample size N are both unknown. This feature prevents application of existing methods for Poisson models.

In Section 2, mixture modeling is introduced and it will be shown how the nuisance parameter N can be eliminated by using a conditional likelihood. In Section 3, a χ^2 goodness-of-fit test and a *dispersion score test* are presented. In Section 4 two graphic diagnostics are developed to assess the existence of heterogeneity. In Section 5, these diagnostics are used to study a cholera infection dataset and a tomato EST dataset.

2 The mixture model

Let $g(x; \lambda)$ be the Poisson density with respect to counting measure and Ω be the set of nonnegative integers, where λ is the mean parameter and $g(x; \lambda) = e^{-\lambda} \lambda^x / x!$, for x in Ω . Suppose X is a Poisson random variable given $\Lambda = \lambda$ and Λ itself is from a latent distribution $P(\lambda)$. The marginal distribution of X has a Poisson mixture density, denoted by $g(x; P)$ and written as $g(x; P) = \int g(x; \lambda) dP(\lambda)$, for x in Ω . There are two simple equalities about the mean and variance of X and the latent variable Λ as follows:

$$EX = E\Lambda \text{ and } Var(X) = Var(\Lambda) + E\Lambda.$$

Note that the variance-to-mean ratio of a Poisson mixture is no less than that of a Poisson distribution. This inflation of variance is called “over-dispersion”.

Let n_j be the number of classes that have exactly j individuals, that is, $n_j = \sum_{i=1}^n I(x_i = j)$, for j in Ω . The n_j 's are called *frequency counts*. Note that n_0 is unobserved but the other n_j 's are seen. It is clear that the joint distribution of the x_i 's is given by

$$\prod_{i=1}^N [g(x_i; P)]^{x_i} = \prod_{j \in \Omega} [g(j; P)]^{n_j}.$$

Therefore, the joint distribution of $\{n_j : j \in \Omega\}$ can be written as

$$N! \times \prod_{j \in \Omega} [g(j; P)]^{n_j} / n_j!,$$

This is the likelihood for the hypothetical data $\{n_j : j \in \Omega\}$. Let n_+ be the number of distinct observed classes, that is, $n_+ = \sum_{i=1}^N I(x_i > 0)$, where $I(E)$ is the indicator function of the event E . Let $\Omega_0 = \Omega - \{0\}$. Since $N = n_+ + n_0$, the likelihood for the observed data $\{n_j : j \in \Omega_0\}$ can be written as

$$L(N; P) = N! \times g(0; P)^{N-n_+} / (N - n_+)! \times \prod_{j \in \Omega_0} [g(j; P)]^{n_j} / n_j!.$$

Note that there are two parameters in the likelihood, the number of classes N and the latent distribution P . Since the issue of interest is whether P is degenerate, N will be treated as a nuisance parameter. We next derive a simple mathematical result, which allows us to focus of the parameter of interest P .

The likelihood $L(N; P)$ can be factored into the marginal likelihood of the statistic n_+

$$L_{n_+}(N; P) = N! \times g(0; P)^{N-n_+} / (N - n_+)! \times [1 - g(0; P)]^{n_+} / n_+!$$

times the conditional likelihood for the data given n_+

$$L_c(P) = n_+! \times \prod_{j \in \Omega_0} [g(j; P)]^{n_j} / n_j!. \quad (1)$$

For fixed P , the statistic n_+ is complete and sufficient for N . Thus it is reasonable to base inference about P on $L_c(P)$, an infinite cell multinomial likelihood with cell probabilities $p_z = g(z; P) / (1 - g(0; P))$. See for example, the theory of optimal unbiased tests in Lehmann (1997).

The following lemma gives further insight into the problem. Let $f(z; \lambda)$ be the density of a zero-truncated Poisson distribution and $f(z; Q)$ be the density of a mixture of zero-truncated Poisson distributions, where $f(z; \lambda) = [e^\lambda - 1]^{-1} \lambda^z / z!$ and $f(z; Q) = \int f(z; \lambda) dQ(\lambda)$, for z in Ω_0 .

Lemma 1 *We have*

$$p_z = f(z; Q) \text{ where } dQ(\lambda) = \frac{(1 - e^{-\lambda}) dP(\lambda)}{\int (1 - e^{-\lambda}) dP(\lambda)}.$$

The proof can be obtained by elementary algebra.

Now suppose the classes are re-indexed such that the observed classes have indexes $1, 2, \dots, n_+$. Let $z_i = x_i | x_i > 0, i = 1, 2, \dots, n_+$. Each z_i follows a zero-truncated Poisson distribution with parameter λ_i . Since the x_i 's follow $g(x; P)$, the z_i 's can be regarded as a random sample arising from $f(z; Q)$. The joint distribution of the z_i 's is given by

$$\prod_{i=1}^{n_+} [f(z_i; Q)]^{z_i} = \prod_{j \in \Omega_0} [f(j; Q)]^{n_j},$$

which gives the joint distribution of $\{n_j : j \in \Omega_0\}$

$$L(Q) = n_+! \times \prod_{j \in \Omega_0} [f(j; Q)]^{n_j} / n_j!. \quad (2)$$

Thus $\{n_j : j \in \Omega_0\}$ come from a multinomial distribution with index n_+ and cell probabilities $f(j; Q)$'s. The nuisance parameter N does not appear in the likelihood given by (2). Note that the conditional likelihood of the observed frequency counts in (1) is exactly the same as the marginal likelihood of a set of n_+ i.i.d. observations from a Q -mixture of zero-truncated Poisson variables in (2). It is also clear, since Q and P have the same support points, that $Q(\lambda)$ is degenerate if and only if $P(\lambda)$ is degenerate. Thus by conditioning and reformulating, we have arrived at a standard heterogeneity question involving i.i.d. variables.

From this point, we restrict attention to the likelihood $L(Q)$ in (2). Thus all tests and diagnostics are effectively conditional tests and plots. Our task is to develop diagnostics to test whether $Q(\lambda)$ is degenerate.

3 The χ^2 test and dispersion score test

One simple approach to the assessment of whether Q is degenerate is to perform a χ^2 test of H_0 : Q is degenerate versus the general multinomial alternative; see Zelterman (1988). However, this test lacks focus. If we reject the null model, there is no reason to conclude that heterogeneity exists; in fact, the data could be underdispersed or have many other model defects.

A more attractive alternative is the dispersion score test. Unlike the χ^2 goodness-of-fit test, the dispersion score test is the locally most powerful test against certain mixture alternatives. It is also a simple and tractable procedure; see Neyman and Scott (1966) and Lindsay (1995).

Let ϕ be the *natural parameter* of zero-truncated Poisson distributions and $f(z; \phi)$ be the density parameterized by ϕ , where $\phi = \log \lambda$. The first and second order score functions $\nu_1(\phi, z)$ and $\nu_2(\phi, z)$ are used to define the dispersion score test, where

$$\nu_1(\phi, z) = [f(z; \phi)]^{-1} \frac{\partial f(z; \phi)}{\partial \phi} \quad \text{and} \quad \nu_2(\phi, z) = [f(z; \phi)]^{-1} \frac{\partial^2 f(z; \phi)}{\partial \phi^2}.$$

Let μ be the mean parameter and σ^2 be the variance for a zero-truncated Poisson distribution. It can be shown that

$$\nu_1(\phi, z) = z - \mu \quad \text{and} \quad \nu_2(\phi, z) = (z - \mu)^2 - \sigma^2.$$

Let $\hat{\phi}$ be ML estimate for ϕ . The test statistic is defined by

$$T = [n_+ \tau(\hat{\phi})]^{-1/2} \sum_{i=1}^{n_+} \nu_2(\hat{\phi}, z_i),$$

where

$$\tau = \tau(\phi) = E_z[\nu_2(\phi, z)]^2 - \frac{E_z^2[\nu_2(\phi, z)\nu_1(\phi, z)]}{E_z[\nu_1(\phi, z)]^2}.$$

Note that τ can be expressed as a function of λ , where

$$\tau = \mu_4 + 8\mu_2\mu^2 - 4\mu\mu_3 - \mu_2^2 - 4\mu^4 - (\mu_2 - \mu^2)^{-1}(\mu_3 - 3\mu\mu_2 + 2\mu^3)^2.$$

Here μ and the μ_i 's are the first four moments of a zero-truncated Poisson distribution, where $\mu = (1 - e^{-\lambda})^{-1}\lambda$, $\mu_2 = (1 - e^{-\lambda})^{-1}(\lambda + \lambda^2)$, $\mu_3 = (1 - e^{-\lambda})^{-1}(\lambda + 3\lambda^2 + \lambda^3)$ and $\mu_4 = (1 - e^{-\lambda})^{-1}(\lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4)$.

Under H_0 , the dispersion score test statistic T is a asymptotically standard normal; see Neyman and Scott (1966) and Lindsay (1995). Rejecting for T indicates a departure from H_0 in the direction of overdispersion. If T is too small, there is a lack of fit in the direction of underdispersion. However, the numerical magnitude of T should not be used as an index of amount of heterogeneity as it reflects sample size as well as heterogeneity.

4 The log ratio plot and residual plot

Graphic diagnostics are often employed as exploratory tools to indicate whether the data come from heterogeneous resources, that is, whether a

mixture model is necessary; see Titterton, Smith and Makov (1985), Lindsay and Roeder (1992) and Lindsay (1995).

We will develop two graphical diagnostics which depend on strikingly different predictive properties of two curves under the simple zero-truncated Poisson model H_0 and a zero-truncated Poisson mixture model. Let $f(z; \mu)$ be the density of zero-truncated Poisson distributions parameterized by μ . Let $Q(\mu)$ be the latent distribution for the mean parameter. The mixture distribution is denoted by $f(z; Q)$, where $f(z; Q) = \int f(z; \mu) dQ(\mu)$. Now define two diagnostic functions as follows:

$$H(z) = \log[z!f(z; Q)] \text{ and } r(z) = \frac{f(z; Q)}{f(z; \mu_0)} - 1, \text{ where } \mu_0 = \int \mu dQ(\mu).$$

The functions $H(z)$ and $r(z)$ are called the *log ratio function* and *residual function* respectively. Note that $H(z)$ differs from Lindsay (1995) where the log ratio function is defined as $\log[f(z; Q)/f(z; \mu_0)]$. Here $H(z)$ does not depend on μ_0 .

The zero-truncated Poisson distributions form a one-parameter exponential family. The following theorem summarizes the results in Shaked (1980), Lindsay and Roeder (1992) and Lindsay (1995).

Theorem 2 *The functions $r(z)$ and $H(z)$ are convex in z . They are linear if and only if Q is degenerate. If Q is not degenerate, then the residual function $r(z)$ has the sign sequence $+, -, +$ when z traverses the real line.*

Let $\hat{\mu}$ be MLE for μ under H_0 , which equals \bar{z} . Let $\hat{H}(z)$ and $\hat{r}(z)$ be the empirical log ratio function and empirical residual function respectively, where

$$\hat{H}(z) = \log \frac{z!n_z}{n_+} \text{ and } \hat{r}(z) = \frac{n_z}{n_+ f(z; \hat{\mu})} - 1.$$

For each z , $\hat{r}(z)$ is called the *residual* at z . The empirical functions $\hat{H}(z)$ and $\hat{r}(z)$ are consistent estimators for $H(z)$ and $r(z)$ respectively, for each z in Ω_0 ; see Lindsay and Roeder (1992), Lindsay (1995) and Mao and Lindsay (2001a). Note that $\hat{H}(z)$ does not require the estimation of $\hat{\mu}$. This is useful because solving for the corresponding $\hat{\lambda}$ requires iterative calculation. The plots $(z, \hat{H}(z))$ and $(z, \hat{r}(z))$ will be called the *log ratio plot* and *residual plot* respectively. If H_a is true, the plots will show strict convexity and the empirical residual function will have a sign sequence $+, -, +$. The plots are linear if and only if Q is degenerate.

Applying the results in Mao and Lindsay (2001a), it is clear that $\hat{H}(z)$ is approximately normally distributed for each z as n_+ goes to infinity. Let $e_H(z)$ be the approximate pointwise standard error of $\hat{H}(z)$ either under H_0 or H_a , where $e_H^2(z) = 1/n_z - 1/n_+$. Also, from Lindsay and Roeder (1992), it is clear that the residual $\hat{r}(z)$ at each z is approximately normally distributed. The approximate pointwise standard error $e_r(z)$ of the residual $\hat{r}(z)$ under H_0 is given by $e_r^2(z) = n_+^{-1}[(1/f(z; \hat{\mu}) - 1 - (z - \hat{\mu})^2/\sigma^2(\hat{\mu}))]$. Let Z_α be the upper $1 - \alpha$ quantile of the standard normal distribution. The α -level pointwise confidence bands for the log ratio plot and the residual plot are given by $\hat{H}(z) \pm Z_{\alpha/2}e_H(z)$ and $\hat{r}(z) \pm Z_{\alpha/2}e_r(z)$ respectively.

In practice, the plots are informative only for values of n_z which are positive. Since n_+ is finite, all but a finite number of frequency counts

are zero. When n_z equal to 0, the empirical residual function takes the lower limit -1 . The empirical log ratio function is undefined at z if n_z is zero. The simplest strategy is to treat these points as missing in the plots. Although one might think it is better to group the data in the right tail into a single cell as done in χ^2 goodness-of-fit test, the above convexity results do not apply in a straightforward way in this device. An alternative strategy is to cut off the plots at the smallest frequency z with n_{z+1} being 0.

The log ratio function can be further processed such that it is easier to recognize linearity or convexity. Note that under H_0 , for all z , we have $r(z) = 0$. So we expect as a flat plot under H_0 . But $H(z) = -\log(e^\lambda - 1) + z \log \lambda$. Under H_0 , the log ratio function is a straight line with possible nonzero slope and nonzero intercept. There is a simple method to take the slope and intercept away from the log ratio function. Suppose $H(z) = a + bz + E(z)$, for some a, b in $(-\infty, +\infty)$, where $E(z)$ has zero intercept and zero slope. Let \hat{a} and \hat{b} be estimates for a and b respectively, which can be obtained by regressing $\hat{H}(z)$ on z with weights n_z using weighted least square estimation. Let $\hat{E}(z) = \hat{H}(z) - \hat{a} - \hat{b}z$. Although $\hat{H}(z)$ and $\hat{E}(z)$ have different intercept and slope, the shape does not change. We will use the confidence bands for the plot of $(z; \hat{H}(z))$, however, so as to eliminate the irrelevant variability in $\hat{a} + \hat{b}z$.

We can interpret these plots as follows. If the population is homogeneous, we expect that a horizontal line to be contained in the confidence bands. For the residual plot the horizontal line is required to be through zero. If there is a clear convexity shown in the plots, then we may conclude that the population is heterogeneous. If neither horizontal line nor convex curve shows up, the Poisson assumption should be suspected.

5 Examples

5.1 Cholera data

The first dataset is about an epidemic of cholera in an India village. There were totally 223 households in the village and some of them were affected by cholera. Let x be the number of cholera cases in a household and n_x be the number of households having x cholera cases. Here the unknown n_0 represents the number of affected households without cholera case. The nonzero frequency counts are $n_1 = 32$, $n_1 = 32$, $n_2 = 16$, $n_3 = 6$ and $n_4 = 1$. This dataset was analyzed by many authors. For example, see Blumenthal, Dahiya and Gross(1978) and references therein.

The MLE for λ is $\hat{\lambda} = 0.972$. The expected frequency counts under the zero-truncated Poisson are $\tilde{n}_1 = 33$, $\tilde{n}_2 = 16$, $\tilde{n}_3 = 5$, $\tilde{n}_4 = 1$. The empirical residuals are $\hat{r}(1) = -0.016$, $\hat{r}(2) = 0.012$, $\hat{r}(3) = 0.171$ and $\hat{r}(4) = -0.197$. The sign sequence is $-, +, -$ instead of $+, -, +$ as expected for a mixture model. We pool the frequency counts less than 5. Let $n_{3+} = n_3 + n_4 = 7$, then $\hat{n}_{3+} = 6$. The χ^2 statistic is 0.197, with p-value 0.906, which suggests a good fit. The dispersion score test has a negative statistics $T = -0.513$, which indicates that there is no over-dispersion nor strong evidence for underdispersion.

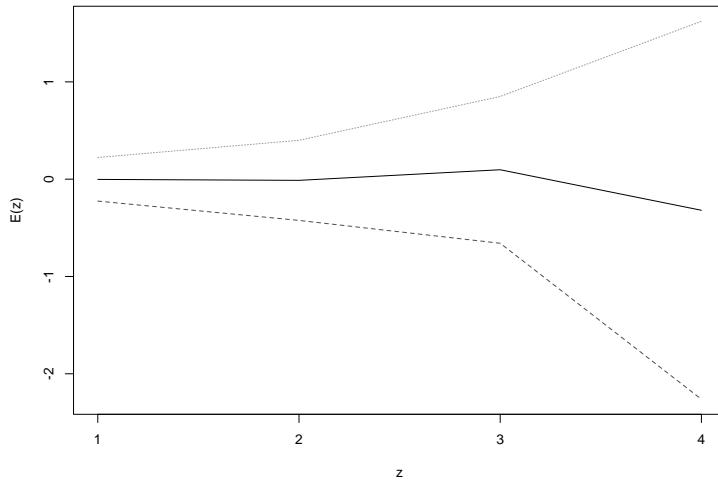


Figure 1: The log ratio plot for cholera data

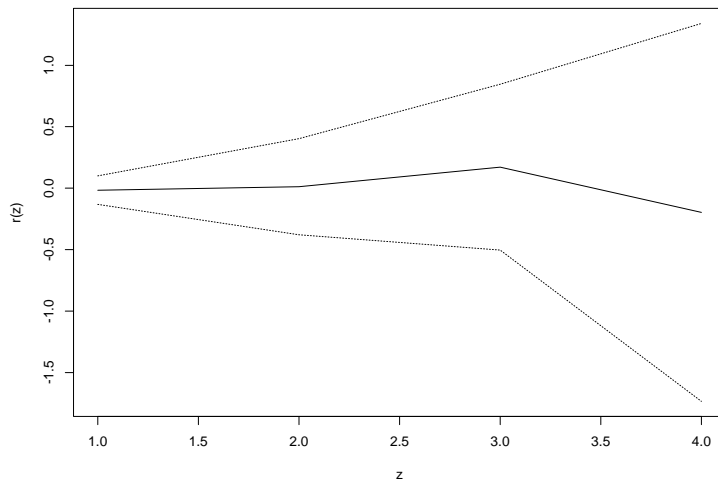


Figure 2: The residual plot for cholera data

Both the confidence bands of the log ratio plot and those of the residual plot contain horizontal line through zero, see Figure 1 and Figure 2. These graphical diagnostics show that a simple Poisson model for the cholera infection cases is appropriate

5.2 Tomato EST data

In the second example, we study a dataset generated from a sequencing project in genomic studies. A typical prepared cDNA library consists of about 10^6 clones, where each clone represents one copy of cDNA from a gene. The cDNA copy numbers of genes may differ by $10^3 \sim 10^4$ fold. It would be prohibitively expensive to count all cDNA copies of each gene, which requires to sequence all the clones. A feasible approach is to sequence a sample of clones and make inference about gene expression from the sample, which is called “digital Northern” or “digital gene expression”; see Audic and Claverie (1997), Sekel, Git and Falciani (2000) and Mao and Lindsay (2000c).

The single-pass cDNA sequences are called expressed sequence tags (ESTs). The ESTs are clustered into unique genes. The number of ESTs from a unique gene can be used to infer expression of that gene. Given a target cDNA library, let N be the number of expressed genes and the genes be indexed by $1, 2, \dots, N$. The number of ESTs from the i th gene is denoted by x_i , which is assumed to be independent Poisson random variable with mean λ_i . Now n_j is the number of genes that have j ESTs in the sample.

One tomato flower cDNA library is studied here. The library was made from $0 \sim 3$ mm buds of tomato flowers. There were 2586 ESTs generated from the library. Here are the nonzero observed frequency counts: $n_1 = 1434$, $n_2 = 253$, $n_3 = 71$, $n_4 = 33$, $n_5 = 11$, $n_6 = 6$, $n_7 = 2$, $n_8 = 3$, $n_9 = 1$, $n_{10} = 2$, $n_{11} = 2$, $n_{12} = 1$, $n_{13} = 1$, $n_{14} = 1$, $n_{16} = 1$, $n_{23} = 1$ and $n_{27} = 1$. The dataset was obtained from the TIGR Tomato Gene Index, see Quackbash et al. (2000). It was analyzed in Mao and Lindsay (2001b, 2001c).

We have $\bar{z} = 1.417$. The MLE $\hat{\lambda}$ for λ equals 0.743. The expected frequency counts decrease quickly. Here are the first five expected frequency counts $\hat{n}_1 = 1230$, $\hat{n}_2 = 457$, $\hat{n}_3 = 113$, $\hat{n}_4 = 21$ and $\hat{n}_5 = 3$. Let $n_{4+} = \sum_{j \geq 4} n_j = 67$. Also $\hat{n}_{4+} = 25$. The χ^2 statistic is 210.94 with degree of freedom 3 and p-value 0, which indicates that the a zero-truncated Poisson distribution does not fit the frequency counts. The dispersion score test gives $T = 90.15$ with p-value 0, which indicates significant over-dispersion.

Since $n_{15} = 0$, we truncate the log ratio plot at $z = 14$. The log ratio plot shows a clear convexity, see Figure 3. The first fourteen residuals are $\hat{r}(1) = 0.166$, $\hat{r}(2) = -0.446$, $\hat{r}(3) = -0.373$, $\hat{r}(4) = 0.570$, $\hat{r}(5) = 2.524$, $\hat{r}(6) = 1.452 \times 10$, $\hat{r}(7) = 4.776 \times 10$, $\hat{r}(8) = 7.867 \times 10^2$, $\hat{r}(9) = 3.180 \times 10^3$, $\hat{r}(10) = 8.565 \times 10^4$, $\hat{r}(11) = 1.268 \times 10^6$, $\hat{r}(12) = 1.024 \times 10^7$, $\hat{r}(13) = 1.793 \times 10^8$ and $\hat{r}(14) = 3.379 \times 10^9$. The sign sequence is $+, -, +$. Note that the residuals at lower frequencies and high frequencies have sharp differences in magnitude. As a result, convexity is not clear in the residual plot truncated at $z = 14$, see Figure 4. However, if we cut the residual

plot at $z = 6$, which is the last frequency with frequency count larger than 5, there is a clear convexity, see Figure 5.

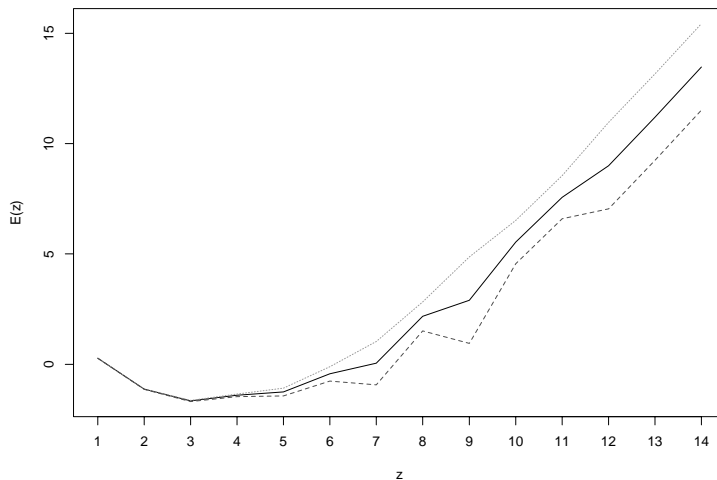


Figure 3: The log ratio plot for tomato EST data

The statistical tests and diagnostics give conclusions that are consistent with experiment experience, which say that gene expressions are highly differentiated.

6 Discussion

We present four methods to assess class abundance homogeneity. The χ^2 test assesses overall fit. The dispersion score test is recommended as assessment of over-dispersion. The residual plot and the log ratio plot provide information about fit or bad thereof, but also indicate by convexity, the adequacy of the mixture model if homogeneity fails. If there are big differences among the frequency counts, the log ratio plot will have better visual effect than the residual plot. The log ratio plot, the simpler of the two, is recommended as the first choice.

Another plot, the gradient plot given by Lindsay and Roeder (1992), was considered in our investigation. However, it tends to be too smooth to pick up interesting deviations from the model. See Lindsay (1995).

There also exist other tests, most importantly, the likelihood ratio test. Unfortunately, the likelihood ratio tests in a mixture model setting are quite complicated in computation and distribution theory. Their actual performance depends on the method of computation, see Lindsay (1995) and Seidel, Mosler and Alker (2000). However, a significant dispersion score test coupled with a convex log ratio plot provides strong evidence

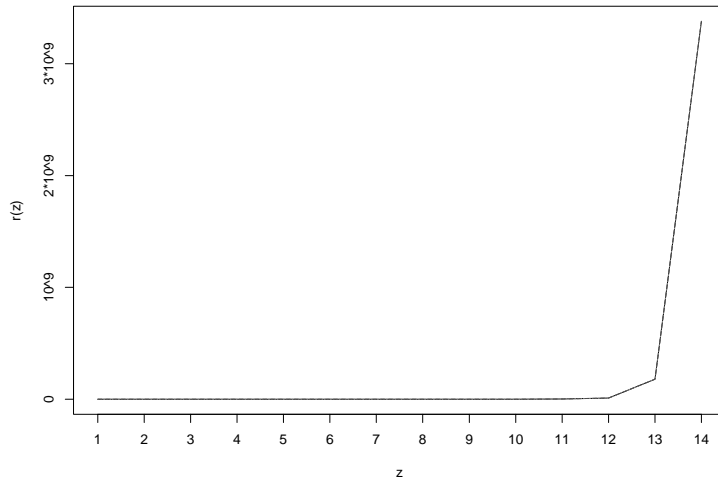


Figure 4: The first residual plot for tomato EST data

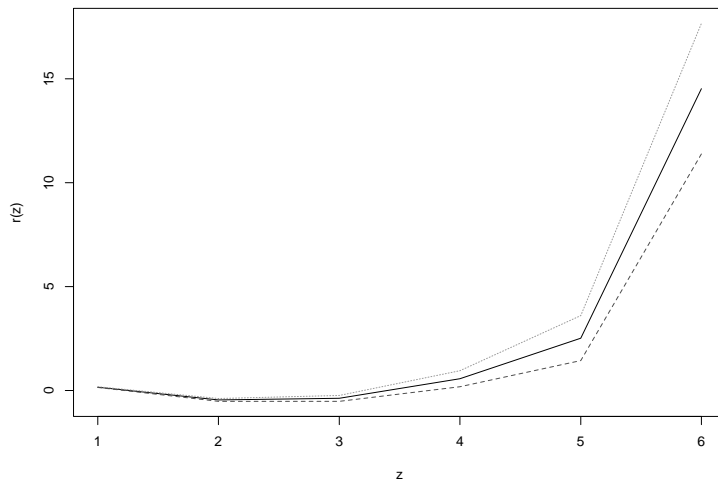


Figure 5: The second residual plot for tomato EST data

that the true distribution is very similar to a mixture distribution with more than one component.

There are models other than a Poisson mixture model to describe class abundance heterogeneity, see Jorgensen (1987) and Gelfand and Dalal (1990) and discussions therein. Although such models are not considered in this paper, we point out that the dispersion score test is not specific for the mixture model setting. In fact, it is sensitive to any overdispersion alternative.

References

- Audic, S. and Claverie, J. M., The significance of digital gene expression profiles, *Genome Research*, 7 (1997) 986-995.
- Blumenthal, S. Dahiya, R. C., and Grosss, A. J., Estimating complete sample-size from an incomplete Poisson sample, *Journal of the American Statistical Association*, 73 (1978) 182-187.
- Bunge, J. and Fitzpatrick, M., Estimating the Number of Species: A Review, *Journal of the American Statistical Association*, 88 (1993) 364-373.
- Chao, A., Nonparametric Estimation of the Number of Classes in a Population, *Scandinavian Journal of Statistics*, 11 (1984) 265-270.
- Chao, A. and Lee, S. M., Estimating the Number of Classes Via Sample Coverage, *Journal of the American Statistical Association*, 87 (1992) 210-217.
- Darroch, J. N. and Ratcliff, D., A Note on Capture-recapture Estimation, *Biometrics*, 36 (1980) 149-153.
- Gelfand, A. E., and Dalal, S. R., A Note on Overdispersed Exponential Families, *Biometrika*, 77 (1990) 55-64.
- Jorgensen, B., Exponential Dispersion Models, *Journal of the Royal Statistical Society. Series B*, 49 (1987) 127-162.
- Lehmann, E. L., *Testing Statistical Hypotheses*, (Springer-Verlag, Second Edition, 1997).
- Lindsay, B. G., *Mixture Models: Theory, geometry and applications*, (NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, 1985).
- Lindsay, B. G. and Roeder K., Residual Diagnostics for Mixture Models, *Journal of the American Statistical Association*, 87 (1992) 785-794.
- Mao, C. X. and Lindsay, B. G., Diagnostics for homogeneity of capture probabilities in capture-recapture experiments, (Technical report, The likelihood center, Department of Statistics, Pennsylvania State University, 2001a).
- Mao, C. X. and Lindsay, B. G., Moment-based nonparametric estimators for the number of classes in a population, (Technical report, The likelihood center, Department of Statistics, Pennsylvania State University, 2001b).

- Mao, C. X. and Lindsay, B. G., A Poisson model for coverage problems with an application in genomic research, (Technical report, The likelihood center, Department of Statistics, Pennsylvania State University, 2001c).
- Neyman, J. and Scott, E. L., On the use of $C(\alpha)$ optimal tests of composite hypothesis, *Bulletin of the Institute of International Statistics*, 41(I) (1966) 477-497.
- Norris, J. L. I. and Pollock, K. H., Non-parametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity Between Species, *Environmental and Ecological Statistics*, 5 (1998) 391-402.
- Ord, J. K. and Whitmore, G., The Poisson-inverse Gaussian Distribution As a Model for Species Abundance, *Communications in Statistics, Theory and Methods*, 15 (1986) 853-871.
- Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J., The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species, *Nucleic Acids Research*, 29 (2000) 159-164.
- Seidel, W., Mosler, K., and Alker, M., A Cautionary Note on Likelihood Ratio Tests in Mixture Models, *Annals of the Institute of Statistical Mathematics*, 52 (2000) 481-487.
- Shaked, M., On mixtures from exponential families, *Journal of the Royal Statistical Society, Series B*, 42 (1980) 192-198.
- Stekel, D. J., Git, Y. and Falciani, F., The comparison of gene expression from multiple cDNA libraries, *Genome Research*, 10 (2000) 2055-206.
- Titterton, D. M. and Smith, A. F. and Makov, U. E., *Statistical analysis of finite mixture distributions* (Wiley, 1985).
- Zelner, D., Robust estimation in truncated discrete distributions with application to capture-recapture experiments, *Journal of statistical planning and inference*, 6 (1988) 225-237.