BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 11 2014, pages 1547–1554
doi:10.1093/bioinformatics/btu079

Gene expression

Advance Access publication February 11, 2014

An efficient algorithm for accurate computation of the Dirichlet-multinomial log-likelihood function

Peng Yu[1,*] and Chad A. Shaw[2,*]

[1]Department of Electrical and Computer Engineering & TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering (CBGSE), Texas A&M University, College Station, TX 77843, USA and [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

Associate Editor: Inanc Birol

ABSTRACT

Summary: The Dirichlet-multinomial (DMN) distribution is a fundamental model for multicategory count data with overdispersion. This distribution has many uses in bioinformatics including applications to metagenomics data, transctriptomics and alternative splicing. The DMN distribution reduces to the multinomial distribution when the overdispersion parameter ψ is 0. Unfortunately, numerical computation of the DMN log-likelihood function by conventional methods results in instability in the neighborhood of $\psi = 0$. An alternative formulation circumvents this instability, but it leads to long runtimes that make it impractical for large count data common in bioinformatics. We have developed a new method for computation of the DMN log-likelihood to solve the instability problem without incurring long runtimes. The new approach is composed of a novel formula and an algorithm to extend its applicability. Our numerical experiments show that this new method both improves the accuracy of log-likelihood evaluation and the runtime by several orders of magnitude, especially in high-count data situations that are common in deep sequencing data. Using real metagenomic data, our method achieves manyfold runtime improvement. Our method increases the feasibility of using the DMN distribution to model many high-throughput problems in bioinformatics. We have included in our work an R package giving access to this method and a vignette applying this approach to metagenomic data.

Availability and implementation: An implementation of the algorithm together with a vignette describing its use is available in Supplementary data.

Contact: pengyu.bio@gmail.com or cashaw@bcm.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 19, 2013; revised on December 10, 2013; accepted on January 31, 2014

1 INTRODUCTION

The analysis of count data (Cameron and Trivedi, 2013; Winkelmann, 2008) or categorical data (Agresti, 2002) is an important topic in statistics and has a wide variety of applications in bioinformatics. The advent of high-throughput sequencing technologies (Metzker, 2010) provides unprecedented opportunities for investigating new and more powerful analysis methods on count data (Anders and Huber, 2010; Robinson *et al.*, 2010).

The Poisson distribution is a basic distribution for modeling count data. An important property of the Poisson distribution is that the mean and variance are the same, which is called equidispersion. However, the mean and the variance of real count data are often not the same; in fact, the variance is often greater than the mean. This makes the Poisson distribution not ideal for analyzing such data because the equidispersion assumption is violated. The phenomenon where a dataset exhibits greater variance than what would be expected in a statistical model is called overdispersion. A commonly used overdispersed model for the Poisson distribution is the negative-binomial distribution. This distribution has been extensively studied in Hilbe (2011) and is an indispensable model for high-throughput sequencing data (Anders and Huber, 2010).

Another fundamental model in count data analysis is the multinomial (MN) distribution, which is useful for analysis of count proportions among multiple categories. One important use case of the MN distribution is Fisher's exact test of contingency tables (Fisher, 1973; Mehta and Patel, 1986), which has been used in the analysis of alternative 3′ UTR utilization (Wan, 2012) and splicing (Lu *et al.*, 2013), as well as metagenomics (Gomez-Alvarez *et al.*, 2012). In the regression context, MN logistic regression is also commonly used (Agresti, 2002). However, real data often exhibit heterogeneity that is usually thought to be caused by dependencies or the similarity of responses of members of the same cluster in cluster sampling (Brier, 1980). This leads to extra-multinomial variation (Haseman and Kupper, 1979), i.e. overdispersion with respect to the MN distribution.

The modeling of overdispersion of the MN distribution has been addressed by extending the MN distribution to the Dirichlet-multinomial (DMN) distribution (Mosimann, 1962; Poortema, 1999). The beta-binomial (BB) distribution—a special case of the DMN distribution with only two categories—has been studied by many (Crowder, 1978; Kleinman, 1973; Skellam, 1948). Because of its flexibility and its mathematical convenience, the DMN distribution is widely applied to diverse fields, such as topic modeling (Mimno and McCallum, 2008), magazine exposure modeling (Leckenby and Kishi, 1984; Rust and Leone, 1984), word burstiness modeling (Madsen *et al.*, 2005), language modeling and (MacKay and Bauman Peto, 1994) multiple sequence alignment (Brown *et al.*, 1993; Sjölander *et al.*, 1996). Bouguila (2008) also considered a generalization of the DMN distribution and applied it to count data

*To whom correspondence should be addressed.

clustering. Another related distribution for handling overdispersion is the Dirichlet negative MN distribution (Mosimann, 1963) allowing the modeling of correlated count data without an upper bound, which has many possible uses in biostatistics and bioinformatics (Farewell and Farewell, 2013).

Likelihood functions play a key role in statistical inference (Casella and Berger, 2002). For example, likelihood functions can be used for parameter estimation, hypothesis testing and interval estimation. In the context of the DMN distribution, there has been recent research to investigate the Fisher information matrix (Paul *et al.*, 2005) and maximum likelihood estimation (MLE) (Neerchal and Morel, 2005). Not all statistical inference methods are based on likelihood functions. For instance Kim and Margolin (1992) developed a method for testing the goodness of fit of the MN distribution against the DMN distribution based on the $C(\alpha)$ test statistic (Tarone, 1979), a flexible framework built on the likelihood approach that enables the analysis of complex experimental designs (McCullagh and Nelder, 1989) that frequently appear in genomic and bioinformatics studies.

In this article, we study the fundamental problem of the evaluation of the DMN log-likelihood function. In Section 2, we demonstrate the instability and runtime problems of two existing methods for computing the DMN log-likelihood function and propose a novel parameterization of the log-likelihood function to allow smooth transition from the overdispersed case (the DMN distribution) to the non-overdispersed case (the MN distribution). For this new parameterized form, in Section 3 we introduce a new formula based on a truncated series consisting of Bernoulli polynomials. In Section 4, a mesh algorithm is devised to increase the applicability of this new formula. In Section 5, we show numerical results of the mesh algorithm, confirm its stability and runtime improvements. Finally, we applied our method to human microbiome data and demonstrated its large performance improvement over the most accurate existing method.

2 DMN DISTRIBUTION

The DMN distribution, a.k.a., the compound MN distribution (Mosimann, 1962), is an extension of the MN distribution. The probability mass function (PMF) of the K categories MN distribution of N-independent trials is given by

$$f_{\mathrm{MN}}(\boldsymbol{x}; N, \boldsymbol{p}) = \frac{N!}{\prod_{k=1}^{K} x_k!} \prod_{k=1}^{K} p_k^{x_k} \tag{1}$$

where $n!$ denotes the factorial of a non-negative integer n; the observations $\boldsymbol{x} = (x_1, ..., x_K)$, satisfying $\sum_{k=1}^{K} x_k = N$, are non-negative integers; and $\boldsymbol{p} = (p_1, ..., p_K)$, satisfying $\sum_{k=1}^{K} p_k = 1$, are the probabilities that these K categories occur.

The DMN distribution can be generated if the probabilities $\boldsymbol{p}$ follow a prior distribution (of the positive parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$) conjugate to the PMF $f_{\mathrm{MN}}(\boldsymbol{x}; N, \boldsymbol{p})$ (Bishop, 2006)

$$f_{\mathrm{Dir}}(\boldsymbol{p}; \alpha) \propto \prod_{i=1}^{K} p_i^{\alpha_i - 1}$$

This distribution is called the Dirichlet distribution whose normalized form is

$$f_{\mathrm{Dir}}(\boldsymbol{p}; \alpha) = \frac{\Gamma(A)}{\prod\limits_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} p_i^{\alpha_i - 1} \tag{2}$$

where $\Gamma(x)$ is the gamma function and $A = \sum_{i=1}^{K} \alpha_i$.

The PMF of the DMN distribution is derived by taking the integral of the product of the Dirichlet prior (2) and the MN likelihood (1) with respect to the probabilities $\boldsymbol{p}$ (Mosimann, 1962),

$$f_{\mathrm{DMN}}(\boldsymbol{x}; N, \alpha) = \frac{N!}{\prod\limits_{k=1}^{K} x_k!} \frac{\Gamma(A)}{\Gamma(A+N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + x_k)}{\Gamma(\alpha_k)} \tag{3}$$

where, same as the MN distribution, $\boldsymbol{x} = (x_1, \ldots, x_K)$ are non-negative integers, satisfying $N = \sum_{k=1}^{K} x_k$. The DMN distribution reduces to the BB distribution when there are only two categories ($K = 2$).

The first term on the right side of (3) does not depend on the parameter α. For common uses of the likelihood function in statistics, e.g. in the maximum-likelihood estimation, we are not interested in the first term but in the product of the remaining two terms, i.e. we are interested in the last two terms of the DMN likelihood function in (3)

$$\mathcal{L}(\alpha; \boldsymbol{x}) = \frac{\Gamma(A)}{\Gamma(A+N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + x_k)}{\Gamma(\alpha_k)} \tag{4}$$

By taking the logarithm of both sizes of (4), we get the log-likelihood function

$$\begin{aligned} \ln \mathcal{L}_{\Gamma}(\alpha; \boldsymbol{x}) = & -(\ln \Gamma(A+N) - \ln \Gamma(A)) \\ & + \sum_{k=1}^{K} (\ln \Gamma(\alpha_k + x_k) - \ln \Gamma(\alpha_k)) \end{aligned} \tag{5}$$

When $A \to \infty$, it can be shown that the DMN distribution is reduced to the MN distribution. As $\psi = 1/A$ becomes 0 under this limit, it is convenient to use the parameter ψ instead of A.

The parameter ψ characterizes how different a DMN distribution is from the corresponding MN distribution with the same category probabilities. The greater the parameter ψ, the greater the difference. This additional parameter gives the DMN distribution the ability to capture variation that cannot be accommodated by the MN distribution. We call ψ the overdispersion parameter in this article, with the understanding that the greater the ψ, the greater the variance. As an example, Figure 1 shows that increasing the dispersion parameter ψ of the BB dispersion increases the variance of the count of the first category x_1. Using $\psi = 1/A$, (5) becomes

$$\begin{aligned} \ln \mathcal{L}(\boldsymbol{p}, \psi; \boldsymbol{x}) = & -(\ln \Gamma(1/\psi + N) - \ln \Gamma(1/\psi)) \\ & + \sum_{k=1}^{K} \left(\ln \Gamma\left(1/\frac{\psi}{p_k} + x_k\right) - \ln \Gamma\left(1/\frac{\psi}{p_k}\right) \right) \end{aligned} \tag{6}$$

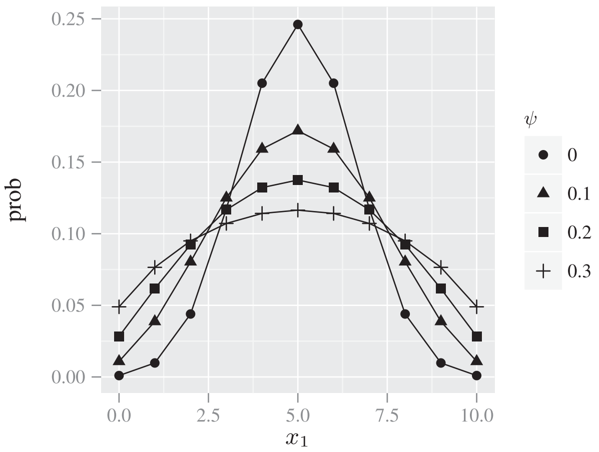

Fig. 1. The PMFs of a family of the BB distributions ($N = 10$) with different dispersion parameters ψ. The spread of the distributions increases with the dispersion parameter, whereas the mean remains constant

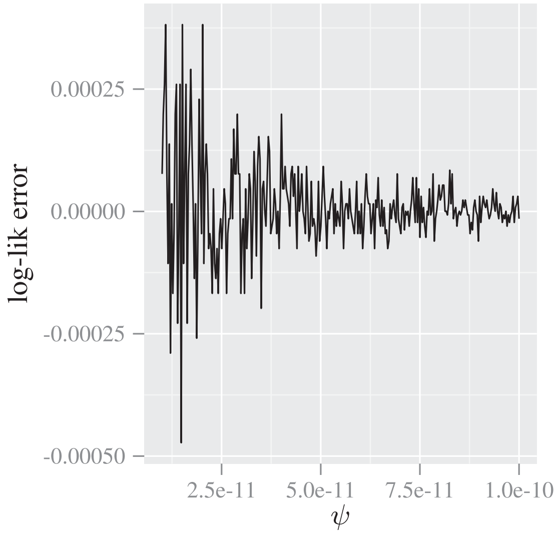

Fig. 2. The function implementing (6) is unstable when ψ approaches 0. The parameters are $\boldsymbol{x} = (2, 3, 1)$ and $\boldsymbol{p} = (.2, .3, .5)$

where $\boldsymbol{p} = \psi\alpha$. One shortcoming of (6) is that it is undefined for $\psi = 0$. Hence, R (R Core Team, 2013) functions implementing (6), such as `dirmult ()` from **dirmult** (Tvedebrink, 2009) and `betabin ()` from **aod** (Lesnoff and Lancelot, 2012), return `NaN` when $\psi = 0$. Another shortcoming of (6) is that as $\psi \to 0$ the function implementing (6) is unstable as shown in Figure 2.

Alternatively, the likelihood representation used in the method in the R package **VGAM** (Yee, 2010, 2012; Yee and Wild, 1996), (4) can be written as

$$\begin{aligned}\mathcal{L}(\alpha; \boldsymbol{x}) &= \frac{\prod_{k=1}^{K}\prod_{r=1}^{x_k}(\alpha_k + (r-1))}{\prod_{r=1}^{N}(A + (r-1))}\\ &= \frac{\prod_{k=1}^{K}\prod_{r=1}^{x_k}(Ap_k + (r-1))}{\prod_{r=0}^{N-1}(A + (r-1))}\\ &= \frac{\prod_{k=1}^{K}\prod_{r=1}^{x_k}\left(\frac{A}{1+A}p_k + \frac{r-1}{1+A}\right)}{\prod_{r=0}^{N-1}\left(\frac{A}{1+A} + \frac{r-1}{1+A}\right)}\\ &= \frac{\prod_{k=1}^{K}\prod_{r=1}^{x_k}(p_k(1-\rho) + (r-1)\rho)}{\prod_{r=1}^{N}((1-\rho) + (r-1)\rho)}\end{aligned}$$

where $\rho = 1/(1 + A)$ is the overdispersion parameter defined therein, which is different from our definition of the overdispersion parameter ψ in (6). The log-likelihood function can be written as

$$\begin{aligned}\ln\mathcal{L}(\boldsymbol{p}, \rho; \boldsymbol{x}) = &\sum_{k=1}^{K}\sum_{r=1}^{x_k}\ln(p_k(1-\rho) + (r-1)\rho)\\ &-\sum_{r=1}^{N}\ln((1-\rho) + (r-1)\rho)\end{aligned} \tag{7}$$

When there is 0 overdispersion ($\rho = 0$), (7) reduces to the MN log-likelihood and it is numerically stable when $\rho \to 0$. But the number of terms on the right side of (7) is proportional to N. When N is large, the runtime is long.

3 APPROXIMATION OF PAIRED LOG-GAMMA DIFFERENCE

We see that (6) consists of paired $\ln\Gamma$ differences, such as $\ln\Gamma(1/\psi + N) - \ln\Gamma(1/\psi)$. When ψ is close to 0, each $\ln\Gamma$ term becomes exceedingly large, but the paired differences become relatively small. Because of the limited precision of the floating-point arithmetic (IEEE Task P754, 2008), the large terms cancel and the result is left with large errors. We solve this large error problem by a new approximation to the $\ln\Gamma$ difference. Let us consider $\ln\Gamma(z + a)$ and $\ln\Gamma(z)$. Rowe (1931) showed that $\ln\Gamma(z + a)$ (z and a are complex numbers) can be asymptotically expanded as

$$\begin{aligned}\ln\Gamma(z+a) = &\left(z + a - \frac{1}{2}\right)\ln z - z + \frac{1}{2}\ln(2\pi)\\ &+\sum_{n=2}^{m}\frac{(-1)^n B_n(a)}{n(n-1)z^{n-1}} + \mathcal{O}(z^{-m}),\ \text{as } z \to \infty\end{aligned} \tag{8}$$

where $B_n(a)$ denotes the nth Bernoulli polynomial. The term $\mathcal{O}(z^{-m})$ means that, for any fixed m, the error of the right side of (8) (with the term $\mathcal{O}(z^{-m})$ removed) as an approximation to $\log\Gamma(z + a)$ is bounded by z^{-m} times some constant as $z \to \infty$. Let $a = 0$,

$$\begin{aligned}\ln\Gamma(z) = &\left(z - \frac{1}{2}\right)\ln z - z + \frac{1}{2}\ln(2\pi)\\ &+\sum_{n=2}^{m}\frac{(-1)^n B_n}{n(n-1)z^{n-1}} + \mathcal{O}(z^{-m}),\ \text{as } z \to \infty\end{aligned}$$

where B_n denotes the nth Bernoulli number ($B_n = B_n(0)$). Let $z = 1/x$ and $a = y$, the difference between the above two equations is

$$\begin{aligned}\ln\Gamma(1/x + y) - \ln\Gamma(1/x) = &-y\ln x + \sum_{n=2}^{m}\frac{(-1)^n\phi_n(y)}{n(n-1)}x^{n-1}\\ &+\mathcal{O}(x^m),\ \text{as } x \to 0\end{aligned} \tag{9}$$

where

$$\phi_n(y) = B_n(y) - B_n \tag{10}$$

is the old type Bernoulli polynomial (Whittaker and Watson, 1927). The infinite series

$$D_\infty(x, y) = \sum_{n=2}^{\infty}\frac{(-1)^n\phi_n(y)}{n(n-1)}x^{n-1} \tag{11}$$

converges absolutely when y is an integer and $|x|\min(|y-1|, |y|) < 1$ (Freitag and Busam, 2009). Note that $x = 0$ is a removable singularity of $\ln\Gamma(1/x + y) - \ln\Gamma(1/x) + y\ln x$. Using the properties of analytic functions, we have (Freitag and Busam, 2009)

$$\ln\Gamma(1/x + y) - \ln\Gamma(1/x) = -y\ln x + D_\infty(x, y) \tag{12}$$

Therefore, for any integer y, we can use the following approximation

$$\ln\Gamma(1/x + y) - \ln\Gamma(1/x) \approx -y\ln x + D_m(x, y) \tag{13}$$

when y is an integer and $|x|\min(|y-1|, |y|) < 1$ and

$$D_m(x, y) = \sum_{n=2}^{m}\frac{(-1)^n\phi_n(y)}{n(n-1)}x^{n-1} \tag{14}$$

The error is bounded by

$$\sum_{n=m+1}^{\infty} \frac{1}{n-1} \delta^n \tag{15}$$

if $|x| \min(|y-1|, |y|) < \delta$, where δ is a constant <1. For the application of computing the DMN log-likelihood function, we have $x \geq 0$ and $y \in \mathbb{N}^+$. So we instead require

$$xy \leq \delta \tag{16}$$

for simplicity.

The error bound (15) can be arbitrarily small for arbitrarily large m without considering the numerical errors in computing the Bernoulli polynomials $\phi_n(y)$. In practice, high order polynomials are difficult to compute using floating-point arithmetic (Lauter and Dinechin, 2008). Because the subscript n equals the order of the Bernoulli polynomial $\phi_n(y)$, if m is too large, the error of each terms of $D_m(x, y)$ may actually be large, which makes $D_m(x, y)$ inaccurate. Hence, we do not want too many terms in (14). So we choose $m = 20$ such that $\phi_n(y)$ $(n \leq m)$ are still numerically accurate. We also do not need the error bound (15) to be smaller than the machine epsilon of the `double` precision data type ($\approx 2.22 \times 10^{-16}$). Therefore, we choose $\delta = 0.2$, which leads to an error bound of $\sim 1.30 \times 10^{-16}$, which is a little less than the machine epsilon.

4 THE MESH ALGORITHM FOR COMPUTING THE DMN LOG-LIKELIHOOD

We apply (13) to compute the DMN log-likelihood function (6). To cope with the requirement (16), by using the idea of analytic continuation (Freitag and Busam, 2009), we introduce a mesh algorithm and allow the computation of the DMN log-likelihood using the approximation (13) in the whole parameter domain of the DMN log-likelihood function. First, we study $\ln \mathcal{L}(\boldsymbol{p}, \psi; \boldsymbol{x})$ in (6) in detail. Let $\boldsymbol{x}^+$ be the vector of the non-zero elements in $\boldsymbol{x}$, $\boldsymbol{p}^+$ be a vector of the corresponding elements in $\boldsymbol{p}$ and K^+ be the length of $\boldsymbol{x}^+$, then (6) becomes

$$\begin{aligned} \ln \mathcal{L}(\boldsymbol{p}, \psi; \boldsymbol{x}) &= -\Bigg(\underbrace{\ln \Gamma(1/\psi + N) - \ln \Gamma(1/\psi)}_{*}\Bigg) \\ &\quad + \sum_{k=1}^{K^+} \Bigg(\underbrace{\ln \Gamma\bigg(1/\frac{\psi}{p_k^+} + x_k^+\bigg) - \ln \Gamma\bigg(1/\frac{\psi}{p_k^+}\bigg)}_{**}\Bigg) \\ &= \ln \mathcal{L}(\boldsymbol{p}^+, \psi; \boldsymbol{x}^+). \end{aligned} \tag{17}$$

When

$$\psi x_k^+ \leq \delta p_k^+, \tag{18}$$

after taking the sum over k on both sides, we have

$$\psi N \leq \delta \sum_{k=1}^{K^+} p_k^+ \leq \delta \tag{19}$$

Therefore, the (*) term and all the K^+ (**) terms in (17) meet the condition (16). Then the approximation (13) can be used for all $K^+ + 1$ paired $\ln \Gamma$ differences in (17),

$$\begin{aligned} \ln \mathcal{L}(\boldsymbol{p}^+, \psi; \boldsymbol{x}^+) &\approx -(-N \ln \psi + D_m(\psi, N)) \\ &\quad + \sum_{k=1}^{K^+} \bigg(-x_k^+ \ln\bigg(\frac{\psi}{p_k^+}\bigg) + D_m\bigg(\frac{\psi}{p_k^+}, x_k^+\bigg)\bigg) \\ &= -D_m(\psi, N) + \sum_{k=1}^{K^+} \bigg(x_k^+ \ln p_k^+ + D_m\bigg(\frac{\psi}{p_k^+}, x_k^+\bigg)\bigg) \end{aligned} \tag{20}$$

When some of the K^+ (**) terms in (17) do not meet the condition (18), we can rewrite the vector $\boldsymbol{x}$ into the sum of L terms choosing the terms to meet this condition

$$\boldsymbol{x} = \sum_{l=1}^{L} \boldsymbol{x}^{(l)} \tag{21}$$

We describe the choice of $\boldsymbol{x}^{(l)}$ below. For convenience, we define

$$\alpha^{(l)} = \alpha + \sum_{i=1}^{l} \boldsymbol{x}^{(i)}, \text{ for } l = 0, \ldots, L \tag{22}$$

Note that we have the following relation between the adjacent $\alpha^{(l)}$'s,

$$\alpha^{(l-1)} + \boldsymbol{x}^{(l)} = \alpha^{(l)}, \text{ for } l = 1, \ldots, L, \tag{23}$$

or

$$\boldsymbol{p}^{(l-1)}/\psi^{(l-1)} + \boldsymbol{x}^{(l)} = \boldsymbol{p}^{(l)}/\psi^{(l)}, \text{ for } l = 1, \ldots, L \tag{24}$$

By taking the sum of all the elements in each vector in (22), we have

$$\frac{1}{\psi^{(l)}} = \frac{1}{\psi} + \sum_{i=1}^{l} N^{(i)}, \text{ for } l = 0, 1, \ldots, L \tag{25}$$

where $1/\psi^{(l)} = \sum_{i=1}^{K} \alpha_i^{(l)}$ and $N^{(l)} = \sum_{i=1}^{K} x_i^{(l)}$. Or we write it as

$$\frac{1}{\psi^{(l)}} = \frac{1}{\psi^{(l-1)}} + N^{(l)}, \text{ for } l = 1, \ldots, L \tag{26}$$

Strictly speaking, (25) is undefined for $\psi = 0$, but when $\psi = 0$, all $\psi^{(l)}$ should be 0 s. To make (25) numerically valid for all $\psi \in [0, +\infty)$, we write

$$\psi^{(l)} = \begin{cases} \frac{1}{\frac{1}{\psi} + \sum_{i=1}^{l} N^{(i)}} & \text{if } \psi \geq 1 \\ \frac{\psi}{1 + \psi \sum_{i=1}^{l} N^{(i)}} & \text{if } 0 \leq \psi < 1 \end{cases} \tag{27}$$

Similarly, for $\boldsymbol{p}$, we have

$$\boldsymbol{p}^{(l)} = \begin{cases} \frac{\frac{\boldsymbol{p}}{\psi} + \sum_{i=1}^{l} \boldsymbol{x}^{(i)}}{\frac{1}{\psi} + \sum_{i=1}^{l} N^{(i)}} & \text{if } \psi \geq 1 \\ \frac{\boldsymbol{p} + \psi \sum_{i=1}^{l} \boldsymbol{x}^{(i)}}{1 + \psi \sum_{i=1}^{l} N^{(i)}} & \text{if } 0 \leq \psi < 1 \end{cases} \tag{28}$$

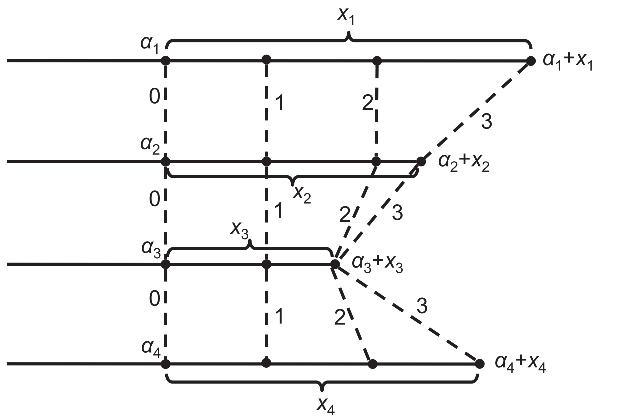

Fig. 3. A graphical depiction of the mesh algorithm for evaluation of the log-likelihood in (29). The count in the ith category is represented by a line segment and can be partitioned into a sum of $L=3$ sub-counts represented by sub-segments. At points connected by the dashed lines ($\alpha^{(l)}$), the DMN log-likelihood can be evaluated using (20), and there are three such evaluations in this example

Then, using (24) and (26), (6) can be broken into a sum of L log-likelihoods,

$$
\begin{aligned}
&\ln \mathcal{L}(\boldsymbol{p}, \psi; \boldsymbol{x}) \\
&= -\textstyle\sum_{l=1}^{L} \left(\ln\Gamma\left(1/\psi^{(l-1)} + N^{(l)}\right) - \ln\Gamma\left(1/\psi^{(l-1)}\right)\right) \\
&+ \textstyle\sum_{l=1}^{L} \sum_{k=1}^{K} \left(\ln\Gamma\left(p_k^{(l-1)}/\psi^{(l-1)} + x_k^{(l)}\right) - \ln\Gamma\left(p_k^{(l-1)}/\psi^{(l-1)}\right)\right) \\
&= \textstyle\sum_{l=1}^{L} \Big(- \left(\ln\Gamma\left(1/\psi^{(l-1)} + N^{(l)}\right) - \ln\Gamma\left(1/\psi^{(l-1)}\right)\right) \\
&+ \textstyle\sum_{k=1}^{K} \left(\ln\Gamma\left(p_k^{(l-1)}/\psi^{(l-1)} + x_k^{(l)}\right) - \ln\Gamma\left(p_k^{(l-1)}/\psi^{(l-1)}\right)\right)\Big) \\
&= \textstyle\sum_{l=1}^{L} \ln\mathcal{L}\left(\boldsymbol{p}^{(l-1)}, \psi^{(l-1)}; \boldsymbol{x}^{(l)}\right) \\
&= \textstyle\sum_{l=1}^{L} \ln\mathcal{L}\left(\boldsymbol{p}^{(l-1)+}, \psi^{(l-1)}; \boldsymbol{x}^{(l)+}\right)
\end{aligned}
\tag{29}
$$

The sum in (29) is used in our algorithm to evaluate the log-likelihood function. We can always increase L and set $\boldsymbol{x}^{(l)}$ intelligently, so that the condition (16) is satisfied for all the terms in the last formula in (29). In this case, each of the L terms in (29) can be computed using (20). This means that the log-likelihood function $\ln \mathcal{L}(\boldsymbol{p}, \psi; \boldsymbol{x})$ can be evaluated incrementally on a mesh (Fig. 3). Hence, we name this method the mesh algorithm. Note that there can be many ways to generate the mesh. We describe below how the mesh is generated in our implementation. We first create an initial mesh with the following scheme

$$x_i^{(l)} = \left\lfloor \alpha_i^{(l-1)} \delta \right\rfloor, \text{ for } l = 1, \ldots, L \tag{30}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. The level of mesh L is chosen so that it is the smallest integer satisfying

$$\sum_{l=1}^{L} x_i^{(l)} \geq x_i, \text{ for all } i = 1, \ldots, K \tag{31}$$

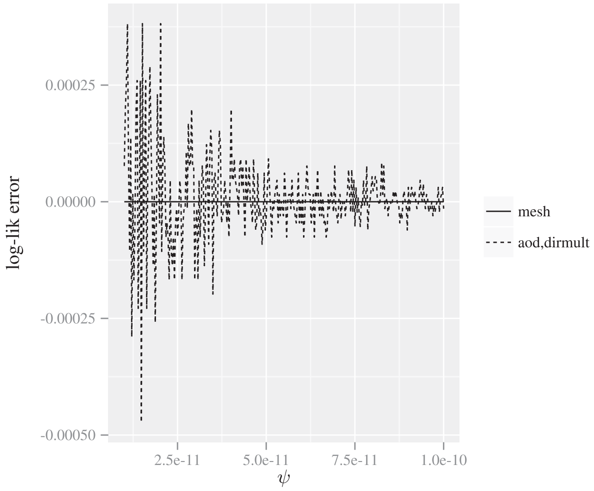

Fig. 4. The figure presents a comparison of methods for evaluation of the DMN log-likelihood function when the dispersion parameter ψ varies. For the mesh algorithm, the evaluation is accurate and stable when the dispersion parameter ψ approaches 0. The aod(dirmult) algorithm is unstable. The parameters are $\boldsymbol{x} = (2, 3, 1)$ and $\boldsymbol{p} = (.2, .3, .5)$

This initial mesh needs to be adjusted because the end of the mesh should total to match x_i exactly. To do so, let L_i' be the smallest number satisfying

$$\sum_{l=1}^{L_i'} x_i^{(l)} \geq x_i, \text{ for } i = 1, \ldots, K \tag{32}$$

We adjust $x_i^{(L_i')}$ so that $\sum_{l=1}^{L_i'} x_i^{(l)} = x_i$. For each i, all the remaining $x_i^{(l)}$ ($l > L_i'$) are set to 0. With this adjusted mesh, we can use the approximation (20) to compute the DMN log-likelihood (29). Figure 3 shows an example with $L=3$. The 3 segments of x_2 is 0; hence, there are only two non-zero segments as shown. x_1, x_3 and x_4 are broken into three non-zero segments. The last non-zero segments of all the four lines are adjusted so the segment sums equal x_i ($i = 1, 2, 3, 4$), respectively.

Note that the time complexity of the mesh algorithm is proportional to $\sum_{i=1}^{K} L_i' \sim \sum_{i=1}^{K^+} \log x_i^+$, which is smaller than $\sum_{i=1}^{K} x_i = \sum_{i=1}^{K^+} x_i^+$, the time complexity of (7) (**VGAM**). The difference becomes especially prominent for high count data (Figs 6 and 8).

5 THE NUMERICAL RESULTS

We implemented the mesh algorithm for computing the DMN log-likelihood in C++. In this section, we demonstrate the accuracy and runtime of the mesh algorithm. All experiments were run on a Linux machine with a 4-core Intel Xeon CPUs E5630@ 3.53 GHz. Each log-likelihood function call is single-threaded.

In contrast to Figure 2, Figure 4 shows that the mesh algorithm is numerically stable when ψ approaches 0.

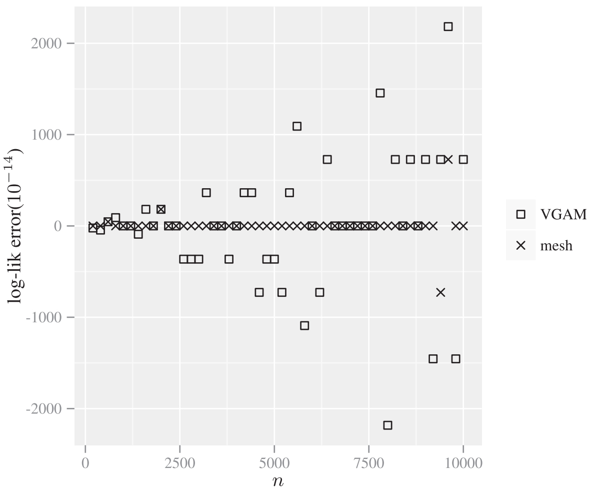

Fig. 5. The error of the mesh algorithm is smaller than the error of the method in **VGAM**. The error of the aod(dirmult) algorithm is larger than the scale presented. The parameters are $\boldsymbol{x} = n(1, 1, 1, 1)$, $\psi = 1/200$ and $\boldsymbol{p} = (.1, .2, .3, .4)$

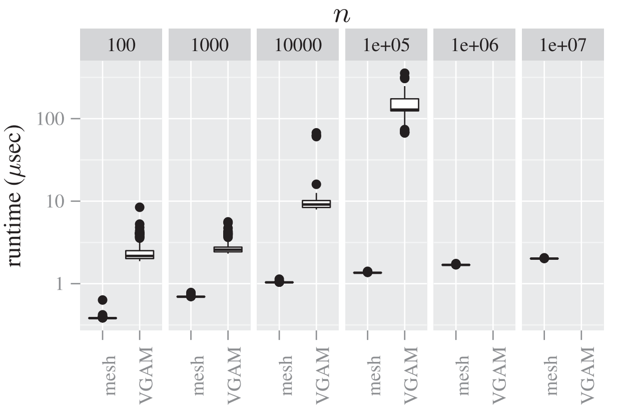

Fig. 6. The mesh algorithm is much faster than the algorithm in **VGAM** for the DMN log-likelihood computation. The parameters are $\boldsymbol{x} = n(1, 2, 3)$, $\boldsymbol{p} = (1/6, 1/3, 1/2)$ and $\psi = 1/60$. The computation using **VGAM** is only up to $n = 1 \times 10^5$, as it takes too much runtime when n is beyond this point. Each boxplot represents 100 DMN log-likelihood evaluations

We compute the error of the mesh algorithm by comparing its results with the results of an implementation of (4) in Sage (Stein *et al.*, 2012), which can achieve arbitrarily high accuracy. Figure 5 compares the error of the mesh algorithm and the error of the method in **VGAM**. We can see the mesh algorithm is more accurate.

Figure 6 shows that the runtime of the mesh algorithm increases more slowly as the counts n increase than the method in **VGAM**. Note that only R code is used to implement the method in **VGAM**, whose speed can be improved by using C++. However, its runtime scalability with respect to the parameter n is intrinsic to the representation of the log-likelihood function (7) and independent of the implementation. The slower increase of runtime is especially important for the high count that is typical in contemporary high-throughput sequencing datasets.

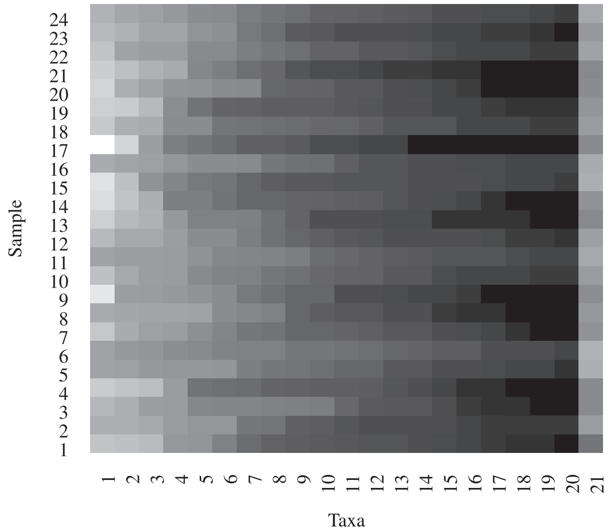

Fig. 7. The heatmap of the saliva 16 S rRNA pyrosequencing dataset. Each row represents a sample and each column represents a taxa. The brighter the color, the higher the proportion of the taxa within the sample

6 BIOINFORMATICS APPLICATION

We demonstrate below the application of our new method in analyzing human microbiome data from the Human Microbiome Project clinical production pilot study (The NCBI BioProject website, 2010). This dataset consists of the pyrosequencing of 16S rRNA genes in samples from four body sites, namely, saliva, throat, tongue and palatine tonsil of 24 human subjects (Rosa *et al.*, 2013). The sequences obtained from the V1–V3 and V3–V5 variable regions of the 16 S ribosomal RNA gene are classified into the 20 most abundant taxa at the genus level and the remaining sequences are classified as the 21st taxa (La Rosa *et al.*, 2012). Figure 7 shows the taxa distribution with each sample of the saliva dataset.

On real datasets, the mesh algorithm is much faster than the algorithm in **VGAM**. For example, the mesh algorithm improves the speed by over 50× on the saliva dataset (Fig. 8).

Because the $C(\alpha)$-based test (Kim and Margolin, 1992) rejects the hypothesis that the data from any of the four body sites are distributed according to the MN distribution (the P-values are 0 for all the four body sites), we use the DMN distribution to model each of the four datasets. Table 1 shows the maximum likelihood estimates of the dispersion parameters ψ for the data from all four body sites.

7 DISCUSSION

Overdispersion is important and needs to be accommodated in modeling count data. To handle overdispersion in MN data, the DMN distribution is commonly used. The numerical computation of the log-likelihood function is important for performing statistical inference using this distribution. Previous work has provided useful methods for this calculation, but the requirements of bioinformatics are difficult to satisfy. Our method solves the accuracy and runtime challenges.

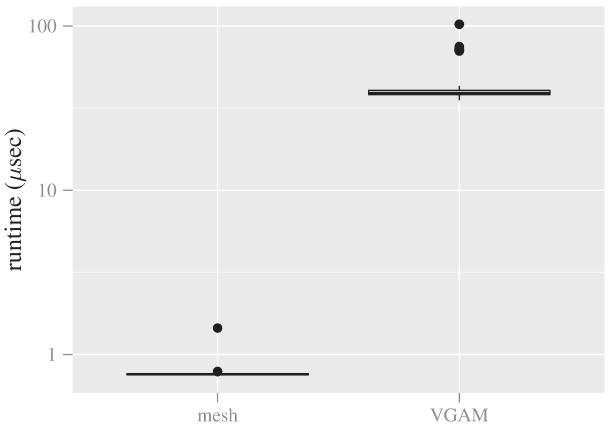

Fig. 8. The mesh algorithm is much faster than the algorithm in **VGAM** on the saliva dataset. The parameters are $\boldsymbol{p} = (0.195, 0.147, 0.117, 0.088, 0.065, 0.054, 0.041, 0.033, 0.026, 0.022, 0.019, 0.017, 0.015, 0.014, 0.012, 0.011, 0.009, 0.009, 0.008, 0.007, 0.090)$ and $\psi = 0$

Table 1. The MLE of the dispersion parameters of the 16S rRNA pyrosequencing data from all four body sites

Body site	ψ
Saliva	0.00389
Throat	0.00639
Tongue	0.00802
Tonsils	0.01039

Overdispersion is commonly found in high-throughput sequencing data. The overdispersed Poisson model (the negative-binomial distribution) has been used to detect differential gene expression. However, the DMN distribution has seen limited use in analyzing high-throughput sequencing data, possibly because the existing methods based on the DMN distribution did not anticipate the high counts and the vast amount of such count tables extracted from the high-throughput sequencing technologies.

To overcome the instability problem and the runtime problem of the existing methods for computing the log-likelihood, we derived a new approximation of the DMN log-likelihood function based on Bernoulli polynomials. Using a novel mesh algorithm, we are able to compute the log-likelihood for any parameters in the domains of the log-likelihood function. Comparing with the existing methods, the mesh algorithm is more accurate and is much faster. We demonstrate the application of the new method in analyzing human microbiome data with a large runtime improvement. This method is generally applicable to other scenarios involving proportions, such as alternative exon utilization (Wang *et al*., 2008) and alternative poly-A utilization (Lutz and Moreira, 2011). For example, suppose we have 10 000 alternative splicing events that need to be tested and each test requires 1000 log-likelihood function evaluations. Our method can reduce the runtime to hours instead of potentially days. This work paves the way for application of the DMN distribution to model overdispersion in large-scale count data available in the high-throughput sequencing era.

Funding: This work was supported by Baylor College of Medicine Computational and Integrative Biomedical Research grant internal funding for CAS (January 2013–June 2013) as well as National Institute Health U54CA149196 to M. Lewis and the startup funding to PY from the ECE department and Texas A&M Engineering Experiment Station/Dwight Look College of Engineering at Texas A&M University.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (2002) *Categorical Data Analysis. Wiley Series in Probability and Statistics*. 2nd edn. Wiley-Interscience, Hoboken, New Jersey.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*., **11**, R106.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer.

Bouguila,N. (2008) Clustering of count data using generalized Dirichlet multinomial distributions. *IEEE Trans. Knowl. Data Eng*., **20**, 462–474.

Brier,S.S. (1980) Analysis of contingency tables under cluster sampling. *Biometrika*., **67**, 591–596.

Brown,M. *et al*. (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. In: *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pp. 47–55.

Cameron,C.A. and Trivedi,P.K. (2013) *Regression Analysis of Count Data. Econometric Society Monographs*. 2nd edn. Cambridge University Press, Cambridge.

Casella,G. and Berger,R. (2002) Duxbury advanced series in statistics and decision sciences. In: *Statistical inference*. Thomson Learning, Pacific Grove, CA.

Crowder,M.J. (1978) Beta-binomial ANOVA for proportions. *Appl. Stat*., **27**, 34–37.

Farewell,D.M. and Farewell,V.T. (2013) Dirichlet negative multinomial regression for overdispersed correlated count data. *Biostatistics*., **14**, 395–404.

Fisher,R.A. (1973) *Statistical Methods for Research Workers*. 14th edn. Hafner Publishing Company, Edinburgh.

Freitag,E. and Busam,R. (2009) *Complex Analysis*. 2nd edn. Springer.

Gomez-Alvarez,V. *et al*. (2012) Metagenome analyses of corroded concrete wastewater pipe biofilms reveal a complex microbial system. *BMC Microbiol*., **12**, 122.

Haseman,J.K. and Kupper,L.L. (1979) Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*., **35**, 281–293.

Hilbe,J.M. (2011) *Negative Binomial Regression*. 2nd edn. Cambridge University Press, Cambridge, UK.

IEEE Task P754. (2008) *IEEE 754-2008, Standard for Floating-Point Arithmetic*. IEEE, New York, NY.

Kim,B.S. and Margolin,B.H. (1992) Testing goodness of fit of a multinomial model against overdispersed alternatives. *Biometrics*., **48**, 711–719.

Kleinman,J.C. (1973) Proportions with extraneous variance: single and independent sample. *J. Am. Stat. Assoc*., **68**, 46–54.

La Rosa,P.S. *et al*. (2012) Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One*, **7**, e52078.

Lauter,C. and Dinechin,F.D. (2008) Optimizing polynomials for floating-point implementation. In: *Proceedings of the 8th Conference on Real Numbers and Computers, Santiago de Compostela, Spain*.

Leckenby,J.D. and Kishi,S. (1984) The Dirichlet multinomial distribution as a magazine exposure model. *J. Mark. Res*., **21**, 100–106.

Lesnoff,M. and Lancelot,R. (2012) *aod: Analysis of Overdispersed Data*. R package version 1.3.

Lu,X. *et al*. (2013) Son connects the splicing-regulatory network with pluripotency in human embryonic stem cells. *Nat. Cell Biol*., **15**, 1141–52.

Lutz,C.S. and Moreira,A. (2011) Alternative mRNA polyadenylation in eukaryotes: an effective regulator of gene expression. *Wiley Interdiscip. Rev. RNA*, **2**, 22–31.

MacKay,D.J.C. and Bauman Peto,L.C. (1994) A hierarchical Dirichlet language model. *Nat. Lang. Eng*., **1**, 1–19.

Madsen,R.E. *et al.* (2005) Modeling word burstiness using the Dirichlet distribution. In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML'05. ACM, New York, NY, pp. 545–552.

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models. Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, New York.

Mehta,C.R. and Patel,N.R. (1986) Algorithm 643: Fexact: a fortran subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Trans. Math. Softw.*, **12**, 154–161.

Metzker,M.L. (2010) Sequencing technologies — the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Mimno,D. and McCallum,A. (2008) *Topic Models Conditioned on Arbitrary Features with Dirichlet-Multinomial Regression*. UAI, Helsinki, Finland.

Mosimann,J.E. (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika.*, **49**, 65–82.

Mosimann,J.E. (1963) On the compound negative multinomial distribution and correlations among inversely sampled pollen counts. *Biometrika.*, **50**, 47–54.

The NCBI BioProject website. (2010) Human Microbiome Project 16S rRNA Clinical Production Pilot (ID: 48335). http://www.ncbi.nlm.nih.gov/bioproject?term=48335.

Neerchal,N.K. and Morel,J.G. (2005) An improved method for the computation of maximum likelihood estimates for multinomial overdispersion models. *Comput. Stat. Data Anal.*, **49**, 33–43.

Paul,S.R. *et al.* (2005) Fisher information matrix of the Dirichlet-multinomial distribution. *Biom. J.*, **47**, 230–236.

Poortema,K. (1999) On modelling overdispersion of counts. *Stat. Neerl.*, **53**, 5–20.

R Core Team. (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.*, **26**, 139–40.

Rosa,P.S.L. *et al.* (2013) *HMP: Hypothesis Testing and Power Calculations for Comparing Metagenomic Samples from HMP*. R package version 1.3.1.

Rowe,C.H. (1931) A proof of the asymptotic series for $\log \gamma(z)$ and $\log \gamma(z+a)$. *Ann. Math., Second Ser*, **32**, 10–16.

Rust,R.T. and Leone,R.P. (1984) The mixed-media Dirichlet multinomial distribution: a model for evaluating television-magazine advertising schedules. *J. Mark. Res.*, **21**, 89–99.

Sjölander,K. *et al.* (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, **12**, 327–45.

Skellam,J.G. (1948) A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *J. R. Stat. Soc. Ser. B Methodol.*, **10**, 257–261.

Stein,W. *et al.* (2012) *Sage Mathematics Software (Version 5.0.1)*. The Sage Development Team.

Tarone,R.E. (1979) Testing the goodness of fit of the binomial distribution. *Biometrika.*, **66**, 585–590.

Tvedebrink,T. (2009) *dirmult: Estimation in Dirichlet-Multinomial Distribution*. R package version 0.1.2.

Wan,J. (2012) Global analysis of alternative polyadenylation regulation using high-throughput sequencing. PhD Thesis, University of Iowa.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature.*, **456**, 470–476.

Whittaker,E.T. and Watson,G.N. (1927) *A Course of Modern Analysis*. 4th edn. Reprinted 1990. Cambridge University Press, Cambridge, UK.

Winkelmann,R. (2008) *Econometric Analysis of Count Data*. 5th edn. Springer, Berlin, Germany.

Yee,T.W. (2010) The VGAM package for categorical data analysis. *J. Stat. Softw.*, **32**, 1–34.

Yee,T.W. (2012) *VGAM: Vector Generalized Linear and Additive Models*. R package version 0.9-0.

Yee,T.W. and Wild,C.J. (1996) Vector generalized additive models. *J. R. Stat. Soc. B*, **58**, 481–493.