Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(12) 3653–3665

DOI: 10.1177/0962280220937324
journals.sagepub.com/home/smm
SAGE

A robust score test of homogeneity for zero-inflated count data

Wei-Wen Hsu[1], David Todem[2], Nadeesha R Mawella[3], KyungMann Kim[4] and Richard R Rosenkranz[5]

Abstract

In many applications of zero-inflated models, score tests are often used to evaluate whether the population heterogeneity as implied by these models is consistent with the data. The most frequently cited justification for using score tests is that they only require estimation under the null hypothesis. Because this estimation involves specifying a plausible model consistent with the null hypothesis, the testing procedure could lead to unreliable inferences under model misspecification. In this paper, we propose a score test of homogeneity for zero-inflated models that is robust against certain model misspecifications. Due to the true model being unknown in practical settings, our proposal is developed under a general framework of mixture models for which a layer of randomness is imposed on the model to account for uncertainty in the model specification. We exemplify this approach on the class of zero-inflated Poisson models, where a random term is imposed on the Poisson mean to adjust for relevant covariates missing from the mean model or a misspecified functional form. For this example, we show through simulations that the resulting score test of zero inflation maintains its empirical size at all levels, albeit a loss of power for the well-specified non-random mean model under the null. Frequencies of health promotion activities among young Girl Scouts and dental caries indices among inner-city children are used to illustrate the robustness of the proposed testing procedure.

Keywords

Early childhood caries indices, model extension, mixture models, model misspecification, health promotion activity frequency, zero inflation

1 Introduction

Zero-inflated models provide a simple parametric framework to describe count data with extra zeros. These models view extra zeros as the result of a heterogeneous process resulting from zero counts being generated from both a random and a non-random source. They have been extensively studied in various substantive applications including engineering, medicine, oral health, and agriculture. Lambert,[1] Ridout et al.,[2] Böhning et al.,[3] Farewell et al.,[4] and references therein provide typical examples of these applications. In typical applications of these models, it is often of interest to evaluate whether the heterogeneity as implied by the mixture is consistent with observed data.[5–10] Due to its ease of implementation resulting from requiring only estimation of the null model regardless of the oftentimes complicated full model, the score test has emerged as a popular approach for this evaluation. It has long been established that this test behaves well in the finite dimensional

[1]Department of Statistics, Kansas State University, Manhattan, KS, USA
[2]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA
[3]Department of Mathematics and Statistics, University of Missouri-Kansas City, Kansas City, MO, USA
[4]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA
[5]Department of Food, Nutrition, Dietetics and Health, Kansas State University, Manhattan, KS, USA

Corresponding author:
Wei-Wen Hsu, Department of Statistics, 108C Dickens Hall, Kansas State University, Manhattan 66506, KS, USA.
Email: wwhsu@ksu.edu

setting under correct specification of the working model.[11–15] However, under model misspecification, it is unclear how the test would perform both asymptotically and in finite sample settings. By misspecification, we mean instances where the working model is restrictive in view of the true but unknown model. In this paper, we are specifically interested in situations where model misspecification arises from the mean component of the model being incorrectly specified (e.g. important variables being ignored in the working model or the functional form relating the mean to covariates being misspecified).

The issue of model misspecification in parametric modeling is always a concern, both from the estimation and inference perspective, due to the true data generating mechanism being unknown in practical settings. In this paper, we investigate the impact of model misspecification on the test of zero inflation for the popular zero-inflated models, and show how this impact can be mitigated in real applications. More specifically, we propose a testing procedure that is robust against the stated misspecifications. Unlike existing tests that require the mean component of the working model to be fully specified, our proposed test can be performed without such specification and hence its protection from this type of misspecification. Conceptually, our approach consists of extending the model to a two-level hierarchical formulation by imposing a layer of randomness on the mean component of the homogeneous model to account for uncertainty in the model specification. We exemplify this approach on the test of zero inflation for the class of zero-inflated Poisson models, where a random term is imposed on the Poisson mean to adjust for relevant covariates missing from the mean model or a misspecified functional form. We consider two scenarios which are often encountered in real applications. The first scenario occurs when the analyst has no knowledge of important covariates for the mean model in which case our approach simply treats the mean as an unobserved random variable. The second scenario occurs when the analyst has partial information on important covariates for the mean model, in which case our approach simply adds an unobserved random term on the mean model being entertained. A computational advantage of our approach is that it imposes a gamma distribution on the unobserved random terms and exploits the Poisson-Gamma conjugacy to avoid numerical integration. Because the marginal homogeneous process (average across the random mean) is negative binomial (NB), it provides an important added value for homogeneity testing vis-a-vis robustness, especially in settings where the true homogeneous distribution is Poisson. From a modeling perspective, a similar approach was proposed by Kassahun et al.[16] to represent correlated count data with extra zeros in addition to overdispersion. These authors focused their effort primarily on the full model fitting, but did not provide any inference on goodness-of-fit (e.g. test of zero inflation) with respect to observed data. Without such evaluation, we argue that the applied analysts may be required to interpret an unnecessarily complex mixture model. This necessitates a need for a critical evaluation of the value of zero-inflated models relative to easily interpretable non mixture-based (one component) models.

The rest of this paper is organized as follows. The zero-inflated models and the classical score test for homogeneity are briefly introduced in Section 2. In Section 3, we conduct a preliminary simulation study to evaluate the impact of two popular misspecifications on the classical score test for zero inflation. In Section 4, we propose a robust homogeneity test for zero inflation in count data models. A simulation study to evaluate the finite sample performance of the proposed test is conducted and presented in Section 5. The proposed test is applied to dental caries indices among inner city Africa-American children in the Detroit Dental Health Project (DDHP) and the frequency of health and nutrition promotion activities among participants in the Scouting Nutrition and Activity Program (SNAP) in Section 6. Discussion and conclusion are given in Section 7.

2 Zero-inflated model and score test of homogeneity

2.1 Zero-inflated model

Suppose that a random sample of n independent subjects with count responses $Y_i, i = 1, \ldots, n$ are drawn from a population governed by a mixture of a point mass at 0 and a discrete distribution with the probability mass function $g_i(.;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. The baseline distribution in zero-inflated models shall refer to $g_i(.;\boldsymbol{\theta})$. The probability mass function of Y_i is

$$P(Y_i = y) = \begin{cases} \omega_i + (1-\omega)\ g_i(0;\boldsymbol{\theta}) & \text{if } y = 0 \\ (1-\omega_i)\ g_i(y;\boldsymbol{\theta}) & \text{if } y > 0 \end{cases}$$

where ω_i is the mixing weight. In general, this weight is a probability constrained between 0 and 1, but a more relaxed constraint $-g_i(0;\boldsymbol{\theta})/(1-g_i(0;\boldsymbol{\theta})) \leq \omega_i \leq 1$, for $i = 1,\ldots,n$, can be entertained to accommodate both zero inflation and zero deflation.[5,6,9] When $\omega_i = 0$, for all i, the mixture model reduces to a homogeneous model which corresponds to the baseline distribution $g_i(.;\boldsymbol{\theta})$. Depending on the baseline distribution, we have the popular two-component mixture models such as zero-inflated Poisson (ZIP) models, zero-inflated negative binomial (ZINB) models and zero-inflated binomial (ZIB) models.[1,17,18]

2.2 Classical score test of homogeneity

As a goodness of fit, it is often of interest to evaluate the mixing weight by setting $\omega_i = 0$, $i = 1,\ldots,n$. For simplicity, many testing procedures in the literature assume a constant mixing weight such that a single ω is evaluated for all i. For example, Van den Broek[5] proposed a score test under ZIP models assuming a constant mixing weight. Deng and Paul[6] also adopted this constancy assumption to perform a score test under ZIB models. Jansakul and Hinde[7,19] further extended the idea to a covariate-adjusted score test by assuming that the mixing weight depends on covariates via an identity link function in order to improve the power of test. Their test can access $H_0 : \omega_i = 0$ for all i versus $H_1 : \omega_i \neq 0$ for some i for ZIP and ZINB models. They have shown that the test adopting a constant mixing weight, such as Van den Broek's test, is just a special case of their approach. Another similar extension to a covariate-adjusted score test in a more general setting is proposed by Todem et al.[9]

In this paper, we adopt the constancy assumption and focus our evaluation on the null hypothesis $H_0 : \omega = 0$ against the two-sided alternative $H_1 : \omega \neq 0$ consistent with the wider support set of ω. The associated score test statistic can be easily derived by following the regular asymptotic theory.[11,20] Specifically, the score test statistic for the class of zero-inflated models is given by

$$S_T = S_\omega(\hat{\boldsymbol{\theta}}, 0)^T \hat{V}_\omega^{-1} S_\omega(\hat{\boldsymbol{\theta}}, 0)$$

where $S_\omega(\hat{\boldsymbol{\theta}}, 0)$ is the partial score function with respect to ω and evaluated at $\omega = 0$ and $\hat{\boldsymbol{\theta}}$ a consistent estimator of $\boldsymbol{\theta}$ under the null hypothesis, and $\hat{V}_\omega$ is the estimated variance of the partial score function $S_\omega(\hat{\boldsymbol{\theta}}, 0)$. This score test statistic follows a χ^2 distribution with one degree of freedom asymptotically under the null hypothesis.

3 Misspecification schemes

Even though the behavior of the score test under proper model specification is well understood, it is amenable to produce unreliable inferences if the working null model $g_i(.;\theta)$ containing the nuisance parameter is misspecified. Misspecifications may be due to the mean function of the homogeneous model not being well specified in terms of important covariates and the functional forms. Table 1 is a matrix representing the profile of the mean model specification both in covariates and the functional form. In this paper, we focus on the practical situations where the practicing statistician has no knowledge of important covariates or the functional form relating covariates to mean response. Because of this lack of knowledge of the true data generating mechanisms, misspecification is a concern that requires a careful examination.

We conduct a preliminary simulation study to evaluate the empirical size of the score test under the misspecification stated in Table 1. As a working example, let X_1 and X_2 be two independent covariates, and assume a Poisson process for which the true mean model is $\lambda^* = \exp\{0.6 + 0.45X_1\}$, but misspecified in working models as $\lambda = \exp\{\beta_0\}$ or $\lambda = \exp\{\beta_0 + \beta_2 X_2\}$. Even when the distribution is well specified, the simulation results indicate that the score test becomes very liberal, rejecting the null hypothesis more often than anticipated (see Table 2). This result holds even when the sample size is large. We observe similar results when the model misspecification arises from wrongly assuming a log-linear model for the mean component (results not shown). This limited simulation study suggests that the test of zero-inflation should be carefully interpreted regarding the potential model misspecifications. This then necessitates the development of a score test of homogeneity that is robust against these popular forms of misspecification.

4 A robust homogeneity test for zero-inflated models

We propose a robust score test for homogeneity for the popular ZI models, focusing on the ZIP models. In its basic formulation, the proposed testing procedure avoids specifying any functional form of the mean response as a function of covariates. Instead, the method relies on a hierarchical model for which the mean component is simply

Table 1. The mean model specification schemes.

	The functional form	
	Correctly specified	Misspecified
Relevant covariates		
Included	Well-specified model[a]	Poor-specified model[b]
Ignored	Poor-specified model[b]	Poor-specified model[b]

[a]The behavior of the score test for the well-specified model is well understood.
[b]The behavior of the score test is poorly understood.

Table 2. Empirical sizes of the score test statistics at the nominal level 0.05 using 1,000 Monte Carlo samples of size n generated from a null model with true mean $\lambda^* = \exp\{0.6 + 0.45X_1\}$.

	Null model		n				
Working mean function	True	Working	50	100	200	500	1000
Misspecification of the mean function							
$\log(\lambda) = \beta_0$	Poisson	Poisson	0.064	0.078	0.082	0.091	0.144
$\log(\lambda) = \beta_0 + \beta_2 X_2$	Poisson	Poisson	0.071	0.087	0.092	0.095	0.137

$X_1 \sim U(0, 1)$ and $X_2 \sim$ truncated normal(0,1) within the interval $(-1, 1)$.

regarded as an unobserved random variable, assuming the analyst has no knowledge of relevant covariates. We impose a gamma distribution on these unobserved random variables, and exploit the Poisson-Gamma conjugacy to avoid numerical integration over the unobserved random mean. For situations where partial knowledge of relevant covariates is available, we extend the proposed test to incorporate covariates for efficiency gain.

4.1 Robust test with no knowledge of covariates

In settings where the analyst has no knowledge of the true mean in terms of important covariates or its functional form, we consider a hierarchical model for which the mean is simply a positive random variable. We assume that $Y_i, i = 1, \ldots, n$ are independent realizations from the hierarchical model

$$Y_i|\Lambda_i \sim \begin{cases} 0, & \text{with weight } \omega_i \\ \text{Poisson}(\Lambda_i), & \text{with weight } 1-\omega_i \end{cases} \quad \text{and} \quad \Lambda_i \sim \text{Gamma}(\alpha, \beta) \tag{1}$$

where ω_i is the mixing weight and α and β are the shape and scale parameters, respectively. In this model, we avoid specifying the mean of the Poisson model, and instead use a general random mean Λ_i. Because the random mean is unobserved, the marginal distribution of $f_Y(y)$ obtained by integrating out Λ is

$$f_Y(y_i) = \int_0^\infty f_{Y|\Lambda}(y_i|\lambda_i) f_\Lambda(\lambda_i) \mathrm{d}\lambda_i = I_{(y_i=0)}\omega_i + (1-\omega_i)\int_0^\infty \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\frac{\lambda_i^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}e^{-\lambda_i/\beta}\mathrm{d}\lambda_i$$

As the result, the zero-inflated model can be re-expressed as

$$P(Y_i = y_i) = \begin{cases} \omega_i + (1-\omega_i)\, f_{Y^*}(0), & \text{if } y_i = 0 \\ (1-\omega_i)\, f_{Y^*}(y_i), & \text{if } y_i = 1, 2, 3, \ldots \end{cases}$$

where the density function of the baseline distribution is

$$f_{Y^*}(y_i) = \frac{\Gamma(y_i+\alpha)}{y_i!\ \Gamma(\alpha)}\left(\frac{\beta}{1+\beta}\right)^{y_i}\left(\frac{1}{1+\beta}\right)^{\alpha}, \quad y_i = 0,1,2,3,\ldots$$

This baseline distribution $f_{Y^*}(y_i)$ is actually a NB with mean $\mathrm{E}(Y^*) = \alpha\beta$ and variance $\mathrm{Var}(Y^*) = \alpha\beta(1+\beta)$. Hence, with an additional dispersion parameter α, a Poisson baseline distribution can be naturally accommodated by this particular zero-inflated model, which indeed is a larger class of model in view of Poisson. Intuitively, we are using a more general and larger model to avoid any potential misspecifications on the mean component.

The log-likelihood function, given independent observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ and assuming $\omega_i = \omega$ for all i, is

$$\begin{aligned} l(\alpha,\beta,\omega;\boldsymbol{y}) = \sum_{i=1}^{n}\{&\mathrm{I}_{(y_i=0)}\log[\omega + (1-\omega)(1+\beta)^{-\alpha}] + \mathrm{I}_{(y_i>0)}[\log(1-\omega) + \log\Gamma(y_i+\alpha) \\ &- \log\Gamma(\alpha) - \log\Gamma(y_i+1) + y_i\log(\beta) - (y_i+\alpha)\log(1+\beta)]\} \end{aligned}$$

Under the null hypothesis $H_0 : \omega = 0$, the proposed robust score test is

$$S_T = S_\omega(\hat{\alpha},\hat{\beta},0)^T \hat{V}_\omega^{\ -1} S_\omega(\hat{\alpha},\hat{\beta},0)$$

where $S_\omega(\hat{\alpha},\hat{\beta},0) = \sum_{i=1}^{n}\left\{\mathrm{I}_{(y_i=0)}[(1+\hat{\beta})\hat{\alpha} - 1]\right\}$ is the partial score function, with $\hat{\alpha}$ and $\hat{\beta}$ being the maximum likelihood estimates of α and β under the null hypothesis. The estimate $\hat{V}_\omega$ is given by $\hat{V}_\omega = I_{\omega\omega}(\hat{\alpha},\hat{\beta},0) - I^{\omega\alpha\beta}(\hat{\alpha},\hat{\beta},0)^T[I^{\alpha\beta}(\hat{\alpha},\hat{\beta},0)]^{-1} I^{\omega\alpha\beta}(\hat{\alpha},\hat{\beta},0)$, where

$$I^{\omega\alpha\beta}(\hat{\alpha},\hat{\beta},0) = \begin{bmatrix} I_{\omega\alpha}(\hat{\alpha},\hat{\beta},0) \\ I_{\omega\beta}(\hat{\alpha},\hat{\beta},0) \end{bmatrix} \quad \text{and} \quad I^{\alpha\beta}(\hat{\alpha},\hat{\beta},0) = \begin{bmatrix} I_{\alpha\alpha}(\hat{\alpha},\hat{\beta},0) & I_{\alpha\beta}(\hat{\alpha},\hat{\beta},0) \\ I_{\beta\alpha}(\hat{\alpha},\hat{\beta},0) & I_{\beta\beta}(\hat{\alpha},\hat{\beta},0) \end{bmatrix}$$

Matrices $I_{\omega\omega}(\hat{\alpha},\hat{\beta},0)$, $I_{\omega\alpha}(\hat{\alpha},\hat{\beta},0)$, $I_{\omega\beta}(\hat{\alpha},\hat{\beta},0)$, $I_{\alpha\alpha}(\hat{\alpha},\hat{\beta},0)$, $I_{\beta\beta}(\hat{\alpha},\hat{\beta},0)$, $I_{\alpha\beta}(\hat{\alpha},\hat{\beta},0)$, and $I_{\beta\alpha}(\hat{\alpha},\hat{\beta},0)$ are submatrices of the Fisher information matrix evaluated under H_0. Under the null hypothesis, this score test statistic S_T follows a χ^2_1 distribution asymptotically. Details are relegated to Supplementary material Appendix A.

4.2 Robust test with partial knowledge of covariates

A limitation of the above proposed testing procedure, albeit its robustness against the stated misspecifications, is that it may not be efficient in settings where the analysts have a good knowledge of important variables. To improve efficiency in this type of settings, a simple extension would consist of entertaining a mixed effects model in which both random effects and fixed effects are entertained in the model in the spirit of Kassahun et al.[16] More specifically, the following hierarchical model is entertained

$$Y_i|\Lambda_i \sim \begin{cases} 0, & \text{with weight } \omega_i \\ \mathrm{Poisson}(\Lambda_i f_\gamma(X_i)), & \text{with weight } 1-\omega_i \end{cases} \text{ and } \Lambda_i \sim \mathrm{Gamma}(\alpha,\beta)$$

Here Y_i $(i = 1, \ldots, n)$ are independent count observations, ω_i is the mixing weight and $f_\gamma(X_i)$ is a non-negative and non-random finite dimensional function in covariate X_i. A good example for such functions is the log-linear model $f_\gamma(X_i) = \exp\{X_i\gamma\}$, where the covariate vector $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and $\gamma = (\gamma_1, \cdots, \gamma_p)^T$. The assumption of multiplicative effect $f_\gamma(X_i)$ is partly driven by the nature of Poisson mean as adopted by Molenberghs et al.[21] and Kassahun et al.[16] The advantage of using the log link for $f_\gamma(X_i)$ is to reflect that with count data the effects of predictors are often multiplicative.[22,23] For model identifiability purposes, no intercept is included in the log-linear model. When there is no knowledge of covariates, this hierarchical model reduces to the simple model in equation (1) by letting $X_i = 0$. Overall, it is straightforward to see that if an important covariate is ignored from the model, the random component Λ_i may help alleviate the impact of such misspecification.

In this formulation, the baseline distribution is again a NB with mean $\mathrm{E}(Y_i^*) = \alpha\beta e^{X_i\gamma}$ and variance $\mathrm{Var}(Y_i^*) = \alpha\beta e^{X_i\gamma}(1 + \beta e^{X_i\gamma})$. As we can see the important covariates X_i are accommodated into its marginal mean and variance to systematically characterize the variability in the data, thus it is anticipated that the testing efficiency can be improved.

The log-likelihood function $l(\alpha, \beta, \boldsymbol{\gamma}, \omega; \boldsymbol{y})$ of this hierarchical model is given as

$$\begin{aligned} l(\alpha, \beta, \boldsymbol{\gamma}, \omega; \boldsymbol{y}) = \sum_{i=1}^{n} \{ & \mathrm{I}_{(y_i=0)} \log[\omega + (1-\omega)(1+\beta e^{X_i\gamma})^{-\alpha}] \\ & + \mathrm{I}_{(y_i>0)} [\log(1-\omega) + \log\Gamma(y_i+\alpha) - \log\Gamma(\alpha) - \log\Gamma(y_i+1) \\ & + y_i \log(\beta e^{X_i\gamma}) - (y_i+\alpha)\log(1+\beta e^{X_i\gamma})] \} \end{aligned}$$

As the score vector and the information matrix can be obtained using the above log-likelihood function, one can follow the standard procedure to conduct the robust score test easily. The formulations of these associated score vector and the information matrix are given in Supplementary material Appendix B. It is expected that this test will be more powerful than the one without any covariates, as long as the important covariates are included, regardless of its functional form, into the working mean component.

5 Simulation study

The empirical performance of the proposed test is evaluated by investigating the sizes and powers of the test under misspecification of the mean function. We generate data from ZIP models with various true mean functions for assessing the empirical sizes and powers of the proposed robust test. We consider the true means $\log(\lambda^*) = 1$, $\log(\lambda^*) = 1 - 0.4X_1$, $\log(\lambda^*) = 1 - 0.4X_1 - 0.25X_2$, and a non-log-linear form $\lambda^* = 3 - 1.2X_1 + 0.24X_1^2$, where $\lambda^* > 0$ with X_1 and X_2 being two independent covariates generated from uniform $U(0, 1)$ and Bernoulli $Ber(0.6)$, respectively. We use the integrated mean squared error (IMSE) of the estimated marginal mean relative to its true counterpart throughout the covariate profile to measure the magnitude of misspecifications. Throughout all simulations, we perform the proposed robust score test with/without covariates and the correctly specified test (i.e. Van den Broek's test[5]) for comparison purposes. All simulations are conducted with 1000 Monte Carlo samples for sample sizes 50, 100, 200, 400, and 800.

We first generate data from homogeneous Poisson models (i.e. the null model) with the stated true mean functions for assessing the empirical sizes of the proposed robust test. The empirical sizes of the tests are given in Table 3. Overall, the size of the proposed test tends to be conservative but remains stable with an increasing sample size. When the working mean is correctly specified in terms of covariates, the proposed robust test can maintain its size at the nominal level of 0.05 in large sample size. This is also true for over-fitted mean models (e.g. an irrelevant covariate is included into the model). For example, the true mean model only involves X_1, but it is over-specified with X_1 and X_2. For such a case, we observe that their associated IMSEs decrease when the sample size increases. This result shows that the impact of misspecification has been relieved as long as an important covariate is incorporated into the working mean model. In contrast, when an important covariate is omitted from the working mean model, the proposed tests tend to be conservative and their associated IMSEs remain large even in a large sample size. A good example in our simulation is that the true mean function is $\log(\lambda^*) = 1 - 0.4X_1$, but the working mean is assumed to be independent of any covariates (i.e. using no covariate). The result indicates that the proposed tests can entertain their robustness but cannot alleviate the impact of misspecification when important covariates are ignored.

We further evaluate the test when the working mean function has a different functional form from the true mean. For instance, the true mean is $\lambda^* = 3 - 1.2X_1 + 0.24X_1^2$ (i.e. a quadratic function) but the working mean is specified as either a function independent of any covariates or a log-linear function of covariates. If the important covariate is totally ignored from the testing procedure, the size of our robust test (without using any covariates) remains stable but slightly conservative. When we include important covariates in the working mean, the proposed robust test (with partial knowledge of covariates) can lessen the effect of misspecification and possibly maintain its size at the nominal level. This is particularly true when the working mean can well approximate the true mean. A good example is that the true mean $\lambda^* = 3 - 1.2X_1 + 0.24X_1^2$ can be well approximated by a working mean function $\exp\{\beta_0 + \beta_1 X_1\}$. This can be easily shown by a Taylor series approximation.

Table 3. Empirical sizes ($\omega^* = 0$) of the robust score test statistics and the Integrated Mean Squared Errors (IMSEs) for estimating marginal mean using 1000 samples generated from Poisson regression models with true mean λ^*, at the nominal level 0.05.

			Sample size n				
True mean function	Test	Working mean function	50	100	200	400	800
$\log(\lambda^*) = 1$	Robust test	Using no covariate	0.031	0.024	0.036	0.046	0.048
			(0.055)	(0.028)	(0.014)	(0.007)	(0.003)
		Using only X_1	0.030	0.026	0.037	0.044	0.049
			(0.111)	(0.055)	(0.027)	(0.014)	(0.007)
		Using only X_2	0.028	0.023	0.039	0.043	0.050
			(0.111)	(0.055)	(0.028)	(0.014)	(0.007)
		Using X_1 and X_2	0.027	0.024	0.037	0.048	0.054
			(0.167)	(0.082)	(0.041)	(0.021)	(0.010)
	Van den Broek's test[a]		0.038	0.042	0.045	0.051	0.054
$\log(\lambda^*) = 1 - 0.4X_1$	Robust test	Using no covariate	0.024	0.021	0.018	0.038	0.030
			(0.108)	(0.088)	(0.078)	(0.072)	(0.069)
		Using only X_1	0.026	0.026	0.026	0.047	0.059
			(0.087)	(0.046)	(0.022)	(0.011)	(0.006)
		Using only X_2	0.028	0.026	0.020	0.041	0.032
			(0.150)	(0.110)	(0.090)	(0.078)	(0.072)
		Using X_1 and X_2	0.027	0.032	0.025	0.047	0.063
			(0.131)	(0.069)	(0.034)	(0.017)	(0.009)
	Van den Broek's test[a]		0.043	0.036	0.043	0.057	0.060
$\log(\lambda^*) = 1 - 0.4X_1 - 0.25X_2$	Robust test	Using no covariate	0.019	0.019	0.013	0.011	0.018
			(0.145)	(0.128)	(0.120)	(0.115)	(0.112)
		Using only X_1	0.020	0.025	0.018	0.024	0.028
			(0.135)	(0.097)	(0.079)	(0.069)	(0.064)
		Using only X_2	0.021	0.026	0.022	0.023	0.028
			(0.126)	(0.090)	(0.071)	(0.061)	(0.056)
		Using X_1 and X_2	0.024	0.030	0.036	0.039	0.059
			(0.119)	(0.060)	(0.031)	(0.015)	(0.007)
	Van den Broek's test[a]		0.033	0.047	0.052	0.044	0.051
$\lambda^* = 3 - 1.2X_1 + 0.24X_1^2$	Robust test	Using no covariate	0.016	0.026	0.025	0.029	0.030
			(0.126)	(0.102)	(0.089)	(0.083)	(0.080)
		Using only X_1	0.020	0.033	0.039	0.044	0.047
			(0.102)	(0.051)	(0.025)	(0.012)	(0.006)
		Using only X_2	0.018	0.027	0.027	0.031	0.028
			(0.172)	(0.125)	(0.102)	(0.089)	(0.083)
		Using X_1 and X_2	0.018	0.041	0.037	0.048	0.045
			(0.151)	(0.075)	(0.038)	(0.019)	(0.009)
	Van den Broek's test[b]		–	–	–	–	–

$X_1 \sim U(0, 1)$ and $X_2 \sim Ber(0.6)$; IMSEs are given in the parentheses.
[a]Van den Broek's test with the correctly specified mean.
[b]Van den Broek's test is under misspecification when the true mean function is not log-linear.

Overall, the size of the proposed test can be well maintained at the nominal level with large sample size, as long as important covariates are included. If there is no knowledge of covariates, the proposed test tends to be slightly conservative but very stable for all sample sizes. Even though the proposed robust test behaves conservative, the test still retains its robustness property against any misspecification of the mean.

Next, we evaluate the powers of the tests by generating the data from ZIP models with the true mixing weight $\omega^* = 0.1$ (i.e. the alternative model). We consider the simulation schemes where the working mean may be well specified, misspecified or over-fitted to evaluate the performance of our robust tests. We also conduct the well-specified score test (i.e. Van den Broek's test) in the simulation for comparison purposes. As expected, the power of the robust test increases as the sample size increases (see Table 4). The proposed robust test gains more power when the working mean is well specified (i.e. important covariates are included). For example, the robust test using covariate X_1 outperforms the test that uses no covariate or only X_2 when the true mean is

Table 4. Empirical powers ($\omega^* = 0.1$) of the robust score test statistics and the Integrated Mean Squared Errors (IMSEs) for estimating marginal mean using 1,000 samples generated from zero-inflated Poisson regression models with true mean λ^*, at the nominal level 0.05.

			Sample size n				
True mean function	Test	Working mean function	50	100	200	400	800
$\log(\lambda^*) = 1$	Robust test	Using no covariate	0.148	0.308	0.534	0.849	0.951
			(0.134)	(0.099)	(0.088)	(0.085)	(0.078)
		Using only X_1	0.153	0.310	0.519	0.842	0.938
			(0.201)	(0.132)	(0.104)	(0.093)	(0.082)
		Using only X_2	0.155	0.315	0.536	0.838	0.951
			(0.191)	(0.129)	(0.104)	(0.093)	(0.082)
		Using X_1 and X_2	0.127	0.283	0.518	0.845	0.956
			(0.259)	(0.162)	(0.120)	(0.101)	(0.086)
	Van den Broek's test[a]		0.494	0.788	0.973	0.999	0.999
$\log(\lambda^*) = 1 - 0.4X_1$	Robust test	Using no covariate	0.083	0.126	0.250	0.455	0.769
			(0.160)	(0.142)	(0.128)	(0.123)	(0.120)
		Using only X_1	0.089	0.131	0.249	0.513	0.848
			(0.144)	(0.103)	(0.074)	(0.064)	(0.057)
		Using only X_2	0.091	0.137	0.250	0.450	0.781
			(0.205)	(0.167)	(0.141)	(0.129)	(0.123)
		Using X_1 and X_2	0.072	0.134	0.256	0.496	0.837
			(0.192)	(0.128)	(0.087)	(0.070)	(0.060)
	Van den Broek's test[a]		0.323	0.581	0.842	0.986	0.999
$\log(\lambda^*) = 1 - 0.4X_1 - 0.25X_2$	Robust test	Using no covariate	0.081	0.108	0.147	0.252	0.477
			(0.182)	(0.163)	(0.157)	(0.151)	(0.150)
		Using only X_1	0.084	0.092	0.147	0.253	0.532
			(0.177)	(0.134)	(0.117)	(0.106)	(0.103)
		Using only X_2	0.074	0.083	0.142	0.281	0.561
			(0.170)	(0.128)	(0.109)	(0.098)	(0.095)
		Using X_1 and X_2	0.064	0.083	0.128	0.292	0.623
			(0.167)	(0.100)	(0.070)	(0.053)	(0.047)
	Van den Broek's test[a]		0.202	0.416	0.683	0.949	0.999
$\lambda^* = 3 - 1.2X_1 + 0.24X_1^2$	Robust test	Using no covariate	0.127	0.170	0.335	0.662	0.878
			(0.190)	(0.164)	(0.148)	(0.146)	(0.143)
		Using only X_1	0.117	0.199	0.387	0.740	0.917
			(0.177)	(0.117)	(0.087)	(0.077)	(0.070)
		Using only X_2	0.113	0.181	0.343	0.663	0.863
			(0.244)	(0.190)	(0.162)	(0.152)	(0.146)
		Using X_1 and X_2	0.096	0.171	0.381	0.719	0.912
			(0.234)	(0.145)	(0.101)	(0.083)	(0.073)
	Van den Broek's test[b]		–	–	–	–	–

$X_1 \sim U(0, 1)$ and $X_2 \sim Ber(0.6)$; IMSEs are given in the parentheses.
[a]Van den Broek's test with the correctly specified mean.
[b]Van den Broek's test is under misspecification when the true mean function is not log-linear.

$\log(\lambda^*) = 1 - 0.4X_1$. It is also true for settings where the working mean is over-fitted. We observe similar results when both X_1 and X_2 are included in the working mean while the true mean is $\log(\lambda^*) = 1 - 0.4X_1$. Notice that the existing classical score tests of homogeneity for zero-inflated models will suffer from model misspecification badly if the true mean function is not a log-linear function. That is because these existing tests are all assuming a log-linear working mean.

Overall, compared to the well-specified test, the robust test could lose some power in detecting the heterogeneity in the data. But the robust test can gain power when important covariates are used. This result is anticipated because when we know more about the important covariates, then we can better alleviate the impact of misspecification. When there is no knowledge of covariates, our robust test becomes conservative as we observed from the simulations. This finding indicates that conservativeness comes as the price for having the robustness

Table 5. Comparison of score test statistics, degrees of freedom, and the associated p-values of homogeneity tests for the girl scouts data.

	Van den Broek's test	The proposed robust test	
		Without covariates	Intervention (X_1)
df	1	1	1
Test statistic	56.856	0.075	1.065
p-Value	<0.001	0.784	0.302

under any mean misspecifications, showing the tradeoff between the statistical power and the robustness when using the proposed test.

6 Real data application

6.1 Girl Scouts data

To illustrate the proposed test, we use the Girl Scouts data from the SNAP study.[24] The objective of this study was to evaluate the effectiveness of an intervention program designed to improve the physical activity and nutrition environment in the Girl Scout troops. In this study, three Girl Scout troops were randomized to the intervention arm and four to the control arm. In the intervention arm, troop leaders were trained to implement policies promoting physical activity and healthful eating opportunities at troop meetings. In the control troops, the leaders were not given any training to implement such promotions. At each of the seven troop meetings taking place between October 2007 and April 2008, a trained research assistant observed and counted the number of times health and nutrition promotions implemented by troop leaders. Research assistants were blinded to the condition of each troop (for more details, see Schlechter et al.[25]).

For our analysis, we are interested in the frequency of nutrition promotion activities occurring every 5 min at the troop meeting. We conduct the proposed robust score test along with Van den Broek's score test in which the baseline distribution is Poisson and the mixing weight is a constant. For Van den Broek's test, the mean is $\lambda = \exp\{\beta_0 + \beta_1 X_1\}$ where X_1 is an indicator variable of the intervention (1 = intervention group, 0 = control group). The test results are presented in Table 5.

Van den Broek's score test rejects the null hypothesis at a 0.05 significance level. However, our proposed robust tests with/without covariates fail to reject the null hypothesis, which are in favor of the model under the null hypothesis. In other words, the robust tests are inclined to suggest a homogeneous model, rather than a more complex model. To further evaluate these test results, we compare several count models using Bayesian Information Criteria (BIC) (see results in Table 6). The mean function in each count model is assumed as $\lambda = \exp\{\beta_0 + \beta_1 X_1\}$ where X_1 is an indicator variable for the intervention. Among these count models, the NB model shows the smallest BIC (BIC = 279.87), suggesting a better model fit compared to the ZIP model which has a BIC of 312.29. Figure 1(a) further supports the relative good fit of the NB model compared to the ZIP model. This result is in agreement with the results of our robust test and that of Jansakul and Hinde's test[19] (test statistic= 0.030, p-value= 0.862) which incorporates covariates. In Table 6, it is interesting to see the opposite signs of the estimated mixing weights for the ZIP and ZINB models. Although the negative mixing weight is not significant in the ZINB model, it does indicate a sign of zero deflation in the data. Compared to the ZINB model, the ZIP model indicates the presence of zero inflation but it does not provide a better model fit to the data. In fact, the ZIP model can incorporate extra zeros in the data but cannot handle overdispersion nicely. The real data exhibit certain overdispersion, thus this may explain why we observe the opposite results in model estimation between the ZIP and ZINB models.

We additionally conduct a stratified analysis to investigate whether both zero inflation and deflation are present in the data. Figure 1(b) and (c) show the observed and predicted proportions with a homogeneous NB model for the intervention group and the control group, respectively. In these figures, the fitted proportions are very close to the observed proportions, suggesting the data could be well fitted by a NB model, showing no evidence for a mixture model. We further examine the mixing weights at zero for both groups using Wald test. The mixing weight is neither significant (p-value = 0.793) for the intervention group nor for the control group (p-value = 0.991); thus, there is no evidence to support the presence of zero inflation or deflation in the data.

Table 6. Fits of different count models for the Girl Scouts data.

Model	Estimated coefficient for intervention $\hat{\beta}_1$	Estimated mixing weight $\hat{\omega}$	BIC
Poisson	1.965 (<0.001)	–	360.36
NB	1.965 (<0.001)	–	279.87
ZIP	1.775 (<0.001)	0.487 (< 0.001)	312.29
ZINB	2.023 (<0.001)	–2.696 (0.864)	283.90

Note: *p*-Value of Wald test is given in the parentheses.

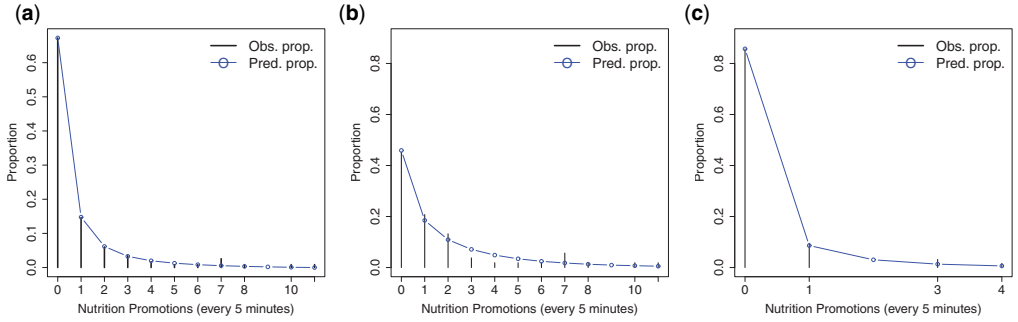

Figure 1. Observed versus fitted proportions by the NB model for all participants (a), the intervention group (b) and the control group (c).

Table 7. Comparison of score test statistics and the associated *p*-values for the dental caries data.

		The proposed robust test			
	Van den Broek's test	Without covariates	Age	SI	Age and SI
df	1	1	1	1	1
Test statistic	740366.9	37.786	38.978	14.737	134.780
p-value	<0.001	<0.001	<0.001	<0.001	<0.001

Clearly, a zero-inflated model is too complex for the Girl Scouts data. These additional results also support our robust test, and indicating our test's robustness and effectiveness.

It is worth mentioning that the coefficient of X_1 in NB is significant (estimated coefficient = 1.965, p-value < 0.001, Table 6), indicating the intervention program has a significant effect on the Nutrition Promotions implemented by Girl Scout leaders.

6.2 Dental caries data

We further illustrate the proposed robust score test with the dental caries data from the DDHP study which was designed to assess dental caries severity of low-income African-American children under age 6 and their main caregivers who resided in Detroit, Michigan.[26] Although the study is longitudinal in nature, we use cross-sectional data of 897 children surveyed in the first wave of examinations conducted between 2002 and 2003. The outcome variable in our analysis is DS which represents the number of decayed tooth surfaces. Sugar intake and age are

used as covariates in our model. For comparison purpose, we perform the proposed robust tests and Van den Broek's test. Van den Broek's test is performed under the assumption of a Poisson distribution with a mean function of covariates Age (child's age in years), SI (the child's sugar intake), and interaction Age*SI. Our robust test does not require the specification of the mean function. By knowing the covariates Age and SI, we can also perform the robust tests that incorporate Age and SI into the mean model. The results are given in Table 7.

Our robust tests with/without covariates and Van den Broek's score test reject the null hypothesis at a 0.05 significance level, supporting the hypothesis of heterogeneity. Even though all tests are not in favor of the null model (i.e. Poisson model), our robust tests provide much smaller test statistics compared to Van den Broek's test statistic. In other words, unlike Van den Broek's test, the proposed tests do not give a very strong evidence in favor of the ZIP model. When observing such a huge discrepancy in test statistic, we suggest that further investigation is needed to ensure the validity of statistical inferences for the study. Especially, rejecting the hypothesis of homogeneity does not give evidence that the ZIP model provides the best fit for the data. In fact, Todem et al.[9] have conducted Jansakul and Hinde's test to assess the inclusion of extra zeros in the data, and the test actually failed to reject the null hypothesis of a homogeneous NB model. On top of the non-significant testing result, Todem et al.[9] further provided additional graphical evidence and indicated that a simple NB model can provide a better fit to the data than a zero-inflated model. This example highlights the need for practicing statisticians to carefully interpret rejection of the null hypothesis in the light of a possible model misspecification. This is a reminder that rejecting the null hypothesis simply implies that the alternative complex model should be further evaluated against other competing models.

7 Discussion

In this paper, we proposed a random-variable approach for evaluating the need of a zero-inflated model when the working model is potentially misspecified. We showed that the test is robust under misspecifications of the associated components of the working model (e.g. important covariates being ignored in the working model). Unlike existing tests that require the mean component to be fully specified in terms of covariates or functional form, our proposed test can be performed without an explicit specification of the mean function. Rather the test uses unobserved random effects as a proxy for the mean function. A gamma distribution was imposed on these random effects to capitalize on the Poisson-Gamma conjugacy and ease computations. This method is particularly important in settings where the analyst has little to no knowledge of covariates and the associated functional form relating these covariates to the mean response. In our view, this approach is the most conservative method in that the practicing statisticians have no knowledge of true underlying model in many applications. We recognized that the gamma distribution may be too restrictive. The reason for using the gamma distribution in our method is its computational simplicity and, more importantly, its popularity among practicing analysts in many applications.[27,28] This restriction may be alleviated by imposing instead a finite mixture gamma distribution for which the number of components may be dictated by the observed data. The advantage of the mixture is that marginally the homogeneous model will become a mixture of negative binomial.[29] Such an approach still enjoys the computational simplicity while providing protection for possible misspecifications. This and other extensions merit further research.

We further extend the working model to incorporate partial information (e.g. some knowledge of covariates) into the testing procedure in order to improve the testing efficiency. Such an extension intrinsically assumes a low-dimensional structure under the null model. However, even under this condition, complications in deriving the information matrix for more than four covariates are expected in practice. Albeit being more complicated, the proposed test should work well. Alternatively, a summary score of many covariates can be computed before performing the proposed test, in which case the estimates under the null model will be only a non-zero regression coefficient associated with the summary score. Such an approach significantly reduces the dimension of covariate space prior to the implementation of the proposed test. One popular and well-known approach for creating a summary score is Principle Component Analysis (PCA). However, the standard PCA may fail to yield consistent estimators of the loading vectors under very high-dimensional settings.[30,31] In such a case, the impact of inconsistent estimators on the asymptotic properties of the resulting test statistics is not clear and this requires a further investigation. For cases where the number of covariates is large in view of the sample size (i.e. high dimensional structure), conducting the proposed test is not straightforward without knowing a priori the important variables. However, guided by the scientific knowledge, a small set of variables that are known to be associated with the phenomenon under study could be chosen to perform the test. And for variables that are associated but not entertained in this selection would then contribute to the random term (latent variable $\log(\Lambda_i)$). In sum, an

important message for practicing statisticians is to use the knowledge from historical studies to include the important covariates when evaluating the zero inflation.

In practice, the test of homogeneity for zero-inflated count data is usually daunting because the true underlying model that generated the data is typically unknown. This then requires analysts to carefully examine whether the features of the data are in agreement with the suggested model. To this end, our recommendation is that the proposed robust score test can be first entertained due to its robustness against the model misspecifications discussed in this paper. Particularly, the proposed score test tends to be conservative and in favor of the homogeneity hypothesis as protection. When the homogeneity hypothesis is rejected, a careful investigation on the zero-inflated models versus other competing models should be conducted in view of the data. That is because the rejection does not necessarily imply that the zero-inflated model gives the best fit to the data. When the proposed test fails to reject the homogeneity hypothesis, we suggest that complex one-component models, such as negative binomial model, can be adopted even when many zeros are apparently observed in the data. As no significant evidence supporting a two-component mixture model, using one-component models can further enjoy the ease of interpretation at the marginal level.

Acknowledgements

The authors thank the two referees for their constructive comments and suggestions. We are also grateful to Amid Ismail for his permission to use the dental caries data.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by NIDCR R03DE027108 and NIMHD U54MD011227.

ORCID iD


Wei-Wen Hsu https://orcid.org/0000-0002-7822-0826

Supplemental material

Supplemental material for this article is available online.

References

1. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**: 1–14.
2. Ridout M, Demétrio CG and Hinde J. Models for count data with many zeros. In: *Proceedings of the XIXth international biometric conference*, Cape Town, South Africa, volume 19, pp.179–192. Cape Town, South Africa: International Biometric Society.
3. Böhning D, Dietz E, Schlattmann P, et al. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J Royal Stat Soc: Series A (Stat Soc)* 1999; **162**: 195–209.
4. Farewell V, Long D, Tom B, et al. Two-part and related regression models for longitudinal data. *Ann Rev Stat Appl* 2017; **4**: 283–315.
5. Van den Broek J. A score test for zero inflation in a Poisson distribution. *Biometrics* 1995; **51**: 738–743.
6. Deng D and Paul SR. Score tests for zero inflation in generalized linear models. *Can J Stat* 2000; **28**: 563–570.
7. Jansakul N and Hinde J. Score tests for zero-inflated Poisson models. *Computat Stat Data Analys* 2002; **40**: 75–96.
8. Xiang L, Lee AH, Yau KK, et al. A score test for zero-inflation in correlated count data. *Stat Med* 2006; **25**: 1660–1671.
9. Todem D, Hsu WW and Kim K. On the efficiency of score tests for homogeneity in two-component parametric models for discrete data. *Biometrics* 2012; **68**: 975–982.
10. Cao G, Hsu WW and Todem D. A score-type test for heterogeneity in zero-inflated models in a stratified population. *Stat Med* 2014; **33**: 2103–2114.
11. Rao CR. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* 1948; **44**(1), 50–57.
12. Cox DR and Hinkley DV. *Theoretical statistics*. Amsterdam, the Netherlands: Chapman and Hall/CRC, 1979.

13. Breusch TS and Pagan AR. The lagrange multiplier test and its applications to model specification in econometrics. *Rev Economic Studies* 1980; **47**: 239–253.
14. Bera AK and Mckenzie CR. Alternative forms and properties of the score test. *J Appl Stat* 1986; **13**: 13–25.
15. Mukerjee R. Rao's score test: Recent asymptotic results. In: *Econometrics, handbook of statistics*, vol. 11. New York, NY: Elsevier, 1993, pp.363–379.
16. Kassahun W, Neyens T, Faes C, et al. A zero-inflated overdispersed hierarchical Poisson model. *Stat Model* 2014; **14**: 439–456.
17. Farewell VT and Sprott D. The use of a mixture model in the analysis of count data. *Biometrics* 1988; **44**(3): 1191–1194.
18. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 2000; **56**: 1030–1039.
19. Jansakul N and Hinde JP. Score tests for extra-zero models in zero-inflated negative binomial models. *Commun Stat-Simulat Computat* 2008; **38**: 92–108.
20. Rao CR. Score test: historical review and recent developments. In: *Advances in ranking and selection, multiple comparisons, and reliability*. New York, NY: Springer, 2005, pp.3–20.
21. Molenberghs G, Verbeke G, Demétrio CG, et al. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Stat Sci* 2010; **25**: 325–347.
22. McLachlan G. On the em algorithm for overdispersed count data. *Stat Meth Med Res* 1997; **6**: 76–98.
23. Maxwell O, Mayowa BA, Chinedu IU, et al. Modelling count data; a generalized linear model framework. *Am J Math Stat* 2018; **8**: 179–183.
24. Rosenkranz RR, Behrens TK and Dzewaltowski DA. A group-randomized controlled trial for health promotion in girl scouts: healthier troops in a snap (scouting nutrition & activity program). *BMC Public Health* 2010; **10**: 81.
25. Schlechter CR, Rosenkranz RR, Guagliano JM, et al. Impact of troop leader training on the implementation of physical activity opportunities in girl scout troop meetings. *Translat Behavior Med* 2018; **8**: 824–830.
26. Tellez M, Sohn W, Burt BA, et al. Assessment of the relationship between neighborhood characteristics and dental caries severity among low-income African-Americans: a multilevel approach. *J Public Health Dentistry* 2006; **66**: 30–36.
27. Congdon PD. *Applied Bayesian hierarchical methods*. Amsterdam, the Netherlands: Chapman and Hall/CRC, 2010.
28. Zhou M, Padilla OHM and Scott JG. Priors for random count matrices derived from a family of negative binomial processes. *J Am Stat Assoc* 2016; **111**: 1144–1156.
29. Gharib M. Two characterisations of a gamma mixture distribution. *Bull Australian Math Soc* 1995; **52**: 353–358.
30. Johnstone IM and Lu AY. On consistency and sparsity for principal components analysis in high dimensions. *J Am Stat Assoc* 2009; **104**: 682–693.
31. Lee YK, Lee ER and Park BU. Principal component analysis in very high-dimensional spaces. *Statistica Sinica* 2012; **22**: 933–956.