Tests for zero-inflation and overdispersion: a new approach based on the stochastic convex order

A. Baíllo, J.R. Berrendero*, J. Cárcamo

Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain

Abstract

A new methodology to detect zero-inflation and overdispersion is proposed, based on the comparison of the expected sample extremes among convexly ordered distributions. The method is very flexible and includes tests for the proportion of structural zeros in zero-inflated models, tests to distinguish between two ordered parametric families and a new general test to detect overdispersion. The performance of the proposed tests is evaluated via some simulation studies. For the well-known fetal lamb data, the conclusion is that the zero-inflated Poisson model should be rejected against other more disperse models, but the negative binomial model cannot be rejected.

Key words: zero-inflated Poisson distribution, binomial distribution, negative binomial distribution, hypothesis testing, convex order, parametric bootstrap

1. Introduction

The Poisson distribution is the standard model for the analysis of count data. However, in many situations this type of observations exhibit a substantially larger proportion of zeros than what is expected for the Poisson model (see [1], [9], [16] and the references therein). For instance, this is often the case with count data coming from medical and public health research

*Corresponding author. Phone: 0034 914976690. Fax: 0034 914974889

Email addresses: amparo.baillo@uam.es (A. Baíllo), joser.berrendero@uam.es (J.R. Berrendero), javier.carcamo@uam.es (J. Cárcamo)

(see [4] and [6]). This phenomenon usually arises when the distribution generating the data is a mixture of two populations, the first of which yields Poisson-distributed counts whereas the second one always contributes with a zero.

One natural model to describe the above situation is the so-called *zero-inflated Poisson* (ZIP) model. We say that the random variable $Y(\theta, p)$ has a ZIP distribution with parameters θ and p ($\theta > 0$ and $0 \leq p < 1$) if

$$\Pr\left(Y(\theta,p)=k\right) = \begin{cases} p + (1-p)e^{-\theta/(1-p)}, & \text{if } k = 0 \\ e^{-\theta/(1-p)} \dfrac{\theta^k}{k!(1-p)^{k-1}}, & \text{if } k = 1, 2, \ldots . \end{cases} \tag{1}$$

Therefore, $Y(\theta, p)$ is a mixture of a degenerate-at-zero distribution (with weight p) and a Poisson distribution of mean $\theta/(1-p)$ (with weight $1-p$). In particular, $Y(\theta, 0)$ is the classical Poisson variable with mean θ. The ZIP distribution has been used in diverse areas such as medicine ([3], [4] and [19]) or biology ([14]), among others.

The expected value of the ZIP distribution is $\mathrm{E}\left(Y(\theta,p)\right) = \theta$ and the variance $\mathrm{Var}(Y(\theta,p)) = \theta + \theta^2 p/(1-p)$ increases as p increases. The zeros coming from the degenerate variable are called *structural zeros* and those from the Poisson model *sampling zeros*. It should be observed at this point that, to keep the mean fixed for different values of p, we do not follow the usual notation for the ZIP models.

If the proportion of atypical zero observations remains undetected, the variability of the population is underestimated and the properties of standard inference techniques are, to some extent, deteriorated. For this reason, in the recent literature there are different proposals to determine whether the Poisson model fits a data set well enough or, alternatively, we should choose a ZIP model that allows for an extra proportion of zero counts. A clear and concise review of several of these tests can be found in [21]. A popular and simple choice with good properties is the score test proposed by [19].

Of course, as pointed out by [8] and [18], the rejection of the Poisson model does not imply that the ZIP distribution is the most appropriate model to fit the data. It may happen that an alternative model that accounts for the observed dispersion could fit the data better. The negative binomial and the zero-inflated negative binomial distributions are examples of reasonable alternatives.

In this work we introduce a new procedure to detect zero-inflation and

overdispersion. The key idea is to link the notion of overdispersion with the concept of *variability stochastic order*. These orders arrange distributions according to their variability (see Section 3 of [17]). Therefore, it is natural to suppose that the observed overdispersion is due to the data actually coming from a different model that dominates the initially assumed distribution in a variability order. The most important variability order is the so-called *convex order*. We use the properties of this order to derive suitable discrepancy measures for tests in which "overdispersion" is understood as "convex domination".

The method we propose is flexible and easy to implement. It is based on the empirical comparison of the expected sample extremes of two ordered models. An important feature is that the main ideas can be readily adapted to cover several different testing problems: tests for the proportion of structural zeros in zero-inflated models; procedures for testing if a parametric model is appropriate against another one with more variability; and a new general test to detect overdispersion. We illustrate in detail the application of the methodology to the case of the ZIP models, but the technique can be analogously applied in other situations.

The definitions and relevant results on stochastic convex dominance are briefly reviewed in Section 2. These results supply the necessary theoretical background for the rest of the paper. In Section 3, we provide a general framework to detect overdispersion in ZIP models, but we note that the proposed method is very general and can be adapted to many other similar scenarios. We find discrepancy measures for tests on the proportion of structural zeros and discuss whether the Poisson model is appropriate or we should opt for a different model with more dispersion. In Section 4 we establish the relationships, in terms of the convex order, for some zero-inflated models usually considered in the literature: the zero-inflated binomial, Poisson and negative binomial model. These results allow to extend the previous ideas to these important discrete models. The choice of a powerful test is discussed in Section 5. Section 6 analyzes the performance of the proposed tests via some Monte Carlo studies. Our proposals are very competitive against the well-known score test in the cases in which the latter can be applied. In Section 7, we analyze the fetal lamb data from [13] using our new procedures. For this data set we conclude that the ZIP distribution should be rejected against other models with more variability. This result is consistent with the previous work by [18]. Moreover, we show that the negative binomial model cannot be rejected. Section 8 includes some final remarks and conclusions.

The proofs of the main results are collected in the appendix.

2. The convex order and overdispersion

In this section, we link the overdispersion phenomenon described in the introduction with the convex stochastic order. Given two integrable random variables X and Y, it is said that X *is less or equal to* Y *in the convex order*, and we denote it by $X \leq_{\text{cx}} Y$, if $\mathrm{E}(\phi(X)) \leq \mathrm{E}(\phi(Y))$ for every convex function ϕ for which the previous expectations are well defined. Notice that, by considering the convex functions $\phi(x) = \pm x$, the condition $X \leq_{\text{cx}} Y$ implies that $\mathrm{E}X = \mathrm{E}Y$. Furthermore, if the variables have finite second moment, applying the definition of the convex order with $\phi(x) = (x - \mathrm{E}X)^2$, we conclude that $\mathrm{Var}(X) \leq \mathrm{Var}(Y)$. Of course, establishing the relation $X \leq_{\text{cx}} Y$ is much more informative than just knowing $\mathrm{Var}(X) \leq \mathrm{Var}(Y)$.

Roughly speaking, since convex functions take larger values when its argument is large, if $X \leq_{\text{cx}} Y$ holds, then Y is more likely to take "extreme values" than X. This idea is clear from the following proposition. The result is a consequence of Corollary 4.A.16 and Theorem 4.A.50 in [17], regarding the expected value of the extreme order statistics of two ordered variables. For $k \geq 1$, if $(X_1, \ldots, X_k)$ is a random sample of size k from X, we denote by $X_{i:k}$ the i-th order statistic of the sample, $i = 1, \ldots, k$. Therefore, $X_{1:k}$ and $X_{k:k}$ stand for the minimum and maximum of the sample.

Proposition 1. *Let X and Y be integrable random variables such that $X \leq_{\text{cx}} Y$.*

(a) *For all $k \geq 1$, $\mathrm{E}Y_{1:k} \leq \mathrm{E}X_{1:k}$ and $\mathrm{E}X_{k:k} \leq \mathrm{E}Y_{k:k}$.*

(b) *If for some $k \geq 2$ $\mathrm{E}X_{1:k} = \mathrm{E}Y_{1:k}$ or $\mathrm{E}X_{k:k} = \mathrm{E}Y_{k:k}$, then X and Y have the same distribution.*

For instance, for the ZIP variables defined as in (1), we can prove (see the appendix) that, for $\theta > 0$,

$$Y(\theta, p_1) \leq_{\text{cx}} Y(\theta, p_2) \quad \text{if and only if} \quad 0 \leq p_1 < p_2 < 1. \tag{2}$$

Hence, Proposition 1 jointly with (2) imply that the ZIP variable $Y(\theta, p_2)$ is expected to take *strictly* larger extreme values than $Y(\theta, p_1)$ whenever $p_1 < p_2$.

3. Tests for overdispersion in ZIP models

In this section we exploit Proposition 1 to derive discrepancy measures useful to test for overdispersion in ZIP models. We emphasize that the same technique, with the obvious modifications, can be applied in a similar way for the zero-inflated binomial and negative binomial models (see Section 4) or, in general, for any pair of ordered distributions.

The discrepancies introduced in this section are defined in terms of the empirical counterparts of the expected extreme order statistics. Therefore, our goal is to detect (significant) differences between the estimates of the expected extremes of two distributions.

Actually, we deal with two different problems. In Subsection 3.1 we propose statistical tests to analyze the proportion of structural zeros in ZIP models. In other situations, we may want to check if the ZIP model cannot account for the dispersion of the data. Then it is adequate to apply the nonparametric procedure of Subsection 3.2.

3.1. Tests for the proportion of structural zeros

Given a random sample $Y_1, \ldots, Y_n$ from a variable $Y(\theta, p)$ with the ZIP distribution (1), we are interested in testing $\mathrm{H}_0 : p \leq p_0$ against $\mathrm{H}_1 : p > p_0$, where p_0 is fixed and belongs to $[0, 1)$ (the left unilateral and bilateral tests may be studied by similar arguments). There are several works in the literature devoted to this testing problem with $p_0 = 0$ (see e.g. [19], [21], [11] and [10]). This particular case is important since it is equivalent to testing the Poisson model against a ZIP model with a positive proportion of structural zeros. However, as far as we know, there are no references in the literature including tests for values of $p_0 \in (0, 1)$.

The method we propose is based on the following simple idea: (2) states that $Y(\theta, p_1) \leq_{\mathrm{cx}} Y(\theta, p_2)$ whenever $0 \leq p_1 < p_2$ and hence according to Proposition 1, the variable $Y(\theta, p)$ is expected to take strictly larger extreme values under H_1 than under H_0. Using the information in $Y_1, \ldots, Y_n$, we can estimate the expectation of the maximum (or minimum) in a generic subsample of size $k \geq 2$ from $Y(\theta, p)$ and $Y(\theta, p_0)$. Then, we reject H_0 whenever the difference between the two estimates is too large.

More precisely, we denote by $\mathrm{E}_{\theta,p}(Y_{k:k})$ and $\mathrm{E}_{\theta,p}(Y_{1:k})$ the expected values of the maximum and minimum of k independent copies of $Y(\theta, p)$, respectively. Given the random sample $Y_1, \ldots, Y_n$ from $Y(\theta, p)$, the maximum

likelihood estimates of the parameters θ and p in the ZIP model satisfy (see [12])

$$\hat{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{and} \qquad \hat{p} = 1 - \frac{1 - n_0/n}{1 - \exp(-\hat{\theta}/(1-\hat{p}))}, \tag{3}$$

where n_0 is the number of zero-counts in the sample. Then, for $k \geq 2$, we compute the discrepancy measures:

$$\Delta_{k:k} = \mathrm{E}_{\hat{\theta},\hat{p}}(Y_{k:k}) - \mathrm{E}_{\hat{\theta},p_0}(Y_{k:k}) \quad \text{and} \quad \Delta_{1:k} = \mathrm{E}_{\hat{\theta},p_0}(Y_{1:k}) - \mathrm{E}_{\hat{\theta},\hat{p}}(Y_{1:k}) \tag{4}$$

and reject H_0 if either $\Delta_{k:k}$ or $\Delta_{1:k}$ is too large. Observe that, from the equalities $\mathrm{E}(Y(\hat{\theta}, p_0)) = \mathrm{E}(Y(\hat{\theta}, \hat{p}))$ and $\mathrm{E}(X_{1:2}) + \mathrm{E}(X_{2:2}) = 2\,\mathrm{E}(X)$ (which holds for any integrable random variable X), it is readily checked that $\Delta_{2:2} = \Delta_{1:2}$.

If we denote by $F_{\theta,p}$ the distribution function of $Y(\theta, p)$, the discrepancies in (4) can be rewritten as:

$$\begin{aligned} \Delta_{k:k} &= \sum_{i=0}^{\infty} \left[\left(F_{\hat{\theta},p_0}(i) \right)^k - \left(F_{\hat{\theta},\hat{p}}(i) \right)^k \right], \\ \Delta_{1:k} &= \sum_{i=0}^{\infty} \left[\left(1 - F_{\hat{\theta},p_0}(i) \right)^k - \left(1 - F_{\hat{\theta},\hat{p}}(i) \right)^k \right]. \end{aligned} \tag{5}$$

In practice, we can always truncate the above series to approximate their value.

To obtain the rejection region of the tests we need to find the distribution of $\Delta_{k:k}$ or $\Delta_{1:k}$ for $k \geq 2$ under H_0. In Theorem 1 we obtain the asymptotic distribution of $\Delta_{2:2}$ when $p_0 = 0$. However, in general, the distribution of these quantities is rather involved and a simple parametric bootstrap scheme can be used instead. The following procedure is described for the discrepancy $\Delta_{k:k}$ but the corresponding one for $\Delta_{1:k}$ is analogous:

(a) Find the estimate $\hat{\theta} = \bar{Y}$.

(b) Extract B parametric bootstrap samples of size n, $Y^*_{1,b}, \ldots, Y^*_{n,b}$, for $b = 1, \ldots, B$, from the distribution of $Y(\hat{\theta}, p_0)$.

(c) For each sample $Y^*_{1,b}, \ldots, Y^*_{n,b}$, obtain the estimates $\hat{\theta}^*_b$ and $\hat{p}^*_b$ using (3).

(d) Compute the discrepancies $\Delta^{*,b}_{k:k} = \mathrm{E}_{\hat{\theta}^*_b,\hat{p}^*_b}(Y_{k:k}) - \mathrm{E}_{\hat{\theta}^*_b,p_0}(Y_{k:k})$, $b = 1, \ldots, B$ using the expressions in (5), where the series are truncated at an appropriate value. In the simulations below, the series have been truncated at 200.

(e) For a significance level α, find $Q^*_{k:k}(\alpha)$, the $(1-\alpha)$-quantile of the values $\{\Delta^{*,b}_{k:k}, b = 1, \ldots, B\}$.

The rejection region for the test $\mathrm{H}_0 : p \leq p_0$ versus $\mathrm{H}_1 : p > p_0$, at significance level α, is approximated by

$$R_\alpha = \{\Delta_{k:k} > Q^*_{k:k}(\alpha)\}. \tag{6}$$

As it was mentioned before, the case $p_0 = 0$ corresponds to testing the Poisson model against a ZIP model with $p > 0$. As argued in Subsection 5.1, in this parametric test, the choice of k is not relevant. Therefore, we propose to use the simplest value $k = 2$ which is shown to have a good behavior in the simulation studies of Subsection 6.1. The use of $\Delta_{2:2}$ means that we compare what we expect to obtain for the maximum (or minimum) of two independent Poisson variables with that of two ZIP variables with $p > 0$. In this case, there is a closed-form expression for $\mathrm{E}_{\theta,0}(Y_{2:2})$ (see [12], p. 166):

$$M_2(\theta) := \mathrm{E}_{\theta,0}(Y_{2:2}) = \theta + \theta e^{-2\theta} \left(I_0(2\theta) + I_1(2\theta)\right), \tag{7}$$

where I_0 and I_1 are modified Bessel functions of the first kind (see e.g. [2]). Using (7) we can rewrite the discrepancy $\Delta_{2:2}$ given in (5) with $p_0 = 0$ as

$$\Delta_{2:2} = 2\hat{p}\hat{\theta} + (1-\hat{p})^2 M_2(\hat{\theta}/(1-\hat{p})) - M_2(\hat{\theta}). \tag{8}$$

This enables us to obtain the asymptotic distribution of $\Delta_{2:2}$ under $\mathrm{H}_0 : p = 0$ (Poissonness). In the following theorem the symbol "$\longrightarrow_d$" stands for "convergence in distribution" and $N(0,1)$ is a standard normal variable.

Theorem 1. *Under* $\mathrm{H}_0 : p = 0$*, it holds that*

$$\sqrt{n}\,\frac{\Delta_{2:2}}{\sigma(\hat{\theta})} \longrightarrow_d N(0,1), \qquad n \to \infty,$$

where

$$\sigma^2(\hat{\theta}) := \frac{\hat{\theta}^2 \left(1 - e^{-2\hat{\theta}} \left[(1+\hat{\theta})I_0(2\hat{\theta}) - I_1(2\hat{\theta}) + \hat{\theta} I_2(2\hat{\theta})\right]\right)^2}{e^{\hat{\theta}} - 1 - \hat{\theta}}, \tag{9}$$

and I_0*,* I_1 *and* I_2 *are modified Bessel functions of the first kind.*

As an immediate consequence of Theorem 1, a critical region with asymptotic significance level α for $\mathrm{H}_0 : p = 0$ against $\mathrm{H}_1 : p > 0$ is

$$R_\alpha = \left\{ \sqrt{n}\, \frac{\Delta_{2:2}}{\sigma(\hat{\theta})} > z_\alpha \right\}, \tag{10}$$

with z_α being the $(1-\alpha)$-quantile of the standard normal distribution. We remark that this test is very simple and easy to implement since the Bessel functions appearing in $\Delta_{2:2}$ and $\sigma(\hat{\theta})$ can be evaluated by any standard mathematical software package.

3.2. A general test to detect overdispersion

Here, we deal with the problem of detecting if a data set comes from a Poisson distribution or there is dispersion that the Poisson model cannot take into account. The same procedure works for the more general ZIP model or the distributions considered in Section 4, but we illustrate the ideas with the Poisson distribution for the sake of simplicity.

Let us consider the family $\mathcal{P} := \{Y(\theta) : \theta > 0\}$, where $Y(\theta)$ is a Poisson variable with mean θ. We denote by $\mathcal{P}_{\mathrm{cx}}$ the set of all integrable random variables, not having the Poisson distribution, that dominate in the convex order a variable in $\mathcal{P}$. Therefore, $\mathcal{P}_{\mathrm{cx}}$ includes distributions with strictly more dispersion than the Poisson variables. In particular, according to (2) and Proposition 3 in Section 4, all the ZIP (with $p > 0$) and the (zero-inflated) negative binomial distributions are included in $\mathcal{P}_{\mathrm{cx}}$. Given a random sample $Y_1, \ldots, Y_n$ from Y, we want to test $\mathrm{H}_0 : Y \in \mathcal{P}$ against $\mathrm{H}_1 : Y \in \mathcal{P}_{\mathrm{cx}}$.

In this new test the alternative hypothesis is not completely specified in the sense that it is not given by a parametric family. However, to handle this problem we can use similar ideas to those in Subsection 3.1. We first estimate the parameter θ, $\hat{\theta} = \bar{Y}$. Then, we compute the expectation of the maximum or minimum of k independent copies of $Y(\hat{\theta})$, $\mathrm{E}_{\hat{\theta}}(Y_{k:k})$ and $\mathrm{E}_{\hat{\theta}}(Y_{1:k})$, as before in Subsection 3.1. On the other hand, since there is no parametric restriction under H_1, we estimate $\mathrm{E}Y_{k:k}$ and $\mathrm{E}Y_{1:k}$ by means of the following nonparametric plug-in estimators:

$$\mathrm{E}_{F_n}(Y_{k:k}) := \sum_{i=1}^{n} \left[\left(\frac{i}{n} \right)^k - \left(\frac{i-1}{n} \right)^k \right] Y_{i:n},$$

$$\mathrm{E}_{F_n}(Y_{1:k}) := \sum_{i=1}^{n} \left[\left(1 - \frac{i-1}{n} \right)^k - \left(1 - \frac{i}{n} \right)^k \right] Y_{i:n},$$

where F_n is the empirical distribution function of the sample $Y_1, \ldots, Y_n$. Hence, for $k \geq 2$, we consider the discrepancies

$$\Lambda_{k:k} := \mathrm{E}_{F_n}(Y_{k:k}) - \mathrm{E}_{\hat{\theta}}(Y_{k:k}) \qquad \text{and} \qquad \Lambda_{1:k} := \mathrm{E}_{\hat{\theta}}(Y_{1:k}) - \mathrm{E}_{F_n}(Y_{1:k}). \quad (11)$$

Under H_0 these discrepancies are close to 0 whereas, if H_1 holds, then $\Lambda_{1:k}$ and $\Lambda_{k:k}$ are (strictly) positive for n large enough. Therefore, we reject H_0 whenever $\Lambda_{1:k}$ or $\Lambda_{k:k}$ are too large. The rejection region of these tests can be derived by using a parametric bootstrap approach similar to the one described in Subsection 3.1.

We finally note that we actually have a different test for each discrepancy. The power of the test may depend on the selection of the statistic. The choice of a test with good power is addressed in Section 5.

4. Extensions to other models

The application of the methodology described in the previous section relies on verifying the convex domination of the involved variables. In this section, we establish all the relationships, according to the convex order, among the zero-inflated versions of some commonly used models for count data: the Poisson, the binomial and the negative binomial models. For these important discrete models, these relationships allow to extend straightaway the ideas developed in the previous section.

We first note that, given a data set, it is sensible to assume that the models that could fit the data have the same mean. Hence, all the parametric distributions considered in this section are selected to have the same expectation θ.

For $m \geq 1$, $0 \leq p < 1$ and $0 < \theta \leq m(1-p)$, let us consider the random variable $X(m, \theta, p)$ which is the mixture between the degenerate-at-zero variable with weight p and a binomial variable of parameters m and $\theta/[m(1-p)]$ with weight $1-p$. In other words, $X(m, \theta, p)$ has the *zero-inflated binomial* (ZIB) distribution with probabilities

$$\Pr\left(X(m,\theta,p) = k\right) = \begin{cases} p + \left(1 - \frac{\theta}{m(1-p)}\right)^m, & \text{if } k = 0 \\ (1-p)\binom{m}{k}\left(\frac{\theta}{m(1-p)}\right)^k \left(1 - \frac{\theta}{m(1-p)}\right)^{m-k}, & \text{if } 1 \leq k \leq m. \end{cases}$$

Furthermore, we also consider the variable $Z(t, \theta)$ with negative binomial

(NB) distribution of parameters $1/t$ and $t\theta$ ($t > 0$ and $\theta > 0$), i.e.,

$$\Pr\left(Z(t,\theta) = k\right) = \binom{k + 1/t - 1}{k} \frac{(\theta t)^k}{(1+\theta t)^{k+1/t}}, \quad k \geq 0.$$

Among the different parametrizations of the NB distribution, we have chosen the *unique* one, $Z(t,\theta)$, with mean θ (for all t) and increasing in t for the convex order, that is, satisfying $Z(t_1,\theta) \leq_{\text{cx}} Z(t_2,\theta)$ whenever $0 < t_1 < t_2$ (see Proposition 3 (d) below).

However, there are infinitely many possibilities to inflate with zeros the variable $Z(t,\theta)$ preserving the mean θ. Among them, we only consider the most representative two. On the one hand, for $t, \theta > 0$ and $0 \leq p < 1$, let $Z_1(t,\theta,p)$ be the mixture between the degenerate-at-zero variable with weight p and the variable $Z(t(1-p), \theta/(1-p))$ with weight $1-p$. On the other hand, for $t, \theta > 0$ and $0 \leq p < 1$ let $Z_2(t,\theta,p)$ be the mixture between the degenerate-at-zero variable with weight p and the variable $Z(t, \theta/(1-p))$ with weight $1-p$. We refer to these two models as the *zero-inflated negative binomial* (ZINB) models.

In order to clarify the notation, Table 1 summarizes the relevant information about the models considered throughout this section. We note that all the variables have a fixed mean θ and a proportion p of structural zeros.

Table 1: Summary of the considered models.

Model	**Notation**	**Variance**
ZIB	$X(m,\theta,p)$	$\theta + \dfrac{\theta^2 p}{1-p} - \dfrac{\theta^2}{m(1-p)}$
ZIP	$Y(\theta,p)$	$\theta + \dfrac{\theta^2 p}{1-p}$
ZINB(1)	$Z_1(t,\theta,p)$	$\theta + \dfrac{\theta^2 p}{1-p} + \theta^2 t$
ZINB(2)	$Z_2(t,\theta,p)$	$\theta + \dfrac{\theta^2 p}{1-p} + \dfrac{\theta^2 t}{1-p}$

The variance of all the zero-inflated variables described before is an increasing function of $p \in [0,1)$. Actually, the next proposition shows that they are convexly ordered for different values of p.

Proposition 2. *Let $X(m,\theta,p)$, $Y(\theta,p)$ and $Z_i(t,\theta,p)$ $(i=1,2)$ be variables with the ZIB, ZIP and ZINB distributions described above. If $0 \le p_1 < p_2 < 1$, then*

(a) $X(m,\theta,p_1) \le_{\text{cx}} X(m,\theta,p_2)$, *for all* $m \ge 1$ *and* $0 < \theta \le m(1-p_2)$.
(b) $Y(\theta,p_1) \le_{\text{cx}} Y(\theta,p_2)$, *for all* $\theta > 0$.
(c) $Z_i(t,\theta,p_1) \le_{\text{cx}} Z_i(t,\theta,p_2)$, *for all* $t > 0$, $\theta > 0$ *and* $i = 1,2$.

Proposition 2 shows that for different values of the parameter p the distributions are (convexly) ordered. Conversely, when the distributions are ordered, their expectations are equal and the variance of the majorizing distribution is greater than the variance of the dominated one. Since the variance is an increasing function of p (see Table 1), the ordering between the distributions also implies that their corresponding values of p are ordered. Therefore, in this situation the order of the parameter p is equivalent to the convex order of the distributions.

The limiting distribution of $X(m,\theta,p)$ (as $m \uparrow \infty$) and of $Z_i(t,\theta,p)$ (as $t \downarrow 0$) for $i = 1,2$ is the ZIP variable $Y(\theta,p)$. The smaller m is, the more the ZIB variable differs from the ZIP one. Also, the larger t is, the more the ZINB variables differ from the ZIP one.

For a fixed proportion of structural zeros, the next proposition presents the relationships among the four models.

Proposition 3. *For a fixed $p \in [0,1)$, we have:*

(a) $X(m,\theta,p) \le_{\text{cx}} X(m+1,\theta,p)$, *for all* $m \ge 1$ *and* $0 < \theta \le m(1-p)$.
(b) $X(m,\theta,p) \le_{\text{cx}} Y(\theta,p)$, *for all* $m \ge 1$ *and* $0 < \theta \le m(1-p)$.
(c) $Y(\theta,p) \le_{\text{cx}} Z_1(t,\theta,p) \le_{\text{cx}} Z_2(t,\theta,p)$, *for all* $\theta > 0$ *and* $t > 0$.
(d) $Z_i(t_1,\theta,p) \le_{\text{cx}} Z_i(t_2,\theta,p)$, *for all* $0 < t_1 < t_2$, $\theta > 0$ *and* $i = 1,2$.

Proposition 2 allows to test on the proportion of structural zeros in all the models of this section. Further, Proposition 3 makes possible the comparison of these parametric families. The nonparametric tests described in Subsection 3.2 can also be adapted to these models. Let us point out that although in these models the variance is a function of the mean, this is not needed to apply the tests. These can be used with models for which the variance is independent of the mean. The only requirement is that the distributions are convexly ordered. An example of the application of these tests to a real data set can be found in Section 7.

5. The choice of the discrepancy measure

The approach discussed in Section 3 generates a family of discrepancies for the addressed testing problems. We actually have a different test if we select the maximum or minimum in the discrepancy: $\Delta_{k:k}$ or $\Delta_{1:k}$ in the tests of Subsection 3.1 and $\Lambda_{k:k}$ or $\Lambda_{1:k}$ in the nonparametric case of Subsection 3.2. Moreover, the test statistics also differ for each $k \geq 2$. Hence, the question of finding a test with good power arises.

We have analyzed through some simulations the choice of the suitable test statistic. The analysis showed that the selection of the discrepancy depends on the testing problem of interest.

5.1. Choice of the discrepancy in the test on the proportion of structural zeros

Regarding the tests on the proportion of structural zeros discussed in Subsection 3.1, observe that both hypotheses assume that the observations follow a parametric (ZIP) distribution. The tests mainly rely on the estimation of the parameters of the model, and the choice of the discrepancy is of secondary importance. Some preliminary simulations showed that different discrepancies yield similar powers. Therefore, in this situation we opt for the simplest one $\Delta_{2:2} = \Delta_{1:2}$ defined in (8), which has computational advantages over the others with larger k's. This stresses the significance of the asymptotic result given in Theorem 1.

5.2. Choice of the discrepancy in the overdispersion test

The test for overdispersion described in Subsection 3.2 is more sensitive to the choice of the discrepancy. Here H_0 is given by a parametric model whereas H_1 includes all the distributions that strictly dominate an element of the initial family. Hence, H_1 is *not* specified by any parametric family. In this case, the power of the tests strongly depends both on the distribution generating the data and on the parametric family assumed in H_0. For a fixed discrepancy, different alternatives could lead to very different powers. Therefore, it is advantageous to have a family of discrepancies since this provides flexibility to select a good test in each situation.

Let us briefly explain how the coefficient of variation (CV) of the discrepancy is useful to choose a test with good properties. Under H_1, an adequate discrepancy to detect deviations from H_0 should have a large mean and low variance, that is, a low CV. The CV of the discrepancy describes well how the corresponding test behaves. In general, under H_1, a low CV is paralleled

by a high power. This is clearly reflected in Figure 1, where, for 1000 Monte Carlo samples, we plot the power of the test for overdispersion for the Poisson family and the inverse of the CV of the discrepancy $\Lambda_{1:k}$ defined in (11), for different values of k. In Figure 1(a), the observations are generated from a ZIP distribution $Y(3, 0.05)$, while in Figure 1(b) they are drawn from the NB distribution $Z(0.05, 3)$.

The choice of k could also depend implicitly on the size, n, of the available sample to carry out the test. When n increases, we have more information on the population and we can estimate better the expectation of the sample extremes of k independent realizations with k large. Therefore, there could be a wider range of values of k for which a good result is obtained. For instance, when the null hypothesis is fixed as a Poisson variable and the alternative is a zero-inflated Poisson, large values of k work well. For sample sizes n from 50 to 200, values around $k = 20$ give good results (see Figure 1(a)). However, for the same hypothesis, if the alternative is negative binomial, then $k = 2$ (the smallest possible value for k) is the one giving the best results in terms of power. In this case, a large value of n does not affect the selection of k (see Figure 1(b)). Taking into account these facts, we use the values $k = 20$ and $k = 2$ in the simulations of Subsection 6.2 for the ZIP and the negative binomial alternatives, respectively.

We finally note that when analyzing only one data set, it also becomes possible to choose a suitable discrepancy by estimating its CV via bootstrap (see Section 7 for details).

6. Simulations

We have carried out a Monte Carlo study to check the performance of the tests described above. The significance level in all cases is fixed as $\alpha = 0.05$.

6.1. Simulations for the test on the proportion of structural zeros

We consider the test on the proportion of structural zeros in a ZIP model (Subsection 3.1). As argued in Subsection 5.1, we select $k = 2$. For the case $p_0 = 0$ (H_0 represents the Poisson distribution), we compare the performance of the *score* test ([19]) and the test methodology that rejects H_0 if the discrepancy $\Delta_{1:2} = \Delta_{2:2}$ in (8) is too large. The rejection region for the latter method is chosen in two ways: via *bootstrap* as in (6) and also using the *asymptotic* distribution of $\Delta_{2:2}$ as in (10). The number of bootstrap samples is $B = 5000$.

In Table 2 we record the proportion of times that $H_0 : p = 0$ is rejected. For each combination of p and θ in the table, we generate 5000 Monte Carlo samples of sizes $n = 50$, 100 and 200 from $Y(\theta, p)$. Note that our proposed procedure has a very competitive performance in comparison to the score test. This is specially apparent for the lowest values of θ, where, when $p > 0$, in general our procedure yields a higher power than the score test.

Table 2: Proportion of times that $H_0 : p = 0$ was rejected.

		p			
n	θ	0	0.05	0.1	
50	3	0.047	0.386	0.784	Bootstrap
		0.056	0.422	0.800	Asymptotic
		0.036	0.313	0.722	Score
50	5	0.041	0.768	0.972	Bootstrap
		0.055	0.794	0.981	Asymptotic
		0.044	0.779	0.978	Score
50	10	0.002	0.923	0.994	Bootstrap
		0.002	0.923	0.994	Asymptotic
		0.002	0.923	0.994	Score
100	3	0.052	0.585	0.964	Bootstrap
		0.059	0.604	0.963	Asymptotic
		0.049	0.494	0.943	Score
100	5	0.043	0.944	0.999	Bootstrap
		0.077	0.966	1.000	Asymptotic
		0.045	0.945	0.999	Score
100	10	0.003	0.994	1.000	Bootstrap
		0.003	0.994	1.000	Asymptotic
		0.003	0.994	1.000	Score
200	3	0.051	0.827	0.999	Bootstrap
		0.054	0.831	0.999	Asymptotic
		0.048	0.762	0.999	Score
200	5	0.050	0.999	1.000	Bootstrap
		0.065	0.999	1.000	Asymptotic
		0.043	0.999	1.000	Score
200	10	0.007	1.000	1.000	Bootstrap
		0.007	1.000	1.000	Asymptotic
		0.007	1.000	1.000	Score

In Table 3 the results for the test $H_0 : p \leq 0.2$ against $H_1 : p > 0.2$ are displayed. In this case we only use the procedure based on $\Delta_{2:2}$ with rejection region (6). The number of Monte Carlo samples is again 5000.

6.2. Simulations for the overdispersion test

We test $H_0 : Y \in \mathcal{P}$ ($\mathcal{P}$ being the Poisson family) against $H_1 : Y \in \mathcal{P}_{\text{cx}}$ following the procedure described in Subsection 3.2. The number of Monte

Table 3: Proportion of times that $\mathrm{H}_0 : p \leq 0.2$ was rejected.

		p		
n	θ	0.2	0.25	0.3
50	3	0.069	0.272	0.581
50	5	0.067	0.253	0.554
50	10	0.061	0.244	0.546
100	3	0.062	0.366	0.781
100	5	0.065	0.346	0.780
100	10	0.061	0.370	0.774
200	3	0.061	0.536	0.953
200	5	0.065	0.557	0.962
200	10	0.069	0.584	0.963

Carlo samples is 5000 and the number of bootstrap samples used to compute the rejection region is $B = 5000$. We generate observations with sample sizes $n = 50$, 100 and 200, from a ZIP distribution $Y(\theta, p)$ and apply the nonparametric procedure based on $\Lambda_{1:20}$. Afterwards, we generate samples from the NB distribution $Z(t, \theta)$ and carry out the test with $\Lambda_{1:2}$. Recall that the justification for selecting such discrepancies was detailed in Subsection 5.2. In Tables 4 and 5 we display the proportion of times that H_0 is rejected. Observe how close the powers in Table 4 are to those of Table 2. We found this property appealing since in this test for overdispersion no parametric model is specified for the alternative hypothesis.

Table 4: Proportion of rejections of $\mathrm{H}_0 : Y \in \mathcal{P}$ when sampling from a ZIP $Y(\theta, p)$.

		p		
n	θ	0	0.05	0.1
50	3	0.043	0.358	0.780
50	5	0.045	0.732	0.966
50	10	0.051	0.901	0.993
100	3	0.047	0.576	0.952
100	5	0.052	0.911	0.999
100	10	0.053	0.982	1.000
200	3	0.054	0.839	0.999
200	5	0.054	0.992	1.000
200	10	0.050	0.999	1.000

Table 5: Proportion of rejections of $H_0 : Y \in \mathcal{P}$ when sampling from a NB $Z(t, \theta)$.

		t	
n	θ	0.05	0.1
50	3	0.183	0.385
50	5	0.309	0.635
100	3	0.269	0.583
100	5	0.479	0.871
200	3	0.409	0.812
200	5	0.710	0.989

7. An example with real data

To illustrate the usefulness of the methods proposed throughout the paper, we analyze a data set from [13]. The number of movements by a fetal lamb observed through ultrasound were recorded. We consider one particular sequence of counts of the number of movements in each of 240 consecutive 5-second intervals (see Table 6).

Table 6: Lamb data set and expected frequencies based on Poisson, ZIP and NB distributions.

Outcome	0	1	2	3	4	5	6	7
Obs. Freq.	182	41	12	2	2	0	0	1
Expect. Freq. (Poisson)	167.7	60.1	10.8	1.3	0.1	0.0	0.0	0.0
Expect. Freq. (ZIP)	182.0	36.9	15.6	4.4	0.9	0.2	0.0	0.0
Expect. Freq. (NB)	182.5	39.0	12.0	4.1	1.5	0.5	0.2	0.1

If we assume that the data follow a Poisson distribution with mean θ, the estimate of θ is $\hat{\theta} = 0.36$. The differences between the observed and the expected frequencies in Table 6 point out that the Poisson model is unsuitable. [7] used the Pearson χ^2 statistics to argue that a ZIP model provides a substantially improved fit. The estimates of the parameters under the ZIP model are $\hat{\theta} = 0.36$ and $\hat{p} = 0.58$. The corresponding expected frequencies are in the fourth row of Table 6. The fit seems indeed better, but we could formalize this statement by testing $H_0 : p = 0$ versus $H_1 : p > 0$. We apply both the asymptotic test (10) and the score test. Both results point out a strong evidence (p-values below 0.0001) against the Poisson model. This leads us to the conclusion that the ZIP distribution fits the data much

better than the Poisson one.

Rejecting the Poisson model does not necessarily imply that the ZIP model provides the best fit. Another model could account better for the observed dispersion. Therefore, using the nonparametric test developed in Subsection 3.2, we now test the null hypothesis that the distribution is ZIP against the alternative that the true model has more variability than the ZIP one. In this case, we have to select the appropriate statistics ($\Lambda_{1:k}$ or $\Lambda_{k:k}$) and a suitable value for k (see Subsection 5.2). For that purpose, we obtain bootstrap estimates (based on 500 bootstrap samples) of the inverse of the CV of $\Lambda_{1:k}$ and $\Lambda_{k:k}$, for different k's. The estimates as a function of k are displayed in Figure 2(a).

According to the results depicted in Figure 2(a), the test based on $\Lambda_{k:k}$ is preferable. Moreover, for $\Lambda_{k:k}$, there is a wide range of k values (between 50 and 200, say) for which the results are fairly similar. For the tests based on $\Lambda_{k:k}$ with $k = 50, 90, 130$ the p-values are under 0.0005. We conclude that the ZIP model is also clearly rejected so that other distributions with higher dispersion (according to the convex order) are more appropriate to fit this data set. Other authors have reached the same conclusion by rather different approaches. For instance, [15] reject the ZIP against the ZINB using a score test in the spirit of [19]. [18] reject the ZIP against general smooth alternatives in the sense of Neyman. A generalized Poisson distribution to fit this data set has also been proposed by [9].

A simpler alternative to model this data is the NB distribution. The estimated parameters are $\hat{\theta} = 0.36$ and $\hat{t} = 1.89$, and the corresponding expected frequencies can be found in the fifth row of Table 6. At first sight it seems the fit provided by the NB is slightly better than the one furnished by the ZIP. To confirm this feature, we adapt the nonparametric procedure described in Subsection 3.2 to test the null hypothesis that the data come from a NB distribution against the alternative that the data come from a distribution that dominates the NB in the convex order.

We have used bootstrap estimates of the inverse of the CV to conclude that in this case $\Lambda_{k:k}$ with $k \approx 8$ yields an appropriate test (see Figure 2(b)). The p-values of the tests for $k = 4, 6, 8, 10, 12$ are all above 0.33. Therefore, we cannot reject the null hypothesis and conclude that the NB distribution accounts for the dispersion of the data better than the ZIP model.

8. Discussion

We introduce a new methodology to detect overdispersion based on the comparison of the expected sample extremes of two variables. The only required ingredient is that the involved models can be compared according to the stochastic convex order, which is a well-known variability order. The proposed methodology is applied to various discrete distributions since we are interested in zero-inflated models. However, the same ideas work analogously for any pair of ordered distributions, not necessarily discrete.

We generate two families of test statistics: one generated through estimators of the expected maximum of a random sample of size $k \geq 2$ and the other one through the corresponding minimum. Therefore, these two families of discrepancies depend on the parameter k, which is the preselected sample size, and can be viewed as a tuning parameter.

Moreover, we deal with different testing problems. On the one hand, we test on the proportion of structural zeros in several zero-inflated models. In this case, the choice of k is not relevant since the null and the alternative hypotheses are determined by a parametric model and the technique mainly relies on an efficient estimation of the parameters of the model under both hypotheses. Hence, the simplest choice $k = 2$ is reasonable and recommended. For the test of the Poisson against the ZIP model, we obtain the asymptotic normality of the discrepancy (for $k = 2$) yielding a simple testing approach, very competitive against the well-known score test.

On the other hand, we also introduce a new general test to detect overdispersion. In this case, the alternative hypothesis includes all the distributions that dominate, in the convex order, a distribution under H_0. Therefore, the distributions under the alternative have more variability than the corresponding ones under the null hypothesis.

In this general overdispersion test the distributions under the alternative are not completely specified. Having a family of tests provides flexibility to select powerful tests under very different alternatives. Hence, the choice of a suitable tuning parameter is important and depends on the distributions considered both under the null and the alternative hypothesis. For instance, if the null is fixed as a Poisson distribution and the alternative is ZIP, then the simulations showed that large values of k work well. However, for the same H_0, if the alternative are negative binomials, $k = 2$ (the smallest possible value for k) gives the highest power.

To distinguish these two possibilities, we have developed an automatic

procedure to select a value of k generating a powerful test. The selection is carried out by taking into account that, under the alternative hypothesis, a discrepancy measure will detect better the differences between the null and the alternative if it has a large expectation and a small variance. Hence, we choose the value of k maximizing an estimate of the inverse of the coefficient of variation of the discrepancy. The ability of this procedure to select a good value of the tuning constant k has been reaffirmed in all the simulation studies we carried out.

The proposed methodology is rather general, flexible and easy to implement. However, it also has some limitations. At the present stage it seems that our methods cannot be easily adapted to the regression setting. Further research is needed in this direction to cover this important model.

Acknowledgements

We thank the two anonymous reviewers of this paper and the associate editor for their valuable comments and suggestions which led to an improved version of the initial manuscript. This research was supported by the Spanish MEC, Grant MTM2007-66632 and MTM2008-06281-C02-02 and by Comunidad de Madrid Grant S-0505/ESP/0158.

A. Proofs

PROOF OF THEOREM 1. We first note that the discrepancy $\Delta_{2:2} = \Delta_{2:2}(\hat{\theta}, \hat{p})$ given in (8) is a smooth function of the maximum likelihood estimates, $\hat{\theta}$ and $\hat{p}$. Therefore, the desired asymptotic distribution can be obtained combining the classical asymptotic theory for maximum likelihood estimators and the delta method.

According to the the asymptotic theory for maximum likelihood estimators, we have that:

$$\sqrt{n}(\hat{\theta} - \theta, \hat{p} - p)^t \longrightarrow_d N((0,0)^t, \Sigma), \quad n \to \infty,$$

where $N((0,0)^t, \Sigma)$ is a bivariate normal distribution centered at the origin with covariance matrix Σ. The matrix Σ is the inverse of the expected Fisher information matrix, that is, $\Sigma^{-1} = -\mathrm{E}_{\theta,p}[\ell''(Y;\theta,p)]$, where $\ell''(y;\theta,p)$ is the 2×2 matrix of second partial derivatives with respect to θ and p of the

log-likelihood function $\ell(y;\theta,p)$. Using this result, after some algebra it is possible to show that, under $\mathrm{H}_0 : p = 0$,

$$\sqrt{n}(\hat{\theta} - \theta, \hat{p})^t \longrightarrow_d \mathrm{N}((0,0)^t, \Sigma_0), \quad n \to \infty,$$

where

$$\Sigma_0 = \begin{pmatrix} \theta & 0 \\ 0 & (e^\theta - 1 - \theta)^{-1} \end{pmatrix}.$$

Now, let $\nabla\Delta_{2:2}(\theta,p)$ be the gradient of $\Delta_{2:2}(\theta,p)$ evaluated at (θ,p). Using the delta method (see e.g. [20], Theorem 3.1., p. 26) we deduce that, under $\mathrm{H}_0 : p = 0$,

$$\sqrt{n}\Delta_{2:2} \longrightarrow_d N(0, \sigma^2(\theta)), \quad n \to \infty, \tag{12}$$

where $\sigma^2(\theta) := \nabla\Delta_{2:2}(\theta,0)^t \cdot \Sigma_0 \cdot \nabla\Delta_{2:2}(\theta,0)$. Now we observe that

$$\left.\frac{\partial\Delta_{2:2}(\theta,p)}{\partial\theta}\right|_{p=0} = 0, \quad \left.\frac{\partial\Delta_{2:2}(\theta,p)}{\partial p}\right|_{p=0} = 2\theta - M_2(\theta) - \theta^2 e^{-2\theta}[I_0(2\theta) - I_2(2\theta)],$$

where the function M_2 is defined in (7). To obtain the last equality above we use the following properties of the modified Bessel functions of the first kind: $I_0'(x) = I_1(x)$ and $I_1'(x) = [I_0(x) + I_2(x)]/2$ (see [2], properties 9.6.27 and 9.6.29, p. 376). Replacing these partial derivatives and the matrix Σ_0 in the expression $\nabla\Delta_{2:2}(\theta,0)^t \cdot \Sigma_0 \cdot \nabla\Delta_{2:2}(\theta,0)$ yields

$$\begin{aligned}\sigma^2(\theta) &= \frac{(2\theta - M_2(\theta) - \theta^2 e^{-2\theta}[I_0(2\theta) - I_2(2\theta)])^2}{e^\theta - 1 - \theta} \\ &= \frac{\theta^2\left(1 - e^{-2\theta}\left[(1+\theta)I_0(2\theta) - I_1(2\theta) + \theta I_2(2\theta)\right]\right)^2}{e^\theta - 1 - \theta}.\end{aligned}$$

Finally, it is obvious that $\sigma(\hat{\theta})$ defined in (9) is a consistent estimator of the standard deviation $\sigma(\theta)$. As a consequence, from (12) we also deduce that the conclusion of Theorem 1 holds.

Proof of Proposition 2. We need to introduce some notation. Given two integrable random variables X and Y, it is said that X *is smaller than* Y *in the increasing convex order*, written $X \leq_{\mathrm{icx}} Y$, if $\mathrm{E}(\phi(X)) \leq \mathrm{E}(\phi(Y))$, for all increasing and convex function ϕ, provided the expectations exist. It is easy to see that

$$X \leq_{\mathrm{cx}} Y \text{ if and only if } X \leq_{\mathrm{icx}} Y \text{ and } \mathrm{E}X = \mathrm{E}Y. \tag{13}$$

Therefore, since all the variables considered in Proposition 2 have the same expectation θ, if suffices to show that they are ordered for the increasing convex order. Moreover, since the proof of parts (a), (b) and (c) with $i = 1$ are similar, we only consider the case of ZIP variables (part (b) of the proposition).

We first note that the family $\mathcal{P} := \{Y(\theta) : \theta \in [0, \infty)\}$, where $Y(\theta)$ is a Poisson random variable of mean $\theta \geq 0$ ($Y(0) \equiv 0$) is *stochastically increasing and convex* (see Example 8.A.2 in [17]). For $0 \leq p_1 < p_2 < 1$, we define the random variables (independent of the variables in $\mathcal{P}$) $\Theta_i = \frac{\theta}{1-p_i} B(1-p_i)$ ($i = 1, 2$), where $B(1-p_i)$ is a Bernoulli variable of parameter $1-p_i$. It is readily checked that $\Theta_1 \leq_{\rm cx} \Theta_2$. Therefore, a direct application of Theorem 8.A.14 (p. 362) in [17] yields $Y(\Theta_1) \leq_{\rm icx} Y(\Theta_2)$, and taking into account (13), we conclude $Y(\Theta_1) \leq_{\rm cx} Y(\Theta_2)$. Therefore, the proof of part (b) is finished since the ZIP variable $Y(\theta, p_i)$ has the same distribution as $Y(\Theta_i)$ ($i = 1, 2$).

The previous argument, based on the properties of stochastically increasing and convex families, cannot be used to prove part (c) with $i = 2$ since it has not been established yet whether the collection of negative binomial variables is stochastically increasing and convex in its second parameter. We therefore need to introduce another technique inspired in the ideas used to prove Lemma 10 in [5]. Fix $t > 0$, $\theta > 0$ and $0 \leq p_1 < p_2 < 1$ and let $Z_2(t, \theta, p_i)$ ($i = 1, 2$) be the ZINB distributions defined in Section 4. Taking into account Lemma 9 in [5] and Theorem 3.A.44 (p. 133) in [17], to prove part (c) (with $i = 2$) it is enough to show that the function

$$p(k) := \Pr(Z_2(t, \theta, p_1) = k) - \Pr(Z_2(t, \theta, p_2) = k), \quad k \geq 0, \tag{14}$$

has two changes of sign, being the sign sequence $-, +, -$. To show this, we first consider the function

$$\varphi(k) := \frac{\Pr(Z_2(t, \theta, p_1) = k)}{\Pr(Z_2(t, \theta, p_2) = k)}, \quad k \geq 0.$$

After some simple computations, it is easy to check that the function $f(p) := \Pr(Z_2(t, \theta, p) = 0)$ is an increasing function of $p \in [0, 1)$. Therefore, $\varphi(0) < 1$. Also, since

$$\frac{\varphi(k+1)}{\varphi(k)} = \frac{1 - p_2 + \theta t}{1 - p_1 + \theta t} =: c < 1, \quad k \geq 1,$$

we have that $\varphi(k) = c^{k-1}\varphi(1)$ $(k \geq 1)$ and this entails $\varphi(k) \downarrow 0$ as $1 \leq k \uparrow \infty$. Moreover, the equality $\sum_{k=0}^{\infty} \Pr(Z_2(t,\theta,p_1) = k) = 1 = \sum_{k=0}^{\infty} \Pr(Z_2(t,\theta,p_2) = k)$ yields $\varphi(1) > 1$. This implies the desired result and the proof is complete.

Proof of Proposition 3. In the case $p = 0$, parts (a)-(d) follow from Lemmas 5 and 10 in [5] and Theorem 3.A.44 (p. 133) in [17]. Therefore, using that the convex order is closed under mixtures (see Theorem 3.A.12 (p. 119) of [17]), we conclude that for any fixed $0 < p < 1$, (a)-(c) and the first stochastic inequality in (d) are valid. To finish, we observe that the distribution of $Z_1(t,\theta,p)$ is the same as the distribution of $Z_2(t(1-p),\theta,p)$ and applying part (c) of Proposition 3, we get $Z_2(t(1-p),\theta,p) \leq_{\text{cx}} Z_2(t,\theta,p)$. This shows that $Z_1(t,\theta,p) \leq_{\text{cx}} Z_2(t,\theta,p)$ and the proof is complete.

References

[1] Aban, I.B., Cutter, G.R. and Mavinga, N., 2009. Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. Computational Statistics and Data Analysis 53, 820–833.

[2] Abramowitz, M. and Stegun, I.A., 1965. Handbook of Mathematical Functions. Dover, New York.

[3] Böhning, D., P. Schlattmann, P. and Lindsay, B., 1992. Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithms. Biometrics 48, 283–303.

[4] Böhning, D., Dietz, E. and Schlattmann, P., 1999. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society Ser. A 162, 195–209.

[5] de la Cal, J. and Cárcamo, J., 2005. Inequalities for expected extreme order statistics. Statistics and Probability Letters 73, 219–231.

[6] Campbell, M.J., Machin, D. and DArcangues, C., 1991. Coping with extra-Poisson variability in the analysis of factors influencing vaginal ring expulsions. Statistics in Medicine 10, 241–251.

[7] Douglas, J. B., 1994. Empirical fitting of discrete distributions. Biometrics 50, 576–579.

[8] El-Shaarawi, A. 1985. Some goodness-of-fit methods for the Poisson plus added zero distribution. Applied and Environmental Microbiology 49, 1304–1306.

[9] Gupta, P.L., Gupta, R.C. and Tripathi, R.C., 1996. Analysis of zero-adjusted count data. Computational Statistics and Data Analysis 23, 207–218.

[10] He, B., Gupta, P.L., Xie, M. and Goh, T.N., 2003. A confidence interval test for testing Poisson model against zero-inflated Poisson model. Journal of Applied Statistical Science 12, 209–220.

[11] Jansakul, N. and Hinde, J.P., 2002. Score tests for zero-inflated Poisson models. Computational Statistics and Data Analysis 40, 75–96.

[12] Johnson, N.L., Kemp, A.W. and Kotz, S., 2005. Univariate Discrete Distributions, 3rd ed. Wiley, New York.

[13] Leroux, B.G. and Puterman, M.L., 1992. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. Biometrics 48, 545–558.

[14] Nie, L., Wu, G., Brockman, F.J. and Zhang, W., 2006. Integrated analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: Zero-Inflated Poisson regression models to predict abundance of undetected proteins. Bioinformatics 22, 1641–1647.

[15] Ridout, M., Hinde, J. and Demétrio, C., 2001. A score test for testing a zero-infalted Poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57, 219–223.

[16] Rigby, R.A., Stasinopoulos, D.M. and Akantziliotou, C., 2008. A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. Computational Statistics and Data Analysis 53, 381–393.

[17] Shaked, M. and Shanthikumar, J.G., 2007. Stochastic Orders. Springer, New York.

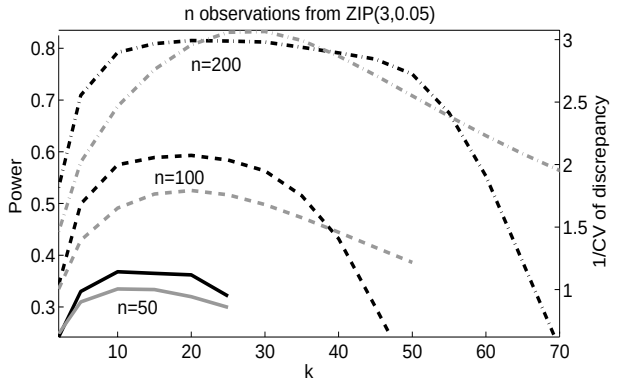
[18] Thas, O. and Rayner, J.C.W., 2005. Smooth test for the zero-inflated Poisson distribution. Biometrics 61, 808–815.

[19] van den Broek, J., 1995. A score test for zero inflation in a Poisson distribution. Biometrics 51, 738–743.

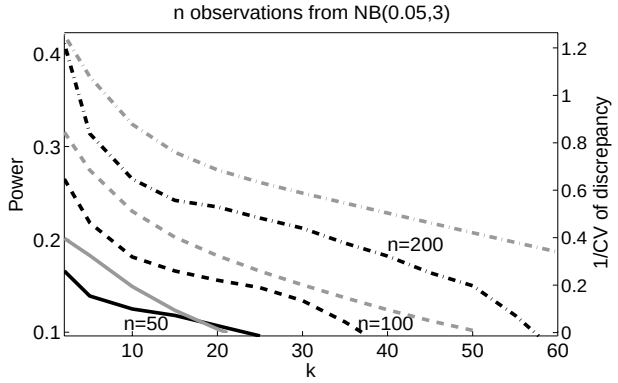
[20] van der Vaart, A.W., 1998. Asymptotic Statistics. University Press, Cambridge.

[21] Xie, M., He, B. and Goh, T.N., 2001. Zero-inflated Poisson model in statistical process control. Computational Statistics and Data Analysis 38, 191–201.

Figure 1: Power (in black) of the overdispersion test for the Poisson family and 1/CV of the discrepancy (in grey).

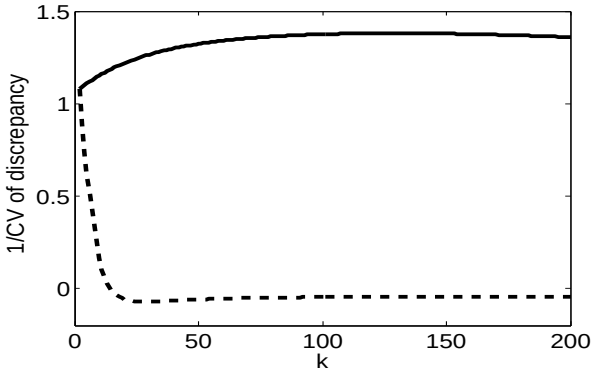


(a)

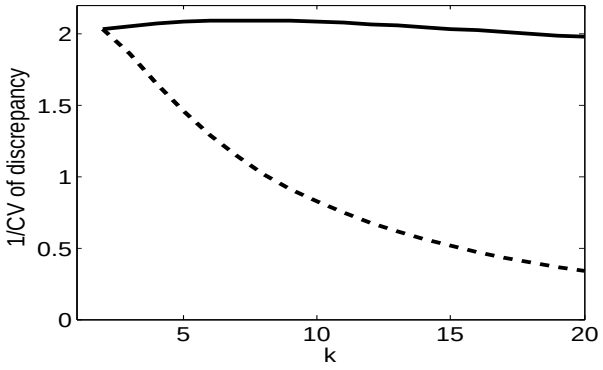


(b)

Figure 2: Bootstrap estimates of CV^{-1} for $\Lambda_{k:k}$ (solid line) and $\Lambda_{1:k}$ (dashed line) for several values of k. Panel (a): Testing ZIP vs. overdispersion. Panel (b): Testing negative binomial vs. overdispersion.



(a)



(b)