AStA Adv Stat Anal (2016) 100:239–257
DOI 10.1007/s10182-015-0255-7

ORIGINAL PAPER

Modelling of count data using nonparametric mixtures

Chew-Seng Chee[1]

Received: 9 February 2015 / Accepted: 9 September 2015 / Published online: 18 September 2015

Abstract Nonparametric modelling of count data is partly motivated by the fact that using parametric count models not only runs the risk of model misspecification but also is rather restrictive in terms of local approximation. Accordingly, we present a framework of using nonparametric mixtures for flexible modelling of count data. We consider the use of the least squares function in nonparametric mixture modelling and provide two algorithms for least squares fitting of nonparametric mixtures. Two illustrations of the framework are given, each with a particular nonparametric mixture. One illustration is the use of the nonparametric Poisson mixture for general modelling purposes. The other illustration is concerned with modelling of count data from some decreasing distribution, in which the Poisson mixture distribution is less appropriate, for its fitted distribution might not be a decreasing distribution. We define a mixture distribution called the discrete decreasing beta mixture distribution that always has fitted probabilities conforming with the assumption of decreasing probabilities. Through numerical studies, we demonstrate the performance of nonparametric mixtures as modelling tools.

Keywords Discrete decreasing distribution · Least squares estimation · Nonparametric mixtures · V-fold cross-validation

1 Introduction

For modelling count data, parametric models have been commonly used as practical instruments, the standard one being that based on the Poisson distribution. As various

✉ Chew-Seng Chee
chee@umt.edu.my

[1] School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia

kinds of counts such as heavy-tailed, multimodal, overdispersed and zero-inflated data are observed from the real-world phenomena, the limitations and inadequacies of parametric models based on basic counting distributions in describing counts have become obvious. A natural means of addressing problems specific to count data would be by considering a mixture distribution on $\mathcal{X} \subseteq \mathbb{N}_0$, the set of nonnegative integers, having a density of the form

$$p(x; G) = \int p(x; \theta)\, \mathrm{d}G(\theta), \quad x \in \mathcal{X}, \tag{1}$$

where $p(x; \theta)$ denotes a component density and G denotes a mixing distribution on $\Theta \subseteq [0, \infty)$. If G belongs to a parametric family of (continuous) distributions, then (1) is referred to as a parametric mixture distribution. See, e.g., Gupta and Ong (2005), Karlis and Xekalaki (2005), Nikoloulopoulos and Karlis (2008) and Rigby et al. (2008) for numerous parametric Poisson mixture models for count data. On the other hand, if G is a discrete distribution with finitely many support points, then (1) is referred to as a finite mixture distribution. As noted by Cameron and Trivedi (2013), finite mixtures have several advantages over parametric mixtures. The former not only relaxes distributional assumptions for the mixing distribution but also may approximate any distribution arbitrarily well. Besides, it is computationally inconvenient to even calculate the probabilities of parametric mixture distributions without closed form density functions, while fast algorithms are available for fitting finite mixture distributions.

The density of a J-component finite mixture distribution can be defined as:

$$p(x; G) = \sum_{j=1}^{J} \pi_j\, p(x; \theta_j), \quad x \in \mathcal{X}, \tag{2}$$

with $G(\theta) = \sum_{j=1}^{J} \pi_j \delta_{\theta_j}(\theta)$, where $\pi_j > 0$ for all j, $\sum_{j=1}^{J} \pi_j = 1$ and δ_ϑ denotes a degenerate distribution at ϑ. Depending on the treatments of the number J of components, the mixture model formulation can be classified as either parametric or nonparametric. When the parameter J is regarded as known (and fixed), called by Böhning et al. (1992) the fixed support size case, we have a parametric mixture formulation. Another case is called by the authors the flexible support size case in which the parameter J is left unspecified and is to be estimated from the data. Observe that the latter case involves an entirely unknown discrete mixing distribution. For a mixture model with such a mixing distribution, we shall refer to it as a nonparametric mixture model owing to the nonparametric nature of the mixing distribution.

Several nonparametric mixture models have been considered in the literature. Yet, the most popular one is the nonparametric Poisson mixture model. A pioneering study on the nonparametric Poisson mixture model and its use for analyzing count data was done by Simar (1976) in which the maximum likelihood approach to estimating the mixing distribution was introduced. On the contrary, Karlis and Xekalaki (2001) employed the minimum Hellinger distance estimation for the nonparametric Poisson mixture model. Later, Böhning and Patilea (2005) studied the nonparametric maximum likelihood estimator of a mixture of power series distributions and mentioned some

applications in which such a mixture is potentially useful. In fact, many common discrete distributions such as the geometric, negative binomial and Poisson distributions are members of the family of power series distributions.

Recently, the usefulness of nonparametric mixtures in the context of estimation of a discrete distribution under some constraint such as monotonicity or convexity was demonstrated. Estimation of a monotone decreasing distribution on $\mathbb{N}_0$ via the method of maximum likelihood was considered by Jankowski and Wellner (2009). Since a monotone decreasing distribution on $\mathbb{N}_0$ may be written as a mixture of the form (2) with component density

$$p(x;\theta_j) = \frac{1}{\theta_j}, \quad x \in \{0,\ldots,\theta_j - 1\},$$

where θ_j is a positive integer, the shape-constrained nonparametric estimation problem becomes that of estimating the mixing distribution in the discrete uniform mixture model. On the other hand, Durot et al. (2013) showed that a convex decreasing distribution on $\mathbb{N}_0$ can be decomposed as a mixture of discrete triangular distributions with component density

$$p(x;\theta_j) = \frac{2(\theta_j - x)}{\theta_j(\theta_j + 1)}, \quad x \in \{0,\ldots,\theta_j - 1\},$$

where θ_j is a positive integer, and studied the least squares estimation of a convex decreasing distribution. We remark that while the least squares estimator of a discrete convex decreasing distribution is unique, there can be a nonuniqueness problem of the nonparametric maximum likelihood estimate of the mixing distribution in the discrete triangular mixture model, which leads to more than one maximum likelihood fitted triangular mixture distribution. In contrast, both the maximum likelihood and least squares estimators of a discrete monotone decreasing distribution are identical.

In this paper, we present a framework of using nonparametric mixture models for modelling count data. See Wang and Chee (2012) for a framework for modelling continuous data based on nonparametric mixture models. Our work here, which investigates the use of the least squares criterion in nonparametric mixture modelling of count data, is distinguished from their work with respect to the estimation method and model type. Also referred to by some authors as the minimum L_2 distance estimation method, the least squares estimation method indeed has already been applied to parametric modelling of data; see, e.g., Scott (2001) and Harris and Shen (2011). The fact that the least squares approach to fitting nonparametric mixture models has received relatively less consideration in the literature motivates us to conduct this investigation. Moreover, the least squares estimator exhibits more robustness to data contamination than the maximum likelihood estimator, which we now illustrate by first fitting nonparametric Poisson mixtures to contaminated data sets, each formed by adding some contaminating value to an uncontaminated sample, and then computing the fitted means. The uncontaminated data set of size 25 we used was generated by Karlis and Xekalaki (1998) from two equally weighted Poisson distributions with means 1 and 3. The frequencies of 0, 1, 2, 3, 4, 5 and 6 are 8, 4, 5, 1, 3, 2 and 2,

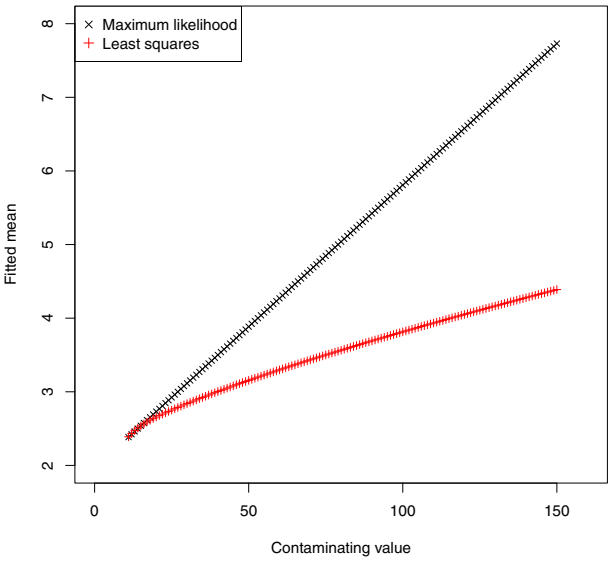

Fig. 1 Means of the maximum likelihood and least squares fitted Poisson mixture distributions for contaminating values of 11 to 150

respectively, resulting in a sample mean of 2.04. Figure 1 clearly shows that the fitted means by the least squares method are affected to a lesser degree as compared with those by the maximum likelihood method. Thus, the least squares approach definitely serves the purpose of offering an alternative to other approaches such as the maximum likelihood approach. After introducing the framework, two specific illustrations, each with a particular nonparametric mixture, are given. First, nonparametric Poisson mixture modelling is briefly mentioned. Next, we describe the case of using a newly defined model here, called the nonparametric discrete decreasing beta mixture model, for modelling data arising from some discrete decreasing distribution. Also, simulated and real-world count data are used to demonstrate the performance of nonparametric mixtures as modelling tools.

2 Modelling of count data using nonparametric mixtures

Suppose we have a random sample of counts from a discrete distribution either with no truncation or with left truncation and we are interested in modelling the distribution of these counts. For accomplishing this modelling task, nonparametric mixture modelling of count data, a technique that describes the distribution of counts based on nonparametric mixture models, is adopted. This technique requires fitting some nonparametric mixture model to the observed counts, which in effect involves estimating the unknown discrete mixing distribution in that model. In this section, we describe the least squares approach, along with fitting algorithms, to estimating the mixing distribution nonparametrically and also mention the case of using the nonparametric Poisson mixture model for modelling count data.

2.1 Least squares fitting of nonparametric mixtures

The least squares criterion function based on the observed counts $x_1, \ldots, x_n$ is defined as:

$$\mathcal{L}(G) = \sum_{x \in \mathcal{X}} \{p(x; G) - \widetilde{p}(x)\}^2 ,$$

where $\widetilde{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{x_i = x\}}$ denotes the relative frequency estimator. We consider the problem of minimizing the least squares function over the set $\mathcal{G}$ of all discrete mixing distributions on Θ. An equivalent minimization formulation for the problem just described is given by

$$\underset{G \in \mathcal{G}}{\text{minimize}} \quad \mathcal{Q}(G) = \sum_{x \in \mathcal{X}} p^2(x; G) - 2 \sum_{x \in \mathcal{X}} \widetilde{p}(x) p(x; G). \tag{3}$$

The minimizing function of the problem (3) is referred to as the nonparametric least squares estimate (NPLSE) of G. Once the NPLSE is available, the mixture distribution corresponding to this NPLSE can be straightaway constructed and then be used for making inferences.

Denoting the probability mass vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)^\top$ and the support point vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^\top$, we also write $\mathcal{Q}(G) \equiv \mathcal{Q}(\boldsymbol{\pi}, \boldsymbol{\theta})$. Now, we express the objective function in (3) more compactly in matrix form as:

$$\mathcal{Q}(\boldsymbol{\pi}, \boldsymbol{\theta}) = \boldsymbol{\pi}^\top \mathbf{D} \boldsymbol{\pi} - 2 \boldsymbol{\pi}^\top \mathbf{b}, \tag{4}$$

where the element in row j' and column j of matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\theta})$ is given by

$$D_{j'j} = \sum_{x \in \mathcal{X}} p(x; \theta_{j'}) p(x; \theta_j),$$

and the jth element of vector $\mathbf{b} = \mathbf{b}(\boldsymbol{\theta})$ is given by

$$b_j = \sum_{x \in \mathcal{X}} \widetilde{p}(x) p(x; \theta_j).$$

An important tool that is of great importance in finding the NPLSE is the gradient function:

$$\begin{aligned} d(\theta; G) &\equiv \left. \frac{\mathrm{d}\mathcal{Q}\{(1-\epsilon)G + \epsilon \delta_\theta\}}{\mathrm{d}\epsilon} \right|_{\epsilon = 0} \\ &= 2 \left\{ \sum_{x \in \mathcal{X}} p(x; \theta) p(x; G) - \sum_{x \in \mathcal{X}} \widetilde{p}(x) p(x; \theta) \right. \\ &\qquad \left. - \sum_{x \in \mathcal{X}} p^2(x; G) + \sum_{x \in \mathcal{X}} \widetilde{p}(x) p(x; G) \right\}. \end{aligned}$$

The gradient function is vital as it characterizes the NPLSE. Specifically, all gradient values evaluated at the NPLSE are at least zero. Furthermore, the set of support points of the NPLSE is a subset of the set of points with zero gradient values. These results for the NPLSE are based on the results mentioned in Chee and Wang (2014). The reader interested in such results in a more general setting may refer to Chee and Wang (2013).

2.2 Fitting algorithms

The fitting of nonparametric mixtures by least squares requires some algorithm for numerical minimization. For a fast computational method for fitting nonparametric mixtures, we mention the work of Wang (2007) in which the constrained Newton method (CNM) with multiple support point inclusion for finding the nonparametric maximum likelihood estimate of a mixing distribution was proposed. Recently, Chee and Wang (2013) developed a variant of the CNM algorithm that can be used to compute the nonparametric minimum quadratic distance estimate of a mixing distribution, which was subsequently applied by Chee and Wang (2014) to compute the NPLSE for a continuous mixture model called the k-monotone model. The CNM algorithm developed by Chee and Wang (2013) is also straightforward applicable to computing the NPLSE for a discrete mixture model whose mixing distribution is defined on a continuous domain. At each iteration, this algorithm enlarges the support point vector by including all local minima of the gradient function, updates the probability mass vector by holding fixed the enlarged support point vector and shrinks the enlarged support point vector according to the updated probability mass vector.

We now describe in detail the computation of the NPLSE via the CNM algorithm. Denote by $\|\cdot\|_2$ the L_2-norm, and let $\mathbf{0} = (0, \ldots, 0)^\top$ and $\mathbf{1} = (1, \ldots, 1)^\top$. With fixed $\boldsymbol{\theta}$, minimizing (4) with respect to $\boldsymbol{\pi}$ is the same as to

$$\text{minimize} \quad \|\mathbf{R}\boldsymbol{\pi} - \mathbf{d}\|_2^2, \ \text{subject to } \boldsymbol{\pi}^\top \mathbf{1} = 1, \ \boldsymbol{\pi} \geq \mathbf{0}, \tag{5}$$

where $\mathbf{R} = \mathbf{R}(\boldsymbol{\theta})$ satisfies $\mathbf{D} = \mathbf{R}^\top \mathbf{R}$ and $\mathbf{d} = \mathbf{d}(\boldsymbol{\theta})$ is the solution of $\mathbf{R}^\top \mathbf{d} = \mathbf{b}$. Problem (5) can be solved numerically by the nonnegative least squares (NNLS) algorithm of Lawson and Hanson (1974) after employing the method of Dax (1990), which transforms it into the least squares problem with only nonnegativity constraints:

$$\text{minimize} \quad ||\mathbf{P}\tilde{\boldsymbol{\pi}}||_2^2 + |\tilde{\boldsymbol{\pi}}^\top \mathbf{1} - 1|^2, \ \text{subject to } \tilde{\boldsymbol{\pi}} \geq \mathbf{0}, \tag{6}$$

where $\mathbf{P} \equiv (\mathbf{r}^1 - \mathbf{d}, \ldots, \mathbf{r}^J - \mathbf{d})$, with $\mathbf{r}^j$ being the jth column of $\mathbf{R}$. Dax established that if $\tilde{\boldsymbol{\pi}}$ solves problem (6), then $\tilde{\boldsymbol{\pi}}/\tilde{\boldsymbol{\pi}}^\top \mathbf{1}$ solves problem (5). After $\boldsymbol{\pi}$ is updated, those support points with zero probability masses are discarded before the next iteration. The support point vector is expanded by adding the set of points that locally minimizes the gradient function. The CNM algorithm is given as follows:

Algorithm 1 Set $s = 0$. From an initial estimate G_0 with a support set of finite cardinality and $\mathcal{Q}(G_0) < \infty$, repeat the following steps:

step 1: Compute all local minima $\theta_{s1}^\star, \ldots, \theta_{sM_s}^\star$ of $d(\theta; G_s), \theta \in \Theta$. Stop, if $\min_{1 \leqslant m \leqslant M_s} \{d(\theta_{sm}^\star; G_s)\} = 0$.

step 2: Set $\boldsymbol{\theta}_s^+ = (\boldsymbol{\theta}_s^\top, \theta_{s1}^\star, \ldots, \theta_{sM_s}^\star)^\top$ and $\boldsymbol{\pi}_s^+ = (\boldsymbol{\pi}_s^\top, \mathbf{0}^\top)^\top$. Find $\boldsymbol{\pi}_{s+1}^+$ by solving problem (5), with $\mathbf{R}$ and $\mathbf{d}$ replaced by $\mathbf{R}_s^+ = \mathbf{R}(\boldsymbol{\theta}_s^+)$ and $\mathbf{d}_s^+ = \mathbf{d}(\boldsymbol{\theta}_s^+)$ respectively.

step 3: Remove all support points with zero probability masses in $\boldsymbol{\pi}_{s+1}^+$, which gives G_{s+1} with $\boldsymbol{\pi}_{s+1}$ and $\boldsymbol{\theta}_{s+1}$. Set $s = s + 1$.

Expanding the support point vector by the technique of including all local minima of the gradient function works very well for the case of a continuous gradient function, but for that of a discrete gradient function, it is less suitable due to the irregularity of the function which can cause too many local minima to be included into the support point vector (see Wang 2008). Thus, we need to make appropriate modifications in Algorithm 1 for the efficient computation of the NPLSE when we have a gradient function which is of discrete type. Suppose that now, for some nonnegative integers L and U, $\mathcal{G}$ is the set of all discrete mixing distributions on $\Theta = \{L, L+1, \ldots, U\}$. Adapting the support set expansion technique of Wang (2008), our new technique includes in the support set the points that have the lowest gradient values between inclusively every two neighbouring points in an increasing ordered set that is formed by the union of the support set and $\{L, U\}$. The discrete-space extension of the CNM algorithm is given below:

Algorithm 2 Set $s = 0$. From an initial estimate G_0 with a support set $\mathcal{S}_0$ of finite cardinality and $\mathcal{Q}(G_0) < \infty$, repeat the following steps.

step 1: Compute $d(\theta; G_s)$ for all $\theta \in \Theta = \{L, L+1, \ldots, U\}$. Stop, if $\min_{L \leqslant \theta \leqslant U}\{d(\theta; G_s)\} = 0$.

step 2: Form the increasing ordered set $\{\theta_{s1}, \ldots, \theta_{sM_s}\}$ by first combining elements of $\mathcal{S}_s$ and $\{L, U\}$ and then sorting them in an increasing order.

step 3: Find $\theta_{sl}^\star = \arg\min_{\theta_{sl} \leqslant \theta \leqslant \theta_{s(l+1)}}\{d(\theta; G_s)\}$ for all $l \in \{1, \ldots, M_s - 1\}$.

step 4: Set elements of $\mathcal{S}_s^+ \equiv \mathcal{S}_s \cup \{\theta_{s1}^\star, \ldots, \theta_{s(M_s-1)}^\star\}$ as components of $\boldsymbol{\theta}_s^+$ and $\boldsymbol{\pi}_s^+ = (\boldsymbol{\pi}_s^\top, \mathbf{0}^\top)^\top$. Find $\boldsymbol{\pi}_{s+1}^+$ by solving problem (5), with $\mathbf{R}$ and $\mathbf{d}$ replaced by $\mathbf{R}_s^+ = \mathbf{R}(\boldsymbol{\theta}_s^+)$ and $\mathbf{d}_s^+ = \mathbf{d}(\boldsymbol{\theta}_s^+)$ respectively.

step 5: Remove all support points with zero probability masses in $\boldsymbol{\pi}_{s+1}^+$, which gives G_{s+1} with $\boldsymbol{\pi}_{s+1}$ and $\boldsymbol{\theta}_{s+1}$. Set $s = s + 1$.

2.3 The case of using the nonparametric Poisson mixture model

Mixture modelling of count data is thus far largely undertaken by Poisson mixtures. Basically, Poisson mixtures are able to cater for a wide range of density shapes, and hence they are well suited for use as general modelling tools. The density of the mixture of Poisson distributions with a discrete mixing distribution on $\Theta = [0, \infty)$ is given by

$$p(x; G) = \sum_{j=1}^{J} \frac{\pi_j e^{-\theta_j} \theta_j^x}{x!}, \quad x \in \mathbb{N}_0.$$

Recently, Harris and Shen (2011) discussed the least squares approach to fitting a finite Poisson mixture model and performed the estimation of G in the fixed support

size case. Since making a priori choice of J for a particular data set is rather arbitrary, we estimate G nonparametrically by the method of least squares. For the case of the Poisson mixture model, the elements of matrix $\mathbf{D}$ and vector $\mathbf{b}$ in (4) are given by

$$D_{j'j} = e^{-(\theta_{j'}+\theta_j)} I_0(2\sqrt{\theta_{j'}\theta_j}),$$
$$b_j = \sum_{x\in\mathbb{N}_0} \frac{\widetilde{p}(x)e^{-\theta_j}\theta_j^x}{x!},$$

where I_0 denotes the modified Bessel function of the first kind of order 0. Since we have a continuous gradient function here, we use Algorithm 1 for computing the NPLSE.

To illustrate the model and method, we choose a set of overdispersed counts studied by Shmueli et al. (2005). This data set, containing 3168 counts of the number of article clothings sold per quarter, was used by the authors to demonstrate the usefulness and flexibility of the Conway–Maxwell–Poisson distribution. With two parameters, this model has the capability for addressing overdispersion and underdispersion with respect to the Poisson model. The least squares fitted Conway–Maxwell–Poisson distribution is plotted in the top left panel of Fig. 2, together with the observed relative frequency distribution and our least squares fitted Poisson mixture distribution. The top right panel of Fig. 2 shows the gradient curve evaluated at the NPLSE for this particular data set. From the residuals of both fits in the bottom panel of Fig. 2, we find that in general our nonparametric fit improves over the parametric fit in the high-density region at the expense of only a negligible loss in the tail region.

We end this subsection by only very briefly mentioning that in situations in which the count variables are truncated, the nonparametric truncated Poisson mixture model can be applied for properly modelling such truncated count data. For example, one can consider a mixture of zero-truncated Poisson distributions for hospital length of stay count data which structurally preclude zeros to be observed.

3 Modelling of count data from decreasing distributions

In count data settings, it is not uncommon in real applications to observe relative frequency distributions which are decreasing (nonincreasing). Thus, discrete distributions, particularly with decreasing shapes on their supports, are attractive and empirically motivated for modelling such data. In practice, discrete decreasing distributions are conceived to be important and useful for fitting zero-truncated count data such as species abundance count data in ecology (Durot et al. 2013) and word frequency count data in linguistics (Baayen 2001). Of course, the motivation for their uses can absolutely be based on prior knowledge or theoretical justification about the population distributions.

Suppose we have a random sample of counts from some unknown decreasing distribution, be it monotone decreasing, convex decreasing, geometric decreasing, etc. Naturally, it would be desirable for a fitted model to conform with the known qualitative information about the population distribution. While the nonparametric Poisson mixture model as a general modelling tool provides a great flexibility in

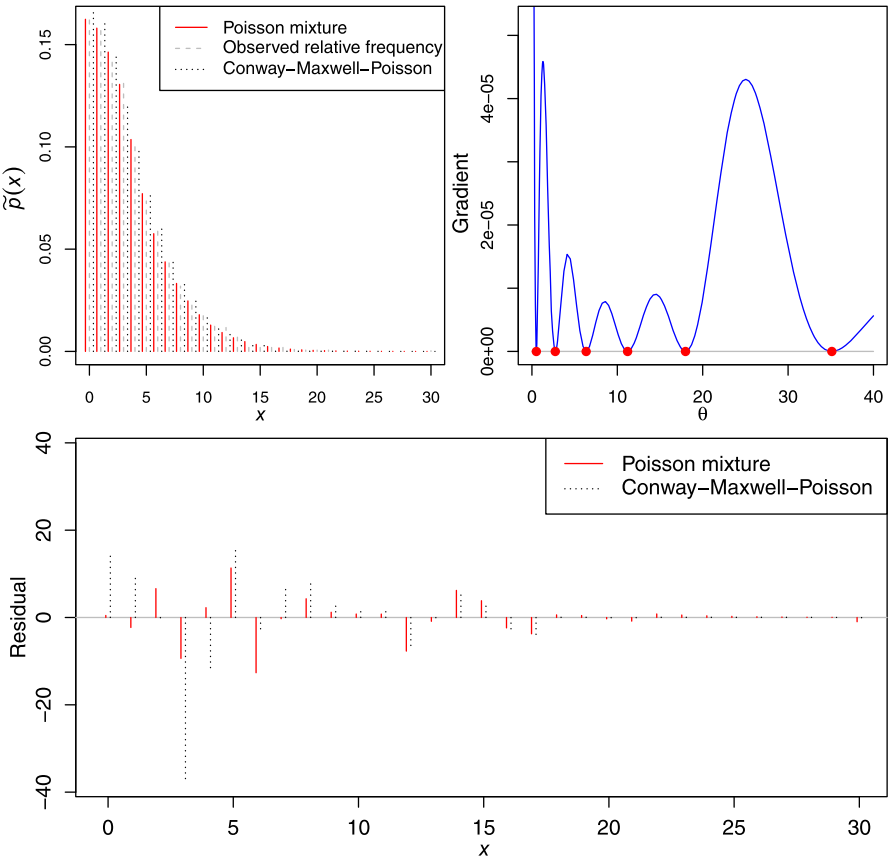

Fig. 2 (*Top left*) The relative frequency distribution of the quarterly sales data (Shmueli et al. 2005), together with the fits of the Conway–Maxwell–Poisson and Poisson mixture distributions. (*Bottom*) The residual plot of both fits. (*Top right*) The gradient function at convergence along with the support points (*red solid dots*) of the NPLSE

describing count data, it does not guarantee the probabilities of its fit to be decreasing. Thus, in this sense, the Poisson mixture model is a less suitable model although still can be used. Motivated by the limitation of the Poisson mixture model and inspired by the work of Balabdaoui and Wellner (2010), we shall define a mixture model with decreasing probability property that is useful and appropriate in this modelling context.

3.1 The discrete decreasing beta distribution and its mixture

The probability density function of the beta distribution of the first kind is

$$f(x) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}\mathrm{B}(\alpha,\beta)}, \quad a < x < b,$$

where $a, b \in \mathbb{R}$, $\alpha, \beta > 0$ and B denotes the beta function. The standard beta density of the first kind equals $f(x)$ with $a = 0$ and $b = 1$. Recently, Punzo and Zini (2012) introduced a discrete analogue of the beta distribution of the first kind. Its density is

$$p(x; \alpha, \beta) = \frac{(x + 1 - a)^{\alpha-1}(b + 1 - x)^{\beta-1}}{\sum_{i=a}^{b}(i + 1 - a)^{\alpha-1}(b + 1 - i)^{\beta-1}}, \quad x \in \{a, \ldots, b\}, \tag{7}$$

where $a, b \in \mathbb{N}_0$ and $\alpha, \beta \in \mathbb{R}$. In an attempt to facilitate the interpretation of the parameters of the discrete beta distribution (7), Punzo (2010) offered a reparameterized version of the distribution, which can conveniently serve as kernel functions in kernel-based methods for estimating a probability mass function (Punzo 2010) and graduating mortality rates (Mazza and Punzo 2011).

Based on (7), by letting $b + 1 = \theta \in \{a + 1, a + 2, \ldots\}$, $\alpha = 1$ and $\beta = k \in \{1, 2, \ldots\}$, we define the discrete decreasing beta distribution, with density

$$p_k(x; \theta) = \frac{(\theta - x)_+^{k-1}}{\sum_{i=1}^{\theta-a} i^{k-1}}, \quad x \in \{a, a + 1, \ldots\},$$

where $(z)_+ \equiv z\mathbb{I}_{\{z \geq 0\}}$. For $k = 1$, this distribution becomes the discrete uniform distribution, whereas for $k = 2$, it reduces to the discrete left triangular distribution. The limiting form of the discrete decreasing beta distribution as $k \to \infty$ is the Dirac distribution located at a. Five different density functions of the discrete beta distribution on $\mathbb{N}_0$ which have strictly decreasing probabilities are shown in the left panel of Fig. 3, from which it can be seen that, as k increases, the majority of the mass is concentrated in the neighbourhood of zero.

The density of the mixture of discrete decreasing beta distributions is defined as:

$$p_k(x; G) = \sum_{j=1}^{J} \frac{\pi_j(\theta_j - x)_+^{k-1}}{\sum_{i=1}^{\theta_j-a} i^{k-1}}, \quad x \in \{a, a + 1, \ldots\}. \tag{8}$$

Here, G is a discrete mixing distribution on $\Theta = \{a + 1, a + 2, \ldots\}$ and k is a common parameter to all component distributions, which controls the smoothness of the mixture distribution. Similarly, five discrete decreasing beta mixtures are depicted in the right panel of Fig. 3.

3.2 Fitting of the mixture of discrete decreasing beta distributions with a fixed k-value

As a flexible means for modelling count data from some decreasing distribution, we consider using the mixture of discrete decreasing beta distributions (8). Due to the fact that the joint estimation of G and k by least squares always yields an estimate of 1 for k, as illustrated later in Fig. 5, we shall fit the mixture of discrete decreasing beta distributions with a fixed k-value but an a priori unspecified number J of component

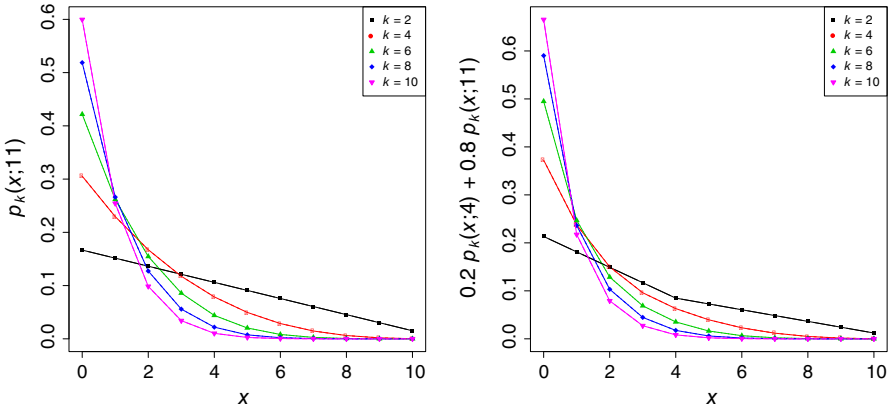

Fig. 3 (*Left*) Density functions of the discrete decreasing beta distribution on $\mathbb{N}_0$ for five different values of k with $\theta = 11$. (*Right*) Five discrete decreasing beta mixtures, each with a different k-value but an identical two-point mixing distribution at $\theta_1 = 4$ and $\theta_2 = 11$ with masses $\pi_1 = 0.2$ and $\pi_2 = 0.8$, respectively. Note that the lines joining the probabilities do not imply continuity but for a better visualization of the shapes of the distributions

distributions to the data. The common parameter k will be treated as a bandwidth parameter, as in the kernel-based density estimation setting, which is subject to selection. Actually, this kind of joint estimation problem is essentially similar to the problem encountered by Wang and Chee (2012) in the maximum likelihood estimation of a density function using the nonparametric normal mixture model.

For a fixed k-value, we consider the following minimization problem:

$$\underset{G\in\mathcal{G}}{\text{minimize}} \quad \mathcal{Q}_k(G) = \sum_{x\geq a} p_k^2(x;G) - 2\sum_{x\geq a} \widetilde{p}(x)\,p_k(x;G). \tag{9}$$

In this particular case, the $D_{j'j}$ and b_j become

$$D_{j'j} = \frac{\sum_{x=a}^{\min\{\theta_{j'},\theta_j\}-1} \{(\theta_{j'}-x)(\theta_j-x)\}^{k-1}}{\sum_{i=1}^{\theta_{j'}-a} i^{k-1} \sum_{i=1}^{\theta_j-a} i^{k-1}},$$

$$b_j = \frac{\sum_{x=a}^{\theta_j-1} \widetilde{p}(x)(\theta_j-x)^{k-1}}{\sum_{i=1}^{\theta_j-a} i^{k-1}}.$$

Since the mixing distribution G considered here is defined on a discrete space, we apply Algorithm 2 for the fitting of the mixture of discrete decreasing beta distributions with a fixed k-value by least squares to the observed counts. In the implementation of Algorithm 2, we replace the infinite discrete space Θ with a finite set $\mathcal{P}$ of positive integers, which is modified if necessary. Specifically, let $\mathcal{P}_s = \{a+1, a+2, \ldots, U_s\}$ at the sth iteration. We enlarge this set by doubling the value of U_s if its value is less than or equal to the largest value of the current support set, and use the updated set

of points for gradient function evaluation and checking whether the convergence is achieved.

3.3 Selection of the optimal k-value

In the realm of data modelling, one is always unsure about the "best" value of k. The idea of cross-validation comes in handy when one wants a data-driven estimate of k, rather than a subjective estimate. See Arlot and Celisse (2010) for a survey of cross-validation procedures for model selection.

For the choice of the optimal k or the selection of the optimal model, we consider the least squares V-fold cross-validation method. In the V-fold cross-validation, the data set is randomly split into V disjoint partitions that are roughly equal in size. Given a set of k-values, the optimal k is the one that minimizes the following criterion:

$$\mathrm{CV}(k) = \frac{1}{V} \sum_{v=1}^{V} \mathcal{Q}_k^{(v)}(\hat{G}_k),$$

where $\mathcal{Q}_k^{(v)}$ is the objective function defined in (9) for the data in the vth subset and $\hat{G}_k$ is the NPLSE of G obtained from the data not in the vth subset while holding k fixed.

3.4 An illustrative simulated example

A random sample of size $n = 100$ was generated from a discrete left triangular distribution on $\mathbb{N}_0$ with $\theta = 11$. We first modelled this decreasing distribution as mixtures of discrete decreasing beta distributions on $\mathbb{N}_0$ with different fixed k-values. In particular, four discrete decreasing beta mixtures with $k \in \{1, 2, 3, 4\}$ were fitted by least squares to the observed distribution which is not monotone decreasing. Each fitted mixture distribution represented by black solid dots is plotted on the relative frequency distribution of the data (see Fig. 4). As k increases, the fitted mixture distribution becomes smoother. We remark that the least squares fitted Poisson mixture probabilities for this particular example are not monotone decreasing. Also shown in Fig. 4 is the gradient function at the NPLSE along with its support points (red solid dots) for each k. Note that all gradient values at convergence are nonnegative.

From the top left panel of Fig. 5, we see that simply minimizing $\mathcal{Q}_k$ with respect to G and k will lead to an estimated value of 1 for k. As indicated in Fig. 4, a small k-value will give an undersmoothed fit with large variance, whereas a large k-value will yield an overly smooth fit, which causes a large modelling bias. To trade off between bias and variance, we then applied the cross-validation method for automatically determining the optimal k-value. Specifically, we considered the least squares V-fold cross-validation, with $V \in \{5, 10, n\}$. All cross-validation functions plotted in Fig. 5 have obvious minimum points. While the cross-validation functions for $V = 5$ and $V = 10$ are minimized at $k = 2$, that for $V = n$ is minimized at $k = 5$. Unfortu-

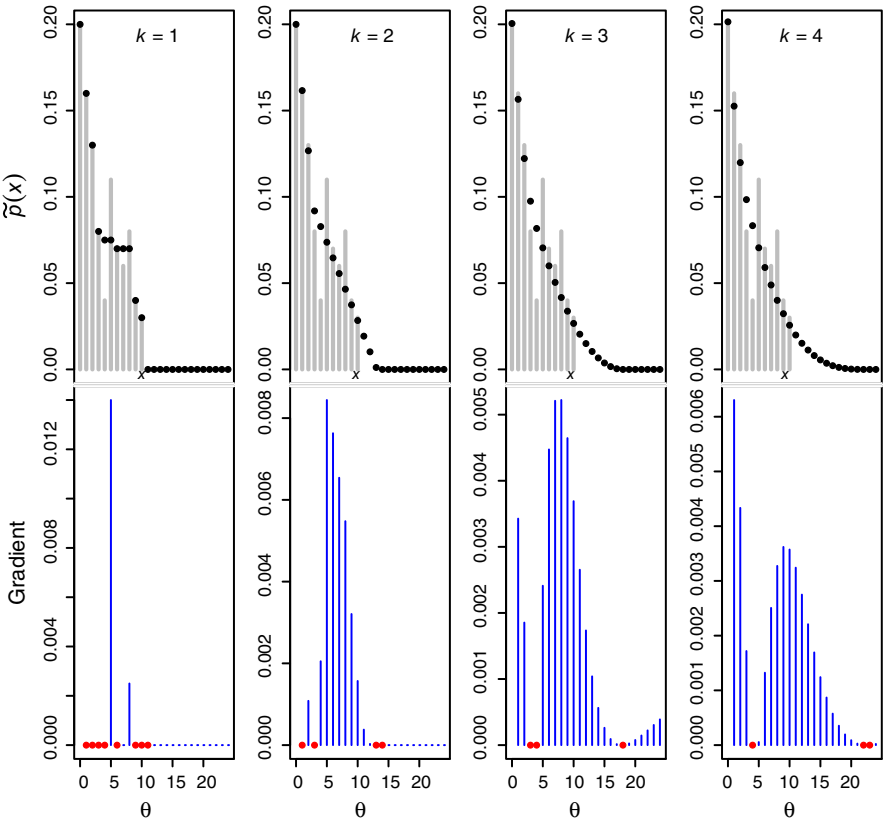

Fig. 4 (*Top*) The relative frequency distribution of a simulated data set of size 100 from a discrete left triangular distribution on $\mathbb{N}_0$ along with the fitted discrete decreasing beta mixture distribution with a fixed k-value (*black solid dots*). Fixed k-values of 1, 2, 3 and 4 are considered. (*Bottom*) The corresponding plot of gradient function at convergence along with the support points (*red solid dots*) of the NPLSE

nately, the optimal k-value is dependent on the value of V. This provides us motivation to investigate the least squares V-fold cross-validation further in a simulation study.

4 Numerical studies

4.1 Simulated count data

In this subsection, we are interested in comparing the performance of nonparametric mixtures as tools for modelling count data from decreasing distributions on $\mathbb{N}_0$. Also of particular interest is the question as to which value of V is most likely to be appropriate for the least squares V-fold cross-validation in selecting the optimal discrete decreasing beta mixture model. Nonparametric mixtures included in this simulation study were the Poisson mixture, the discrete uniform mixture (Jankowski and Wellner 2009), the discrete left triangular mixture (Durot et al. 2013) and the discrete decreasing beta mixture selected by the least squares V-fold cross-validation for each value of

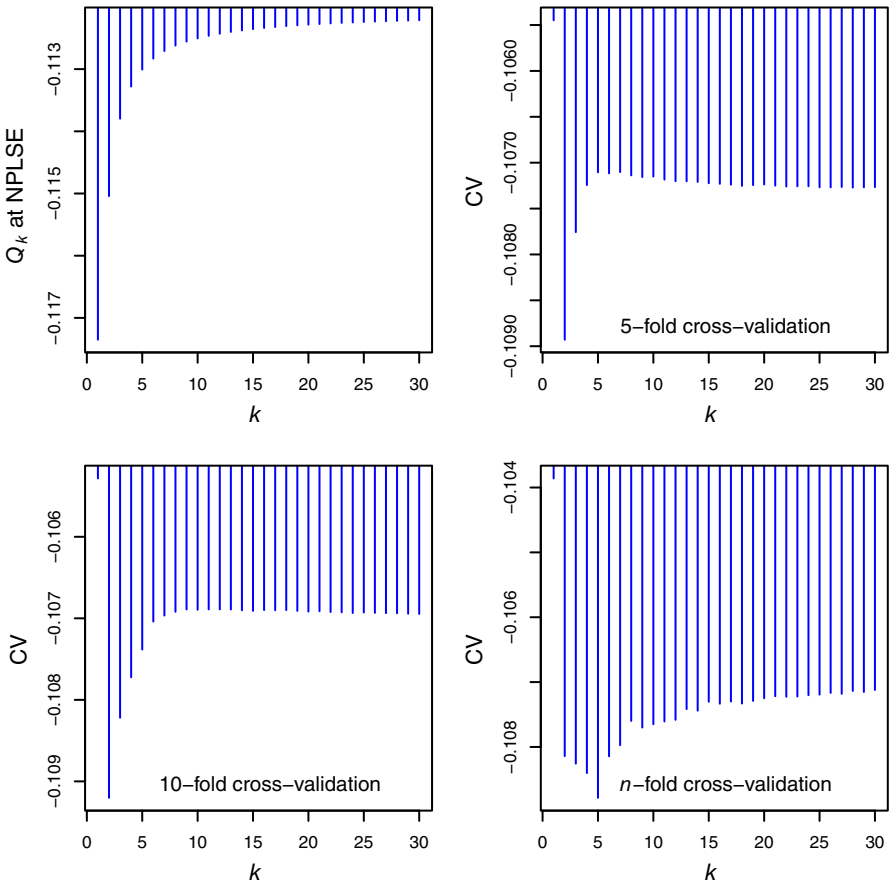

Fig. 5 (*Top left*) Objective function values evaluated at the NPLSE's for different k-values. (*Others*) Least squares V-fold cross-validation plots for V-values of 5, 10 and n

$V \in \{5, 10, n\}$. The cross-validation selection of the optimal k-value was among a set of k-values from 1 to 30.

Four decreasing distributions on $\mathbb{N}_0$ were considered as data generating distributions. Denoted by PO(1), DB(11,2), DB(8,5) and GE(0.75), they were, respectively, the Poisson distribution with mean 1, the discrete left triangular distribution with $\theta = 11$, the discrete decreasing beta distribution with $\theta = 8$ and $k = 5$ and the geometric distribution with density $p(x) = 0.75(0.25)^x$. Four sets of 200 samples were drawn from each of the four data generating distributions with, respectively, sizes 100, 500, 1000 and 5000. For performance evaluation, we computed the average sum of squared errors given by

$$\frac{1}{R}\sum_{r=1}^{R}\left[\sum_{x\geq 0}\left\{\hat{p}^{(r)}(x) - p(x)\right\}^2\right],$$

where R is the total number of samples, $p(x)$ denotes a data generating density and $\hat{p}^{(r)}(x)$ denotes a fitted density based on the rth sample.

The simulation results are shown in Table 1, from which it is clear that no model is strictly dominated by any other model. However, the cross-validation selected discrete decreasing beta mixtures are never the worst. For the PO(1) setting in which the distribution is not convex, the triangular mixture is really worse than the others even with a sample size as large as 5000. For the DB(11,2) setting, the Poisson and uniform mixtures fare poorly as compared with the other mixtures. For the DB(8,5) setting, the uniform mixture performs rather badly for small sample sizes. For the GE(0.75) setting, all mixtures show a rather similar performance.

As mentioned earlier, in using the least squares V-fold cross-validation to aid the selection of k, different V-values may lead to different optimal k-values. Based on the simulation results in Table 1, the issue of as to which V-value should be used seems not very critical. Since larger V-values typically require more computational time, we suggest using $V = 5$ for a pragmatic solution to the selection problem.

4.2 Real-world count data

It is well known that mixtures have a broad range of applications in many contexts. For our real applications, the interest is to apply nonparametric mixtures as convenient tools for providing adequate descriptions of the data. To demonstrate the use of nonparametric mixtures, we consider two sets of real-world data from Deb and Trivedi (1997). As given in Table 2, the two data sets contain, respectively, the number of visits to the emergency room and the number of visits to a physician in a hospital outpatient setting for a sample of 4406 individuals aged 66 and over. A notably salient feature of both the emergency room visit and physician outpatient visit count data is that they have a high proportion of zeros, about 82% and 77%, respectively. Apart from showing zero inflation, the observed frequency distribution of the number of physician outpatient visits also exhibits a long tail, with a few extreme visits beyond 40, i.e., 55, 61, 71 and 141.

While no information about the shapes of the two underlying distributions is available, fitting the Poisson mixture model to these data sets would be a reasonable choice. However, on seeing that both observed frequency distributions decrease rapidly for some small count values and remain about constant thereafter, assuming decreasing shapes for the underlying distributions and considering fitting the discrete decreasing beta mixture model to these data sets would not seem inappropriate.

The top row of Fig. 6 shows the results of the least squares fivefold cross-validation for the selection of the "best" discrete decreasing beta mixture distribution, suggesting that the optimal k-values for the emergency room visit and physician outpatient visit count data are 2 and 4, respectively. To assess the mixture fits to each data set, we provide their residuals in the bottom row of Fig. 6. Note that the range of the residual plot for the physician outpatient visit count data is from 0 to 70 instead of to the largest observed count value. On the whole, the Poisson mixture does not fit the emergency room visit count data as closely as the discrete decreasing beta mixture. On the other hand, both mixtures fare similarly for the physician outpatient visit count data, but the

Table 1 Average sum of squared errors ($\times 10^4$) with the standard error in parentheses

	Poisson mixture	Uniform mixture	Triangular mixture	Discrete decreasing beta mixture		
				Fivefold CV	Tenfold CV	n-fold CV
PO(1)						
$n = 100$	33.70 (3.05)	47.58 (3.46)	89.47 (2.10)	66.19 (3.78)	66.40 (3.78)	65.98 (3.90)
$n = 500$	7.20 (0.54)	8.89 (0.64)	70.92 (0.26)	11.69 (1.21)	11.40 (1.17)	10.93 (1.10)
$n = 1000$	4.69 (0.37)	5.42 (0.37)	70.06 (0.20)	5.83 (0.57)	5.64 (0.49)	5.61 (0.47)
$n = 5000$	0.75 (0.07)	0.92 (0.07)	68.02 (0.03)	0.92 (0.07)	0.92 (0.07)	0.92 (0.07)
DB(11,2)						
$n = 100$	42.32 (2.07)	38.56 (1.89)	18.41 (1.72)	28.34 (2.02)	28.61 (2.06)	28.01 (2.01)
$n = 500$	11.31 (0.52)	11.89 (0.41)	3.56 (0.36)	5.92 (0.49)	6.35 (0.52)	6.09 (0.51)
$n = 1000$	5.98 (0.22)	6.67 (0.21)	1.69 (0.15)	2.76 (0.26)	3.00 (0.26)	3.09 (0.27)
$n = 5000$	2.29 (0.05)	1.72 (0.06)	0.37 (0.03)	0.63 (0.07)	0.58 (0.07)	0.58 (0.07)
DB(8,5)						
$n = 100$	48.07 (3.19)	62.18 (3.45)	39.15 (2.60)	40.85 (3.12)	39.30 (3.09)	38.96 (3.15)
$n = 500$	13.03 (0.73)	15.24 (0.75)	12.79 (0.65)	11.25 (0.68)	11.43 (0.67)	11.21 (0.65)
$n = 1000$	5.35 (0.35)	6.34 (0.38)	5.83 (0.33)	4.66 (0.31)	4.67 (0.31)	4.58 (0.32)
$n = 5000$	1.26 (0.07)	1.38 (0.07)	1.38 (0.07)	1.15 (0.07)	1.11 (0.07)	1.15 (0.07)
GE(0.75)						
$n = 100$	28.91 (2.26)	31.89 (2.35)	31.33 (2.32)	27.02 (2.17)	26.95 (2.20)	26.47 (2.14)
$n = 500$	7.89 (0.71)	8.21 (0.71)	8.19 (0.71)	7.31 (0.69)	7.26 (0.69)	7.14 (0.68)
$n = 1000$	3.93 (0.32)	4.06 (0.32)	4.05 (0.32)	3.64 (0.31)	3.64 (0.31)	3.60 (0.31)
$n = 5000$	0.74 (0.06)	0.75 (0.06)	0.75 (0.06)	0.69 (0.06)	0.69 (0.06)	0.68 (0.06)

Table 2 Deb and Trivedi's (1997) count data

Number of emergency room visits	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed frequency	3602	588	137	54	11	7	2	1	2	0	0	1	1
Number of physician outpatient visits	0	1	2	3	4	5	6	7	8	9	10	11	$\geq$12
Observed frequency	3397	526	204	76	51	36	25	10	13	7	7	7	47

discrete decreasing beta mixture is visually smoother than the Poisson mixture which has multiple modes in the tail region (see Fig. 7). These modes could be spurious and do not genuinely present in the underlying distribution; as often in practice, one would expect tails of distributions to be monotone decreasing.

5 Conclusion

Parametric modelling of count data has become a commonly accepted practice among practitioners, facilitated by the availability of a range of parametric count models

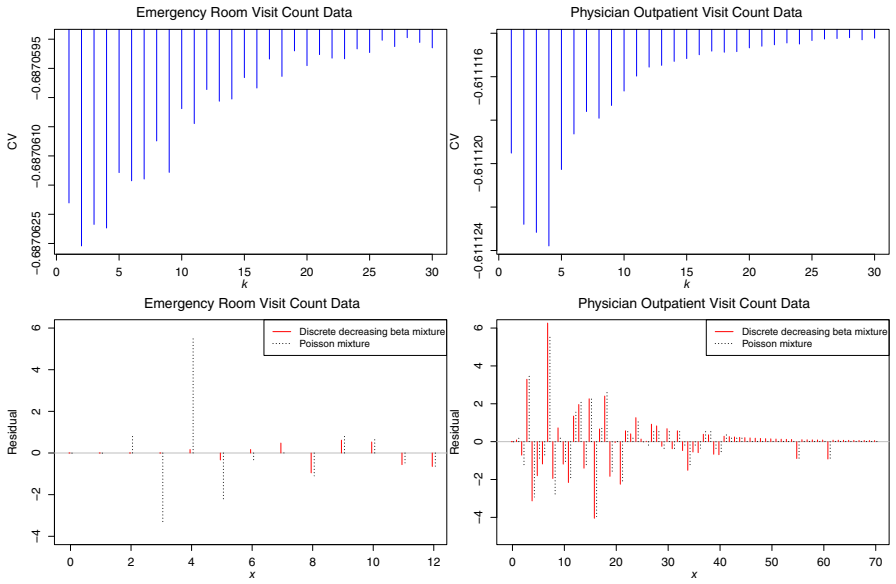

Fig. 6 (*Top row*) Least squares fivefold cross-validation plots for the two real data examples. (*Bottom row*) Residual plots of two mixture fits

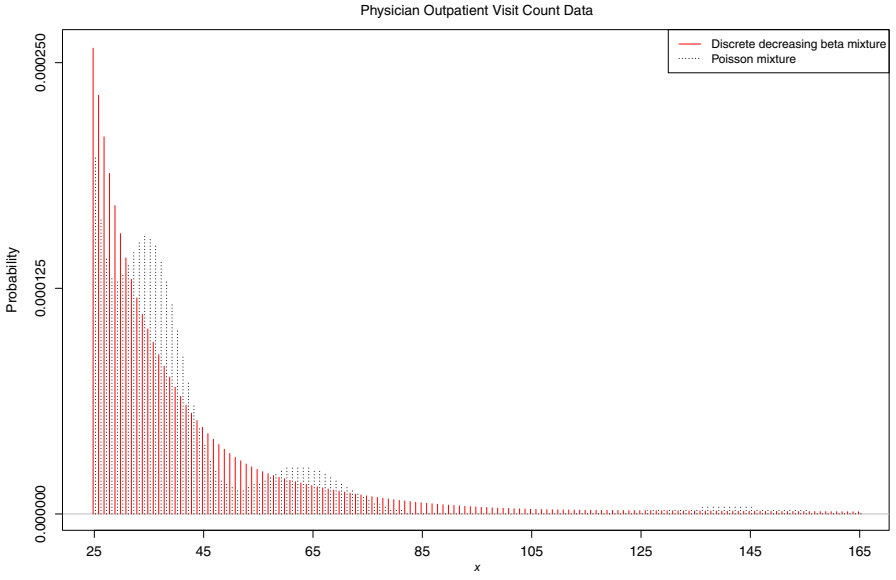

Fig. 7 Two fitted mixture distributions, only shown on the discrete interval [25,165], for the physician outpatient visit count data

and easy-to-use fitting software packages. Nonetheless, on observing a distribution of counts which is more complex than the assumed one, formulating a suitable parametric

count model for the data at hand has not always been straightforward and perhaps might be a difficulty for practitioners. To allow for more flexibility in capturing different shapes of distributions, we present a framework for modelling count data based on nonparametric mixture models, which have capabilities to cope with various problems pertinent to count data such as multimodality, overdispersion and zero inflation. Fast algorithms for least squares fitting of nonparametric mixture models to count data are also provided. This framework potentially has a wider applicability, although we have only offered two specific illustrative cases, of which in the first case the competitiveness of the nonparametric Poisson mixture model with respect to an existing parametric count model is demonstrated using a real data set.

When one has some prior knowledge that the underlying distribution of interest is decreasing or a decreasing distributional shape assumption seems tenable based on empirical evidence, then a model that is able to yield a fitted distribution with decreasing probabilities is conceived necessary. In the second case, we define the discrete decreasing beta mixture model, which includes the uniform and left triangular mixture models as special cases, and show its ability in accommodating possible shapes of discrete decreasing distributions. The method of least squares is employed for fitting the discrete decreasing beta mixture model with a fixed k-value. Viewing the parameter k as a bandwidth parameter, the cross-validation procedure for the choice of its value is suggested. As shown by two sets of real-world data, there are situations in which the discrete decreasing beta mixture model can be a very competitive alternative to the Poisson mixture model.

Nonparametric mixture models can serve the purposes of inference and exploratory. They can be used as exploratory models for discovering the structure of the data and as inferential models for statistical inferences of interest and, as such, they deserve to be included in the modelling toolbox of practitioners. We hope that our work will attract more practitioners to consider modelling of count data using nonparametric mixture models.

Acknowledgments The author is grateful to the associate editor and two reviewers for their insightful and valuable comments. The author also acknowledges and thanks the Universiti Malaysia Terengganu for providing the Research Incentive Grant (No. 68007/2013/121).

References

Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Stat. Surv. **4**, 40–79 (2010)

Baayen, R.H.: Word Frequency Distributions. Springer, Dordrecht (2001)

Balabdaoui, F., Wellner, J.A.: Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. Stat. Neerl. **64**, 45–70 (2010)

Böhning, D., Patilea, V.: Asymptotic normality in mixtures of power series distributions. Scand. J. Stat. **32**, 115–131 (2005)

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

Cameron, A.C., Trivedi, P.K.: Regression Analysis of Count Data, 2nd edn. Cambridge University Press, Cambridge (2013)

Chee, C.-S., Wang, Y.: Minimum quadratic distance density estimation using nonparametric mixtures. Comput. Stat. Data Anal. **57**, 1–16 (2013)

Chee, C.-S., Wang, Y.: Least squares estimation of a k-monotone density function. Comput. Stat. Data Anal. **74**, 209–216 (2014)
Dax, A.: The smallest point of a polytope. J. Optim. Theory Appl. **64**, 429–432 (1990)
Deb, P., Trivedi, P.K.: Demand for medical care by the elderly: a finite mixture approach. J. Appl. Econom. **12**, 313–336 (1997)
Durot, C., Huet, S., Koladjo, F., Robin, S.: Least-squares estimation of a convex discrete distribution. Comput. Stat. Data Anal. **67**, 282–298 (2013)
Gupta, R.C., Ong, S.H.: Analysis of long-tailed count data by Poisson mixtures. Commun. Stat. **34**, 557–573 (2005)
Harris, I.R., Shen, S.: The minimum L_2 distance estimator for Poisson mixture models. J. Stat. Plan. Inference **141**, 1088–1101 (2011)
Jankowski, H.K., Wellner, J.A.: Estimation of a discrete monotone distribution. Electr. J. Stat. **3**, 1567–1605 (2009)
Karlis, D., Xekalaki, E.: Minimum Hellinger distance estimation for Poisson mixtures. Comput. Stat. Data Anal. **29**, 81–103 (1998)
Karlis, D., Xekalaki, E.: Robust inference for finite Poisson mixtures. J. Stat. Plan. Inference **93**, 93–115 (2001)
Karlis, D., Xekalaki, E.: Mixed Poisson distributions. Int. Stat. Rev. **73**, 35–58 (2005)
Lawson, C.L., Hanson, R.J.: Solving Least Squares Problems. Prentice-Hall Inc, Englewood Cliffs (1974)
Mazza, A., Punzo, A.: Discrete beta kernel graduation of age-specific demographic indicators. In: Ingrassia, S., Rocci, R., Vichi, M. (eds.) New Perspectives in Statistical Modeling and Data Analysis Studies in Classification, Data Analysis, and Knowledge Organization, pp. 127–134. Springer, Berlin (2011)
Nikoloulopoulos, A.K., Karlis, D.: On modeling count data: a comparison of some well-known discrete distributions. J. Stat. Comput. Simul. **78**, 437–457 (2008)
Punzo, A.: Discrete beta-type models. In: Locarek-Junge, H., Weihs, C. (eds.) Classification as a Tool-for Research, Studies in Classification, Data Analysis, and Knowledge Organization, pp 253–261. Springer, Berlin, Heidelberg (2010)
Punzo, A., Zini, A.: Discrete approximations of continuous and mixed measures on a compact interval. Statistical Papers **53**, 563–575 (2012)
Rigby, R.A., Stasinopoulos, D.M., Akantziliotou, C.: A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. Computational Statistics and Data Analysis **53**, 381–393 (2008)
Scott, D.W.: Parametric statistical modeling by minimum integrated square error. Technometrics **43**, 274–285 (2001)
Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P.: A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. J. R. Stat. Soc. **54**, 127–142 (2005)
Simar, L.: Maximum likelihood estimation of a compound Poisson process. Ann. Stat. **4**, 1200–1209 (1976)
Wang, Y.: On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. J. R. Stat. Soc. **69**, 185–198 (2007)
Wang, Y.: Dimension-reduced nonparametric maximum likelihood computation for interval-censored data. Comput. Stat. Data Anal. **52**, 2388–2402 (2008)
Wang, Y., Chee, C.-S.: Density estimation using non-parametric and semi-parametric mixtures. Stat. Model. **12**, 67–92 (2012)