Statistics and Computing. Revised manuscript 28 September 2003.

Semi-parametric extended Poisson process models for count data

HEATHER M. PODLICH*, MALCOLM J. FADDY† and GORDON K. SMYTH**

* *School of Land and Food Sciences, University of Queensland, Australia*
† *School of Mathematics and Statistics, University of Birmingham, UK*
** *Walter and Eliza Hall Institute of Medical Research, Australia*

A general framework for the analysis of count data (with covariates) is proposed using formulations for the transition rates of a state-dependent birth process. The form for the transition rates incorporates covariates proportionally, with the residual distribution determined from a smooth non-parametric state-dependent form. Computation of the resulting probabilities is discussed, leading to model estimation using a penalized likelihood function. Two data sets are used as illustrative examples, one representing under-dispersed Poisson-like data and the other over-dispersed binomial-like data.

Keywords: count data, over- and under-dispersion, covariate effects, extended Poisson process model, penalized likelihood

1 Introduction

This paper proposes a very flexible approach to the analysis of count data in which almost no assumptions other than smoothness are made about the distribution of the data. Non-parametric smoothing techniques are used to allow the fitted models to adapt to arbitrary response distributions.

The most common models for count data are those based on the binomial and Poisson distributions (McCullagh and Nelder, 1989). Data often exhibit departures from these models and so various alternatives and generalizations have been proposed including quasi-likelihood methods, random effect models and mixture models. See for example Dean (1998), Lindsey (1999, Chapter 7), McCullagh and Nelder (1989, Sections 4.5, 5.5 and 6.2) and the references cited therein. The bibliography in Lindsey (1999) is particularly extensive. In this paper we use extended Poisson process models (EPPMs), a new class of models that enable a more general approach to the analysis of count data than any of the existing methods (Faddy, 1997a, 1997b, 1998a, 1998b; Faddy and Bosch, 2001; Faddy

and Fenlon, 1999; Toscas and Faddy, 2003). EPPMs represent any discrete distribution as the distribution of the number of events occurring in a finite time interval of a state-dependent Markov birth process, also called a pure birth process. It is the nature of the state-dependence of the transition rates of the Markov process which determines the dispersion properties of the resulting distribution. Standard distributions correspond to linearly increasing or decreasing transition rates while over or underdispersion relative to these standard distributions in manifested as increasing or decreasing, convex or concave transition rate curves. EPPMs can be usefully applied to almost any regression problem with count responses but are especially appropriate when the counts can be thought of as accumulating over time. In such cases the shape of the state-dependence of the transition rates is often directly interpretable and may provide an intuitive explanation for why any over- or under-dispersion has occurred.

Previous papers on EPPMs have used simple non-linear parametric functions for describing the transition rates as a function of the number of events. This has the effect of describing a parameterized class of response distributions which includes the Poisson, negative binomial and binomial distributions as special cases. Although this approach is appealing, the choice of the specific parametric form for the transition rates is inevitably somewhat ad hoc. This paper takes a more computational, data analytic approach. The transition rate sequence is instead estimated non-parametrically, allowing the data to determine the form of the transition rate sequence. The transition rates are unconstrained and a penalty function approach is taken to impose smoothness on the sequence. In this way the method is able to adapt, given adequate data, to quite arbitrary response distributions.

Covariates are allowed to affect the response distribution through a proportional rate model in which the transition rates depend parametrically on the covariates but non-parametrically on the number of events. We call the resulting regression models *semi-parametric extended Poisson process models.*

The penalty function approach taken in this paper to non-parametric smoothing is analogous to penalty function methods used for smoothing probability densities (Tapia and Thompson, 1978) or for scatterplot smoothing (Eubank, 1998; Green and Silverman, 1994). There are close links with spline theory here, especially with cubic smoothing

splines which arise from an integrated second derivative squared penalty function (Reinsch, 1967). We propose an adjusted version of this penalty function using both first and second differences which allows for modelling departures from the Poisson, negative binomial and binomial distributions within the one framework.

The most serious limitation of semi-parametric EPPMs is that they are computationally intensive. The probabilities which define the response distribution are defined only indirectly in terms of the parameters defining the transition rates. Except in special cases, there are no closed form expressions even for the mean or variance of the response distribution. Although there exist numerical techniques for computing the probabilities based on matrix exponentials (Sidje, 1998), these methods require $O(y_{\max})$ operations where $y_{\max}$ is the largest response count and so are suitable only for small to moderate counts. For large n, very good saddle point approximations are available for the probabilities (Daniels 1982; Smyth and Podlich, 2002). The use of saddlepoint approximations together with exact calculations for the smallest n makes it possible to perform accurate calculations while keeping the computational load manageable.

Section 2 of this paper reviews the basic concepts of EPPMs and surveys briefly results which relate dispersion to the shape of the transition rate sequence. Section 3 introduces semi-parametric proportional rate models and Section 4 discusses numerical strategies for computing probabilities and the likelihood function. Section 5 explains the penalized likelihood approach to non-parametric smoothing. Sections 6 to 9 consider two data examples in detail, one Poisson-like and one binomial-like. Section 7 considers how to choose the degree of smoothing and Section 9 discusses tests for extra-dispersion and for the covariates, in the context of the two data examples. The paper concludes with discussion in Section 10.

2 Extended Poisson process models

2.1 *Transition rate representations of count distributions*

An extended Poisson (or pure birth) process is a continuous time process $\{X(t); t \geq 0\}$ on the non-negative integers satisfying $X(0) = 0$ and

$$P\{X(t+\delta t) = n+1 \,|\, X(t) = n\} = \lambda_n \delta t + o(\delta t) \tag{1}$$

Table 1: Numbers of surviving foetal implants

Number of implants	Frequency
1	10
2	4
3	4
4	3
5	8
6	12
7	16
8	31
9	66
10	184
11	363
12	458
13	444
14	296
15	157
16	47
17	16
18	10
19	1
20	0
21	1

for $n \geq 0$ where the λ_n are non-negative transition rates. If the $\lambda_n = \lambda$ for all n, then this is an ordinary Poisson process and $X(t)$ follows a Poisson distribution with mean λt. Consider the process at a given time, which we will take to be one without loss of generality, and define $\pi_n = P\{X(1) = n\}$. It is obvious that any sequence λ_n determines a sequence of probabilities π_n. Faddy (1997a) has shown that the reverse is also true. For any probability distribution π_n on the non-negative integers, there is a transition rate sequence λ_n such that $\pi_n = P\{X(1) = n\}$ for all n. For any n such that $P\{X(1) \geq n\} > 0$, λ_n is uniquely determined. We call the sequence λ_n the transition rate representation of the count distribution defined by the π_n.

To see how the λ_n sequence corresponds to a probability distribution consider the data in Table 1. Each count gives the number of sites where fertilized eggs have implanted in a mouse utero. McCaughran and Arnold (1976) found that none of the normal, Poisson, binomial or negative binomial distributions provided an acceptable fit to these data. The data are in fact considerably under-dispersed relative to the Poisson, with sample variance 4.77 much less sample mean 12.18. It is not clear what theoretical distribution could be

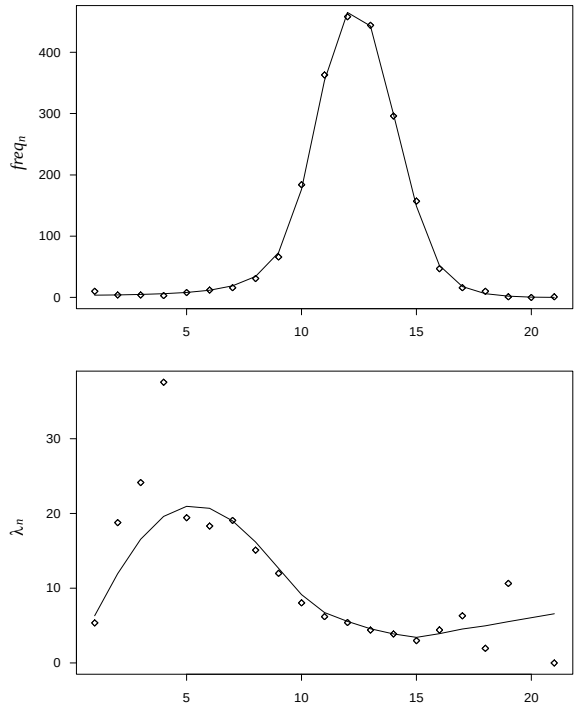

Figure 1: Foetal implants empirical probability distribution (above) and empirical transition rates (below). Symbols denote observed values and solid lines denote estimates from smoothing

used to model this data. The plot symbols in the upper plot in Figure 1 show the empirical distribution of these data while those in the lower plot give the λ-sequence corresponding to this empirical distribution. (No symbol is shows for λ_{20} in the lower plot as it is infinite, corresponding to zero observed frequency.) The transition rates are far from constant which would correspond to a Poisson distribution. The rise then fall in the transition rates as n increases shows that the distribution is left skew relative to the Poisson. The solid line on the lower plot gives the results of a smooth non-parametric fit to the transition rates, explained in Section 6. The solid line on the upper plot gives the probabilities arising from the smoothed transition rates. The smoothed probabilities show close agreement with the observed frequencies. The transition rate representation together with non-parametric smoothing provide an excellent fit the observed data.

2.2 *Models for overdispersion and underdispersion*

As already noted, a constant transition sequence $\lambda_n = \lambda$ for all n corresponds to the Poisson distribution with mean λ. Other linear transition rate sequences also result in standard distributions. A linearly increasing sequence, $\lambda(n) = a(N + n)$ with $a > 0$, gives the negative binomial distribution with success probability $p = \exp(-a)$, mean $N(1-p)/p$ and variance $N(1-p)/p^2$. A linearly decreasing sequence, $\lambda(n) = a(N - n)$ with $a > 0$, gives the binomial distribution with success probability $p = 1 - \exp(-a)$ and size parameter N. It is well known that the negative binomial and binomial distributions are overdispersed and underdispersed respectively relative to the Poisson. Ball (1995) has shown that any increasing transition rate sequence leads to an overdispersed distribution with $\text{var}\{X(t)\} > E\{X(t)\}$ and any decreasing sequence to an underdispersed distribution with $\text{var}\{X(t)\} < E\{X(t)\}$.

Further, the results of Ball and Donnelly (1987) and Brown and Donnelly (1993) show that $X(t)$ can show overdispersion or underdispersion relative to the binomial distribution depending on the convexity or concavity of the λ_n sequence. For increasing sequences λ_n, the work of Donnelly, Kurtz and Marjoram (1993) can be used to show that the variance of $X(t)$ is as a function of the mean greater than predicted by the quadratic function corresponding to a negative binomial distribution for convex λ_n sequences and less for concave sequences. Further, numerically comparing distributions suggests that convex increasing λ_n sequences result in distributions with greater skewness, and concave increasing sequences result in distributions with less skewness, than negative binomial distributions with the same mean and variance.

There is thus a wide range of possible probability distributions to be constructed from monotone λ_n sequences and the dispersion properties of these distributions relative to the standard Poisson, negative binomial and binomial distributions can be broadly interpreted in terms of the rate of increase or decrease of the λ_n sequence. Non-monotone sequences provide even more possibilities such as that arising from the data in Table 1. The complete generality of the methodology allows any count distribution to be constructed from an equivalent λ_n sequence.

3 Semi-parametric proportional models

Assume that independent responses y_i are observed together with covariate vectors $\mathbf{x}_i$, $i = 1, \ldots, I$. Each y_i takes values on the non-negative integers. Let $\lambda_{i,n}$, $n \geq 0$, be the transition rate sequence corresponding to the distribution of y_i given $\mathbf{x}_i$.

In previous work on EPPMs, parametric models have been proposed for $\lambda_{i,n}$. In Faddy (1997a) deviations from the Poisson distribution were allowed using the functional form,

$$\lambda_{i,n} = a_i(b+n)^c$$

with $a_i > 0$, $b > 0$ and $c \leq 1$. The special case $c = 0$ corresponds to a constant sequence and the Poisson distribution while $c = 1$ results in a linear sequence and the negative binomial distribution. In Faddy and Fenlon (1999) deviations from the binomial distribution were allowed using

$$\lambda_{i,n} = a_i e^{bn}(N-n), \;\; n = 0, 1, \ldots, N$$

with $a_i > 0$. The special case $b = 0$ gives the binomial distribution while $b > 0$ and $b < 0$ result in overdispersion and underdispersion relative to the binomial distribution respectively. Covariate effects can be incorporated into these models by setting $a_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a vector of regression parameters. The arising models are proportional transition rate models in which the transition rates are a product of an n-dependent factor and a covariate-dependent factor.

Parametric proportional rate models are a flexible and useful class of EPPMs. However, to allow the data to play a greater role in determining an appropriate distribution and describe the residual variation more accurately, the n-dependence factor of the transition rates may be estimated non-parametrically. Suppose that the y_i are unconstrained counts. Let

$$\lambda_{i,n} = a(\mathbf{x}_i^T \boldsymbol{\beta}) h(n) \tag{2}$$

where $a()$ is a known link function taking positive values and the function $h()$, rather than taking on a parametric functional form, is to be estimated non-parametrically. We use the term 'semi-parametric' to describe the model (2) since the covariate or systematic effects are incorporated parametrically while the n-dependence or distributional form is

determined non-parametrically. Non-constancy of $h(n)$ is associated with deviations from Poisson variation while non-linearity in $h(n)$ is associated with departures from negative binomial variation. If $h(n)$ is constant and $a()$ is the exponential function then (2) is equivalent to the well known log-linear model for Poisson counts (McCullagh and Nelder, 1989).

Suppose now that y_i is the observed number of cases out of a maximum possible N_i, $0 \leq y_i \leq N_i$. An adjusted form of (2) appropriate for modelling binomial-like data is

$$\lambda_{i,n} = a(\mathbf{x}_i^T \boldsymbol{\beta}) h(n)(N_i - n), \quad n = 0, 1, \ldots N_i, \tag{3}$$

Departures from binomial variation are characterized by non-constancy of $h(n)$. If $h(n)$ is constant and $a()$ is the exponential function then (3) is equivalent to the well known complementary log-log model for binomial counts.

The function $h()$ in (2) and (3) should take positive values. It is convenient to write $h(n) = \exp\{g(n)\}$ so that the sequence $g(n)$ is unconstrained. In many cases it will be natural to assume $a()$ to be exponential also. In that case the proportional rate model becomes an additive model on the log-scale in terms of the linear predictors $\mathbf{x}_i^T \boldsymbol{\beta}$ and the log-transition rates $g(n)$.

Practical experience shows that the proportional transition rate model is a useful one for a wide range of data sets. It is wise though to check whether the assumption of proportionality is reasonable for any given data set. If $a()$ is the exponential function then a test can be developed analogous to Tukey's (1949) one-degree-of-freedom test for non-additivity. On the log-scale (2) becomes

$$\log(\lambda_{i,n}) = \mathbf{x}_i^T \boldsymbol{\beta} + g(n),$$

where $g(n) = \log\{h(n)\}$. The Tukey one-degree-of-freedom idea is to add an extra term which is multiplicative in the additive effects. A test for non-additivity can be made by testing H_0: $\delta = 0$ in the model

$$\log(\lambda_{i,n}) = \mathbf{x}_i^T \boldsymbol{\beta} + g(n) + \delta\, \mathbf{x}_i^T \boldsymbol{\beta}\, g(n)$$

In all of the examples considered in this paper, a likelihood ratio test of $\delta = 0$ has been carried out to check the assumption of proportional transition rates. In all data sets

encountered by the authors so far the assumption of proportionality has been judged to be satisfactory.

4 Computation of probabilities

The greatest practical obstacle to the use of EPPMs is the efficient and accurate computation of the response distribution probabilities and related quantities. This section discusses numerical strategies available for computation.

Write Y_i for the theoretical random variable of which y_i is a realization. The log-likelihood of the observed data is $\sum_{i=1}^{I} \log(p_i)$ where $p_i = P(Y_i = y_i \,|\, \boldsymbol{\beta}, h(n), n \geq 0)$. Likelihood computations require that we compute the p_i and perhaps also derivatives of the p_i with respect to any parameters. The recursive nature of (1) ensures that each p_i is a function of all the transition rates from $n = 0$ to $n = y_i$, i.e., $P(Y_i = y_i \,|\, \boldsymbol{\beta}, h(n), n \geq 0) = P(Y_i = y_i \,|\, \lambda_{i,0}, \ldots, \lambda_{i,y_i})$.

For the remainder of this section we will drop the subscript i from y_i and $\lambda_{i,n}$ and consider the problem of computing the probability that $Y = y$ given a specified lambda sequence $\lambda_0, \ldots, \lambda_y$. Write $p_n(t) = P\{X(t) = n\}$ for the probabilities defined by (1). In this notation, the probability we seek to compute is $p_y(1)$. The classical approach to computing the probabilities of pure birth processes is through the Chapman-Kolmogorov differential equations (Cox and Miller, 1965, Chapter 4). The probabilities satisfy the equations

$$\begin{aligned} p_0'(t) &= -\lambda_0 p_0(t) \\ p_n'(t) &= -\lambda_n p_n(t) + \lambda_{n-1} p_{n-1}(t), \quad n \geq 1 \end{aligned}$$

subject to the initial condition $p_0(0) = 1$. Although an analytic solution to these equations can be written down (Bartlett, 1978, Chapter 3), this solution is extremely ill-conditioned in finite precision arithmetic. The analytic solution is therefore not suitable for numerical computation.

Given that we need consider only $n = 0, \ldots, y$, the above system of differential equations may be written in matrix notation as

$$\frac{\partial \mathbf{p}(t)}{\partial t} = \mathbf{Q}\,\mathbf{p}(t) \tag{4}$$

where $\mathbf{p}(t)$ is the vector of probabilities $\{p_0(t), \ldots, p_y(t)\}^T$ and $\mathbf{Q}$ is the transfer matrix

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & 0 & 0 & \cdots & 0 & 0 \\ \lambda_0 & -\lambda_1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_{y-1} & 0 \\ 0 & 0 & 0 & \cdots & \lambda_{y-1} & -\lambda_y \end{pmatrix}.$$

The value of $\mathbf{p}(t)$ which solves (4) at $t = 1$ can be written

$$\mathbf{p} = \exp(\mathbf{Q})\, \mathbf{e}_1$$

where $\exp(\mathbf{Q})$ is the matrix exponential of $\mathbf{Q}$, defined to be the matrix resulting from exponentiating the eigenvalues of $\mathbf{Q}$, and $\mathbf{e}_1$ is the unit vector $(1, 0, \ldots, 0)^T$ (Stewart, 1994). The probability required to be computed is

$$p_y(1) = \mathbf{e}_s^T \exp(\mathbf{Q})\, \mathbf{e}_1, \tag{5}$$

i.e., the last element of the first column of $\exp(\mathbf{Q})$.

Although the $\mathbf{Q}$ is a sparse matrix, $\exp(\mathbf{Q})$ is a dense $N \times N$ matrix and computing it is in general an $O(N^2)$ operation (Moler and van Loan, 1978). Computing (5) though requires only the first column of $\exp(\mathbf{Q})$. Efficient algorithms for performing vector-matrix exponential operations of this sort have been developed recently by Sidje (1998). Sidje's algorithm uses Krylov space methods to compute $\exp(\mathbf{Q})\, \mathbf{e}_1$ without computing $\exp(\mathbf{Q})$ itself and gives good results providing y is relatively small. Although our need is in fact for only one element of $\mathbf{p}$, there are no existing computational methods that the authors are aware of which make it possible to compute this element without computing the complete vector $\mathbf{p}$.

The vector-matrix exponential approach can be extended to compute derivatives of the probabilities. Suppose that the λ_n depend on a parameter a and suppose that the derivative $\partial\mathbf{p}/\partial a$ is required. Differentiating both sides of (4) with respect to a and reversing the order of differentiation yields

$$\frac{\partial}{\partial t}\left(\mathbf{p} \quad \frac{\partial \mathbf{p}}{\partial a}\right) = \mathbf{Q}_2 \left(\mathbf{p} \quad \frac{\partial \mathbf{p}}{\partial a}\right)$$

where

$$\mathbf{Q}_2 = \begin{pmatrix} \mathbf{Q} & 0 \\ \frac{\partial \mathbf{Q}}{\partial a} & \mathbf{Q} \end{pmatrix}$$

This shows that both $\mathbf{p}$ and its derivative respect to a may be computed using a vector-matrix exponential operation where the matrix $\mathbf{Q}_2$ is of dimension $2N \times 2N$. Podlich *et al.* (1999) show how this approach may be extended to compute second derivatives as well using matrices of dimension $4N \times 4N$.

In practice the matrix exponential approach to computing the probabilities is limited to small y for reasons of accuracy as well as computational efficiency. The efficiency limitation arises from the fact that the complete vector $\mathbf{p}$ of length y needs to be computed for every y even though only the last element is required. The accuracy limitation arises from the fact that all elements of $\mathbf{p}$ are computed with the same additive precision. The last element $\mathbf{p}$ will not be computed to full machine accuracy if it is smaller than other elements of the vector. If the last element is very much smaller than other elements then it will be computed with very poor relative precision and may even be negative.

Fortunately, there are saddlepoint approximations to the probabilities which are very accurate when y is large. The saddlepoint approximation of Daniels (1982) gives

$$P(Y = y) \approx \frac{\prod_{j=0}^{y-1} \lambda_j e^{-\tilde{s}}}{\prod_{j=0}^{y} (\lambda_j - \tilde{s}) \left\{ 2\pi \sum_{j=0}^{y} \frac{1}{(\lambda_j - \tilde{s})^2} \right\}^{1/2}} \left\{ 1 + \frac{1}{8}\rho_4 - \frac{5}{24}\rho_3^2 \right\}, \tag{6}$$

with $\tilde{s}$ satisfying

$$1 = \sum_{j=0}^{y} \frac{1}{\lambda_j - \tilde{s}},$$

and with

$$\rho_r = \left\{ \sum_{j=0}^{y} \frac{1}{(\lambda_j - \tilde{s})^2} \right\}^{-r/2} (r-1)! \sum_{j=0}^{y} \frac{1}{(\lambda_j - \tilde{s})^r}.$$

This approximation maintains a small relative error over the entire range of count values which is important when the approximation to compute log-likelihood functions. An even more accurate saddlepoint approximation is developed by Smyth and Podlich (2002). The approximation of Smyth and Podlich (2002) has the added advantage of being exact for the Poisson, binomial and negative binomial special cases. Even using these approximations,

evaluation of $P(Y = y)$ is still an $O(y)$ operation so that computation becomes more expensive when the observed counts are large.

5 Non-parametric model fitting

This section considers the problem of estimating non-parametrically the smooth function $h(n)$ which determines the shape of the fitted distribution for each y_i. There are several smoothing methods which could be applied. In this section we propose a roughness penalty approach.

Suppose that a set of q equally spaced knot points has been chosen between 0 and $y_{\max} = \max y_i$ and let $\mathbf{h} = (h_1, \ldots, h_q)^T$ be the vector of values taken by $h()$ evaluated at the knot points. If $y_{\max}$ is not very large then the knot points will be all the integers from 0 to $y_{\max}$ inclusive. Otherwise q will be chosen smaller than $y_{\max}$. The number of knot points will not materially affect the smoothed curve provided that it is large compared with the effective degrees of freedom associated with the fitted curve. It is assumed that $h(n)$ for any n between 0 and $y_{\max}$ can be expressed as a function of $\mathbf{h}$. Given values for $\boldsymbol{\beta}$ and the $h(n)$ it is possible to evaluate $\lambda_{i,n}$, $n = 1, \ldots, y_i$, and hence $p_i = P(Y_i = y_i)$ for each i.

Write $\ell(\boldsymbol{\beta}, \mathbf{h}; \mathbf{y}) = \sum_{i=1}^{I} \log p_i$ for the log-likelihood function as a function of $\boldsymbol{\beta}$ and $\mathbf{h}$. The regression parameters and h-curve are estimated by maximizing with respect to $\boldsymbol{\beta}$ and $\mathbf{h}$ the penalized log-likelihood function

$$\ell_p(\boldsymbol{\beta}, \mathbf{h}; \mathbf{y}) = \ell(\boldsymbol{\beta}, \mathbf{h}; \mathbf{y}) - \alpha \, \text{Penalty}(\mathbf{h}) \tag{7}$$

where $\alpha > 0$ and $\text{Penalty}(\mathbf{h})$ is a non-negative roughness penalty function. The smoothing parameter α controls the trade-off between goodness of fit as measured by the log-likelihood and smoothness of the n-dependent form as measured by the roughness penalty. Small values for α will result in irregular values for the h_i. The limiting case $\alpha = 0$ corresponds to maximizing the ordinary log-likelihood and therefore to a q-parameter model for the n-dependence. Larger values of α will see the penalty term forced downwards and hence result in a smoother form for $\mathbf{h}$ using fewer effective parameters. As $\alpha \to \infty$, $\mathbf{h}$ converges to the smoothest possible form for which the penalty term is zero.

It is appropriate to choose the penalty function such that it is zero only if $\mathbf{h}$ is constant, corresponding to the Poisson distribution in (2) or to the binomial distribution in (3). To achieve this, consider penalty functions of the form,

$$\text{Penalty}(\mathbf{h}) = \mathbf{h}''^T\mathbf{K}\mathbf{h}'' + \gamma\mathbf{h}'^T\mathbf{K}\mathbf{h}', \tag{8}$$

where $\mathbf{h}'$ is the vector first differences $h'_j = h_{j+1} - h_j$, $j = 1, \ldots, q-2$, $\mathbf{h}''$ is the vector of second differences $h''_2 = h'_{j+1} - h'_j$ $j = 1, \ldots, q-2$, and $\mathbf{K}$ is a suitable weight matrix. One well known penalty function of this form is that leading to a natural cubic spline form for $h()$. This would arise from $\gamma = 0$ and $\mathbf{K} = \mathbf{R}^{-1}$ where $\mathbf{R}$ is the symmetric banded matrix with 2/3 on the diagonal and 1/6 on the first off-diagonals (Reinsch, 1967). This penalty function is equivalent to imposing an integrated squared second derivative penalty on $h()$ (Silverman, 1985; O'Sullivan *et al.*, 1986; Hastie and Tibshirani, 1990; Green and Silverman, 1994). The smoothing application considered here is essentially discrete, in that values taken by h at non-integer arguments are immaterial, so integrated derivative penalty functions seem less relevant. This suggests that we may as well choose $\mathbf{K} = \mathbf{I}$ so that $\mathbf{h}''^T\mathbf{K}\mathbf{h}''$ is simply the sum of squared second differences. In the limited practical experience of the authors, the fitted curve for h is relatively insensitive to the particular weight matrix used and $\mathbf{K} = \mathbf{I}$ has proved satisfactory.

The use of first differences as well as second differences in penalty functions has also been suggested by Good and Gaskins (1971) in the context of non-parametric density estimation. Good and Gaskins used $\gamma = 1$. Here γ needs to be much smaller, e.g., 10^{-4} or 10^{-8}, so that the second difference term in the penalty tends to impose linearity on h for moderate α values while the first difference term comes into play to force constancy for larger values of α. An ideal for γ will yield a full range of forms for $h(n)$ as α increases before h is forced to a constant value. In our experience, $\gamma = 10^{-4}$ is satisfactory when the counts are below around 20 while a smaller value, $\gamma = 10^{-8}$ say, may be required if larger counts are observed. For moderate values of α the first difference term in the penalty function should have negligible effect.

6 Example: Foetal implants re-visited

This section returns to the foetal implant count data introduced in Section 2. The counts represent the number of sites where fertilized eggs have implanted in utero in control mice. There are no zero counts since an animal has to have at least one site where a fertilized egg has implanted in order to contribute. A fitted distribution to these control data is of interest to provide a benchmark distribution for comparison with experimental data under different environmental conditions. It is natural take a non-parametric approach to these data because the empirical transition rate sequence shown in Figure 1 is non-monotonic and does not follow any obvious parametric shape.

There no covariates for these data so $\lambda_{i,n} = \lambda_n$ for all i. Figure 2 shows the λ_n sequences obtained using non-parametric estimation for various values of the smoothing parameter, α. Knots have been placed at each integer between $n = 1$ and $n = 21$ and the multiplier of the penalty component involving the first differences was $\gamma = 10^{-4}$. The maximum penalized log-likelihood, ℓ_p, and corresponding log-likelihood, ℓ, are also shown for each smoothing parameter value. Note that with a knot placed at every integer within the range of the data, $\alpha = 0$ yields fitted λ_n equal to the empirical sequence plotted in Figure 1 corresponding to the observed frequency distribution of counts. Visual inspection of the n-dependent fits and log-likelihoods in Figure 2 suggests that either the $\alpha = 0.5$ or $\alpha = 1$ fits are reasonable. The $\alpha = 0.1$ fit is perhaps still a little rough for larger n while the $\alpha = 2$ fit has a large decrease in the log-likelihood relative to the increase in smoothness. The smoothed sequence and probabilities shown in Figure 1 were those for $\alpha = 0.5$.

There are very few parametric distributions available to model under-dispersion relative to the Poisson distribution. Table 2 compares the fit of the non-parametric EPPM with that of three such distributions and with the Poisson itself. The table gives log-likelihoods and Pearson chi-squared goodness of fit statistics. When computing the Pearson statistics, the counts were grouped so that the expected count was at least five in each bin. Consul's generalized Poisson distribution (Consul, 1989) does sum to unity in this case but does not in general and may require arbitrary truncation (Nelson, 1975). For these data it fails to correctly model the left-tail and peak of the distribution. The multi-

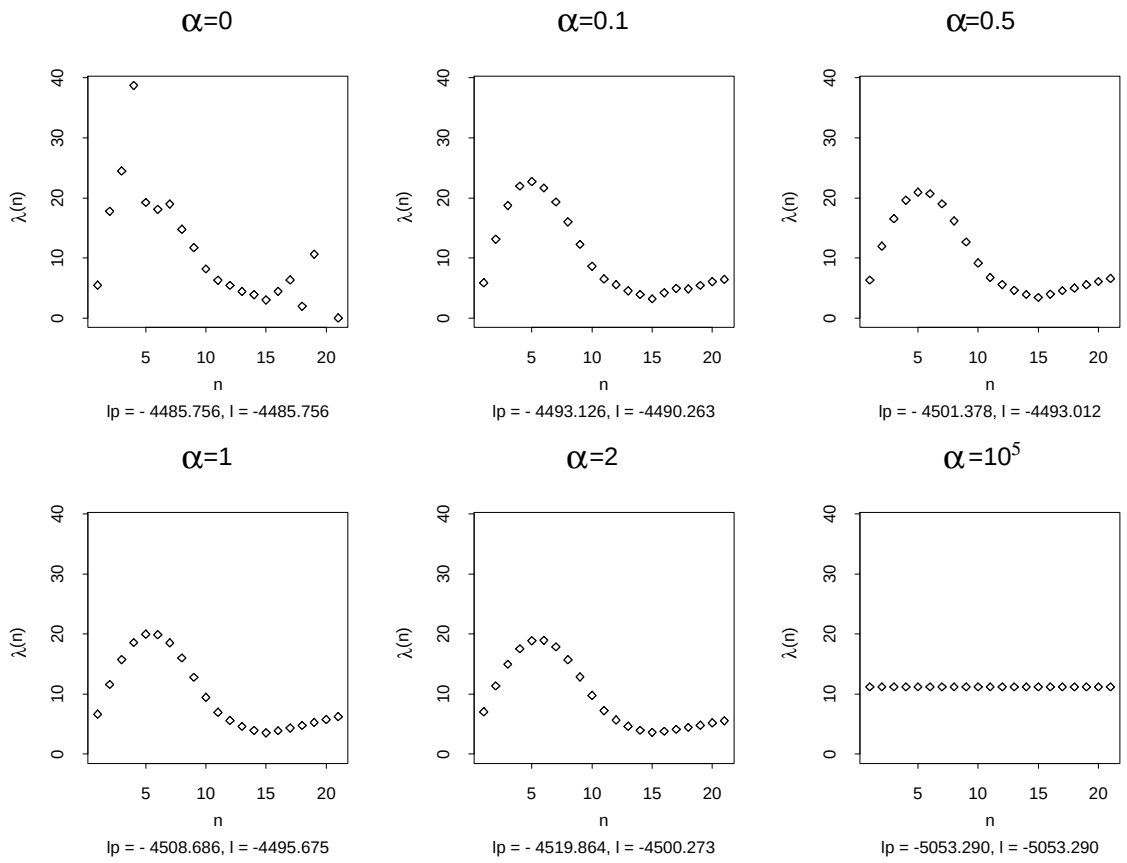

Figure 2: Non-parametric n-dependent fits for the foetal implants data.

Table 2: Comparison of model fits for the foetal implants data.

Model	Log-likelihood	No. parameters	χ^2_{gof}	df_{gof}
Poisson	-5053.2	1	889.6	17
Gamma count	-4828.7	2	358.9	12
Consul	-4718.0	2	270.3	11
Multiplicative Poisson ($\theta = 3$)	-4704.1	2	248.7	11
Non-parametric EPPM	-4493.0	6.2*	9.7	9.8

* See Section 9 for calculation of this value.

plicative Poisson (Lindsey, 1999, Section 7.2.3) and gamma count distributions (Winkelman, 1997) are less well known and also give unacceptable fits. The best performing parametric model is a 4-parameter EPPM proposed in Faddy (1998a) which, using saddlepoint probabilities, gives $\ell = -4493.96$ and $\chi^2 = 13.0$ on 13 degrees of freedom. The EPPMs are the only models to provide acceptable fits to these data.

It is tempting to interpret the increasing then decreasing fitted form of $h(n)$ in terms of biological feedback mechanisms. Successful implants in a host mouse indicate a healthy environment so further implants become increasingly likely. Indefinitely large litter sizes though are not biologically sustainable so, when an optimal litter size has been reached in any mouse, further implants become increasingly discouraged.

7 Choice of smoothing parameter

The subjective, graphical approach taken in the previous section to choosing the smoothing parameter is similar to the subjective methods often used to selection the degree of non-parametric smoothing in regression settings (Green and Silverman, 1994, Chapter 3). Automatic selection of the smoothing parameter is desirable but is computationally expensive, especially so in our setting since model fitting in itself is already intensive. In this section some preliminary investigations are presented towards automatic selection of the smoothing parameter.

The most common method used for automatic smoothing in non-parametric regression is cross-validation. Green (1987) discusses a cross-validation score for general regression problems derived from a 'delete-one' operation where each observation is deleted in turn and then predicted from the resulting fit. A similar approach can be taken in our setting.

Our cross-validation measure of predictive ability is based on twice the change in log-likelihood on including the deleted observation,

$$CV(\alpha) = \sum_i 2\{\ell(\hat{\boldsymbol{\beta}}^{(-i)}, \hat{\mathbf{h}}^{(-i)}|\mathbf{y}^{(-i)}) - \ell(\hat{\boldsymbol{\beta}}^{(-i)}, \hat{\mathbf{h}}^{(-i)}|\mathbf{y})\}, \tag{9}$$

where the $(-i)$ notation indicates that the ith observation has been removed. Here $\hat{\boldsymbol{\beta}}^{(-i)}$ and $\hat{\mathbf{h}}^{(-i)}$ are the maximum penalized likelihood estimators of $\boldsymbol{\beta}$ and $\mathbf{h}$ based on $\mathbf{y}^{(-i)}$.

The computational demands of the leave-one-out operation, requiring as many model re-fits as there are observations, are extremely onerous for general use, even more so here than in the regression setting. The computation can be reduced substantially by taking a single Newton-Raphson step from the complete fit in the delete-one operation rather than full likelihood maximization. Preliminary investigations using the data examples in this paper and several others suggest that this 'one-step' method approximates the full cross-validation well. For the foetal implants data, using a grid of α values at every tenth, the full cross-validation and this one-step method both give $\alpha = 0.3$ as the smoothing parameter value giving the minimum value of the cross-validation function. Of the values of α shown in Figure 2, $\alpha = 0.5$ gave the smallest value for the score function (9), agreeing with value chosen by visual inspection.

Another method of estimating the smoothing parameter which is sometimes used for non-parametric regression with normal errors is REML, as in Wecker and Ansley (1983) and Verbyla *et al.* (1999). It is possible to extend the idea of REML to this setting, by treating $\boldsymbol{\beta}$ and $\mathbf{h}$ as nuisance parameters and computing an approximate conditional penalized likelihood in a way analogous to Cox and Reid (1987). This is potentially less computer intensive than the cross-validation approach but experience so far suggests that it gives n-dependent forms that might be considered too rough on visual inspection.

8 Example: Toxoplasmosis data

Efron (1978) gave data on the proportion of subjects testing positive for toxoplasmosis in 34 cities of El Salvador. Interest lies in using the annual rainfall of each city to predict the proportion testing positive. Efron (1978) originally used a logistic regression model for these data and found that a cubic polynomial function of rainfall was highly

significant for predicting toxoplasmosis. However he noted the inadequacy of the model fit, with the chi-squared goodness of fit statistic being suggestive of over-dispersion relative to the binomial. He attributed this extra variation to imperfect random sampling and consequently unrepresentative sample sizes. Efron (1986) subsequently proposed a double logistic exponential family model in which the dispersion parameters were modelled as a quadratic regression on sample size.

Consider a semi-parametric EPPM model for these data,

$$\lambda_{i,n} = \exp(b_1 x_i + b_2 {x_i}^2 + b_3 {x_i}^3) h(n)(N_i - n),$$

where $x_i, {x_i}^2, {x_i}^3$ are the values of linear, quadratic and cubic orthonormal polynomials of rainfall and N_i is the number of patients tested in the relevant city. Owing to the presence of over-dispersion, it is expected that $h()$ will be an increasing function.

Figure 3 shows the fitted $h(n)$ obtained for various values of the smoothing parameter α with $\gamma = 10^{-8}$ and with a knot placed at each integer between $n = 0$ and the maximum observed count $n = 53$. The maximized penalized log-likelihoods and corresponding log-likelihoods are also given on the plot. Visual inspection of the curves suggests that a linear increasing form for $h(n)$ with $\alpha \approx 10^6$ is appropriate since the less smooth curves with $\alpha < 10^6$ do not represent significantly smaller log-likelihoods. The choice of $\alpha = 10^6$ is confirmed by the one-step cross-validation method.

The EPPM with $\alpha = 10^6$ uses one fewer parameter than Efron's (1986) double exponential family model, however the fitted means compared in Figure 4 show that the EPPM generally represents an improved fit. In particular the two cities with the largest counts are much improved. The Pearson chi-squared goodness of fit statistic for the EPPM is 37.91 on 29 degrees of freedom indicating an acceptable fit.

The steadily increasing transition rates suggest an interpretation in terms of inter-dependencies between the subjects during the testing process. As the number of subjects testing positive increases in each city, symptoms of the disease may become more well-known in the population and more positive subjects may as a result present themselves for testing. This type of positive feedback results in data over-dispersed relative to the binomial.

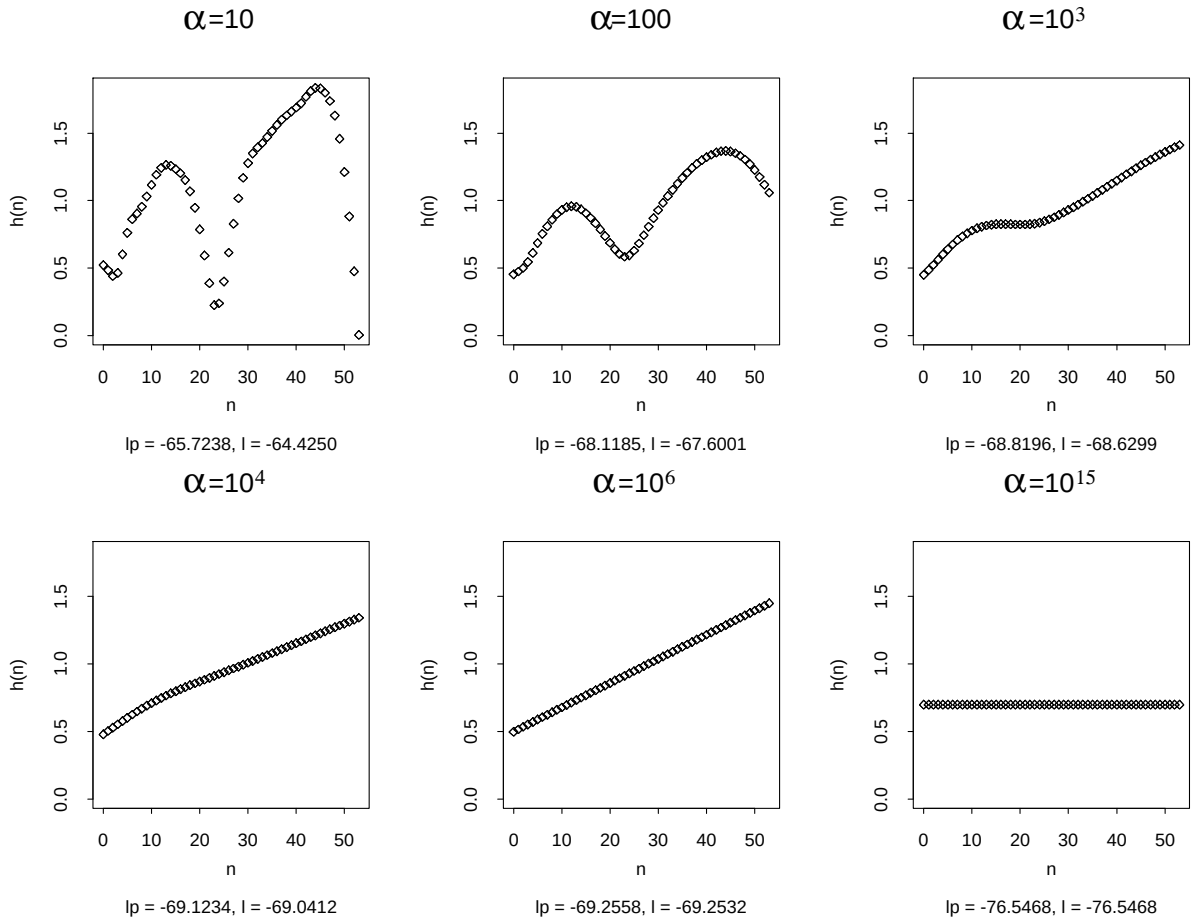

Figure 3: Non-parametric n-dependent fits for the toxoplasmosis data.

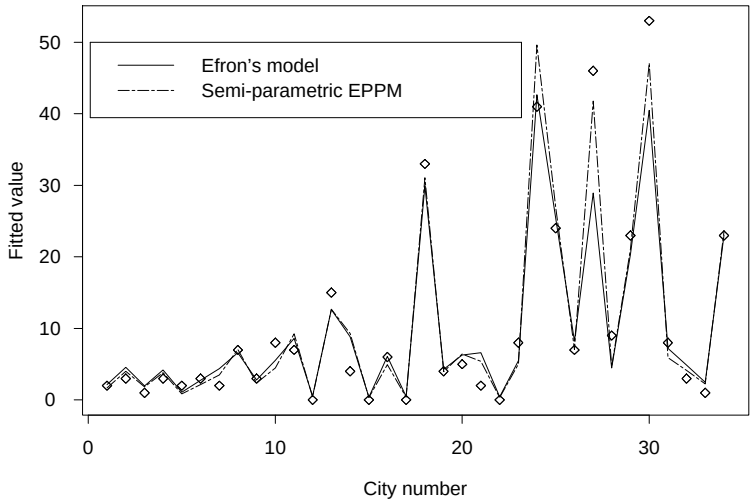

Figure 4: Comparison of fitted means for the toxoplasmosis data.

9 Inference in the semi-parametric framework

This section addresses the problems of assessing significance of $\hat{\mathbf{h}}$ and $\hat{\boldsymbol{\beta}}$ for non-parametric EPPMs. The discussion is undertaken in the context of the two data examples.

Our approach to testing hypotheses abut $\hat{\mathbf{h}}$ and $\hat{\boldsymbol{\beta}}$ is to undertake likelihood ratio tests with the smoothing parameter fixed at the selected value. If α is fixed at 10^6 then testing for deviations from binomial variation for the toxoplasmosis data is equivalent to testing for constant $h(n)$ versus linear $h(n)$. Twice the change in log-likelihood between the binomial and the $\alpha = 10^6$ EPPM model is $2 \times \Delta\ell = 14.59$ on one degree of freedom. The usual asymptotic chi-square approximation to log-likelihood changes gives a p-value of 0.0001. To assess the significance of the polynomial rainfall covariates, twice the change in log-likelihood is 6.57 on 3 degrees of freedom giving $p = 0.087$. It appears that, once the over-dispersion in the data has been accommodated into the model, the significance of rainfall for predicting toxoplasmosis is marginal at best.

Simulations verify the chi-squared approximations used in the two log-likelihood two tests above. Both the binomial and linear $h(n)$ models were fitted to 1000 simulated binomial data sets. Figure 5 shows a quantile-quantile plot comparing twice the change in log-likelihood between the two models with theoretical deviates from a chi-squared distribution on one degree of freedom. The diagonal line is the line of equality and the

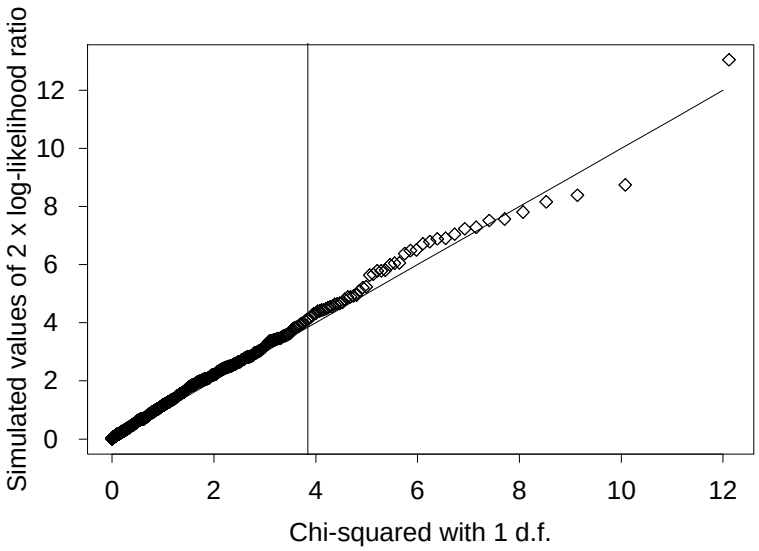

Figure 5: Chi-square q-q plot of simulated values of the log-likelihood ratio for testing departures from binomial variation for the toxoplasmosis data.

vertical line represents the 95% quantile of the chi-squared distribution. The plot shows good agreement. This is further supported by the mean likelihood change which at 1.11 is close to the nominal value of one and the variance which at 2.15 is close to the nominal value of two. Figure 6 shows a similar quantile-quantile plot from 1000 simulations of twice the change in log-likelihood between the $\alpha = 10^6$ model with and without the inclusion of rainfall covariates, simulated under the null model. This indicates that the chi-squared approximation used for assessing the significance of the rainfall effect is also in order, with the mean of 3.10 close to the nominal 3 degrees of freedom and variance of 6.28 close to twice this value. For these data the shape of the fitted $h(n)$ sequence does not change appreciably on removing the covariates since their effect is small. Treating the smoothing parameter as fixed as we have done here may not be so appropriate for data where the addition or removal of covariates dramatically changes the level of residual variation.

When a large number of degrees of freedom are involved, the chi-square approximation to the likelihood ratio statistic cannot be expected to hold as well as above. Results suggest that null distribution of the likelihood ratio statistic tends to have a larger mean than the degrees of freedom would suggest. For example, if no smoothing is used then the non-parametric EPPM model used for the foetal implants data is known to have 21 parameters, equal to the number of knots. Simulations comparing this model to a Poisson

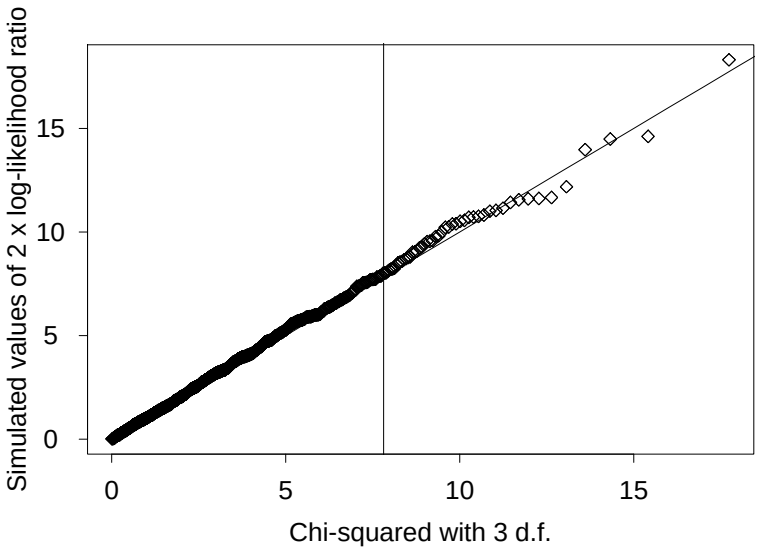

Figure 6: Chi-square q-q plot of simulated values of the log-likelihood ratio statistic for covariate assessment for the toxoplasmosis data.

model with simulated Poisson data gives a mean likelihood ratio statistic of 27.5 whereas the degrees of freedom are 20. The same simulation with $\alpha = 0.5$ gave a mean likelihood ratio statistic of 10.5, which also seems high. On the other hand, the Pearson goodness of fit statistic should hold its nominal chi-squared distribution well provided that the data are grouped appropriately to ensure that expected counts are never less than five say.

We have used Pearson goodness of fit statistics to give a rough estimates of equivalent degrees of freedom associated with the semi-parametric fits. Consider for example the semi-parametric EPPM fitted to the foetal implants data with $\alpha = 0.5$. Data was simulated from the fitted model, the EPPM refitted to each data set and the Pearson goodness of fit statistic computed. To compute the Pearson statistics, the counts 2–3, 4–5 and 18 and over were grouped into single bins to give a total of 16 bins. The mean Pearson statistic was 9.8 suggesting that equivalent degrees of freedom are 6.2 for the non-parametric $h(n)$ curve with $\alpha = 0.5$. This value is a little higher than the four parameters used in Faddy's (1998a) parametric model for the same data although the fitted $h(n)$ sequences are very similar. Nevertheless we give 6.2 as a conservative figure for the equivalent number of parameters for this model in Table 2.

In scatterplot smoothing it is common to compute equivalent degrees of freedom for non-parametric curves from the trace of the projection matrix (Hastie and Tibshirani,

1990). That method is not available to us here because the non-parametric curve is fitted to a latent rather than an observed sequence. A direct method for computing equivalent degrees of freedom for semi-parametric EPPMS is clearly desirable but is not yet available.

10 Conclusion

EPPMs provide applied statisticians with a much wider class of count distributions than earlier methods. Increasing birth rate sequences give rise to over-dispersed distributions including but not limited to distributions arising from mixture models. Decreasing sequences give rise to under-dispersed distributions which are otherwise not straightforward to construct. Non-monotonic sequences give rise to distributions which are distinctly non-Poisson or non-binomial but which do not necessarily show marked over- or under-dispersion. This paper has widened the class of EPPMs available by demonstrating how an appropriate distributional shape can be determined non-parametrically from the data using penalized likelihood. The model proposed incorporates covariate effects proportionally into the transition rate model. The ability to accommodate non-standard variation allows for more realistic assessment of the significance of covariate effects than otherwise could be made.

The semi-parametric models provide a particularly promising framework within which to approach data such as the foetal implants example, where data give rise to non-monotonic birth rate patterns. Such a birth rate profile is interpretable in terms of the process which produced the counts. In this example, it is natural to imagine that for small n the birth rates would increase with n due to positively associated implantation. However, the rate profile would eventually turn down for larger n because the mouse is a finite biological system with a limited capacity to host implants. With this interpretation, the non-monotonic rate sequence provides a graphical explanation for the non-Poisson nature of the counts. Existing methods of regression analysis are not well suited to this type of data because the non-Poisson nature of the counts is not necessarily reflected in their mean-variance relationship, and because standard count distributions will not correctly model the short right tail of the distribution. Even when the $h(n)$ sequence doesn't admit a natural interpretation like this, non-parametric EPPMs remain

useful for empirical modelling of count data. The $h(n)$ sequence may be seen simply as a convenient re-parameterization of the probability distribution representing the residual variation.

Although the non-parametric methods considered in this paper can no doubt be refined considerably, particularly from the point of view of computational efficiency, they provide as they stand an practical methodology for allowing the data to determine the dispersion properties of the response distribution. The methodology is also useful as an exploratory tool for verifying or demonstrating the lack of fit of standard models such as the Poisson, negative binomial and binomial models. In this exploratory role, semi-parametric EPPMs may be used to suggest other suitable parametric functions for the n-dependent form of the transition rates. In both these parametric and semi-parametric forms, EPPMs provide a flexible, promising and distinctively different framework for the analysis of count data.

Software for fitting EPPMs has been developed for the popular statistical package S-Plus and is available from `http://www.maths.uq.edu.au/~hmp/`.

References

Ball, F. 1995. A note on variation in birth processes, Math. Scientist 20: 50–55.

Ball, F. and Donnelly, P. 1987. Interparticle correlation in death processes with application to variability in compartmental models, Adv. Appl. Prob. 18: 755–766.

Bartlett, M.S. 1978. An Introduction to Stochastic Processes. Cambridge University Press, Cambridge.

Brown, T.C. and Donnelly, P. 1993. On conditional intensities and on interparticle correlations in nonlinear death processes, Adv. Appl. Prob. 25: 255–260.

Consul, P.C. 1989. Generalized Poisson Distributions. Marcel Dekker Inc., New York.

Cox, D.R. and Miller, H.D. 1965. The Theory of Stochastic Processes. Methuen, London.

Cox, D.R. and Reid, N. 1987. Parameter orthogonality and approximate conditional inference (with discussion), J. R. Statist. Soc. B 49: 1–39.

Daniels, H.E. 1982. The saddlepoint approximation for a general birth process. J. Appl. Prob. 19: 20–28.

Dean, C. B. 1998. Over-dispersion. In: Encyclopedia of Biostatistics (eds P. Armitage

and T. Colton) pp. 3226–3232. Wiley, London.

Donnelly, P., Kurtz, T. and Marjoram, P. 1993. Correlation and variability in birth processes, J. Appl. Prob. 30: 275–284.

Efron, B. 1978. Regression and ANOVA with zero-one data: measures of residual variation, J. Am. Statist. Ass. 73: 113–121.

Efron, B. 1986. Double exponential families and their use in generalized linear regression, J. Am. Statist. Ass. 81: 709–721.

Eubank, R.L. 1988. Spline Smoothing and Nonparametric Regression. Marcel Dekker Inc., New York.

Faddy, M.J. 1997a. Extended Poisson process modelling and analysis of count data, Biometrical Journal 39: 431–440.

Faddy, M.J. 1997b. On extending the negative binomial distribution and the number of weekly winners of the UK national lottery, Math. Scientist 22: 77–82.

Faddy, M.J. 1998a. Markov process modelling and analysis of discrete data. Appl. Stochastic Models and Data Analysis 13: 217–223.

Faddy, M.J. 1998b. Stochastic models for analysis of species abundance data, In: Statistics in Ecology and Environmental Monitoring 2 (eds D.J. Fletcher, L. Kavalieris and B.F.J. Manly) pp. 33–40. University of Otago Press, Dunedin.

Faddy, M.J. and Bosch, R.J. 2001. Likelihood-based modelling and analysis of data under-dispersed relative to the Poisson distribution, Biometrics 57: 620–624.

Faddy, M.J. and Fenlon, J.S. 1999. Stochastic modelling of the invasion process of nematodes in fly larvae, J. R. Statist. Soc. C 48: 31–37.

Good, I.J. and Gaskins, R.A. 1971. Nonparametric roughness penalties for probability densities, Biometrika 58: 255–277.

Green, P.J. 1987. Penalized likelihood for general semi-parametric regression models, Int. Statist. Rev. 55: 245–259.

Green, P.J. and Silverman, B.W. 1994. Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London.

Hastie, T. J. and Tibshirani, R. J. 1990. Generalized Additive Models. Chapman and Hall, London.

Lindsey, J. K. 1999. Models for Repeated Measurements, Second Edition. Clarendon

Press, Oxford.

McCaughran, D.A. and Arnold, D.W. 1976. Statistical models for numbers of implantation sites and embryonic deaths in mice, Toxicol. and Appl. Pharmacol. 38: 325–333.

McCullagh, P. and Nelder, J.A. 1989. Generalized Linear Models. 2nd edn. Chapman and Hall, London.

Moler, C. B., and van Loan, C. F. 1978. Nineteen dubious ways to compute the exponential of a matrix. SIAM Review, 20: 801–836.

Nelson, D.L. 1975. Some remarks on generalizations of the negative binomial and Poisson distributions, Technometrics 17: 135–136.

O'Sullivan, F., Yandell, B.S. and Raynor, J. Jr. 1986. Automatic smoothing of regression functions in generalized linear models, J. Am. Statist. Ass. 81: 96–103.

Podlich, H.M., Faddy, M.J. and Smyth, G.K. 1999. Likelihood computations for extended Poisson process models, InterStat, September no. 1, 15 pages.

Reinsch, C.H. 1967. Smoothing by spline functions, Numer. Math. 10: 177–183.

Sidje, R.B. 1998. EXPOKIT: Software package for computing matrix exponentials, ACM Transactions on Mathematical Software 24: 130–156.

Silverman, B. W. 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting, J. R. Statist. Soc. B 47: 1–52.

Smyth, G.K. and Podlich, H.M. 2002. An improved saddlepoint approximation based on the negative binomial distribution for the general birth process. Computational Statistics 17: 17–28.

Tapia, R.A. and Thompson, J.R. 1978. Nonparametric Probability Density Estimation. Johns Hopkins University Press, Baltimore.

Toscas, P.J. and Faddy, M.J. 2003. A likelihood-based analysis of longitudinal count data using a generalised Poisson model. Statistical Modelling. In press.

Tukey, J.W. 1949. One degree of freedom for nonadditivity. Biometrics 5: 232–242.

Stewart, W. J. (1994). Introduction to the numerical solution of Markov chains. Princeton University Press, New Jersey.

Verbyla, A. P., Cullis, B. R., Kenward, M. G., and Welham, S. J. 1999. The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion), J. R. Statist. Soc. C 48: 269–311.

Wecker, W. E., and Ansley, C. F. 1983. The signal extraction approach to nonlinear regression and spline smoothing, J. Am. Statist. Ass. 78: 81–89.

Winkelman, R. 1997. Econometric Analysis of Count Data. Springer, Berlin.