Statistical Modelling 2007; **7**(3): 239–254

Ill-posed problems with counts, the composite link model and penalized likelihood

Paul HC Eilers
Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands

Abstract: Certain data sets with distributions or counts can be interpreted as indirect observations of latent distributions or (time) series of counts. The structure of such data matches elegantly with the composite link model (CLM). The parameters can be estimated with iteratively re-weighted linear regression. Unfortunately, the estimating equations generally are singular or severely ill-conditioned. An effective solution is to impose smoothness on the solution, by penalizing the likelihood with a roughness measure. The optimal smoothing parameter is found efficiently by minimizing Akaike's Information Criterion (AIC). Several applications are presented.

Key words: back-calculation; mixtures; negative binomial distribution; over-dispersion

Received September 2006; **revised** January 2007; **accepted** February 2007

1 Introduction

Time series of counts, histograms and line transect samples are common examples of ordered count data. In many applications, it is sufficient to fit a smooth pattern to the raw counts for characterizing a trend or estimating a density. Density smoothing, especially with kernels, is a very well-developed field (Simonoff 1996; Wand and Jones, 1995). Many filtering and smoothing techniques for time series of counts are also available (Fahrmeir and Tutz, 2001; Gersch and Kitagawa, 1996; Pole *et al.*, 1994). These are all solutions to what one might call the 'direct observation' problem, to smoothly represent the original counts.

In this paper I present a modeling technique for a harder class of problems, which we might call 'indirect observation'. The expected values of observed counts are linear compositions (like partial sums or weighted averages) of a latent sequence representing a distribution or intensity. That sequence is to be estimated. Examples that will be analyzed in some detail in this paper are:

Address for Correspondence: Paul HC Eilers, Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands. E-mail: P.H.C.Eilers@uu.nl

 10.1177/1471082X0700700302

- Histograms with coarse intervals, as might be caused by the limitations of a measuring instrument, infrequent observation or deliberate grouping before reporting the data. The underlying density is to be estimated on a detailed grid.
- Discrete frequency distributions that do not fit to a standard distribution, like Poisson or binomial. A continuous mixture of standard distributions is a possible model. The mixing density is to be estimated.
- Back-calculation of HIV infections from counts of AIDS cases (Bachetti *et al.*, 1993; Heisterkamp, 1995). Time series of reported AIDS cases are given, as well as the (possibly time-varying) distribution of incubation times. From these observations the time series of HIV infections is to be estimated.
- Present status data with counts. Sun and Kalbfleisch (1993) analyzed a special type of survival data. Rats that died in the course of an experiment were examined for the number of tumors they had developed. For each animal, survival time and number of tumors are given. A possible model is to assume an unknown gradually changing intensity of occurrence of tumors. The expected number of tumors in an animal is the integral of this intensity over its lifetime. The goal is to estimate this intensity.
- The Wicksell corpuscle problem (Baddeley and Vedel Jensen, 2005; Stoyan *et al.*, 1995). A piece of solid material containing spherical inclusions is cut. Where spheres were cut, disks are seen. The goal is to estimate the distribution of the diameters of the spheres from (a histogram of) the measured disk diameters.

Indirect observation problems are also known as inverse problems. They are generally ill-posed, which means that a possible solution has too many degrees of freedom to be uniquely defined by the data. The example of the coarse histogram makes this clear. We observe, say, 10 numbers, the counts in 10 coarse intervals. We wish to estimate a discrete density on a grid of five times narrower intervals, or 50 numbers. There are infinitely many possible solutions that exactly fit the data.

When the data do not provide enough information to lead to well-behaved and unique estimates, we have to bring in prior knowledge, constraints and common sense to make the problem well-posed. It seems natural to exploit the natural order in the solution and to impose smoothness. As we will see, this can be implemented elegantly with a roughness penalty. It is a powerful tool to tame many ill-posed problems.

One can interpret the smoothness assumption as just a computational trick, but probably it cuts deeper. Smoothness seems a natural property to ask for, both on physical and philosophical grounds. If we are estimating, say, an unknown distribution from grouped observations, it is reasonable to assume it to be smooth, because we know little about it. A smooth curve has little detail. Only if sufficient data indicate details, we should allow them. On the other hand, we cannot expect to estimate fine details from limited data. By lowering our demands, by accepting a smooth result, we can get something useful from data that at first sight may seem inadequate for any answer.

The composite link model (CLM) is an elegant framework for modeling indirect observations of counts. It was proposed by Thompson and Baker (1981) as an extension of the generalized linear model (GLM) (McCullagh and Nelder, 1989; Nelder and Wedderburn, 1972). The CLM can easily be extended with a penalty on the roughness of the parameter vector.

The penalized CLM (PCLM) will be presented in two forms. One form estimates a discrete latent distribution directly. The second form uses B-splines, with a penalty on the coefficients, following the P-splines methodology, as advocated by Eilers and Marx (1996).

The parameters of the CLM are estimated by maximum likelihood, leading to an iteratively re-weighted least squares algorithm that is very similar to the one used for fitting a GLM. To add a penalty, only minor modifications are necessary.

The paper is constructed as follows: after this introduction, the essence of the CLM is introduced in an applied context. In Section 3, the derivation of the estimating algorithm is presented. Section 4 introduces difference penalties for smoothness and their integration in the CLM. The optimization of the penalty, using Akaike's Information Criterion (AIC), is treated in Section 5. Section 6 describes several applications. A discussion concludes the paper.

This paper is an extension of the work in Eilers (1995). More details and applications are presented and P-splines are added.

2 The composite link model

A flow diagram illustrates the composite link model in the context of this paper (see Figure 1). We assume a smooth unobservable vector γ. It is formed as $\gamma = \exp(X\beta)$, with β smooth, to guarantee non-negatives values of γ. A given matrix C 'composes' the vector μ from γ. The observations y are interpreted as realizations from Poisson processes with $E(y) = \mu$.

Note that this model can be quite different from a process where samples from Poisson distributions with expectations γ are generated and then linearly combined with C. In the latter case, when rows of C overlap, independence of the elements of y cannot be assumed.

The common element in the models to be presented is the unobservable distribution γ. The way it is modeled is largely independent of the observations. The

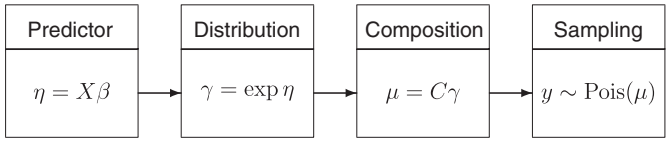

Figure 1 Flow diagram of the composite link model

second part is the 'composition matrix' C in $\mu = C\gamma$, where μ is the expectation of the observed y. C describes how the latent distribution was mixed before generating the data. In some applications C might aptly be called a convolution or filtering matrix. It is characteristic for the process that generates the data. The examples should make this clear.

Coarse histogram. In (2.1), C is shown for a histogram with three coarse and 15 narrow intervals. Generally, when m coarse (equally wide) intervals are given and the density is to be estimated on r times narrower intervals, C is m by mr, and the elements of C are zero, except that $c_{ij} = 1$ if $r(i-1) < j \leq ri$.

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{2.1}$$

HIV back-calculation. In a simplified model of the AIDS epidemic, the distribution of incubation times is assumed to be constant over time. It is denoted by $q = q_0, q_1, q_2, \ldots$. If γ is the distribution of yearly infections over time, then

$$C = \begin{bmatrix} q_0 & 0 & 0 & 0 \\ q_1 & q_0 & 0 & 0 \\ q_2 & q_1 & q_0 & 0 \\ q_3 & q_2 & q_1 & q_0 \end{bmatrix}. \tag{2.2}$$

We observe y, the number of AIDS diagnoses in each year. In a more complicated model, with a changing distribution of incubation times, the lower triangle of C would loose its Toeplitz structure, but the triangle above the diagonal would stay empty.

Poisson mixture. Consider a mixture of one-parameter discrete distributions. The distributions are indexed by j and the possible discrete outcomes by i, so q_{ij} is the probability of observing outcome i under distribution j. If p_j represents the probability of sampling from distribution j, the probability of outcome i is $\sum_j q_{ij} p_j$. Let a discrete distribution y be observed, with $\sum y_i = t$. The expected value of outcome i then is $\mu_i = t \sum_j q_{ij} p_j = \sum_j q_{ij} \gamma_j$ if we set $\gamma_j = t p_j$. This shows that we can apply the CLM if we fill the columns of C with the discrete distributions which are mixed, that is, $c_{ij} = q_{ij}$.

In the case of Poisson distributions, the columns of C are Poisson distributions, each having a different parameter λ:

$$c_{ij} = \lambda_j^{y_i} e^{-\lambda_j} / y_i! \tag{2.3}$$

The row index i is understood to start at zero here.

Present status data. The number of columns of C is determined by the unit of time and the maximum survival time. In the example of Sun and Kalbfleisch (1993), units of one week seem appropriate, with 150 columns, as the maximal observed survival time was 150 weeks. For each individual there is a row in C, with $c_{ij} = 1$ if $j \leq t_i/\Delta t$, if t_i is the survival time of individual i and Δt the unit of time. This choice of C models the expected value of y_i, the observed number of tumors in individual i, as the sum over its lifetime of the unknown intensity γ.

In many applications it can occur that I, the number of observations, is less than J, the number of parameters. Then the likelihood equations will be singular, with no useful solution. Even if $I > J$, the condition of the equations may be so bad that the solution is extremely sensitive to small changes in the model or the data. A way out is to constrain the solution vector to be smooth.

3 Estimation of the composite link model

Thompson and Baker (1981) present the CLM and the estimation algorithm very succinctly. Because we will extend the model, it is useful to give a more detailed derivation. The context of counts and the Poisson distribution will be used exclusively.

Assume that we observe counts z_j, $j = 1 \ldots J$ and explanatory variables x_{jk}, $j = 1 \ldots J$, $k = 1 \ldots K$. Assume further that z_j has a Poisson distribution with expectation γ_j:

$$P(z_j) = \gamma_j^{z_j} e^{-\gamma_j}/\gamma_j! \tag{3.1}$$

The GLM introduces a link function $g(\gamma)$ and a 'linear predictor' η_j:

$$g(\gamma_j) = \eta_j = \sum_{k=1}^{K} x_{jk}\beta_k. \tag{3.2}$$

The method of maximum likelihood (ML) is used to estimate β. McCullagh and Nelder (1989, equations 2.12 and 2.13) showed that the ML equations are

$$\sum_{j=1}^{J} \frac{(z_j - \gamma_j)}{v(\gamma_j)} \frac{\partial \gamma_j}{\partial \beta_k} = \sum_{j=1}^{J} \frac{(z_j - \gamma_j)}{v(\gamma_j)} \frac{\partial \gamma_j}{\partial \eta_j} x_{jk} = 0, \tag{3.3}$$

where $v(\gamma_j)$ is the variance when $E(z) = \gamma_j$. With the Poisson distribution and the log link, $\eta_j = \ln \gamma_j$, we have that $\partial \gamma_j/\partial \eta_j = \gamma_j$ and $v(\gamma_j) = \gamma_j$. Thus (3.3) simplifies to

$$\sum_{j=1}^{J} (z_j - \gamma_j) x_{jk} = 0. \tag{3.4}$$

These equations are nonlinear in β and an iterative procedure is needed to solve them.

Assume that approximate values $\tilde{\beta}_k$, and corresponding $\tilde{\gamma}_j$, are known. For small changes in β we find

$$\Delta\gamma_j = \gamma_j - \tilde{\gamma}_j \approx \sum_{k=1}^{K} \frac{\partial \gamma_j}{\partial \beta_k} \Delta\beta_k = \gamma_j \sum_{k=1}^{K} x_{jk}(\beta_k - \tilde{\beta}_k), \tag{3.5}$$

and

$$\sum_{j=1}^{J}\sum_{l=1}^{K} \tilde{\gamma}_j x_{jk} x_{jl} \Delta\beta_l = \sum_{j=1}^{J} x_{jk}(z_j - \tilde{\gamma}_j). \tag{3.6}$$

Add $\sum_j \sum_l \tilde{\gamma}_j x_{jk} x_{jl} \tilde{b}_l = \sum \tilde{\gamma}_j x_{jk} \tilde{\eta}_j$ to both sides to get

$$\sum_{l=1}^{K}\sum_{j=1}^{J} \tilde{\gamma}_j x_{jl} x_{jk} \beta_l = \sum_{j=1}^{J} \tilde{\gamma}_j x_{jk} \left(\frac{z_j - \tilde{\gamma}_j}{\tilde{\gamma}_j} + \tilde{\eta}_j\right), \tag{3.7}$$

with $\tilde{\eta}_j = \sum_{k=1}^{K} x_{jk} \tilde{\beta}_k$.

It is easy to recognize the last system of equations as weighted linear regression of a 'working variable' $(z_j - \tilde{\gamma}_j)/\tilde{\gamma}_j + \tilde{\eta}_j$ on X, with weights $\tilde{\gamma}_j$. This is a special case of the iteratively reweighted least squares (IRWLS) algorithm for the estimation of GLMs (Nelder and Wedderburn, 1972). Written as a matrix equation it is

$$X'\tilde{W}X\beta = X'\tilde{W}\{\tilde{W}^{-1}(y - \tilde{\gamma}) + X\tilde{\beta}\}, \tag{3.8}$$

with $\tilde{W} = \mathrm{diag}(\tilde{\gamma})$.

Suppose that we do not observe z, but other variables y, with $\mu = E(y) = C\gamma$, or $\mu_i = \sum_j c_{ij}\gamma_j$. Adapting (3.3), we find that

$$\sum_{i=1}^{I} \frac{(y_i - \mu_i)}{v(\mu_i)} \frac{\partial \mu_i}{\partial \beta_k} = 0 \tag{3.9}$$

We assume y to have a Poisson distribution. Because

$$\frac{\partial \mu_i}{\partial \beta_k} = \sum_{j=1}^{J} c_{ij} \frac{\partial \gamma_j}{\partial \beta_k} = \sum_{j=1}^{J} c_{ij} x_{jk} \gamma_j, \tag{3.10}$$

we get the likelihood equations

$$\sum_{i=1}^{I} (y_i - \mu_i) \breve{x}_{ik} = 0, \tag{3.11}$$

where $\breve{x}_{ik} = \sum_j c_{ij} x_{jk} \gamma_j / \mu_i$. We can interpret the matrix with elements $\breve{x}_{ik}$ as a 'working X' and proceed as for the GLM. The IWRLS equations become

$$\sum_{k=1}^{K} \sum_{i=1}^{I} \tilde{\mu}_i \breve{x}_{ik} \breve{x}_{il} \tilde{\beta}_l = \sum_{i=1}^{I} \tilde{\mu}_i \breve{x}_{ik} \left(\frac{y_i - \tilde{\mu}_i}{\tilde{\mu}_i} + \sum_{k=1}^{K} \breve{x}_{ik} \tilde{\beta}_k \right), \tag{3.12}$$

and

$$\breve{X}' \tilde{W} \breve{X} \beta = \breve{X} \tilde{W} \{ \tilde{W}^{-1} (y - \tilde{\mu}) + \breve{X} \tilde{\beta}) \} \tag{3.13}$$

where $\tilde{W} = \mathrm{diag}(\tilde{\mu})$.

In many applications the identity matrix can be chosen for X. This is the case when J is moderate or small, say $J \leq 100$. This choice has been made for all applications in Section 6. In other cases it may be advantageous to use a B-spline basis for X. The moderate size of the B-splines basis keeps the size of the estimating equations small.

Direct application of (3.13) generally will not work for ill-posed data. The equations will be singular, when $K > I$, or severely ill-conditioned otherwise. To solve this problem the following section presents the introduction of a roughness penalty on the parameter vector β.

4 Smoothness and difference penalties

When X in $\log \gamma = \eta = X\beta$ is the identity matrix, it is clear that smoothness of β implies smoothness of γ. Although it is perhaps less obvious, the same is true when X is a B-spline basis. Eilers and Marx (1996) proposed to combine B-splines with a (discrete) roughness penalty on the coefficients. The idea is not to try to optimize the number of B-splines and the positions of their knots, but to use an 'over-dimensioned' basis with an ample number of B-splines. Such a basis would give too flexible a fit to the data, but the difference penalty corrects this. In this section I present the essential background on discrete roughness penalties based on differences.

The roughness of a vector β can be measured with differences. The simplest form is

$$S_1 = \sum_{k=2}^{K} (\Delta \beta_k)^2 = \sum_{k=2}^{K} (\beta_k - \beta_{k-1})^2. \tag{4.1}$$

A rough vector β will show large differences between neighbouring elements and hence give a high value of S_1, while a smooth vector will make S_1 low. Ultimate smoothness is obtained when all elements of β are equal and $S = 0$.

Higher order differences can be used:

$$S_2 = \sum_{k=3}^{K} (\Delta^2 \beta_k)^2 = \sum_{k=3}^{K} (\beta_k - 2\beta_{k-1} + \beta_{k-2})^2 \tag{4.2}$$

or

$$S_3 = \sum_{k=4}^{K} (\Delta^3 \beta_k)^2 = \sum_{k=4}^{K} (\beta_k - 3\beta_{k-1} + 3\beta_{k-2} - \beta_{k-3})^2. \tag{4.3}$$

We get $S_2 = 0$ when $\beta_k = c_1 k + c_0$, for arbitrary c_0 and c_1, while S_3 will be zero for any β that is a quadratic in k.

It is useful to introduce matrix notation for differencing. Let D_d be the matrix that computes dth differences: $\Delta^d \beta = D_d \beta$. As examples, D_1 and D_2 are as follows, when $K = 5$

$$D_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}; \quad D_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix}. \tag{4.4}$$

High-level languages like S-plus and Matlab have functions to apply the difference operator to a matrix; construction of D_d then is trivial, by (repeated) differencing of the identity matrix. The roughness measure with differences of order d can now be written as

$$S_d = \beta^T D_d^T D_d \beta = |D_d \beta|^2 \tag{4.5}$$

The partial derivatives are

$$\frac{\partial S_d}{\partial \beta} = 2 D_d^T D_d \beta. \tag{4.6}$$

In a GLM or CLM we can force the solution vector β to be smooth by subtracting a roughness penalty from the log likelihood:

$$L^* = L - \alpha \| D_d \beta |^2 / 2. \tag{4.7}$$

The system of equations for the GLM become

$$(X' \tilde{W} X + \alpha D_d' D_d) \beta = X' \tilde{W} \{ \tilde{W}^{-1} (y - \tilde{\gamma}) + X \tilde{\beta} \}, \tag{4.8}$$

while those of the CLM change to

$$(\breve{X}' \tilde{W} \breve{X} + \alpha D_d' D_d) \beta = \breve{X} \tilde{W} \{ \tilde{W}^{-1} (y - \tilde{\mu}) + \breve{X} \tilde{\beta}) \}. \tag{4.9}$$

5 Optimizing the penalty

To find a good value for α, we minimize AIC, which is equivalent to

$$\text{AIC} = \text{Dev}(y|\mu) + 2\text{Dim} = 2 \sum_{i=1}^{I} y_i \ln(y_i / \mu_i) + 2\text{Dim}, \tag{5.1}$$

where $\text{Dev}(y|\mu)$ is the deviance and Dim is the effective dimension of the model. For the latter we follow the suggestion of Hastie and Tibshirani (1990) to take the trace of the smoother or 'hat' matrix H that is implicit in the linearized smoothing problem (4.9):

$$\hat{u} = \breve{X}\hat{\beta} = \breve{X}(\breve{X}^T W \breve{X} + \alpha D_d' D_d)^{-1}(\breve{X}^T W)u = Hz. \tag{5.2}$$

A simple search algorithm is sufficient: $\log\alpha$ is varied on a grid and the value that minimizes AIC is picked.

6 Applications

This section applies the PCLM to some of the problems that were presented in the introduction and in Section 2.

Figure 2 shows yearly counts of reported AIDS cases among homosexual and bisexual men in France (data courtesy of Siem Heisterkamp). AIDS is a manifestation of infection by the HIV virus. Column j of C (13 rows by 16 columns) gives the distribution of times of diagnosis, following infection in year j. For the purpose of illustrating the PCLM, this distribution is assumed to be identical for all j; it is shown in Figure 3. Note the large spread: the average incubation time is about 10 years. In a more extensive study, one would use quarterly data and a C matrix that reflects changes in medical care, as well as delays and seasonal effects in reporting of AIDS cases.

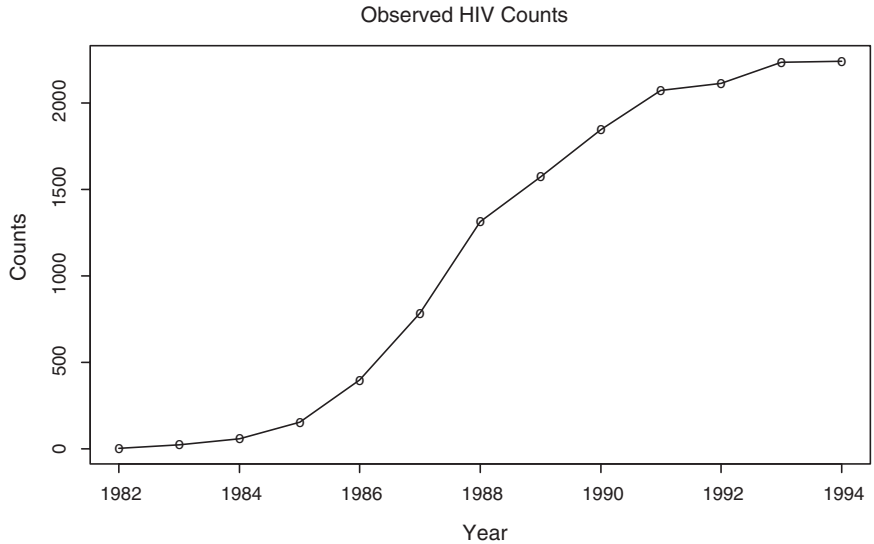

Figure 2 Yearly counts of AIDS cases among French homosexual and bisexual men

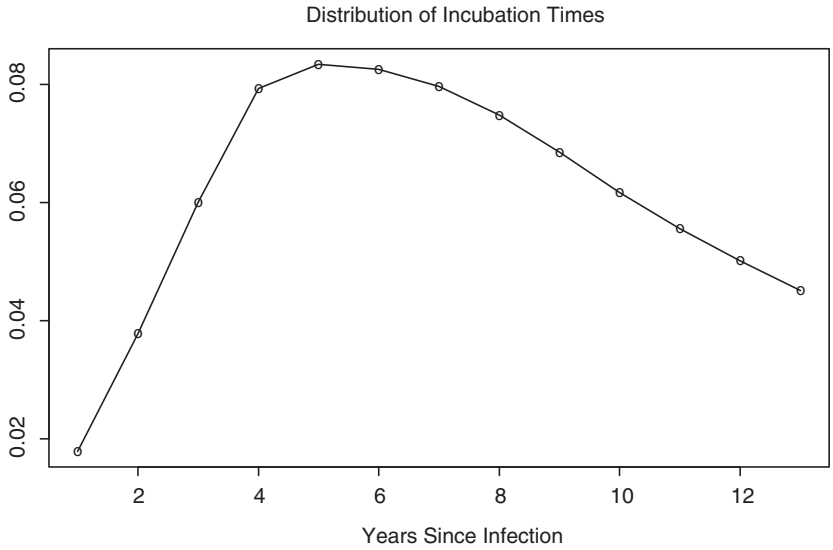

Figure 3 Distribution of AIDS incubation times

For estimating the time series of HIV infections, γ, Heisterkamp (1995) proposed to use a difference penalty to stabilize the back-calculation problem. He used a general optimization routine in S-plus and encountered long computation times. The PCLM algorithm converges readily in about 15 iterations and takes a few seconds to find a solution. For smaller values of α convergence is slowest. To find an optimal value for α, it was varied on a 'nice' grid (down from 1000, 500, 200, 100, and so on) and the value that gave the minimum of AIC was found to be 10 (AIC = 26.89, $d = 2$). The corresponding time series γ is drawn in Figure 4. Around the minimum, rather large changes of α give small changes in AIC. For $\alpha = 2$ (AIC = 26.93) and $\alpha = 50$ (AIC = 27.08), the corresponding series are plotted too. The peak around 1983 does not change much, but the most recent estimates of infections are strongly influenced by the amount of smoothing. This makes sense: because of the long average incubation times, there is very weak information about recent infections.

Figure 5 shows a coarse histogram from a paper by Hasselblad *et al.* (1980). In that paper a technique was developed to estimate a log-normal distribution from grouped data. Note that the leftmost group of the histogram is wider than the other ones. Here we estimate a non-parametric distribution with the PCLM. Six wide intervals have been observed so C has six rows. The observations cover the range from 0 to 70. For the latent distribution γ narrow intervals of width 1 are chosen, so C has 70 columns. With $d = 2$, the minimum of AIC (6.56) occurs for $\alpha = 10^4$; the corresponding detailed distribution γ is shown in Figure 5. With $d = 3$, we find that AIC decreases monotonically with increasing α. For $\alpha > 10^5$ the fitted distribution is drawn with a broken line in Figure 5.

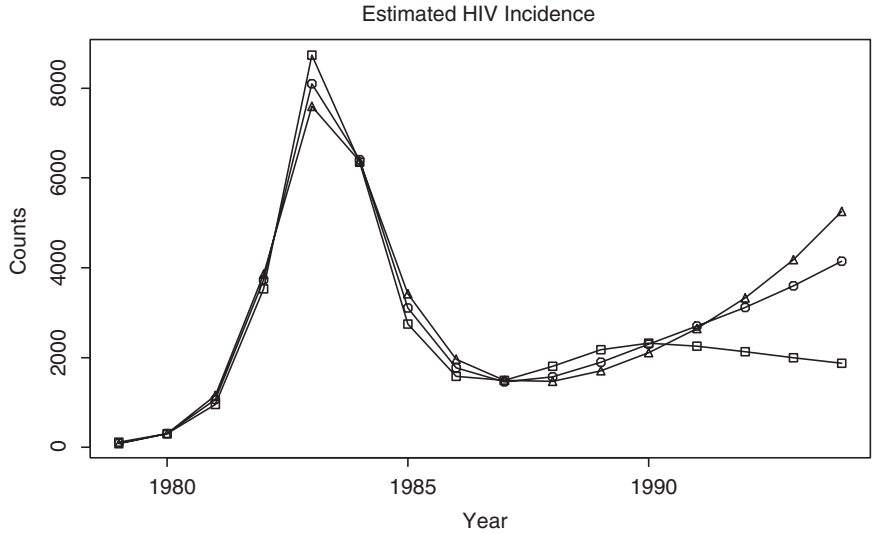

Figure 4 Estimated time series of yearly HIV infections. squares: $\alpha = 2$; circles: $\alpha = 10$; triangles: $\alpha = 50$

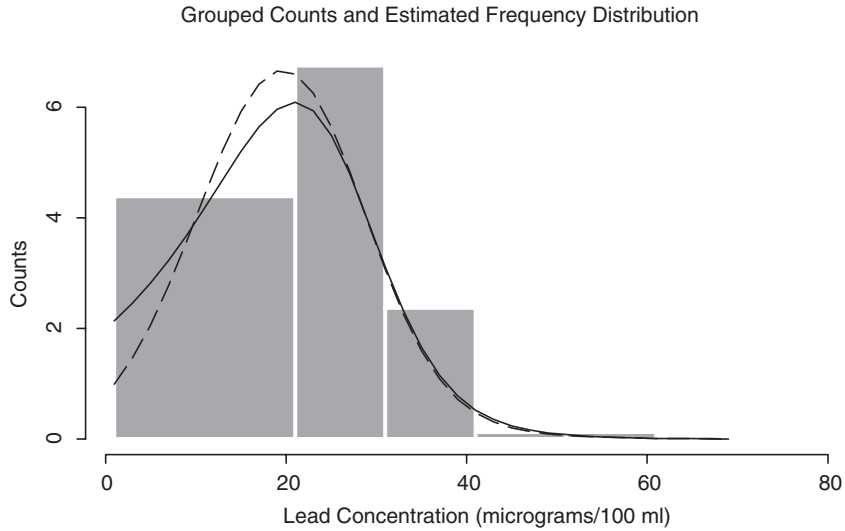

Figure 5 Histogram of lead concentrations in blood and estimated distributions; unbroken line: $d = 2, \alpha = 10^4$, AIC = 6.56; broken line: $d = 3, \alpha = 10^6$, AIC = 8.16

An interesting result is that AIC is higher when $d = 3$. It appears that $\mathrm{tr}(H)$ is near 2, when α is changed over a large range and γ changes a lot in appearance. Ideally, the value of $\mathrm{tr}(H)$ should indicate the 'complexity' of the curve of γ. One can show that the minimum of this trace is equal to d. Increasing d by one increases AIC by two, if the fit to the data stays the same.

The last example is an over-dispersed discrete distribution. The Rijnmond Environmental Agency registers complaints from inhabitants of the area about annoying odours, dust, noise and other blessings of modern life, about 20 000 each year. The number of complaints per day is very variable, as can be seen from Figure 6, which shows data for 1988. A Poisson distribution gives a deviance of 2382. A negative binomial distribution performs much better with a deviance of 121.8, but, as can be seen from Figure 6, it shows systematic departures, especially at the left tail. We now try to model the data as a mixture of Poisson distributions.

The large spread of the observed distribution suggests taking a linear grid for $\log_{10}(\lambda)$, instead of λ itself. A grid was chosen that runs from -1 to 1, in steps of 0.1. Thus the expected values of the mixed Poisson distributions run from 0.1 to 10 (complaints per day). This gives 21 columns for C and 100 rows, because the bins for the counts of complaints run from 0 to 99.

As before, we seek the minimum of AIC by varying α on a 'nice' grid (steps 1, 2, 5, 10, 20, ...). With second order differences in the penalty ($d = 2$), $\alpha = 50$ is best, with AIC = 101.4. For the negative binomial distribution in Figure 6, AIC = $2 * 2 + 121.8 = 125.8$. The mixture model gives a marked improvement. With $d = 3$,

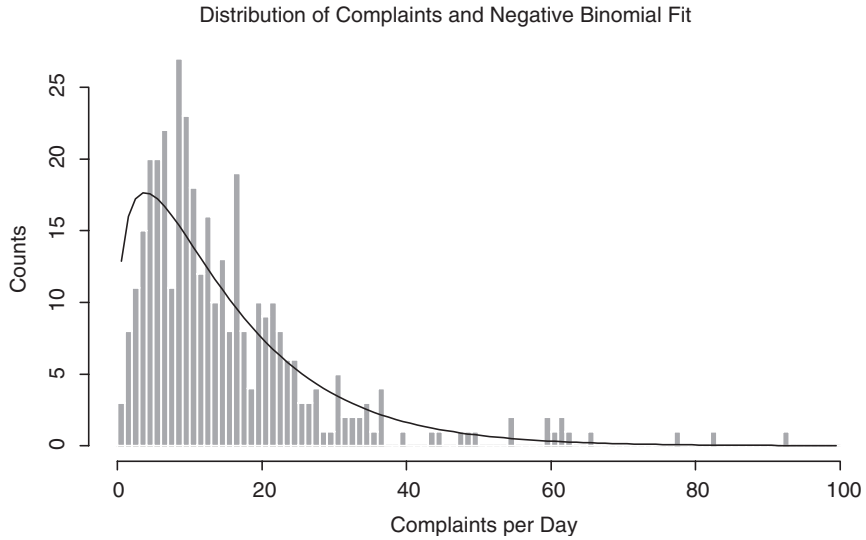

Figure 6 Distribution of daily number of environmental complaints about annoying odours (vertical bars) and the estimated negative binomial distribution

AIC shows a monotone decrease with increasing α, approaching the value 98.2. This means that effectively the mixing distribution of $\ln \lambda$ is (a discrete approximation to) a normal distribution. The reason for this is that with a heavy penalty and third order differences, $\Delta^3 \beta$ is very near to zero. For a second order polynomial in k, $\Delta^3 \beta_k$ is zero for all k. Increasing the penalty pushes the solution towards such a polynomial (the one that minimizes the deviance). Figure 7 shows the estimated distribution of $\log_{10} \lambda$ and Figure 8 shows the data distribution, the estimated mixed distribution ($d = 3$, $\alpha = 10^6$, AIC $= 98.2$).

Rival techniques for modelling mixed discrete distribution do exist, as proposed by Böhning *et al.* (1993). Their programme C.A.Man gives the mixing distribution as a set of (unequally spaced) discrete probability masses. Even though this may result in essentially the same fit to the data, a smooth distribution makes more sense, both from a theoretical and from a practical point of view. It is unlikely that λ is drawn from a set of point probability masses, and a smooth γ gives more sensible results in pseudo-Bayes estimates of λ for individual days.

7 Discussion

In this paper the PCLM was developed and applied to ill-posed problems with counts (or frequencies). The heart of the model is the composition matrix C. It is a flexible

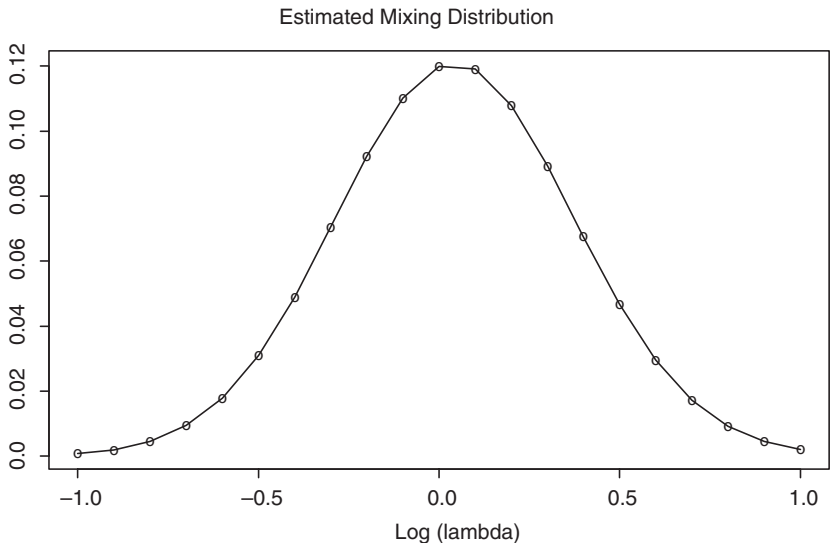

Figure 7 The estimated distribution of (the base 10 logarithm of) the Poisson parameter λ for the odour complaints distribution

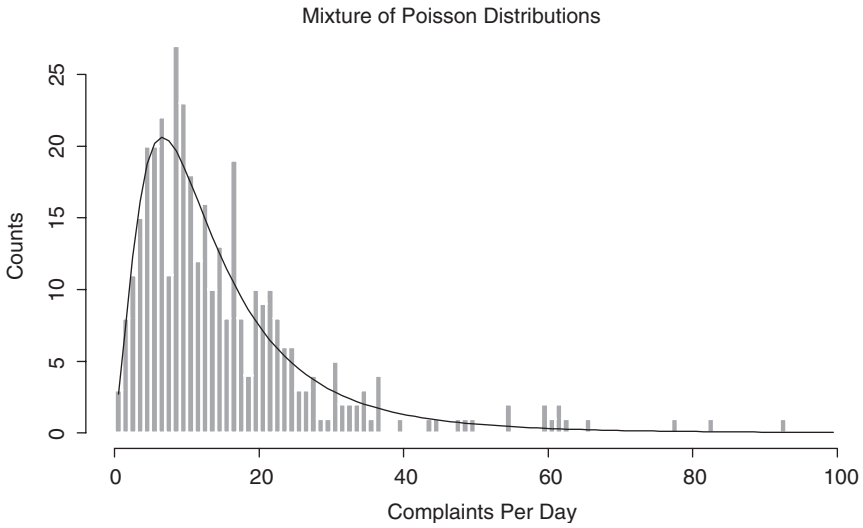

Figure 8 Distribution of daily number of environmental complaints about annoying odours (bars) and fitted mixture of Poisson distributions (line)

tool for the description of many types of observation processes. I will sketch some additional possibilities.

To estimate densities from arbitrarily censored or truncated individual observation, construct a C for which row i reflects the interval in which observation i lies, and take all $y_i = 1$. The coarse histogram is a special case, in which groups of rows of C are the same. Because the expected value of a sum of Poisson variates is the sum of the expectations, equal rows can be combined to one and y_i be taken as their number.

When data are observed with errors that have a known distribution, their distribution can be estimated with the PCLM. Then C is a real convolution matrix, its columns reflecting the distribution of the errors.

Eilers and Borgdorff (2004) use the PCLM to model digit preference in the context of tuberculosis surveys. Human observers estimate the size of the 'induration', the skin reaction after a Mantoux test. Depending on the skill of the observer, appreciable preference for even numbers and multiples of five will occur. A model for the transfer of counts from their actual ending digit to a preferred one fits in the PCLM framework.

The PCLM was presented here in a discrete framework: γ is a discrete distribution. When B-splines are used, γ can be interpreted as a continuous distribution as well. The elements of the composition matrix C then represent (weighted) integrals of γ (over subdomains).

The PCLM is not limited to the analysis of counts. A possible application is the analysis of overdispersed binomial data, assuming a mixing distribution is assumed on the logit scale (Eilers, 1994).

It also possible to extend the approach to the estimation of mixture distributions in (generalized linear) mixed models. These results will be reported elsewhere.

Several details of the computations deserve further investigation. In the applications presented here, AIC was minimized for values of α for which convergence occurred. But divergence can occur for small values of α. This is especially the case for higher values of d, the order of the differences in the penalty. In some sense, this is to be expected: the weaker the penalty, the weaker the stabilizing effect it will have. But it was also found that with damping schemes the convergence was slower, but smaller values of α were allowed. An interesting subject for study is to determine where the PCLM really breaks down, independent of the algorithm for maximizing the penalized likelihood.

It is remarkable that the CLM has received very little attention in statistical literature. In their monograph on latent variables, Skrondal and Raabe-Hesketh (2004a) mention it briefly and use it in one application. Skrondal and Raabe-Hesketh (2004b), and Raabe-Hesketh and Skrondal (2007) present a fuller treatment. See also Rindskopf (1992). Hopefully, the applications described in this paper will lead to a larger popularity of this very flexible model.

The penalties used in this paper exploit smoothness. In application without a natural order of the latent probabilities, this will not be possible. Still there are opportunities to use ridge penalties. Rindskopf (1992), when modelling contingency tables for observations with missing data, notes that the CLM may not have a unique solution, making it unidentifiable. This will manifest itself as large negative values of logarithms of latent probabilities. A ridge penalty can constrain their size. This approach is presently being applied in joint research with HW Uh, in the context of haplotype frequency estimation from genotype data (Uh and Eilers, 2005). The results will be presented elsewhere.

References

Bachetti P, Segal MR and Jewell NP (1993) Backcalculation of HIV infection rates. *Statistical Science*, **2**, 82–119.

Baddeley A and Vedel Jensen EB (2005) *Stereology for statisticians*. Chapman and Hall/CRC.

Böhning D, Schlattmann P and Lindsay B (1993) Computer-assisted analysis of mixtures. *Biometrics*, **48**, 283–304.

Eilers PHC (1994) Nonparametric estimation of mixed discrete distributions with penalized likelihood. Paper presented at the XVIIth International Biometric Conference, Hamilton.

Eilers PHC (1995) Indirect observations, composite link models and penalized likelihood. In Seeber GUH, Francis BJ, Hatzinger R and Steckel-Berger G eds *Proceedings of the 10th International Workshop on Statistical Modelling*. New York: Springer-Verlag, 91–98.

Eilers PHC and Borgdorff MW (2004) Modeling and correction of digit preference in tuberculin surveys. *International Journal of Tuberculosis and Lung Diseases*, **8**, 232–39.

Eilers PHC and Marx BD (1996) Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statistical Science*, **11**, 89–121.

Fahrmeir L and Tutz G (2001) *Multivariate statistical modelling based on generalized linear models, 2nd edition*. New York: Springer-Verlag.

Gersch W and Kitagawa G (1996) *Smoothness priors analysis of time series*. New York: Springer-Verlag.

Hasselblad V, Stead AG and Galke W (1980) Analysis of coarsely grouped data from the lognormal distribution. *Journal of the American Statistical Association*, **75**, 771–78.

Hastie TJ and Tibshirani RJ (1990). *Generalized additive models*. London: Chapman and Hall.

Heisterkamp S (1995) *Quantitative analysis of AIDS/HIV: development of methods to support policy making for infectious disease control*. PhD Dissertation, Leiden.

McCullagh P and Nelder JA (1989) *Generalized linear models (2nd edition)*. London: Chapman and Hall.

Nelder JA and Wedderburn RWM (1972) Generalized linear models. *Journal of the Royal Statistical Society A*, **135**, 370–84.

Pole A, West M and Harrison J (1994) *Applied Bayesian forecasting and time series analysis*. London: Chapman and Hall.

Raabe-Hesketh S and Skrondal A (2007) Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika*, **72**, 123–40.

Rindskopf D (1992) A general approach to categorical data analysis with missing data, using generalized linear models with composite link. *Psychometrika,* **57**, 29–42.

Simonoff JS (1996) *Smoothing methods in statistics*. New York: Springer-Verlag.

Skrondal A and Raabe-Hesketh S (2004a) *Generalized latent variable modeling*. Chapman and Hall/CRC.

Skrondal A and Raabe-Hesketh S (2004b) Generalised linear latent and mixed models with composite links and exploded likelihoods. In Biggeri A, Dreassi E, Lagazio C and Marchi M eds *19th International Workshop on Statistical Modeling*. Florence: Firenze University Press.

Stoyan D, Kendall WS and Mecke J (1995) *Stochastic geometry and its Applications, 2nd edition*. New York: Wiley.

Sun JG and Kalbfleisch JD (1993) The analysis of current status data on point-processes. *Journal of the American Statistical Association*, **88**, 1449–54.

Thompson R and Baker RJ (1981) Composite link functions in generalized linear models. *Applied Statistics*, **30**, 125–31.

Uh HW and Eilers PHC (2005) Haplotype estimation with the penalized composite link model (abstract). *Proceedings of the 25th European Meeting of Statisticians*, Oslo.

Wand MP and Jones MC (1995) *Kernel smoothing*. London: Chapman and Hall.