Generalized linear mixed models for over-dispersed counts

Paul H. C. Eilers

Abstract Over-dispersion can be a nuisance when modeling count data with a generalized linear model, when it is assumed that the Poisson distribution holds. A solution may be to use the negative binomial distribution, but there are limitations to the dispersion it can handle. This paper presents a first look at modeling over-dispersed counts using a semi-parametric smooth mixing distribution. The basis is the penalized composite link model. Several applications are presented.

Key words: composite link model, penalty, negative binomial distribution

1 Introduction

The generalized linear model [2] is a beautiful extension of the linear model. It allows us to model and analyze counts and proportions in a very general way. In its standard form it assumes the Poisson or binomial distribution. This is fine if these strong assumptions hold, but in practice one often encounters over-dispersed data. That means that the variance is larger than what the theory says and that standard errors based on these variances can be very wrong.

One solution is to use a distribution that allows over-dispersion. For binary data the beta-binomial distribution is popular. It does not assume a constant probability (conditional on the same values of covariates) but imagines that it is drawn from a beta distribution. Similarly the negative binomial distribution assumes a latent gamma distribution generating the expected value of a Poisson distribution. This often works well, but we will see examples where the over-dispersion is too large to be described well by the negative binomial distribution.

As an alternative I propose to estimate the mixing distribution from the data, using a semi-parametric model. This leads to the penalized composite link model

Paul Eilers
Ersmus University Medical Centre, Rotterdam, The Netherlands, e-mail: p.eilers@erasmusmc.nl

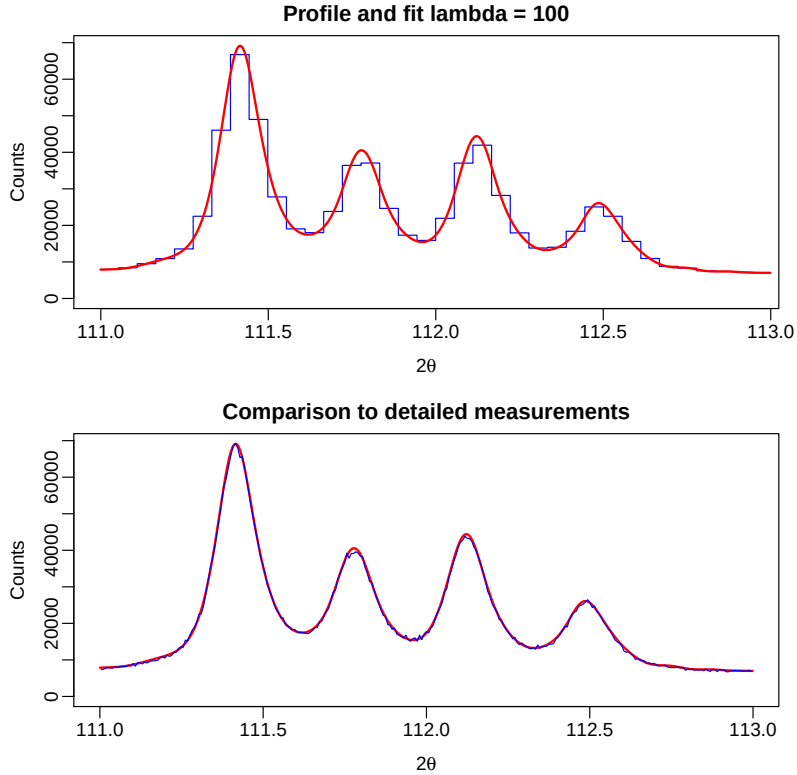

Fig. 1 Fitting a detailed distribution to a very coarse X-Ray diffraction scan. Upper panel: photon counts in very wide bins (blue) and the estimated smooth detailed profile (red). Lower panel: the estimated profile (red) compared to detailed measurements in a second dat set (blue).

(PCLM), a generalization of the GLM. The core of the PCLM is a "composition" matrix that connects observed counts to a latent GLM. Thompson and Baker [3] proposed the composite link model (CLM) and used it in applications with relatively few parameters. In our applications we will have relatively many parameters and a penalty is needed for stability [1].

In simple applications the composition matrix is constant. More interesting things can be done if we extend the PCLM to situations in which expected values depend on covariates.

I will use a dataset on environmental complaints to illustrate the PCLM and its practical use.

2 The penalized composite link model

The generalized linear model for counts commonly assumes a Poisson distribution and a log link function:

$$y_i = \mathrm{Po}(\mu_i); \quad \text{with} \quad \mu_i = \exp(\eta_i) \quad \text{and} \quad \eta_i = \sum_j x_{ij}\alpha_j.$$

Here y_i, $i = 1 : m$, are observed counts and X is an m by n matrix of explanatory variables. To fit this model to data, one iteratively solves

$$X'\tilde{M}X\alpha = X'(y - \tilde{\mu} + \tilde{M}\tilde{X}\tilde{\alpha}), \tag{1}$$

where a tilde indicates an approximation and $M = \mathrm{diag}(\mu)$. With proper starting values, like $\tilde{\eta} = \log(y+1)$, this algorithm generally converges quite quickly and a handful of iterations is sufficient for six-digit precision.

The composite link model [3] adds an extra layer: $\mu = C\gamma$, with $\gamma = \exp(\eta)$. The matrix C is called the composition matrix, as it composes the expected values of the observations as a linear combination of the elements in γ, which has a GLM structure.

Thompson and Baker presented an algorithm for estimating the CLM. Let $Q = M^{-1}C\Gamma X$, where $\Gamma = \mathrm{diag}(\gamma)$. Again iteratively reweighted regression is used:

$$\tilde{Q}'\tilde{M}\tilde{Q}\alpha = \tilde{Q}'(y - \tilde{\mu} + \tilde{M}\tilde{Q}\alpha). \tag{2}$$

It can happen that α has more elements than y and μ. Then the model is ill-conditioned an a solution cannot be found, because the equations will be singular. Figure 1 illustrates such a case. The data are counts of photons in 29 wide bins; they come from an X-Ray diffraction (XRD) scan of Corundum powder obtained at Delft University. We like to find a good estimate of the underlying curve, at a high resolution. That means determining 290 unknowns if we increase the resolution tenfold. There is no parametric model available, so it seems that we are stuck. A solution is to constrain the vector α in such a way that it shows desirable properties. In the present case we demand that α be smooth and we use a penalty to achieve that.

To make α smooth we have to make sure that it is not rough. That is not just playing with words, because we can measure roughness of a series in a simple way. Consider differences $\alpha_k - \alpha_{k-1}$. They have relatively large values when α is rough, because then it changes much from one element to the next. A global measure of roughness is

$$R_1 = \sum_k (\alpha_k - \alpha_{k-1})^2. \tag{3}$$

A somewhat more sophisticated measure uses double differences

$$R_2 = \sum_k ((\alpha_k - \alpha_{k-1}) - (\alpha_{k-1} - \alpha_{k-2}))^2 = \sum_k (\alpha_k - 2\alpha_{k-1} + \alpha_{k-2})^2. \tag{4}$$

By introducing matrices D_1 with the pattern $(-1,1)$ in its rows, or D_2, with the pattern $(1,-2,1)$, we can conveniently express these measures as $||D_1\alpha||^2$ or $||D_1\alpha||^2$, where $||.||^2$ indicates the sum of squares norm.

We minimize the penalized deviance $\mathrm{Dev}(\alpha;y)+\lambda||D\alpha||^2$, where D is D_1 or D_2 and λ is a parameter to regulate the influence of the penalty. The penalty changes the CLM algorithm only marginally:

$$(\tilde{Q}'\tilde{M}\tilde{Q}+\lambda D'D)\alpha = \tilde{Q}'(y-\tilde{\mu}+\tilde{M}\tilde{Q}\alpha). \tag{5}$$

Figure 1 shows the results of applying the penalized composite link model (PCLM) to the XRD data. We can check our results by comparing them to a more detailed measurement of Corundum. It looks rather good. The physical interest is in the location and the heights of the peaks. They can be determined surprisingly well from the coarse data.

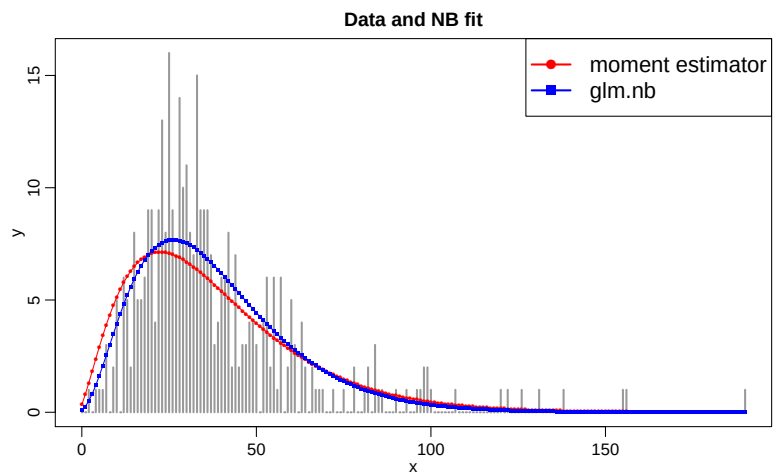

Fig. 2 Distribution of the daily number of complaints about annoying odors in the Rijnmond area in 1994 (grey bars). Estimated negative binomial distributions using the method of moments (red) or the function `glm.nb()` (blue).

3 Applications

The XRD data in the example serve to illustrate the potential of the PCLM in a situation where we can judge the results using additional data. We now turn to more realistic scenarios, where such a check is not available. The data were collected by the Rijnmond Central Environmental Agency (Dutch abbreviation DCMR) in the Rijnmond region around Rotterdam, the Netherlands. This region contains residential

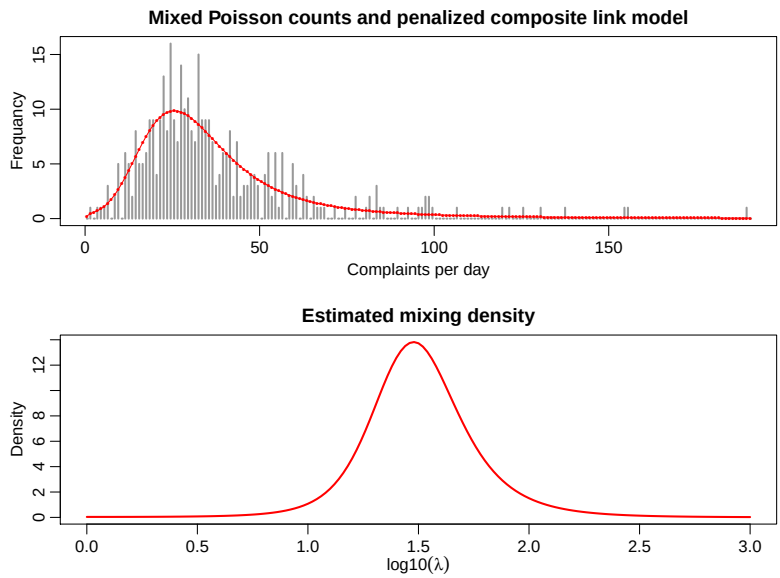

Fig. 3 Top: Distribution of the daily number of complaints about annoying odors in the Rijnmond area in 1994 (grey bars) and the fit of a semi-parametric mixture of Poisson distributions (red line). Bottom: the estimated mixing distribution.

areas, with 1.2 million inhabitants in total, but it is also the scene of extensive (petrochemical) industrial activities, stimulated by the large Rotterdam harbor. Large industrial areas can lead to environmental disturbances. DCMR has as monitoring and control task. Citizens can complain about environmental problems around the clock and their complaints are registered and reported. When possible, specially trained officers can go into the field and take action.

In 1994, the year we analyze here, the majority of the complaints was about annoying odors. Figure 3 shows the distribution of the numbers of complaints per day. As can be seen it is very variable. Actually so much that a negative binomial distribution does not give a good fit. This might not be directly obvious, but we will see it once a better model has been fit.

The model we will consider is a PCLM with column j of the composition matrix C being a Poisson distribution with expected value $\exp\theta_j$. The elements of the vector θ, of length n, form a linear sequence. In the present case θ has 151 elements with minimum 0 and maximum 3. With this matrix C we get the distribution of a mixture of Poisson distributions as $C\gamma$, if γ is the mixing distribution. Of course, the goal of the exercise is to estimate γ from the available data. We set $X = I$, the identity matrix, and thus $\gamma = \exp\alpha$. With a second order penalty and $\lambda = 10^6$ we get the results shown in Figure 3.

Visual judgment already gives the impression that the fit of this model is better than that of the negative binomial. In the right and left tail, as well as in the peak,

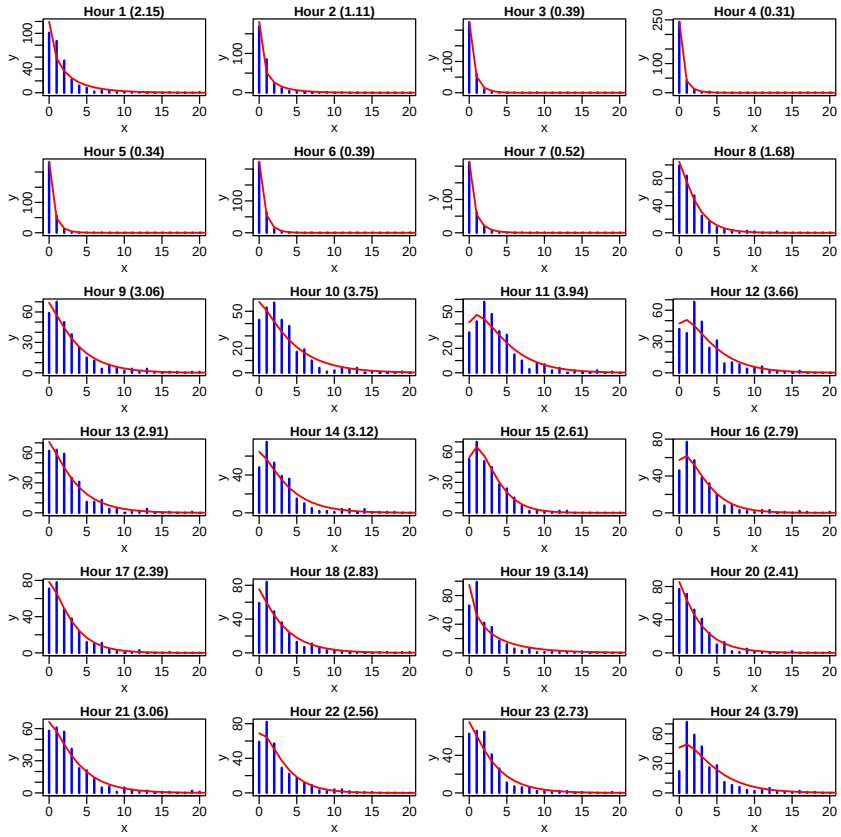

Fig. 4 Distributions of the number of complaints per hour of the day in 1994 (blue bars), and the fit of the negative binomial distribution (red line). The numbers between parentheses in the titles are average values.

the estimated distribution is more faithful to the data. A more objective measure is the deviance, which is 205.5 for the better NB model and 171.4 for the estimated mixture.

This is interesting and useful, but it has been done before [1] and we like to see more advanced applications. Let us assume that we have a number of groups and that the mixing distribution has the same shape for each group, but a different location. As an example we take the hour of the day for the grouping. As Figure 4 shows, there is a strong pattern, which makes sense: during the night most people will be asleep at home and will less easily notice obnoxious odors.

For each hour the fit of a negative binomial distribution is displayed. It is not great, so we will try to model a semi-parametric mixing distribution. Let the counts now be indicated by y_{hi}, where h stand for the hour and i for the histogram bin. We introduce offsets ϕ_h and a three-dimensional array C such that

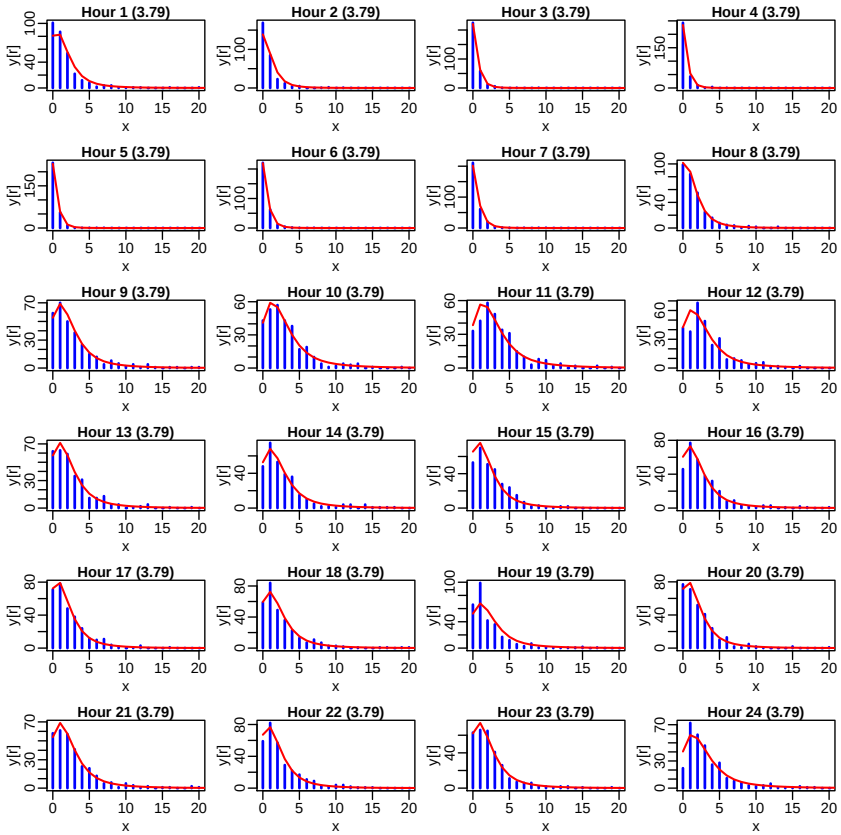

Fig. 5 Distributions of the number of complaints per hour of the day in 1994 (blue bars), and the fit of mixed Poisson distributions with a common latent semi-parametric mixing distribution. The numbers between parentheses in the titles are average values.

$$c_{hij} = \nu_{hj}^{x_i} \exp(-\nu_{hj}) \quad \text{with} \quad \nu_{hj} = \exp(\phi_h + \theta_j). \tag{6}$$

Given the vector ϕ, we get a large PCLM, with a composition matrix that consists of 24 matrices of the type we have already seen, one for each hour, placed on top of each other. The vector y is a concatenation of al 24 hourly distributions. The mixing distributions γ will have the same length as before.

We have to come up with numerical values for ϕ. Let $\bar{x}_h = \sum_i x_i y_{hi} / \sum_i y_{hi}$ be the average number of complaints in hour h. Then we set $\phi_h = \log \bar{x}_h$. It is desirable to have $\sum_i \hat{\mu}_{hi} x_i / \sum \hat{\mu}_{hi} = \bar{x}_h$. It turns out that this is not automatically the case after estimation. The differences are not dramatic, only a few percent, but they are there.

A solution is to update ϕ iteratively, using the ratio of the observed and estimated means per hour. It is not a really fast procedure, but it works. Figure 5 shows results. Visual inspection already tells us that it is an improvement compared to the negative binomial distribution. We now see that peaked distributions can be modeled, which

was not the case before. The change in the deviance is large: it drops from 1350.5 for the negative binomial model to 760.6 for the semi-parametric model.

4 Discussion

A semi-parametric mixing distribution is a powerful tool for modeling over-dispersed data. Thanks to the penalized composite link model [1], estimation is feasible and relatively fast. We can get a good fit to our data even in situations where the negative binomial distribution fails.

The applications I described cover only a small part of the terrain where the PCLM can be useful. Many more opportunities exist. I will show some more examples in my presentation during the conference. Here I'll only present a short description of current research.

In all applications the data were counts in bins, of which the expected values were modeled. The explanatory variable, like the hour of the day for the environmental complaints, could be handled by forming distributions within groups. To exploit explanatory variables in a more general way, individual observations without any grouping should be handled. This can be done by switching to estimating probabilities under a constraint.

The example of the hourly counts of complaints showed the principle of shifting the mixing distribution per group to work well. More generally, we can imagine a common distribution, that is shifted per individual observation, the amount of shifting determined by a (linear) combination of the explanatory variables.

More general than shifting would be to combine shifting and scaling of a common "prototype" mixing distribution.

In this paper I did not touch on inference. But to pick the fruits of the semi-parametric model it will be necessary to develop proper procedures to compute confidence intervals. Also the proper comparisons of models using AIC is not yet described.

A reviewer commented that a comparison to a finite mixture for the latent distribution would be worthwhile. I believe that it will depend on the situation at hand. If the data lead to an estimated bimodal latent distribution, a mixture (of normal distribution) probably will be an attractive alternative, and it will not be too hard to estimate. But fitting a mixture that gives a slightly skewed long-tailed result as in Figure 3 might be hard, because we only observe the latent distribution indirectly.

References

1. Eilers, P.H.C.: III-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, **7**, 239 – 254 (2007).
2. Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. *Journal of the Royal Statistical Society*, A **135**, 370-84 (1972).
3. Thompson, R., Baker, R.J.: Composite Link Functions in Generalized Linear Models. *Applied Statistics*, **30**, 125–131 (1981).