Comput Stat (2010) 25:341–361
DOI 10.1007/s00180-009-0180-x

ORIGINAL PAPER

# Penalized regression with individual deviance effects

**Aris Perperoglou · Paul H. C. Eilers**

Received: 21 April 2009 / Accepted: 13 November 2009 / Published online: 3 December 2009


**Abstract** The present work addresses the problem of model estimation and computations for discrete data when some covariates are modeled smoothly using splines. We propose to introduce and explicitly estimate individual deviance effects (one for each observation), constrained by a ridge penalty. This turns out to be an effective way to absorb model excess variation and detect systematic patterns. Large but very sparse systems of penalized likelihood equations have to be solved. We present fast and compact algorithms for fitting, estimation and computation of the effective dimension. Applications to counts, binomial, and survival data illustrate practical use of this model.

**Keywords** Generalized linear models · Smoothing · Effective dimension · Penalized regression

## 1 Introduction

Generalized linear models (GLM) have made regression and smoothing with counts or binary observations a standard tool of statistics. In contrast to a normal response, the variance follows implicitly from the Poisson or binomial distribution and, given the data, it is completely determined by the estimated expected values. In many applications, however, data show excess variability that is not easily captured by the model.

A. Perperoglou (✉)
Department of Statistics and Actuarial Financial Mathematics, University of the Aegean, 83200 Samos, Greece
e-mail: perperoglou@aegean.gr

A. Perperoglou · P. H. C. Eilers
Department of Biostatistics, Erasmus Medical Center, 3015GE Rotterdam, The Netherlands
e-mail: p.eilers@erasmusmc.nl

This variability is attributed sometimes in overdispersion, when the observed variance is larger than the theoretical one, or in some hidden patterns in the data.

A similar problem occurs in smoothing. When the effective bandwidth is chosen by cross-validation or with an information criterion like AIC (Akaike 1974), it will generally come out too small. Formally this makes sense: optimal cross-validation detects systematic high-frequency components in the data, which should be exploited when predicting left out observations. However, from the subject matter we may know that it is reasonable to assume a smooth trend and we would like to have more or less objective guidance on the amount of smoothing needed to compute it.

There have been several proposals for dealing with overdispersion, the simplest one being correction of the covariance matrix by a constant $\phi$, assuming $\text{var}(y_i) = \phi u_i$ with $\phi$ estimated by equating the Pearson $X^2$ statistic from a binomial fit to its degrees of freedom (Williams 1982), and $u_i$ the theoretical variance under the assumed model. Another way is to assume a parametric form for $\phi$ which will lead to a mixing distribution. For example, in binomial data, the variance of the response probability $\pi_i$ is defined as $\text{var}(\pi_i) = \phi p_i(1 - p_i)$. The variability of $\pi_i$ can also be modeled by a beta distribution with parameters $\alpha_i$ and $b_i$ and $\phi_i = 1/(\alpha_i + b_i + 1)$ which leads to the beta binomial model (Crowder 1978). When data come from a Poisson distribution the mean equals the variance. In such a case, the mean could follow a gamma distribution with mean $\mu$ and variance $\phi\mu$. This mixture leads to the negative binomial model. Efron (1986) introduced the use of double exponential families in generalized linear regression, in which a second parameter is introduced to control the variance. A different approach to deal with overdispersion is to assume a more general form for the variance function using additional parameters. These models are using quasi-likelihood methods for estimation and are described by several authors Hinde and Demetrio (1998), McCullagh and Nelder (1989). For a general discussion on overdispersion refer to Collet (2003, Chap. 6), Agresti (1996) and Morgan (1992).

Overdispersion may also rise as a result of unexplained heterogeneity. To account for this heterogeneity a random effects model can be fitted to the data. Generalized linear mixed models (GLMM) were proposed as a general framework by Breslow and Clayton (1993). They include an unobserved vector of random effects in a GLM, assumed to arise from a normal distribution, and use an approximation of the marginal quasi-likelihood based on Laplace's method, leading to equations based on penalized quasi-likelihood. Lin (1999) extended the idea by using smoothing cubic splines to propose generalized additive mixed models, in the spirit of Hastie and Tibshirani (1990). To avoid the complex numerical integration required to estimate the model, they proposed a double penalized marginal quasi-likelihood also based on a Laplace approximation. Schall (1991) proposed a general algorithm for the estimation of random effects and dispersion parameters applicable in GLMs, regardless of the structure of the linear predictor, and without the need to specify the distribution of the random effect. In his application section, he used random effects to explain extra-binomial variation, however, he did not examine this case in detail. Lee (1996) proposed a broader class of models, in which the random vector is not restricted to be normal, and a hierarchical likelihood to estimate it, without the integration that is

needed in the marginal likelihood techniques; they broadened this class of models in Lee and Nelder (2001). All of the above approaches deal with the problem of overdispersion, depending on different backgrounds of the same problem. However, some of them are computationally hard to apply, especially in large datasets and some other involve complicated mathematical procedures.

The present work addresses the problem of model estimation for discrete data when some covariates are modeled smoothly using splines. We introduce an extra term in the model to capture excess variability. Our approach is based on penalized likelihood, using individual deviance effects as an extra parameter in the linear predictor for each observation. This makes the number of parameters in the model larger than the number of observations. In order to be able to estimate such a large number of parameters, we add a ridge penalty on the deviance effects. This removes collinearity in the estimating equations and at the same time reduces the effective model dimension drastically. To optimize the weight of the penalty, AIC, AICc (Hurvich and Tsai 1989; Hurvich et al. 1998) or REML methods can be used. This setting provides a tool to deal with a range of problems, including hierarchical structures and smoothing.

An important merit of our proposal is simplicity. In contrast to random effects modelling, no assumptions are made for the distribution of the deviance effects, and the ridge penalty provides a way of avoiding integration and complex approximation of a marginal likelihood. We consider individual deviance effects not as a device for absorbing overdispersion; these effects serve as an explanatory tool in complex statistical applications, where other approaches are becoming computationally demanding or theoretically too complicated. As an explanatory tool, deviance effects should be examined, in order to reveal hidden patterns. In a sense, deviance effects can be seen as residuals. In most cases, these effects will reveal possible bias in the model and indicate the source and nature of increased variation or they might indicate whether there is overdispersion present in the data. However, deviance effects are not just residuals, since their inclusion in the model might improve the fit and the behavior of smoothing parameters.

Implementation of individual deviance effects is straightforward, but it leads to large systems of equations. However, they are extremely sparse and structured in such a way that we can use explicit shortcuts. These shortcuts not only improve the speed of computation by orders of magnitude, but (in the case of Poisson regression) also reveal interesting relationships with the negative binomial distribution.

The paper is structured as follows. In Sect.2, we introduce the individual deviance effects for regression and smoothing for counts, binomial data and survival analysis, followed by a section on inference and the the choice of penalty weights. In Sect. 4, we discuss an algorithm for efficient computation. Applications and simulation studies are presented in Sect. 5 and a discussion follows in the last section. Details of the sparse matrix calculations are presented in the Appendix.

As an acronym for our approach we have invented PRIDE: Penalized Regression with Individual Deviance Effects. Note that individual here means unit of observation, like an observed count; it does not mean that a parameter is connected to each individual counted.

## 2 Penalized regression with individual deviance effects

Count data are often encountered in applications. It is natural to assume that numbers of events can be fitted with a Poisson model. This model relates the expected value of $Y$, $\mathrm{E}(Y) = \mu$, to the systematic component $\eta$ by the canonical link, $\log(\mu) = \eta$. Let counts $y_i$, $i = 1, \ldots, m$ be a realization of a Poisson distribution. Then the probability of $y_i$ is given by:

$$p_i = \mu_i^{y_i} e^{-\mu_i} / y_i!$$

and the log-likelihood is proportional to:

$$l = \sum_{i=1}^{m} (y_i \eta_i - \mu_i) \tag{1}$$

Consider the $X_{m \times p}$ matrix of $p$ covariates and the systematic component of the model $\log(\mu) = \eta = X\beta$, with $\beta$ the vector of unknown but estimable coefficients.

The optimization of (1) leads to a system of linear equations which can be solved with iterative weighted linear regression as:

$$(X' \tilde{W} X)\hat{\beta} = X'(y - \tilde{\mu}) + X' \tilde{W} \tilde{\beta}$$

which is equivalent to $(X' \tilde{W} X)\hat{\beta} = X' \tilde{W} \tilde{z}$, where $W$ is a diagonal matrix containing the weights $\mu$ and $\tilde{z} = \tilde{W}^{-1}(y - \tilde{\mu}) + \eta$ and tilde denotes an approximate solution, i.e., the values that are computed at the intermediate steps before final convergence of the iterative algorithm.

To account for potential model bias and randomness, we propose to include a vector of 'deviance' effects $\gamma$ to the systematic component $\eta$ such as:

$$\eta = X\beta + \gamma \tag{2}$$

Equation (2) is the central idea of PRIDE regression. The systematic part of a generalized linear model $X\beta$ is enriched by a vector of effects, one for each observation. Once the covariate information is in the model, we suggest to estimate a set of effects that will describe deviances of the model estimates from the real data. These deviance effects are model parameters that will model variation that is not explained by the covariates and detect sources of potential bias. To maintain identifiability, we subtract a ridge penalty term from the log-likelihood:

$$l^* = \sum_{i=1}^{m} (y_i \eta_i - \mu_i) - \kappa \sum_{i=1}^{m} \gamma_i^2 / 2. \tag{3}$$

A ridge penalty is necessary in (3) to constrain the deviance effects. With the inclusion of the $\gamma$ vector the model becomes overparameterized and some of the parameters will be overestimated. The ridge penalty handles the increased number of parameters that

are controlled by the penalty weight. In the case of a well defined model, in which all the important covariates are included in the model and no systematic bias is present, then all variation of the model will be well explained by the systematic part $X\beta$. Then the deviance effects should be small -or even zero. Inclusion of the ridge penalty will be shrunk towards zero, as the penalty weight gets larger. On the other hand, when the covariates do not fully describe the data, or there are hidden patterns in the data, then the extra variation should be described by the individual effects. In such cases the penalty weight should be small.

Setting the partial derivatives equal to zero gives the following system of penalized equations:

$$X'(y-\mu)=0, \quad y-\mu=\kappa I$$

where $I$ is an identity matrix of the proper dimensions. One then iteratively solves the following system of weighted regression, with $W=\text{diag}(\mu)$:

$$\begin{pmatrix} X'\tilde{W}X & X'\tilde{W} \\ \tilde{W}X & \tilde{W}+\kappa I \end{pmatrix}\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}=\begin{pmatrix} X'\tilde{W}\tilde{z} \\ \tilde{W}\tilde{z} \end{pmatrix}. \tag{4}$$

This is a large but sparse system: its size is equal to the size of $\beta$ plus the number of observations. However, with some matrix algebra we can avoid computational problems. For details see Sect. 4. Moreover, we can eliminate $\gamma$ quite easily:

$$\hat{\gamma}=(\tilde{W}+\kappa I)^{-1}\tilde{W}(\tilde{z}-X\hat{\beta}).$$

If we introduce $W^*=\kappa(\tilde{W}+\kappa I)^{-1}\tilde{W}$, we have $\kappa\hat{\gamma}=W^*(\tilde{z}-X\hat{\beta})$. With this result we can derive, via simplification of

$$(X'\tilde{W}X)\hat{\beta}+X'\tilde{W}\hat{\gamma}=(X'\tilde{W}X)\hat{\beta}+X'\tilde{W}(W^*(\tilde{z}-X\hat{\beta})/\kappa)=X'\tilde{W}\tilde{z},$$

that

$$(X'W^*X)\hat{\beta}=X'W^*\tilde{z}.$$

These are the same equations as for fitting a generalized linear model without overdispersion, with a change of weights and the addition of $\gamma$ to $z$.

A common method of dealing with overdispersion in count data is by a mixture model. The assumption is that the mean of a given individual, say $Z$, arises from a gamma distribution in the population, with $\text{E}(Z)=\mu$ and the variance proportional to the square of its mean. This mixture of Poisson and gamma distributions leads to a negative binomial model, where the mean value of $Y$ is $\text{E}(Y)=\mu$ as in a Poisson, and the variance is $\text{var}(Y)=\mu+\mu^2/\psi$ for some nonnegative constant $\psi$. Note that, for large $\psi$ the model approaches the Poisson model. McCullagh (McCullagh and Nelder, 1989, Chapter 9), describe how to fit such a model via quasi-likelihood, and Thurston et al. (2000) discuss an extension for negative binomial additive models. McCullagh and Nelder write the canonical parameter as $\log(\mu/(\mu+\kappa))$

McCullagh and Nelder (1989, p. 326, Table 9.1) and Thurston et al. (2000) describe an algorithm to fit the model with weights $\kappa\mu/(\mu+\kappa)$. This bears a striking similarity with our approach where the weight matrix $W^*$ can be used, which is given as a diagonal of $w_i^* = \kappa w_i/(w_i+\kappa)$ and $w_i = \mu_i$.

### 2.1 Smoothing with P-splines and PRIDE

Eilers (1996) proposed generalized linear smoothing with penalized B-splines for data pairs $(x_i, y_i)$, with non-normal $y$. The linear predictor is $\eta = B\alpha$, where $B = [b_{ij}]$ is a matrix of B-splines, $b_{ij} = B_j(x_i)$. The log-likelihood is modified by a penalty based on differences of $\alpha$. This model can also be extended with individual deviance effects as before. In the case of Poisson regression this leads to the penalized log-likelihood

$$l^* = \sum_{i=1}^{m}(y_i\eta_i - \mu_i) - \lambda\sum_k(\Delta^d\alpha_k)^2/2 - \kappa\sum_i\gamma_i^2/2. \tag{5}$$

Here $\eta = B\alpha + \gamma$ and $d$, the order of the differences, generally will be 2 or 3. The weighted regression equations are very similar to (4), with $B$ taking the place of $X$ and $X'WX$ replaced by $B'WB + \lambda D'D$, where $D$ is a matrix such that $D\alpha = \Delta^d\alpha$.

### 2.2 Binomial data

The scoring algorithm in (4) applies to a whole class of generalized linear models, as detailed by McCullagh and Nelder (1989). Suppose, we have binomial data $(y_i, t_i)$, where $y$ denotes the number of "successes" and $t$ the number of trials. Let $E(Y_i) = \mu_i = t_i p_i$, the canonical link $p_i = 1/(1+\exp(-\eta_i))$, with $p_i$ the probability of success. The weights are $w_i = t_i p_i(1-p_i)$. Again individual deviance effects can be introduced by setting $\eta = X\beta + \gamma$, in the case of regression, or $\eta = B\alpha + \gamma$, in the case of P-spline smoothing.

### 2.3 Smoothing of life tables

Survival data can come as pre-grouped data, when there is a natural unit of accounting, like years. When individual survival times and censoring status are given, we can follow Efron (1988) and introduce (narrow) time intervals. In each interval the number of subjects at risk is counted, as well as the number of events. The relationship between time and probability of an event can then be estimated with a parametric or semi-parametric model.

Let $r_j$ be the number of people at risk in interval $j$ and let $y_j$ be the number of events in the same interval. Then we can write a generalized linear model for the probability of an event $p_j$ as:

$$\log\left(\frac{p_j}{1-p_j}\right) = \eta_j = B\alpha$$

In practice the probabilities are small and then it will be advantageous to switch to a Poisson model, in which we model the expectation, $\mu_j$, of $y_j$

$$\log \mu_j = \eta_j = B\alpha + \log(r_j)$$

where $\log(r_j)$ is an offset term. Here $B$ is a B-splines basis and a difference penalty is put on $\alpha$.

### 2.4 Optimal penalty weights

A common technique for finding an optimal value of the smoothing parameter $\lambda$ is to combine the deviance and effective degrees of freedom of a fitted model in Akaike Information Criterion (AIC). We have found that AIC served us well in many applications, although AIC has a reputation for under-smoothing, especially in models with large numbers of parameters. Once individual deviance effects are included in models, optimization of AIC generally indicates a relatively small effective dimension (compared to the nominal number of parameters, which includes the deviance effects). The use of corrected AIC does not change results much.

Another approach comes from generalized linear mixed models (GLMM). A general algorithm for the estimation of the fixed and random effects and components of dispersion in GLMMs was proposed by Schall (1991). The proposed algorithm can be adapted here to estimate the optimal values of the penalties. Consider the model in Sect. 2.1 with log-likelihood function given by (5), let $H$ denote the hat matrix and $H_d$ the lower right submatrix of the hat matrix, corresponding to the individual deviance effects. Then the optimal value of the ridge penalty can be computed as:

$$\hat{\kappa} = tr(H_d)/\gamma'\gamma$$

Similarly, the weight of the penalty for the smoothing splines can be given as:

$$\hat{\lambda} = tr(H_s)/\alpha D'_\alpha D_\alpha \alpha$$

with tr($H_s$) the trace of the upper left submatrix of the hat matrix. Throughout this work, we will refer to this approach for computing the optimal weight as Schall's algorithm.

## 3 Efficient computation

The penalized likelihood equations and the iterative solution algorithm lead to large linear equation systems. Unless one tries very small values of $\kappa$, numerical stability problems do not occur, even though the number of equations is larger than the number of observations. The ridge penalty stabilizes the computation, as is borne out by the effective dimension, which turns out to be much smaller than the number of equations.

Solving the system (4) can lead to efficiency problems. If the number of observations becomes larger than, say, 1,000, the demands on memory and computation time

could become a problem, if one would simply store and repeatedly solve the system. However, using our proposed algorithm the computations can become efficient even in very large data sets.

On convergence we also need the inverse of the matrix on the left-hand side of (4), to compute the standard errors. Furthermore we need an additional matrix product to compute the effective dimension. In the Appendix, we describe how to exploit the extreme sparseness of the equations to speed up the computations, without explicitly forming the matrices. Note that we compute the diagonal of the inverse of a sparse matrix; standard sparse matrix software will not work here.

## 4 Applications

### 4.1 Number of faults in fabric rolls

Bissel (1972) reported a data set on the number of faults in rolls of fabric. Assuming that the number of faults is proportional to the length of a roll, Poisson regression on the logarithm of length of roll ($x$) as the explanatory variable should provide a reasonable fit, see Hinde (1982). The estimated intercept is $-4.173$ ($se = 1.135$) and coefficient of $\log(x)$ is 0.997 ($se = 0.176$). The residual deviance of the model is 64.5 with 30 degrees of freedom, indicating overdispersion. A negative binomial model gives $-3.795$ (1.457) for the intercept and 0.938 (0.228) for the coefficient of $\log(x)$. The residual deviance of the negative binomial model was reduced to 30.67 while the dispersion parameter was estimated to be 8.667.

To illustrate the mechanism behind our methodology consider the simple model, where only a constant is added to the model and there is no information available on the length of the fabric rolls. Then the fit will be a straight line (as shown in upper left plot of Fig. 1) with deviance effects corresponding to the distance of each point from the fitted line. The weight of the penalty for that model is 3.981. When the fabric length is included in the model the weight of the penalty becomes 9.549, and the deviance effects are smaller this time (1, middle right plot). However, the model can be further improved by taking the logarithm of the fabric length. The optimum weight of the penalty was $\kappa = 8.912$. With the inclusion of the deviance effects and $\log(x)$ the intercept is estimated as $-3.651(1.436)$ and the coefficient of $\log(x)$ as 0.910 (0.225). These results are very similar to those obtained with the negative binomial model. However, further comparison could be made on the basis of AIC. The simple Poisson model has AIC = 191.84. The negative binomial model has AIC = 181.39 while the PRIDE model has AIC = 179.339.

In the bottom graph the fit has become better, and the deviance effects even smaller. A visual investigation of the deviance effects may indicate which model provides a better fit to the data. A better fit should result in smaller deviance effects.

### 4.2 Simulation studies

In order to assess how the PRIDE models perform in cases that the data arise from a specific theoretical model, a series of simulation studies was performed. We simulated

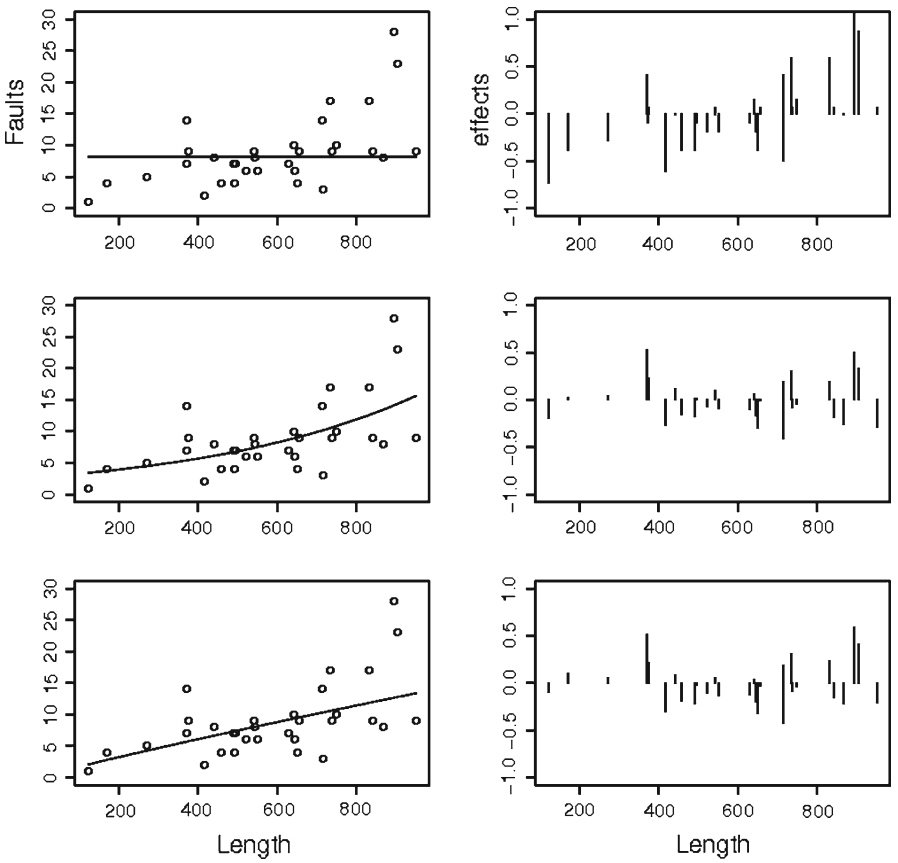


**Fig. 1** Fabric fault data. Results of three models; *upper graph* data and fitted line $\eta = \beta_0 + \gamma$ and a plot of deviance effects, *middle graph* data and fitted line $\eta = \beta_0 + \beta_1 X + \gamma$ and a plot of deviance effects, *bottom graph* data and fitted line $\eta = \beta_0 + \beta_1 \log(X) + \gamma$ and a plot of deviance effects

data coming from a negative binomial model. The framework within which the data were simulated was similar to the example of the fabric data. We simulated 100 counts arising from a negative binomial model, based on an explanatory variable, and variance $\text{var}(Y) = \mu + \mu^2/\psi$ with parameter $\psi$ chosen from the set of different values $\{2, 4, 6, 8, 10, 20\}$. For each different parameter the data were created on the theoretical model with $\mu = 1 + 0.5 \log(x)$ and each setting was repeated a thousand times. Three different models were fitted on the data, a simple Poisson model, a negative binomial and a PRIDE model. The results are presented in Table 1.

As expected, a simple Poisson model does not perform well, especially for small values of the $\psi$ parameter, where it underestimates the standard errors, and the number of cases where the true value of the coefficient was in the interval created from the estimated coefficient plus or minus two times the standard errors, was small. On the other hand, the negative binomial model corrected the standard errors and gave

**Table 1** Results of 1,000 simulations on 100 cases simulated from a negative binomial model with $\mu = 1 + 0.5\log(x)$ where $\beta_0 = 1$ and $\beta_1 = 0.5$ are the true values of coefficients and $\psi$ (the parameter of the negative binomial distribution) was chosen from the set {2, 4, 6, 8, 10, 20}

| | | $\hat{\beta}_0$ | $\hat{\beta}_1$ | $\hat{\psi}$ | Coverage (%) |
|---|---|---|---|---|---|
| $\psi = 2$ | Poisson | 0.974(0.141) | 0.509(0.075) | | 74.4 |
| | neg. bin. | 0.969(0.246) | 0.512(0.137) | 2.12 | 94.8 |
| | PRIDE | 0.789(0.314) | 0.495(0.176) | 1.46 | 94.1 |
| $\psi = 4$ | Poisson | 0.989(0.140) | 0.504(0.074) | | 82.8 |
| | neg.bin | 0.987(0.200) | 0.506(0.110) | 4.38 | 94.9 |
| | PRIDE | 0.933(0.209) | 0.493(0.112) | 4.19 | 95.3 |
| $\psi = 6$ | Poisson | 0.986(0.140) | 0.506(0.074) | | 86.8 |
| | neg.bin. | 0.985(0.183) | 0.506(0.101) | 6.77 | 95.4 |
| | PRIDE | 0.958(0.185) | 0.497(0.102) | 6.78 | 95.5 |
| $\psi = 8$ | Poisson | 0.991(0.140) | 0.501(0.074) | | 87.8 |
| | neg.bin. | 0.991(0.172) | 0.501(0.094) | 9.73 | 94.7 |
| | PRIDE | 0.976(0.173) | 0.494(0.094) | 9.81 | 95.0 |
| $\psi = 10$ | Poisson | 0.987(0.139) | 0.506(0.074) | | 90.7 |
| | neg.bin. | 0.987(0.166) | 0.506(0.090) | 71.28 | 94.9 |
| | PRIDE | 0.976(0.167) | 0.501(0.091) | 113.19 | 94.9 |
| $\psi = 20$ | Poisson | 0.995(0.140) | 0.502(0.074) | | 93.7 |
| | neg.bin. | 0.994(0.154) | 0.503(0.083) | 2174.9 | 95.9 |
| | PRIDE | 0.991(0.154) | 0.500(0.083) | 2951.2 | 96.2 |

Standard errors of the estimated coefficients are given in parentheses. The last column presents the number of times that the true value of the constant lay within the estimated confidence interval of the coefficient (95% nominal coverage)

estimates for the coefficients closer to the real ones. Even though the negative binomial is the true model from which the data rise, the PRIDE model performs very satisfactory and in some cases better (with regard to coverage). The PRIDE model corrects the estimated standard errors, gives better estimates for the coefficients but also estimates the $\psi$ better than the negative binomial model, with the only exception when $\psi = 2$. It has been noted at the last paragraph of Sect. 2, when fitting a PRIDE model to overdispersed count data the weights that are used are the same with the ones suggested by Thurston et al. (2000). Thus, the estimated $\kappa$ from PRIDE models should be similar to $\psi$ parameter estimated from the negative binomial model. In the last row of the table, we also present a case when $\psi = 20$ to simulate a case where the model approximates a Poisson model. As one should expect, when $\psi$ gets larger, the coverage and estimates from a simple Poisson model improve.

### 4.3 Comparison of gynaecological practices

The data arise from a project on quality comparison of gynecological practices in the Netherlands. The study monitors the performance of about 140 centers from 1988 up to recent date, with respect to different aspects of childbirth. In this section, we only

**Table 2** Estimated coefficients for fixed effects and their standard errors (left) from a linear mixed model

Data were fitted using `R` and `lmer4` library. The right part of the table presents estimated coefficients (and standard errors) from a PRIDE model with $\kappa = 1.17$

| | Coef | St.err | Coef | St.err |
|---|---|---|---|---|
| Constant | 12.733 | 1.125 | 11.991 | 1.094 |
| Xweek | −0.317 | 0.042 | −0.300 | 0.041 |
| Xblood | −0.009 | 0.003 | −0.009 | 0.004 |
| Xweigth | −0.002 | 0.000 | −0.002 | 0.000 |
| Xsex | −0.196 | 0.128 | −0.191 | 0.125 |
| Xzek | −0.302 | 0.239 | −0.274 | 0.235 |

consider data from 1998 and concentrate on the mortality of pre-term infants (from 32 up to 37 weeks). The covariates are: weight of the child (Xweight), pregnancy length in weeks (Xweek), gender of the child (Xsex), blood pressure (Xblood) and a binary indicator of whether the mother had some sort of illness before giving birth (Xzek).

In 1998, in 114 centers, 2,212 infants were born prematurely. We only considered cases with full records, leaving a data set of 2,067 births which contained 561 deaths. The mean number of births per center is 18.13 and the overall mortality rate is 27.1%.

First we checked whether an individual deviance effect per child made sense. This was not the case: AIC indicated an essentially infinitely high value of $\kappa$. This is a fundamental issue, since in the binomial case with clusters of size 1 the individual deviance effects are not identifiable, and that forces the penalty to infinity. This is equivalent to what McCullagh and Nelder (1989, page 125) describe, that overdispersion cannot be fit to binomial data with $n_i = 1$.

We introduced deviance effects for the centers, leading to the linear predictor $\eta = X\beta + C\theta$, where $C$ is an indicator matrix connecting a child to a center, and $X$ the matrix of covariates. According to AIC the optimal value of $\log_{10}\kappa$ is 1.17.

To compare the results we fitted the data using a linear mixed effect model (Pinheiro and Bates 2000, p. 146). The variance of the random effect was estimated 1.14. Table 2 presents the estimated coefficients of the mixed model with their standard errors, along with the estimated coefficients (and standard errors) from the PRIDE fit.

It is instructive to compare the results of a simple regression model, which implies fixed center effects, with the results of the PRIDE fit and the mixed model. As Fig. 2 shows, strong shrinking takes place, especially for the more extreme center effects. By visual inspection, the deviance effects appear to shrunk towards zero more than the corresponding random center effects, however, the differences are subtle.

Lack of space does not allow a further analysis of these data. We note, however, that the estimated deviance effects and their standard errors allow the implementation of probabilistic ranking procedures (van Houwelingen et al. 2004; Spiegelhalter 1999; Goldstein and Spiegelhalter 1996; Thomas et al. 1994). We will report on this elsewhere.

### 4.4 Digit preference in demographic data

Age heaping is a common phenomenon in demography, caused by age misstatement in data registration when reliable records are not available. Many people tend to misstate

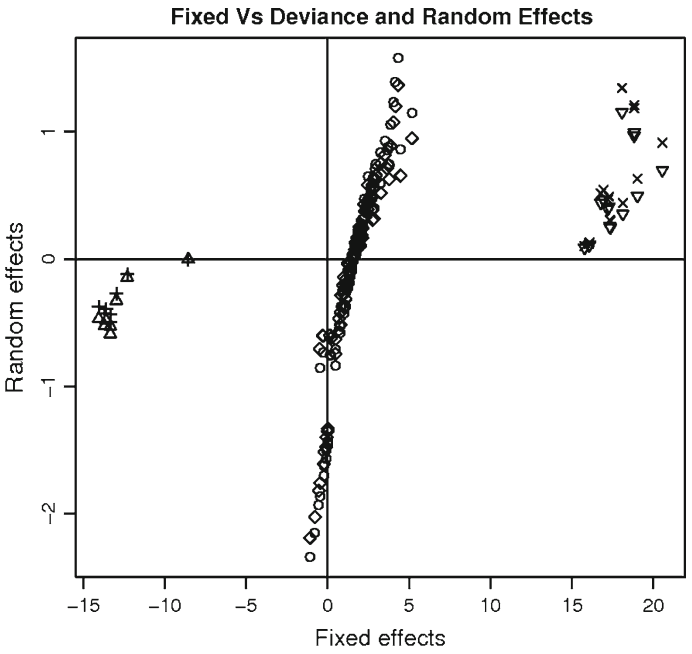


**Fig. 2** Random center effects versus fixed center effects for the gynecological practices data. *Triangle* centers with death rate lower than 0.08, estimated by mixed model. *Times* centers with death rate upper than 0.85, estimated by mixed model. *Circ* centers with death rates in between 0.08 and 0.85, estimated by mixed model. *Doplus* centers with death rate lower than 0.08, estimated by PRIDE model. *Bigtriangledown* centers with death rate upper than 0.85, estimated by PRIDE model. *Diamondsuit* centers with death rates in between 0.08 and 0.85, estimated by PRIDE model

their age (or their year of birth) in favor of numbers ending in multiples of five. To illustrate this, Fig. 3 shows empirical data of the observed deaths of the male Greek population in 1960. The raw data are presented in the upper right histogram (as vertical narrow bars). For ages over 45, we observe large heaps every 5 years.

The Poisson smoother was constructed as follows. Define $y_i$ the number of death at age $i$, and $E(y_i) = \mu_i$, then the model is $\eta = \log(\mu) = B\alpha$ where $B$ is a B-spline bases. The size of $y$ is small and intervals have equal widths, so if we evaluate a zero-degree B-spline basis $B$ at midpoints we get the identity matrix $I$. A difference penalty $\lambda|D\alpha|^2$ on $\alpha$ controls the amount of smoothness. The upper left graph shows the graph of AIC, indicating a small value of $\lambda$, leading to the quite rough line in the upper right graph, which essentially follows the data. A first indication that the problem stems from the counts at ages that are multiples of five, can be seen in the lower right graph. The counts at multiples of five have been replaced by the average of the preceding and the following age. The optimal smooth curve already looks better, but it still shows spurious detail.

This phenomenon, also known as digit preference, can lead to complicated and misleading patterns. Eilers (2004) describe systematic ways of dealing with the problem, accounting for transfers of counts from "unpopular" to "popular" ending digits. Here we take the simple route of adding a deviance effect: $\log \mu = \eta + \gamma$.

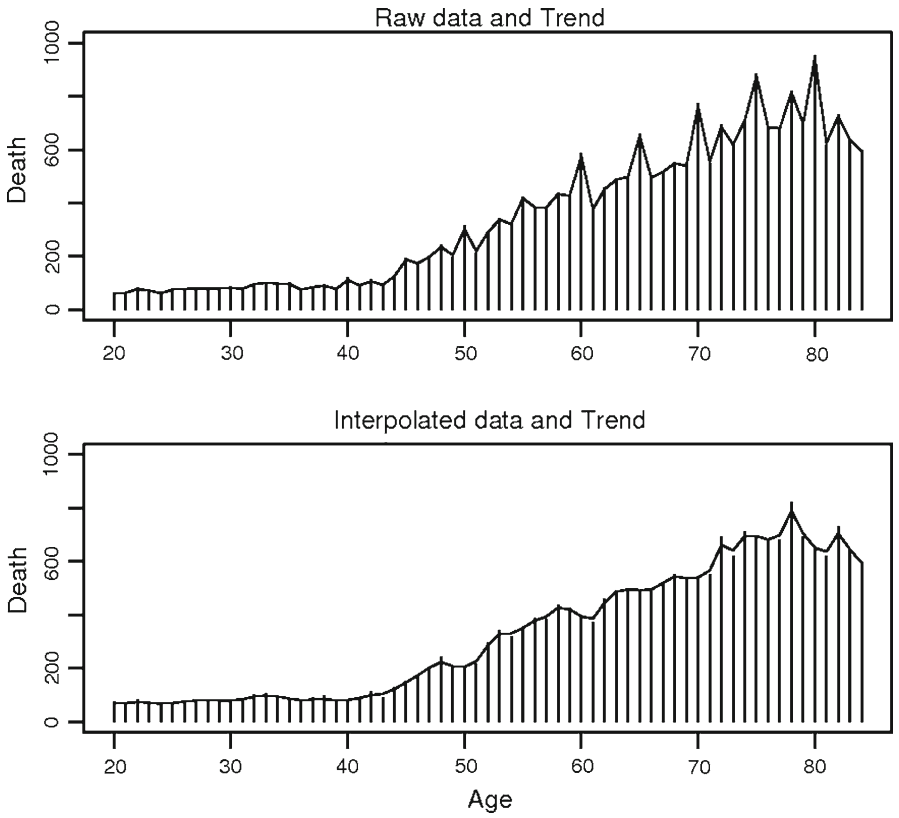


**Fig. 3** Number of deaths versus Age of the Greek male population in 1960. Optimal weight of the penalty for raw data $\lambda = 3.98$ (*upper graph*), and for interpolated (*lower graph*) $\lambda = 50.11$, based on AIC criterion

We fitted the data using both the AIC and Schall's algorithm to compute the weight of the penalty. Results are presented in Fig. 4. A contour plot illustrates the dependence of AIC on $\lambda$ and $\kappa$. The best choice is $\log_{10} \lambda = 3.4$ and $\log_{10} \kappa = 1.8$, based on a two dimensional grid search. The profile plots show the behavior of AIC for optimal values of the parameters. Following the AIC indicated weight the smoothed histogram now looks much more realistic. On the other hand, the smoother from Schalls algorithm was still influenced by the digit preference. The pattern of the deviance effects emphasizes digit preference: large positive values at multiples of five flanked by negative values.

Another approach of modelling the data is by the use of Generalized Additive Models (GAMs) as suggested by Wood (2008). We used the library **mgcv** in **R** to fit the data. When specifying the model, the degrees of freedom for the smoothing spline has to be given as well. In Fig. 5 we present the GAM fits for a different number of degrees of freedom. The GAM model follows the digit preference when the number of knots is larger than 40 knots. In all cases, the PRIDE fit is added to the graphs for comparisons. In conclusion, the PRIDE fit either outperforms the GAM fit or, when the number of knots is 'correctly' specified, it performs equally well.

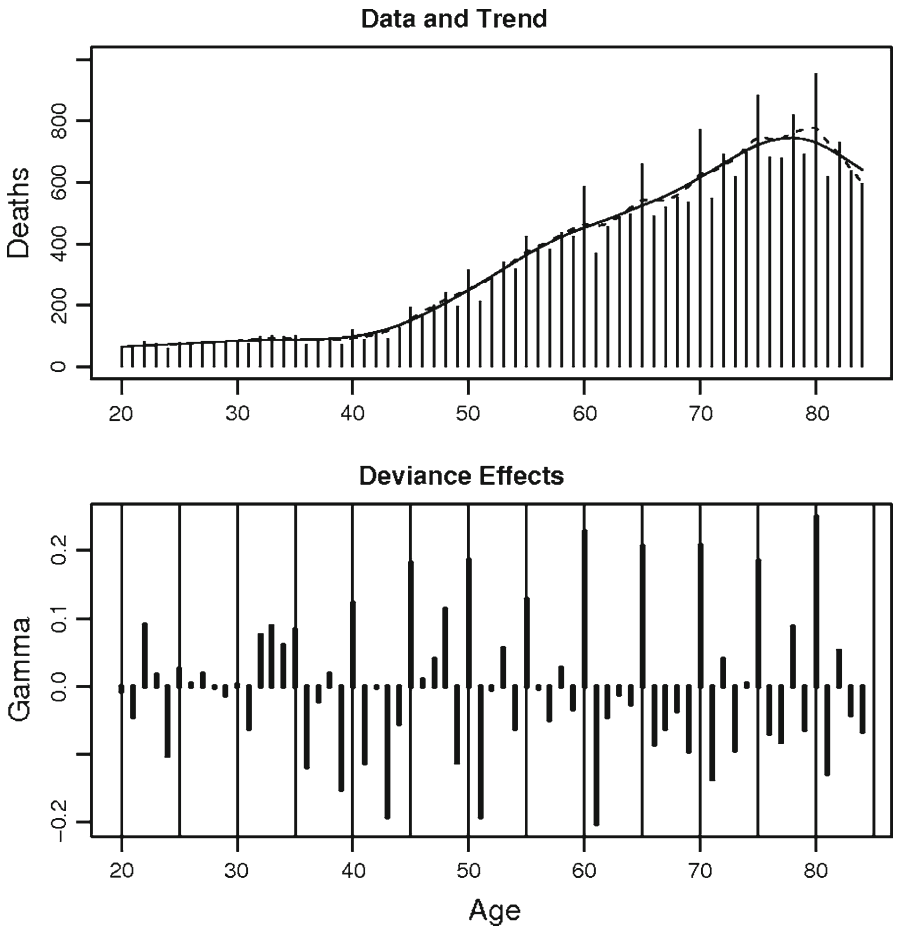


**Fig. 4** **a** Histogram of empirical data and smoother, *solid line* AIC smoother, *dashed line* smoother based on Schalls algorithm, **b** Values of individual deviance effects $\gamma$ for different ages

## 4.5 Survival of Mediterranean flies

We will now extend the idea of adjusting for overdispersion in Poisson counts to the field of survival analysis. As an example consider data which consist of lifetables for 46 cohorts of female Mediterranean flies (Ceratitis capitata). Each cohort consisted of about 4,000 flies which were put in a cage and for each cage, the number of flies alive at the beginning of each day was recorded. The flies were observed for up to 174 days in some cohorts, and the number of deaths for each cohort was recorded at the end of each day. For a detailed analysis of the data see Müller et al. (1997). We restrict our analysis in two cohorts from the study chosen at random.

The model is essentially the same as for the age distribution that we discussed before. The response is the number of flies dying per day. The number at risk, $r$, is introduced as an offset $E(y) = B\alpha + \log(r) + \gamma$. We used both AIC and Schall's algorithm to determine the optimal value of the penalty weights. Figure 6 (right) shows

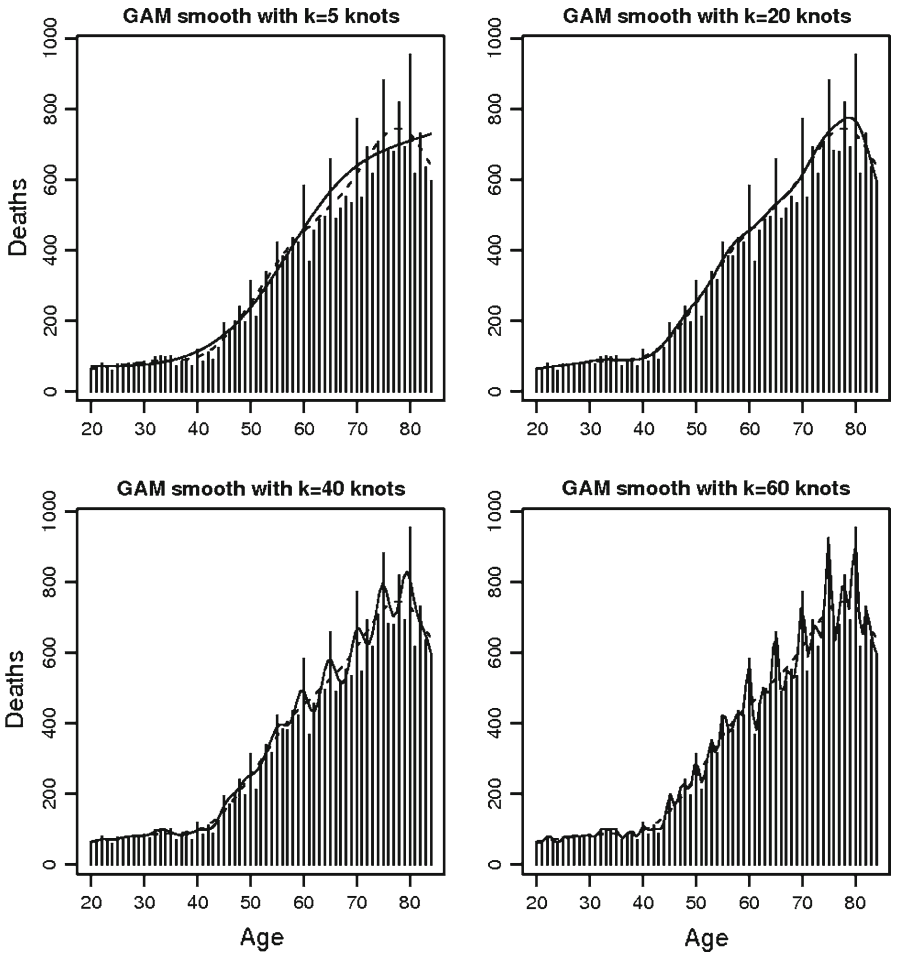


**Fig. 5** Histogram of empirical data and smoother for different number of knots in a GAM fit. *Dashed line* denotes the smoother coming from a PRIDE fit

an example where the size of the deviance effects are small, as is also indicated by the large value of $\kappa$ (251, determined by AIC), while the difference amongst the fit using AIC and Schall's algorithm are only visible in the last few days of the follow up. In cohort 2 one can see quite large deviance effects ($\kappa = 39.81$) with an absolute value up to 0.45. Apparently, there is clustering in dying (and not dying) of the medflies. This means that on certain episodes the hazard increases or decreases by a factor of almost $1.6(\exp(0.45) = 1.57)$.

## 5 Discussion

We have introduced a simple device, individual deviance effects, to model overdispersion, account for model bias and randomness in generalized linear regression and

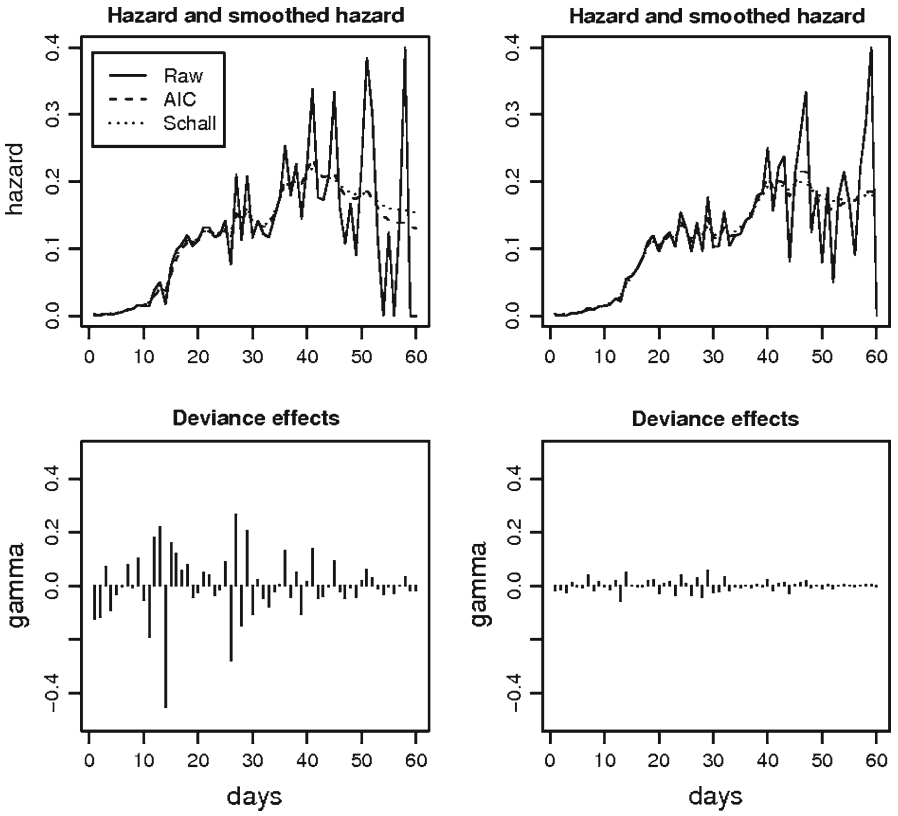


**Fig. 6** Hazard and smoothed hazard of flies in cohort 2 (*left side*) and 5 (*right side*), along with histograms of the corresponding deviance effects

smoothing. Although the nominal number of parameters is increased enormously this way, a ridge penalty makes all parameters identifiable, reduces the effective model dimension, and stabilizes computations. A very large system of estimating equations results from our model, but it is extremely sparse and we have shown how to solve it efficiently, deriving explicit formulas for components of partitioned matrices.

We have considered a number of simple, but realistic, applications. We have shown that PRIDE models can work as an approximation of the negative binomial distribution, in the example of the fabric data. Experiments in large life tables (over 100 years, 100 ages) have also shown good results (Iain Currie, personal communication). Fitting the model is no more complicated than for the Poisson model, because only the effective weights change. On convergence, the fast algorithm we describe in the Appendix allows efficient computation of effective dimension and standard errors of the fitted values.

For larger problems one would run into problems, unless one uses very smart generalized linear mixed model software. Our approach has been used on life tables with 100 by 100 cells. The sparse algorithm keeps memory use and computation time small. Most standard software will not be able to handle 104 random effects. The

computational approach used in this paper provides a useful tool in a wide area of applications. A nice application of our algorithm can be found in Eilers et al. (2008).

We propose individual deviance effects mainly as an exploratory tool. After fitting, one should study plots of its elements, to detect local patterns their size and direction. This might suggest patterns in the data that can be caught by modified models. Successful modification should lead to a stronger weight of the penalty, with correspondingly smaller deviance effects.

The improvements of estimated standard errors are obtained at relatively low computational costs. One could set up full-scale (generalized linear) mixed model machinery, specify a distribution for and use any of the established algorithms to estimate its variance. The deviance effects will then, of course, become bona fide random effects. Our $\kappa$ is the inverse of their variance. For exploration little would be gained, and changes in estimated standard errors will be small too.

When presenting this work to colleagues, we sometimes experienced that the adjective "individual" in PRIDE caused confusion, especially in the context of counts or proportions. We emphasize that it does not point to the subjects (faults, flies or men) that make up the counts, but the individual observational units (fabric rolls, days or ages intervals) to which the counts are connected. In other words: the individual rows in the regression model $\eta = B\alpha$ for the linear predictor.

One of the referees suggested to produce standard errors for the individual effects in the fabric faults example in order to demonstrate whether these are due to randomness in the data and use a hypothesis test to check. Le Cessie and van Houwelingen (1995) specified a score test for testing the fit of a regression model with random effects, without the need of specifying a distribution for the random vector. Consider the regression model with individual deviance effects and assume that the vector $\gamma$ is a random effects vector with mean 0 and covariance $\sigma^2 R$, with matrix $R$ to describe the dependency structure amongst the random effects. To test whether a simple model without the random effect is adequate we use the score statistic for testing $H_1{:}\sigma^2 = 0$ versus $H_A{:}\sigma^2 > 0$. The derivation is based on the quadratic form $Q = (Y-\mu_1)'R(Y-\mu_1)$, which leads to the goodness of fit statistic:

$$T = \frac{(Y-\mu_1)'R(Y-\mu_1) - \text{trace}(RV)}{\left[\sum_i R_{ii}^2(\mu_{4i} - 3\mu_{2i}^2) + 2\text{trace}(RVRV)\right]^{1/2}} = \frac{Q - E(Q)}{s.e.(Q)} \tag{6}$$

with $\mu_{ji}$ the $j$th central moment of $Y_i$ and $V = \text{cov}(Y)$. The distribution of $Q$ can be approximated with a scaled chi-squared distribution, $c\chi_\nu^2$ where $\chi^2$ is a chi-square distribution with $\nu$ degrees of freedom, and the constants $c$ and $\nu$ are obtained by equating the mean and variance of $Q$ and $c\chi_\nu^2$, yielding $c = \text{var}(Q)/[2E(Q)]$ and $\nu = 2[EQ]^2/\text{var}[Q]$.

This statistic is valid in an ideal world where the true value of the parameters is known. In reality one has to adjust the test for the estimation of the parameters. It can be verified that $(Y - \hat{\mu})$ equals to first order $Y - \hat{\mu}_1 = (I-H)(Y-\mu_1)$ with $H$ the hat matrix, leading to the statistic

$$\hat{Q} = (Y-\hat{\mu}_1)'R(Y-\mu_1) \approx (Y-\mu_1)'(I-H)'R(I-H)(Y-\mu_1)$$

One then has to adjust for the estimation of the parameters by using $(I - \hat{H})' R (I - \hat{H})$ instead of $R$ to compute the mean and standard error of $\hat{Q}$. This simple test can be used for testing whether the deviance effects should be included in the model. In the fabric faults example the score test gave a scaled Chi-square statistic with a value of 66.46, with 25.66 degrees of freedom and scale parameter $c = 10.39$, giving a $p$-value less than $<0.0001$.

Although our approach shows similarities with mixed modelling we stress that PRIDE models do not estimate random effects. In contrast to the quasi-likelihood approach, we prefer the appropriate exponential family distribution, like Poisson or binomial. Established information criteria, like AIC, corrected AIC or BIC can be computed, because the proper likelihood is available. Of course our proposed methodology could be translated to mixed model methodology, and use for instance REML methods to estimate the variance of the deviance effect.

Mixed models treat the random effects as parameters and require modelling and distributional assumptions for their estimate. These assumptions are part of the overall modelling of the data, and as such, they should be checked whether they hold or not. In our approach, we have to deal with a penalty which is chosen for modelling convenience and it is not open to the usual model criticism using tests on the significance of the random effects.

The estimating strategy of PRIDE models can be closely related to penalized quasi likelihood (PQL). In fact the penalized likelihood defined in (1) is actually an extended likelihood (Pawitan 2001, p 429) and can be written in a more general form as:

$$L(\theta, y) = p_\theta(x|y)\, p_\theta(y)$$

where $p_\theta(x|y)$ is the pure likelihood term and $p_\theta(y)$ is the information that $y$ is random. In our penalized likelihood, the penalty term is equivalent to $p_\theta(y)$ and is derived by assuming normality for the deviance effects. This likelihood is essentially the same as the $h-$likelihood, defined by Lee and Nelder (1996), while in smoothing literature it is known as *quasi-likelihood* (Green and Silverman 1993). However, Lee and Nelder chose to estimate the variance of the random effect using restricted maximum likelihood estimates (REML) whereas we can also use AIC for optimizing a penalty which is related to deviance effects. Moreover, Lee and Nelder defined their likelihood to work in a special class of conjugate hierarchical models where the distribution of the random effect is conjugate to the conditional distribution of $y$ given that random effect. In our approach, although the likelihood is like being derived on the assumption of normality of the deviance effects, in practice normality need not to hold and no distributional assumptions have to be met.

The proposed methodology could easily be extended to handle hierarchical structures. Whatever the linear component of the model would be, individual parameter vectors could be added to account for overdispersion due to different causes. Such an extended model would involve multiple ridge penalties, one for each set of deviance effects. Methods for extending the proposed methodology on correlated and multivariate deviance effects can be derived, as well interactions of the fixed with the deviance effects, and is currently a topic of research.

One can look at PRIDE as taking conditional modelling to the limit. The analysis of the fabric fault data illustrates this. We get essentially the same results as from a negative binomial (NB) fit, which is a marginal model, without the complications of the NB likelihood. There the deviance effects showed no obvious pattern. We could have used NB smoothing for the Greek mortality data and perhaps we would have found a pleasing trend. However, we could only look at residual plots and we would not have isolated the digit preference pattern that the deviance effects represent.

For the time, we suggest PRIDE modelling as an explanatory tool for applied statisticians. However, we have pointed the reader to the similarities of our approach with existing literature, mainly in the area of overdispersion modelling. There is still work to be done in order to provide the theoretical basis of our approach on the basis of inference. On small simulation studies the results suggest that PRIDE modelling provide better estimates for standard errors than simple GLMs or joint–likelihood models. However, a more detailed study is needed to illustrate whether PRIDE models are actually better in estimating standard errors than using conventional quasi-likelihood techniques, or whether they produce less bias.

**Acknowledgments** The authors would like to thank the two referees for their helpful suggestions and comments on the original manuscript.



## Appendix: Efficient computation

Consider a PRIDE model with systematic component $\eta = B\alpha + \gamma$ where $B$ is the basis matrix, $\alpha$ the corresponding coefficients, a penalty $\alpha' P\alpha$ and individual deviance effects $\gamma$. We have to invert a partitioned information matrix:

$$\begin{bmatrix} B'WB + P & B'W \\ WB & W + \kappa I \end{bmatrix} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

It follows that

$$S_{11} = \left[ (B'WB + P) - B'W(W + kI)^{-1}WB \right]^{-1} = (B'W^*B + P)^{-1},$$

with $W^*$ a diagonal matrix having $w^*_{ii} = \kappa w_{ii}/(\kappa + w_{ii})$. This is a small matrix with size equal to the number of basis functions. We also have:

$$S_{22} = \left[ (W + \kappa I) - WBS_{11}B'W \right]^{-1} = (W + \kappa I)^{-1} + (W^*/\kappa)BS_{11}B'(W^*/\kappa),$$

where we have used the Morrison–Woodbury matrix inversion lemma:

$$(A + PQR)^{-1} = A^{-1} - A^{-1}P(P'A^{-1}R + Q^{-1})^{-1}RA^{-1}$$

The off-diagonal submatrices follow directly:

$$S_{21} = S'_{12} = -(W^*/\kappa)BS_{11}.$$

For the estimation of the effective dimension of the model the trace of the hat matrix is needed. That means multiplying the inverse of the information matrix, with the information matrix without the penalties as given by:

$$H = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} B'WB & B'W \\ WB & W \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

Working in the same way as before:

$$H_{11} = S_{11}B'WB + S_{12}WB = S_{11}(B'WB - B'W^*WB) = S_{11}B'W^*B.$$
$$H_{22} = S_{21}B'W + S_{22}W = (W^*/\kappa) - (W^*/\kappa)BS_{11}B'W^*$$

In the practical implementation one should handle large diagonal matrices as vectors. Pre-multiplication, as in $WB$, with such a matrix should be implemented as scaling of the rows of $B$ by the corresponding elements of the vector $w$ that forms the diagonal of $W$. The code fragment below, for *R*or *S*+, uses these devices.

```
v <- kappa * w * (Fm <- 1/(w+kappa))
G1 <- rep(v, ncol(X)) * X
S11 <- solve(t(X) %*% G1 + P)
G2 <- rep((v / kappa), ncol(X)) * X
G3 <- G2 %*% S11
L1 <- rowSums(G3*G2)
S22 <- Fm + L1
R11 <- S11 %*% t(X) %*% G1
R22 <- (v / kappa)- L1 * kappa
tr <- sum(diag(R11)) + sum(R22)
```

## References

Agresti A (1996) An introduction to categorical data analysis. Wiley, New York

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control AC-19:716–723

Bissell AF (1972) A negative binomial model with varying elements sizes. Biometrika 59:435–441

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88(421):9–25

Collet D (2003) Modeling binary data. Chapman and Hall/CRC, London

Crowder M (1978) Beta-binomial ANOVA for proportions. Appl Stat 27:34–37

Efron B (1986) Double exponential families and their use in generalized linear regression. J Am Stat Assoc 81:709–721

Efron B (1988) Logistic regression, survival analysis, and the Kaplan Meier curve. J Am Stat Assoc 83(402):414–425

Eilers P, Gampe J, Marx B, Rau R (2008) Modulation models for seasonal time series and incidence tables. Stat Med 27:3430–3441

Eilers PHC, Borgdorff MW (2004) Modeling and correction of digit preference in tuberculin surveys. Int J Tuberc Lung Dis 8(2):232–239
Eilers PHC, Marx BD (1996) Flexible smoothing with b-splines and penalties. Stat Sci 11(2):89–121
Goldstein H, Spiegelhalter DJ (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc A 156:385–409
Green P, Silverman B (1993) Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman and Hall, London
Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrica 76:297–307
Hurvich CM, Simonof JS, Tsai CL (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J R Stat Soc B 60:271–293
Hastie T, Tibshirani R (1990) Generalized additive models. Chapman and Hall, London
Hinde J, Demetrio CGB (1998) Overdispersion: models and estimation. Comput Stat Data Anal 27:151–170
Hinde JP (1982) Compound Poisson regression models. In: Gilchrist R (ed) GLIM82. Springer, New York pp 109–121
Lee Y, Nelder JA (1996) Hierarhical generalized linear models. J R Stat Soc B 58(4):619–678
Lee Y, Nelder JA (2001) Hierarhical generalized linear models: a synthesis of generalized linear models, random effects models and structured dispersions. Biometrika 88(4):987–1006
le Cessie S, van Houwelingen HC (1995) Testing the fit of a regression model via score tests in random effects models. Biometrics 51:600–614
Lin X, Zhang D (1999) Inference in generalized additive mixed models by using smoothing splines. J R Stat Soc B 61(2):381–400
McCullagh P, Nelder JA (1989) Generalized linear models. Chapman and Hall, London
Morgan BJT (1992) Analysis of quantal response data. Chapman and Hall, London
Müller H-G, Wang J-L, Capra WB (1997) From lifetables to hazard rates: the transformation approach. Biometrika 84(4):881–892
Pawitan Y (2001) In all likelihood: statistical modellind and inference using likelihood. Oxford Science Publications, Oxford
Pinheiro J, Bates D (2000) Mixed effects models in S and S-plus. Springer, New York
Schall R (1991) Estimation in generalized linear models with random effects. Biometrika 78(4):719–727
Spiegelhalter DJ (1999) Surgical audit: statistical lessons from nightingale and codman. J R Stat Soc A 162:45–58
Thomas N, Longford NT, Rolph JE (1994) Empirical bayes methods for estimating hospital-specific mortality-rates. Stat Med 13:889–903
Thurston SW, Wand MP, Wiencke JK (2000) Negative binomial additive models. Biometrics 56:139–144
van Houwelingen HC, Brand R, Louis TA (2004) Empirical bayes methods for monitoring health care quality. Technical report, Department of Medical Statistics, LUMC
Williams DA (1982) Extra binomial variation in logistic linear models. Appl Stat 31:144–148
Wood S (2008) Fast stable direct fitting and smoothness selection for generalized additive models. J R Stat Soc B 70:495–518