Chapter 6

Overdispersion Modelling with Individual Deviance Effects and Penalized Likelihood

Abstract

Overdispersion is common when modelling discrete data like counts or fractions. 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 overdispersion, to get correct standard errors and to 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 of standard errors and computation of the effective dimension. Applications to counts, binomial, and survival data illustrate practical use of this model.

6.1 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. Standard errors are commonly computed based on this theoretical variance. Unfortunately, real data often show overdispersion: the observed variance is (much) larger than the theoretical one. Consequently GLM standard errors will be too small and the significance of effects will be overestimated.

A similar problem occurs in smoothing. When the effective bandwidth is chosen by cross-validation or with an information criterion like AIC [6], 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 ϕ, assuming $\text{var}(y_i) = \phi u_i$ with ϕ estimated by equating the Pearson X^2 statistic from a binomial fit to its degrees of freedom [105], and u_i the theoretical variance under the assumed model. Another way is to assume a parametric form for ϕ which will lead to a mixing distribution. For example, in binomial data, the variance of the response probability π_i is defined as $\text{var}(\pi_i) = \phi p_i(1 - p_i)$. The variability of π_i can also be modelled by a beta distribution with parameters α_i and b_i and $\phi_i = 1/(\alpha_i + b_i + 1)$ which leads to the beta binomial model [26]. 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 μ and variance $\phi\mu$. This mixture leads to the negative binomial model. 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 [46, 68]. For a general discussion on overdispersion refer to Collet [22] Chapter 6, [5] and [69]).

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 [17]. 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 and Zhang [63] extended the idea by using smoothing cubic splines to propose generalized additive mixed models, in the spirit of [42]. 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 [84] 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 and Nelder [60] 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 [61]. 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 overdispersion both in GLMs and GAMs. 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. To maintain identifiability, 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 or AICc [87] 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 only as a device for absorbing overdispersion; we emphasize that it is worthwhile to study their pattern. In most cases these effects will reveal possible bias in the model and indicate the source and nature of increased dispersion. Inference will also improve because standard errors will be more realistic.

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 Chapter is structured as follows. In Section 6.2 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 Section 6.4 we discuss an algorithm for efficient computation. Applications and simulation studies are presented in Section 6.5 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 unit. Software, for S-PLUS and R, is available from the first author.

6.2 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 η by the canonical link, $\log(\mu) = \eta$. Let counts y_i, $i = 1, ..., m$ 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:

$$l = \sum_{i=1}^{m} (y_i \eta_i - \mu_i) \tag{6.1}$$

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

The optimization of (6.1) 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 μ and $z = W^{-1}(y - \mu) + \eta$ and the tilde denotes an approximate solution.

To account for potential model bias and randomness, we propose to include a vector of 'deviance' effects γ: $\eta = X\beta + \gamma$. To maintain identifiability, we subtract a ridge penalty term from the log-likelihood:

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

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

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

One then iteratively solves the following system of weighted regression, with $W = \text{diag}(\mu)$:

$$\begin{pmatrix} X'\tilde{W}X & X'\tilde{W} \\ \tilde{W}X & \tilde{W}+\kappa I \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} X'\tilde{W}\tilde{z} \\ \tilde{W}\tilde{z} \end{pmatrix} \tag{6.2}$$

This is a large but sparse system: its size is equal to the size of β plus the number of observations. However, with some matrix algebra we can avoid computational problems. For details see Section 6.4. Moreover, we can eliminate γ 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 γ 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 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 E$(Y) = \mu$ as in a Poisson, and the variance is var$(Y) = \mu + \mu^2/\psi$ for some constant ψ. Note that, for large ψ the model approaches the Poisson model. McCullagh ([68], Chapter 9), describe how to fit such a model via quasi-likelihood, and [94] discuss an extension for negative binomial additive models. McCullagh and Nelder write the canonical parameter as $\log(\mu/(\mu+\kappa))$ ([68], page 326, table 9.1) and Thurston et al. [94] 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 $w_i = \mu_i$.

Smoothing with P-splines and PRIDE

Eilers and Marx [31] 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 α. 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

$$l^* = \sum_{i=1}^{m}(y_i\eta_i - \mu_i) - \lambda \sum_k (\Delta^d \alpha_k)^2/2 - \kappa \sum_i \gamma_i^2/2. \qquad (6.3)$$

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 (6.2), with B taking the place of X and $X'WX$ replaced by $B'WB + \lambda D'D$, where D is a matrix such that $D\alpha = \Delta^d\alpha$.

Binomial data

The scoring algorithm in (6.2) applies to a whole class of generalized linear models, as detailed by McCullagh and Nelder [68]. 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.

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 [28] 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:

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

In practice the probabilities are small and then it will be advantageous to switch to a Poisson model, in which we model the expectation, μ_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 α.

6.3 Inference

We propose individual deviance effects mainly as an exploratory tool. After fitting γ, 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. We could call this inference in a wide sense.

To appreciate the usefulness of PRIDE for inference in a narrow sense, we can study bias and standard errors in a simulation setting. We do that on a limited scale in the applications Section, for Poisson regression. The main advantage of PRIDE is that is leads to more realistic standard error estimates, which generally will be (much) more conservative than those obtained under a model that neglects overdispersion.

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 γ and use any of the established algorithms to estimate its variance. The deviance effects will then, of course, become bona fide random effects. Our κ is the inverse of their variance. For exploration little would be gained, and changes in estimated standard errors will be small too.

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 10^4 random effects.

Optimal penalty weights

A common technique for finding an optimal value of the smoothing parameter λ 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 [84]. The proposed algorithm can be adapted here to estimate the optimal values of the penalties. Consider the model in Section 6.2.1 with log-likelihood function given by (6.3), 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 Chapter, we will refer to this approach for computing the optimal weight as Schall's algorithm.

6.4 Efficient computation

The penalized likelihood equations and the iterative solution algorithm lead to large linear equation systems. Unless one tries very small values of κ, 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 (6.2) can lead to efficiency problems. If the number of observations becomes larger than, say, 1000, 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 of 10000 cases or more.

On convergence we also need the inverse of the matrix on the left-hand side of equation (6.2), 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.

If one is willing to accept an approximate solution, say in an exploratory phase of research, a very simple fast algorithm is available for regression problems. One first estimates β by a standard GLM and keeps it fixed. Then only a diagonal system of equations for the deviance effects remains, which is trivial to handle.

6.5 Applications

Number of faults in fabric rolls

Bissell [14] 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 [47]. The estimated intercept is -4.17 (se = 1.14) and coefficient of $\log(x)$ is 0.99 (se = 0.17). The deviance is 64.5 with 30 degrees of freedom, indicating overdispersion. A negative binomial model gives -3.79 (1.42) for the intercept and 0.937 (0.225) for the coefficient of $\log(x)$.

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 figure 6.1) 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 (6.1, 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 $\kappa = 8.709$. With the inclusion of the deviance effects and log(x) the intercept is estimated as -3.647 (1.442) and the coefficient of $\log(x)$ as 0.909 (0.225). These results are very similar to those obtained with the negative binomial model. In the bottom graph the fit has become better, and the deviance effects even smaller. Some of the bias of the previous models has been eliminated and what is left described in the deviance effects plot is due to randomness in the data.

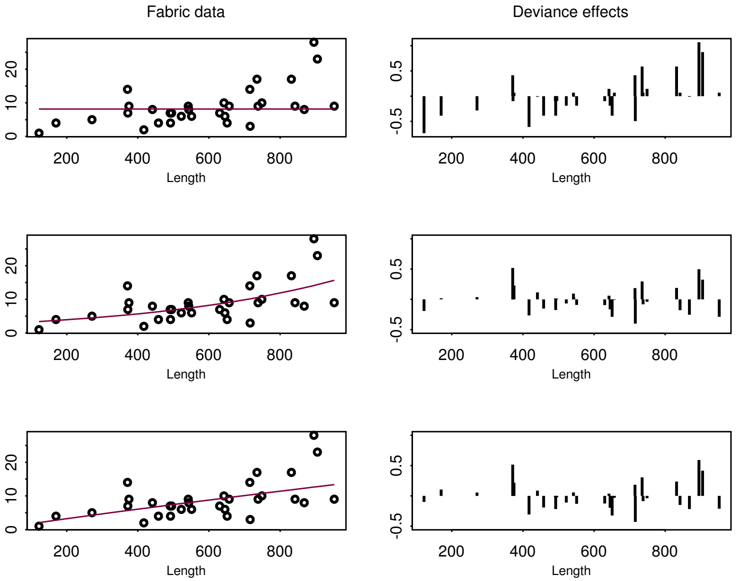

Figure 6.1: Fabric fault data. Results of three models; upper graph: data and fitted line $\eta = \beta_0 + \gamma$ and a plot of deviance effects, middle graph: data and fitted line $\eta = \beta_0 + \beta_1 X + \gamma$ and a plot of deviance effects, bottom graph: data and fitted line $\eta = \beta_0 + \beta_1 \log(X) + \gamma$ and a plot of deviance effects.

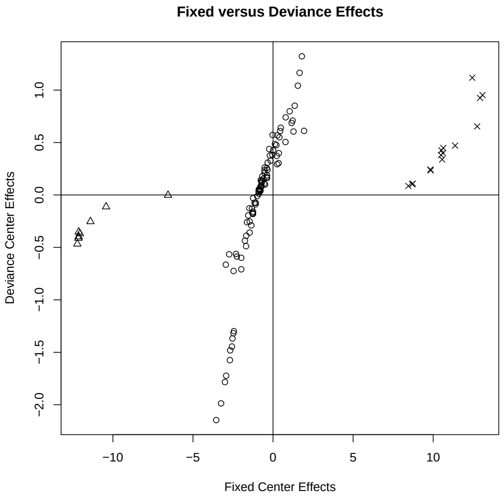

Figure 6.2: Deviance versus fixed center effects for the gynecological practises data. (△) represent centers with rate of death less than 0.08 and (×) centers with death rate more than 0.85.

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 Chapter we only 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, pregnancy length, sex, blood pressure and a binary indicator of data quality.

In 1998, in 114 centers, 2212 infants were born prematurely. We only considered cases with full records, leaving a data set of 2067 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 κ. 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 infinitive.

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 1.25.

It is instructive to compare $\hat{\theta}$ when $\kappa = 0$, implying fixed center effects, with the results of PRIDE. As Figure 2 shows, strong shrinking takes place, especially for the more extreme center effects. 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([98, 88, 36, 93]). We will report on this elsewhere.

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 their age (or their year of birth) in favor of numbers ending in multiples of five. To illustrate this, Figure 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 five 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 α controls the amount of smoothness. The upper left graph shows the graph of AIC, indicating a small value of λ, 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 and Borgdorff [30] 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: $\log\mu = \eta + \gamma$.

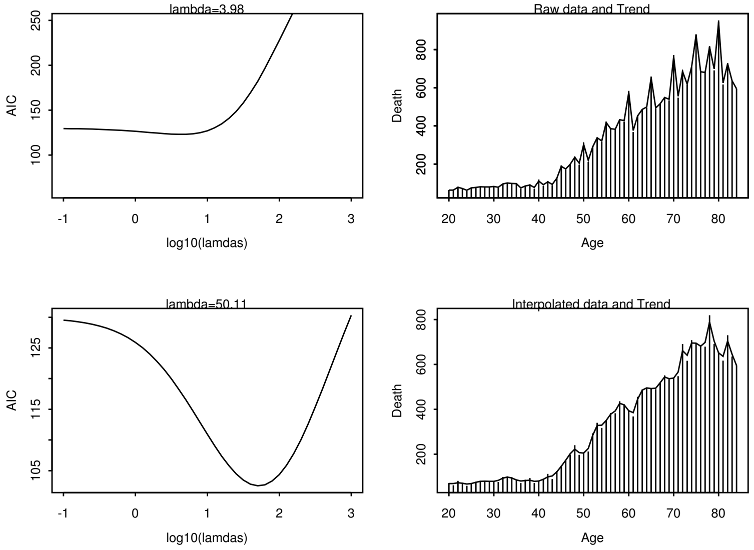

Figure 6.3: Graphs of optimal AIC, and smoothed data of the Greek male population in 1960, for the raw and interpolated data set

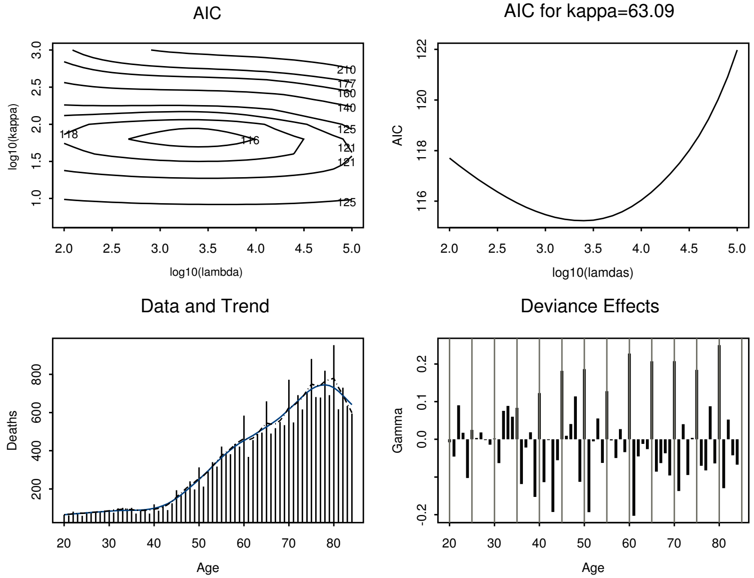

Figure 6.4: (a). Change in AIC for different values of λ and κ (b). Change in AIC for $\log_{10}(\kappa) = 1.8$ and varying λ(c). Histogram of empirical data and smoother, -.- smoother based on Schalls algorithm (d). Values of the overdispersion γ for different ages.

We fitted the data using both the AIC and Schall's algorithm to compute the weight of the penalty. Results are presented in Figure 4. A contour plot illustrates the dependence of AIC on λ and κ. 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.

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 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 ψ 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 $\eta = 2.7172 + x$ 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 6.5.4. As expected, a simple Poisson model does not perform well, especially for small values of the ψ 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 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 outperforms it in most of the cases. The pride model corrects the estimated standard errors, gives better estimates for the coefficients but also estimates the ψ better than the negative binomial model, with the only exception when $\psi = 2$. In the last row of the table we also present a case when the "true" model is Poisson (when $\psi = 20$).

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

Table 6.1: Results of 1000 simulations on 100 cases, simulated from a negative binomial model with $\eta = 2.7172 + x$ and ψ (the parameter of the negative binomial distribution) was chosen from the set $\{2, 4, 6, 8, 10\}$. 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 lied within the estimated confidence interval of the coefficient (95% nominal coverage).

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\psi}$	coverage (%)
$\psi = 2$	Poisson	1.132 (.125)	0.591 (.066)		55.4
	neg. bin.	1.215 (.236)	0.544 (.132)	2.09	80.7
	PRIDE	1.016 (.321)	0.527 (.181)	1.12	96.4
$\psi = 4$	Poisson	1.160 (.124)	0.578 (.065)		55.1
	neg.bin	1.229 (.190)	0.539 (.105)	4.16	72.5
	PRIDE	1.163 (.202)	0.525 (.112)	3.88	83.2
$\psi = 6$	Poisson	1.142 (.125)	0.586 (.065)		59.4
	neg.bin.	1.201 (.172)	0.552 (.094)	6.33	70.5
	PRIDE	1.169 (.175)	0.540 (.096)	6.03	76.5
$\psi = 8$	Poisson	1.141 (.125)	0.589 (.065)		60.8
	neg.bin.	1.194 (.162)	0.559 (.088)	8.50	68.2
	PRIDE	1.174 (.163)	0.550 (.089)	8.38	72.5
$\psi = 10$	Poisson	1.138 (.124)	0.589 (.065)		65.0
	neg.bin.	1.186 (.156)	0.562 (.085)	10.57	71.0
	PRIDE	1.171 (.156)	0.555 (.085)	10.48	73.7
$\psi = 20$	Poisson	1.098 (.126)	0.612 (.066)		82.2
	neg.bin.	1.391 (.105)	0.118 (.014)	621.98	5.6
	PRIDE	1.387 (.105)	0.118 (.014)	931.06	6.5

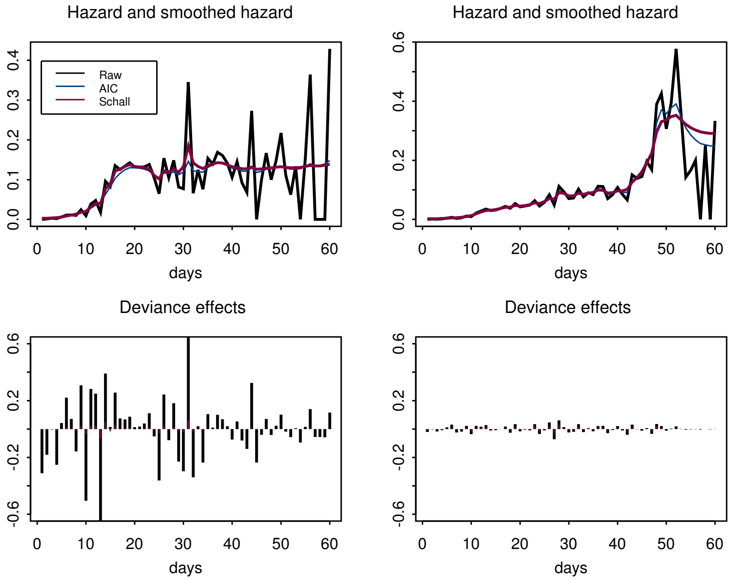

Figure 6.5: Hazard and smoothed hazard of flies in cohort 2 (left side) and 5 (right side), along with histograms of the corresponding deviance effects.

lifetables for 46 cohorts of female Mediterranean flies (Ceratitis capitata). Each cohort consisted of about 4000 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 [71]. 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 6.5 (right) shows an example where the size of the deviance effects are small, as is also indicated by the large value of κ (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 ($\kappa = 15.8$) with an absolute value up to 0.6. 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 2 ($\exp(0.6) = 1.8$). In this cohort however, the smoother chosen by Schall's algorithm follows the fluctuations of the data somewhat closer than required. Based on that, the AIC will be preferred in determining the optimal value of the penalty.

6.6 Discussion

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

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. However, our experience from generalized linear smoothing and semi-parametric modelling has shown that AIC serves well.

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 (6.1) is actually an extended likelihood ([73], 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 [60], while in smoothing literature it is known as *quasi-likelihood* [41]. However, Lee and Nelder chose to estimate the variance of the random effect using restricted maximum likelihood estimates (REML) whereas we 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.

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.

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.

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.

APPENDIX: Efficient Computation

Consider a PRIDE model with systematic component $\eta = B\alpha + \gamma$ where B is the basis matrix, α the corresponding coefficients, a penalty $\alpha' P\alpha$ and individual deviance effects γ. 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 / (w + kappa)
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 <- (G3 * G2) %*% rep(1, ncol(X))
S22 <- Fm + L1
R11 <- S11 %*% t(X) %*% G1
R22 <- (v / kappa)- L1 * kappa
tr <- sum(diag(R11)) + sum(R22)
```