INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

University of Alberta

GENERALIZATIONS OF ALB MODELS

by

Irina Dinu

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences

Edmonton, Alberta

Fall 2005

To my husband Adi, and my parents, Emil and Valeria

ABSTRACT

This thesis extends Adaptive Logistic Basis (ALB) models to accommodate the exponential family of distributions. ALB models were introduced by Hooper (2001) as a flexible family of regression models for multi-dimensional data. For both ALB models and their extensions to exponential family of distributions, the model is expressed as a linear combination of logistic basis functions, parameterized using reference points in the covariate space. The method is adaptive, selecting simple or more complex models as appropriate. The number, location and, to some extent, shape of the basis functions are automatically determined from the data.

Various extensions of ALB models to the exponential family of distributions include the case where the conditional distribution of the response given the predictors is Poisson, the case where extra-Poisson variation is present, Poisson counts observed over time, and the case where the conditional variance of the response given the predictors is a smooth function of the conditional mean of the response given the predictors. The original ALB methodology (Hooper, 2001) employs squared error and absolute error loss functions. Generalizations for count data are achieved by introducing a log-link function and an appropriate likelihood or quasi-likelihood function. While the idea is straightforward, several technical complications arise in the implementation.

Accuracy, interpretability, and computational speed are important when comparing regression methods for multi-dimensional data. The comparative accuracy of various methods is generally dependent on the target function f. I will report results of experiments comparing ALB with Generalized Additive Models (GAM), Multivariate Adaptive Regression Splines (MARS), and Generalized Projection Pursuit, using both real and simulated data. ALB models are equivariant under linear transformations of the predictors. GAM and MARS are typically constructed from additive effects and low-order interactions. Consequently, GAM and MARS tend to provide superior performance in examples exhibiting additive structure while ALB is superior in some ex-

amples with more complex interactions.

Acknowledgements

My sincerest thanks go to my supervisor Professor Peter Hooper. Without him, this thesis would not have come to light. Thanks are due to him for supporting me, from his NSERC funds for living and attending conferences. Throughout my Ph.D. programme, I always had in Dr. Peter Hooper a source of profound knowledge and sustained guidance and help.

Thanks are also due to Prof. S. Lele; for his valuable suggestions to improve this thesis; Prof. D. P. Wiens, who is one of the best teachers I ever had; Prof. N. G. Prasad, for his time spent to help and teach students choosing and applying the appropriate statistical methods. I want to thank Prof. K.C. Carriere for giving me the opportunity to gain important statistical consulting experience through the Training and Consulting Centre, within the Department of Mathematical and Statistical Sciences. If today, I have a global vision in this discipline, my vision is only due to them.

I thank Profs, N. E. Heckman and J. Roland for agreeing to be external members on my exam committee and for their careful reading of my thesis. I also thank Prof. I. Mizera for reading my thesis and for his valuable comments.

Thanks to the Department of Mathematical and Statistical Sciences for the financial support I received during my stay. Thanks are also due to the support staff members and especially Dona Guelzow and Marion Benedict.

Finally, I would like to thank my family. Their continuous encouragement, and their love, made me strong enough to pursue my goal to completion.

Irina

Contents

List of Tables

List of Figures

Conventions and notations:

Throughout the thesis, we will use lower-case letters to indicate random variables as well as their observed copies. Also, the following notations and acronyms will be used:

Symbol	Meaning
f_K	ALB predictor, $\sum_{k=1}^{K} \delta_k \phi_k(\mathbf{x})$
ϕ_k	k-th logistic basis function
GAM	Generalized Additive Models
GLM	Generalized Linear Models
GPP	Generalized Projection Pursuit model
WLS	Weighted Least Squares criterion
WQL	Weighted Quasi-Likelihood criterion

Chapter 1

Introduction

1.1 Background of the thesis

This thesis consists of various generalizations of Adaptive Logistic Basis (ALB) models, introduced by Hooper (2001) as a flexible family of regression models for multidimensional data. There is a vast literature on regression models. The main objective of regression analysis is to model the relationship between a response variable and one or more predictor variables. Linear regression models assume that the relationship between the mean response and the predictors is linear in the parameters and that the response variable is normally distributed with constant variance. Generalized linear models extend these assumptions: the mean response is related to the linear predictor through a link function, and the normal distribution is extended to the exponential family of distributions. This thesis focuses on extending ALB to accommodate exponential family of distributions. In the next two sub-sections I will summarize existing regression models and ALB models respectively.

1.1.1 Background on regression models

Regression methods are used to model the relationship between a response variable and one or more predictors. The simplest regression model is the linear model, where the relationship between the mean response and the predictors is assumed to be linear in the parameters. The investigator starts with a tentative parametric model. Once the parameters are estimated, diagnostics are applied to see if the model should be modified.

Nonparametric regression provides a flexible alternative with the regression equation determined adaptively from the data. Hastie and Tibshirani (1987) contrast the two approaches, noting that residual and partial residual plots are used to detect departures from linearity and often suggest parametric fixes. An attractive alternative to this indirect approach is to model the regression function non-parametrically and let the data decide on the functional form. Silverman (1985) states that nonparametric models give the data more of a chance to speak for themselves in choosing a model to be fitted. Altman (1992) defines nonparametric regression as a collection of techniques for fitting a curve when there is little a priori knowledge about its shape.

Nonparametric regression is often described as smoothing, especially in the case of a single predictor. The conceptual basis for smoothing is averaging the response values of observations having predictor values located in a neighbourhood of a target value. Several examples of smoothers are the running mean smoother, the running line smoother, and various kernel smoothers including the Gaussian, the Epanechinikov (1969) and the minimum variance

kernel. To generalize these smoothers to several dimensions, one can use Euclidian distance to define the neighbourhoods, or an alternative distance involving the covariance matrix of the predictors.

Splines combine polynomial regression with local fitting by representing the fit as a piecewise polynomial. The regions defining the pieces are bounded by knots. The piecewise polynomials are chosen so that they join smoothly at the knots. The number of subregions and the lowest order derivative allowed to be discontinuous at knots controls the tradeoff between smoothness and flexibility of the approximation.

Nonparametric regression is reasonably well understood in one dimension. Many methods are difficult to generalize to higher dimensions because of "the curse of dimensionality" (Bellman, 1961); i.e , "local neighbourhoods" must be large if they are to contain an adequate number of data points. In response to the dimensionality problem, strategies that attempt to approximate general functions in high dimensions based on adaptive computation have been devised. Adaptive computation adjusts its strategy to take into account particularities of the function to be approximated. Adaptive algorithms have been developed based on two paradigms: projection pursuit (Friedman and Stuetzle 1981; Friedman, Grosse and Stuetzle 1983; and Friedman 1985), and recursive partitioning (Morgan and Sonquist 1963; Breiman, Friedman, Olshen and Stone 1984).

Projection pursuit regression uses approximations of the form:

$$\hat{f}(x) = \sum_{m=1}^{M} f_m \left(\sum_{i=1}^{d} \alpha_{im} x_i \right).$$

The univariate functions f_m are required to be smooth. The function and the coefficients of the linear combinations are jointly optimized to produce a good fit to the data based on some distance criterion. Diaconis and Shahshahani (1984) showed that any smooth function of n variables can be represented by this approximation for large enough M. Also, for small to moderate M, many classes of functions can be closely fit by approximations of this form (Donoho and Johnstone, 1989). Disadvantages of projection pursuit are that there exist some simple functions that require large M for good approximations (Huber, 1985), interpretation is difficult for large M, and the approximation is computationally time consuming.

A recursive partitioning regression model is defined on subsets that form a partition $(R_m)_1^M$ of $\Re^d$:

$$\hat{f}(x) = g_m(x), \text{ if } x \in R_m.$$

The functions $(g_m)_1^M$ are taken to be of simple parametric form, the most common choice being a constant function (Morgan and Sunquist 1963; Breiman et al., 1984). The partitioning is accomplished through the recursive splitting of previous subregions. The subregions are then recombined in a reverse manner until an optimal set is reached based on a criterion that penalizes both for lack-of-fit and increasing number of regions. Recursive partitioning uses a forward stepwise algorithm. The basis functions produced by this algorithm are step functions. Two weaknesses of recursive partitioning are a lack of continuity at subregion boundaries and an inability to capture simple relationships such as linear, additive or lower-order interactions. Breiman and

Meisel (1976) and Friedman (1979) propose linear functions. Hansen, Kooperberg and Sardy (1998) introduced the triogram method, a natural approach for modeling data when the domain of the prediction variables is a polygon region in the plane. The estimates are continuous, piecewise linear functions defined over adaptively selected triangulations in the plane.

Friedman (1991) introduced the Multivariate Adaptive Regression Splines (MARS) approach to multivariate nonparametric regression. The main goal is to overcome the two limitations of recursive partitioning outlined above. MARS produces continuous models by replacing the step function by truncated power spline basis functions. To overcome the second limitation of recursive partitioning, MARS permits the recursive splitting of all basis functions in the model and not just those that are currently terminal.

The additive model,

$$E(y|x) = \alpha + \sum_{1}^{p} f_j(x_j)$$

is a special case of projection pursuit. A general algorithm to fit additive models is the backfitting algorithm. This is an iterative fitting procedure that updates the functions f_j by smoothing partial residuals $y - \alpha - \sum_{k \neq j} f_k$ against x_j. The cycle stops when the individual functions f_j converge.

A generalized additive model has the form:

$$g(\mu(x)) = \alpha + \sum_{1}^{p} f_j(x_j) \ , \ \mu(x) = E(y|x),$$

and represents an extension of the additive model to the exponential family. The local scoring algorithm is used to fit this model (Hastie and Tibshirani, 1986).

Finally, we discuss variable selection and shrinkage methods for linear regression models. Linear models are simple and often provide an adequate and interpretable description of how predictors affect the response. In general, there are two reasons why we are not satisfied with least square estimates: prediction accuracy and interpretation. We should clarify that our discussion here assumes that the true model follows a linear model. Departures from this assumption have motivated researchers to develop flexible models that give a better approximation to the true model without using too many parameters.

First we talk about prediction accuracy. The Gauss-Markov theorem asserts that, when the true model is a linear model, the least squares estimator has the smallest variance among all linear unbiased estimators. This result does not imply that we should necessarily restrict our search to unbiased estimators. The mean squared error of an estimator can be decomposed into the sum of the variance of the estimator and the squared bias. There may well exist a biased estimator with smaller mean squared error. Such an estimator would trade a little bias for a larger reduction in variance. Mean squared error is closely related to prediction accuracy. In fact, the expected prediction error at a new observation and the mean squared error differ only by an amount that equals the variance of the new observation. Shrinkage methods produce estimates that permit some bias to reduce the variance of the predicted values, and hence improve the overall prediction accuracy.

The second reason why we are not satisfied with least squares estimates is interpretation. With a large number of predictors, we often would like to

select a smaller subset that exhibit the strongest effects.

Variable selection algorithms include forward selection, backward elimination and stepwise selection. Forward selection starts with the intercept, then sequentially adds into the model the predictor that most improves the fit, stopping when no predictor produces a significantly better fit when added to the model. Backward elimination starts with the full model, then sequentially deletes predictors that are not improving the fit. The algorithm stops when there are no further deletions that produce a significantly better fit. Stepwise selection considers both forward and backward steps at each stage and makes the "best" move.

Although subset selection techniques produce interpretable models with possibly lower prediction error than the full model, the discrete nature of the algorithm exhibits high variance. Shrinkage methods are more continuous alternatives to subset selection and do not suffer as much from high variance. We mention here two shrinkage methods: ridge regression and least absolute shrinkage and selection operator, or shortly "lasso". Ridge regression shrinks the regression coefficients by imposing the L_2 size constraint:

$$\hat{\beta}^{ridge} = \text{argmin}_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq s.$$

The lasso shrinks the regression coefficients by imposing the L_1 size constraint:

$$\hat{\beta}^{lasso} = \text{argmin}_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2, \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq s.$$

Both methods use cross-validation to select an optimal value for s. Ridge regression is a continuous process that shrinks the coefficients, and hence is

more stable than subset selection. However, ridge regression does not set any coefficient to zero, and hence does not give an easily interpretable model. Because of the nature of the LASSO constraint, making s sufficiently small will cause some of the coefficients to be exactly zero. Thus, LASSO does a kind of continuous selection retaining the good features of both subset selection and ridge regression.

Least Angle Regression(LARS), a model selection algorithm proposed by Efron, Hastie, Johnstone and Tibshirani (2004), is a useful and less greedy version of traditional forward selection methods. A modification in the LARS algorithm implements the LASSO. A different LARS modification implements another recent model selection method called Forward Stagewise Linear Regression.

1.1.2 Summary of ALB methodology

Hooper (2001) introduced a flexible family of regression models called Adaptive Logistic Basis (ALB) models. The ALB methodology is summarized in this section.

Consider the problem of estimating a regression function $f(\mathbf{x}) = E\{y|\mathbf{x}\}$, where y is a response variable and $\mathbf{x}$ is a vector of d covariates. Estimators often approximate f by a linear combination of basis functions:

$$f(\mathbf{x}) \approx \sum_{k=1}^{K} \delta_k \phi_k(\mathbf{x}) \tag{1.1}$$

Examples include tensor product splines (Gu, Bates, Chen and Wahba, 1989; Friedman, 1991), thin-plate splines (Wahba, 1990) and ridge functions (Fried-

man and Stuetzle, 1981). ALB models are defined by logistic basis functions:

$$\phi_k(\mathbf{x}) = \exp(\alpha_k + \beta_k'\mathbf{x}) / \sum_{m=1}^{K} \exp(\alpha_m + \beta_m'\mathbf{x}). \tag{1.2}$$

Note that $\sum \phi_k(\mathbf{x}) = 1$, so approximation (1.1) does not require a constant term. The redundancy in the parameterization can be handled by dividing the numerator and denominator of (1.2), by $\exp(\alpha_K + \beta_K'\mathbf{x})$. In other words, α_K and β_K can be set to zero. After removing redundancies, the effective number of parameters in approximation (1.1) is

$$p = 1 + (K-1)(d+2).$$

An alternative parameterization can be obtained by expressing the basis functions in terms of Euclidean distance from reference points ξ_k in the covariate space:

$$\phi_k(\mathbf{x}) = \exp(\gamma_k - \tau^{-2}\|\mathbf{x} - \xi_k\|^2) / \sum_{m=1}^{K} \exp(\gamma_m - \tau^{-2}\|\mathbf{x} - \xi_m\|^2).$$

The two parameterizations can be related as follows. Starting with the linear parameterization and the constraints $\alpha_K = 0$ and $\beta_K = 0$, one could set

$$\begin{aligned} \tau &= 1 \\ \xi_k &= (1/2)\beta_k \\ \gamma_k &= \alpha_k + (1/4)\|\beta_k\|^2, \end{aligned} \tag{1.3}$$

obtaining $\gamma_K = 0$ and $\xi_K = 0$. Conversely, starting with the reference point parameterization, one could set

$$\begin{aligned} \alpha_k &= \gamma_k - \tau^{-2}\|\xi_k\|^2 - (\gamma_K - \tau^{-2}\|\xi_K\|^2) \\ \beta_k &= 2\tau^{-2}(\xi_k - \xi_K), \end{aligned}$$

obtaining $\alpha_K = 0$ and $\beta_K = 0$.

The method of estimation is adaptive, selecting simple or more complex models as appropriate. The number, location and (to some extent) shape of the basis functions are automatically determined from the data. ALB estimators $\hat{f}$ are defined for a family of location measures, including the conditional mean and median. Suppose $(\mathbf{x}, y)$ is a random vector, choose $q \geq 1$ and let f be a function minimizing $E|y - f(\mathbf{x})|^q$. It is assumed that the expectation is finite. Conditional mean and median functions are obtained by taking $q = 2$ and $q = 1$, respectively. The ALB function is estimated as follows. Suppose we have a sample $(\mathbf{x}_i, y_i), i = 1, \ldots, n$. For given K, an ALB estimator $\hat{f}_K$ is calculated by minimizing the training risk $(1/n) \sum |y_i - f_K(\mathbf{x}_i)|^q$ over $\{(\delta_k, \gamma_k, \xi_k), k = 1, ..., K\}$, with τ set to a convenient value. The parameter values defining $\hat{f}_K$ are determined separately for different numbers K, and a generalized cross-validation technique is used to select K.

ALB has connections with two types of neural networks: radial basis functions and feed-forward back-propagation networks. If the γ_k are set to zero, then $\hat{f}$ is a radial basis function network (Moody and Darken, 1989). Radial basis functions are related to kernel regression estimators (Xu, Kryzak and Yuille 1994). In most applications of radial basis functions, the basis function parameters $\{\tau, \xi_1, ..., \xi_K\}$ are selected using only the covariates, and the δ_k are estimated by a least squares fit of the linear model. ALB shares the approximation properties of the radial basis functions (Ripley, 1996) but, by employing a richer family of basis functions and by using the response to

estimate all parameters, ALB typically uses far fewer basis functions. Under the linear parameterization, ALB can be represented much like a feedforward neural network with a single hidden layer (Cheng and Titterington, 1994), but with softmax applied to the hidden layer. The term "softmax", introduced by Bridle(1990), describes the conversion of a set of outputs z_k into probabilities $\exp(z_k)/\sum \exp(z_m)$. The idea stems from multiple logistic regression. Softmax is normally applied to the final output layer of a network. Hidden units are normally transformed individually, often with a sigmoidal function. Applying softmax to the hidden layer violates the "feedforward" property, but allows a spatial interpretation of the initial connection weights through the reference point parameterization.

ALB has the property of affine invariance, a result that one can easily prove using the linear parameterization. ALB possesses the universal approximation property; i.e., continuous functions f can be approximated uniformly on compact sets as $K \to \infty$. ALB is expected to work best in comparison with other methods when f is well approximated using a small number of basis functions.

1.1.3 Generalized Linear models and ALB models

Linear models approximate a regression function $f(\mathbf{x}) = E(y|\mathbf{x})$, where y is a response variable and $\mathbf{x}$ is a vector of d covariates, by a linear combination of the covariates. Suppose we have a sample of size n, $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n$ represent the covariates at each observation and $y_1, \ldots, y_n$

represent the responses. In matrix notation the model can be written as:

$$E(y_i|\mathbf{x}_i) = \beta'\mathbf{x}_i$$

where β represents the vector of parameters and has dimension d. The vector of parameters is estimated by minimizing the squared error:

$$\sum_{i=1}^{n}(y_i - \beta'\mathbf{x}_i)^2.$$

If the conditional distribution of y_i given $\mathbf{x}_i$ is normal with constant variance, then the least squares estimator is the maximum likelihood estimator.

Generalized linear models extend linear models in two ways. First, the conditional distribution of y given $\mathbf{x}$ is a member of the exponential family:

$$f_{Y|\mathbf{x}}(y|\mathbf{x};\theta,\phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\},$$

for some specific functions $a(.)$, $b(.)$ and $c(.)$. The two parameters θ and ϕ may be functions of $\mathbf{x}$. Second, the linear predictor $\eta = \beta'\mathbf{x}$ can be any monotonic differentiable function of the conditional mean $\mu = E(y|\mathbf{x})$:

$$\eta = g(\mu),$$

where $g(.)$ is called the link function. There are special link functions for which there exists a minimal sufficient statistic for the canonical parameter θ. These links are called canonical links and they occur when $\theta = \eta$. We enumerate some distributions from the exponential family together with their corresponding canonical links: Normal distribution with the *identity* link, Poisson distribution with the *log* link, Binomial distribution with the *logit* link,

Gamma distribution with the *reciprocal* link. The maximum likelihood method is typically used to fit the model. One way to solve the maximum likelihood equations is using the iteratively reweighted least squares algorithm.

ALB methods can be applied in the context of generalized linear models (GLM). Suppose the conditional distribution of y given $\mathbf{x}$ comes from an exponential family with distribution form $p(y|\mathbf{x}, \mu)$ and a specified monotonic differentiable function of the conditional mean $\mu = E(y|\mathbf{x})$ is approximated by an ALB function f_K:

$$g(\mu) = f(\mathbf{x}) \approx f_K(\mathbf{x}) = \sum_{k=1}^{K} \delta_k \phi_k(\mathbf{x}),$$

where $g(.)$ is the link function. The linear function $\eta = \beta'\mathbf{x}$ in GLM is thus replaced by the more flexible ALB function f_K.

To estimate the above model, we can maximize the log-likelihood. Below we display the factors involved in the derivative of a single term of the log-likelihood:

$$\frac{\partial \log p(y|\mathbf{x}, \mu)}{\partial \boldsymbol{\theta}} = \frac{\partial \log(p(y|\mathbf{x}, \mu))}{\partial \mu} \frac{\partial g^{-1}(f)}{\partial f} \frac{\partial f}{\partial \boldsymbol{\theta}}, \tag{1.4}$$

where here $\boldsymbol{\theta}$ denotes the vector of ALB parameters, and not the canonical parameter described on the previous page. We note that, on the right-hand side, the first derivative involves the exponential model, the second derivative involves the link function, and the last derivative involves the ALB function. We will refer to the above equation when developing estimation for ALB models in the GLM context.

The work in Hooper (2001) provides a basis for much of our work, and is referenced throughout the remainder of the thesis as H01. When extending

ALB models to accommodate the exponential family of distributions, some results in H01 can be applied directly, while others require appropriate modification.

1.2 Overview of the thesis

In Chapter 2, we discuss ALB models for the case where the conditional distribution of the response given the predictors is Poisson. We describe the model and the method of estimation and conduct a comparison between ALB models and GAM. We present functions where the performance of ALB models is substantially better than that of GAM, even if the sample size is large, and also functions where GAM has advantage over ALB. We derive approximate standard errors for the fit and present estimated coverage probabilities of a 95% confidence interval for the mean using several simulation studies. We also discuss ALB models for Poisson counts observed over time.

In Chapter 3. we discuss ALB models for the case where there is more variation in the responses than that expected from Poisson sampling theory. We discuss the constant coefficient of variation case in Poisson over-dispersed data and illustrate the ALB model with simulated data and a real life example. We discuss the case where the variance is a function of the mean. We summarize the ideas behind the heteroscedastic linear model, combine ALB and parametric variance function estimation to model heteroscedasticity, and use ALB to model the variance function.

In Chapter 4, we illustrate ALB models using two examples. We indicate

ALB strengths and weaknesses in comparison to other models.

In Chapter 5, we give background on Projection Pursuit(PP) and the extension to the exponential family, Generalized Projection Pursuit(GPP). We compare predictive accuracy of ALB, GPP and GAM, for the case where the conditional distribution of the response given the predictors is Poisson using simulation studies on a variety of functions previously employed by researchers to test performance of other flexible regression methods such as Projection Pursuit, Automatic Smoothing Splines Projection Pursuit, Multivariate Adaptive Regression Splines, Additive Models.

Finally, Chapter 6 summarizes the main contributions of this thesis to the literature. Some possible future research is suggested to expand and improve on the strategies suggested in this thesis.

Chapter 2

ALB models for Poisson counts

2.1 Introduction

Classical linear models began with the work of Gauss and Legendre on astronomical data, usually measurements of continuous quantities. The variability in the observations was largely the effect of measurement error. Gauss introduced the Normal distribution of errors as a device for describing variability.

Another important direction in the history of statistics is the development of methods for dealing with discrete events rather than with continuously varying quantities. In the context of rare events, the basic distribution for counts of events is the Poisson distribution. This distribution has been applied to diverse kinds of events, such as annual number of traffic accidents, number of outbreaks of an infectious disease in a county, number of typographical errors on a page.

In this chapter, we discuss ALB models for the case where the conditional distribution of the response given the predictors is Poisson. The organization of this chapter is as follows. In Section 2, we describe the model and the method of estimation. In Section 3, we conduct a comparison between ALB

models and GAM. We also present functions where the performance of ALB models is substantially better than that of GAM, even if the sample size is large. In Section 4, we derive approximate standard errors for the fit. In Section 5, we discuss ALB models for Poisson counts observed over time.

2.2 Estimation

2.2.1 Estimation of f_K using stochastic approximation

In the following, the conditional distribution of Y given $\mathbf{x}$ is assumed to be Poisson with mean μ, that is $P(Y = y \mid \mathbf{X} = \mathbf{x}) = e^{-\mu}\mu^y/y!$, where $\mu > 0$ is a function of $\mathbf{x}$. In this case our ALB model takes the form:

$$E(Y|\mathbf{x}) = \mu,$$

$$\log(\mu) = f_K(\mathbf{x}) = \sum_{k=1}^{K} \delta_k \phi_k(\mathbf{x}).$$

In this subsection we consider the estimation of f using a fixed number of basis functions. The estimation of f follows H01, but instead of minimizing a predictive risk corresponding to a constant variance assumption, we minimize the negative log-likelihood corresponding to a Poisson distribution. Further modifications needed are explained as we derive the estimation algorithm. We will use the reference point parameterization:

$$\phi_k(\mathbf{x}) = \exp(\gamma_k - \tau^{-2}\|\mathbf{x} - \boldsymbol{\xi}_k\|^2) / \sum_{m=1}^{K} \exp(\gamma_m - \tau^{-2}\|\mathbf{x} - \boldsymbol{\xi}_m\|^2).$$

The reference points parameterization is easier to interpret than the linear parameterization. The location of ϕ_k can be controlled by $\boldsymbol{\xi}_k$, the relative influence of ϕ_k can be controlled by γ_k, and the smoothness of ϕ_k can be

controlled by τ. This interpretation is useful when initializing parameter values for estimation. The roles of the parameters are actually not so clearly separated due to redundancies among the parameters. For example, τ can be fixed without limiting the generality of the parameterization, and smoothness can be controlled by adjusting the remaining parameters.

The log-likelihood of the conditional distribution of y given $\mathbf{x}$ is:

$$l(\boldsymbol{\theta}|\mathbf{x}, y) = -\mu + y \log \mu - \log y!, \tag{2.1}$$

where $\boldsymbol{\theta}' = (\delta_1, \gamma_1, \boldsymbol{\xi}_1', \ldots, \delta_K, \gamma_K, \boldsymbol{\xi}_K')$. The last term in the log-likelihood is treated as a constant. Our underlying aim is to minimize the negative log-likelihood for a given sample $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$:

$$-l(\boldsymbol{\theta}|\mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n) = \sum_{i=1}^{n} (\mu_i - y_i \log \mu_i) + \text{const}, \tag{2.2}$$

where $\mu_i = \mu(\mathbf{x}_i) = \exp(f_K(\mathbf{x}_i))$.

Let $\hat{f}_K$ denote the estimator that minimizes the negative log-likelihood and let $\hat{\boldsymbol{\theta}}'$ denote a parameter vector defining $\hat{f}_K$. Since the parameters are not uniquely determined, $\hat{\boldsymbol{\theta}}$ is not regarded as an estimator, but as one of the many equivalent parameterizations of $\hat{f}_K$.

For fixed K, the log-likelihood can be maximized using stochastic approximation, which was introduced by Robbins and Monro (1951). Alternative non-stochastic optimization methods, such as Newton Raphson, are potentially available. The stochastic approximation algorithm has been constructed so that the number of steps in the algorithm does not depend on the sample size, making this algorithm competitive for large data sets. Nevertheless,

the number of iterative steps can always be increased for larger sample sizes. The following brief review of the stochastic approximation algorithm follows Benveniste, Metivier and Priouret (1990). Consider minimizing a function $Q(\boldsymbol{\theta})$ using an iterative algorithm driven by a sequence of independent and identically distributed vectors z_m,

$$\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m-1} + a_m \mathrm{H}(\boldsymbol{\theta}_{m-1}, z_m) \tag{2.3}$$

In stochastic algorithms, the updating function H is defined so that $-\mathrm{E}\{\mathrm{H}(\boldsymbol{\theta}, z)\}$ is proportional to the gradient of $Q(\boldsymbol{\theta})$. Let $\boldsymbol{\theta}_m = \boldsymbol{\theta}(t_m)$, where $t_m = \sum_{i=1}^{m} a_i$. After an initial transient phase, the behavior of the above process is represented to a first approximation by that of the differential equation $d\boldsymbol{\theta}(t)/dt = E[\mathrm{H}\{\boldsymbol{\theta}(t), z\}]$. The following conditions on the gain function:

$$\begin{aligned} a_m &> 0, \\ \sum a_m &= \infty, \\ \sum a_m^{\alpha} &< \infty, \text{ for some } \alpha > 1 \end{aligned} \tag{2.4}$$

are sufficient for the convergence of the above sequence toward a local minimum for $Q(\boldsymbol{\theta})$, provided that the sequence $\{\boldsymbol{\theta}_m\}$ is bounded, (Kushner and Clark 1978). An example of a gain function satisfying these conditions would be $a_m = 1/m$.

Initial parameter values are motivated by the following Proposition (H01, Proposition 4):

Proposition 2.1 *Set* $\zeta_k = \tau^2 \gamma_k$ *and define*

$$A_k = \{\mathbf{x} : \|\mathbf{x} - \boldsymbol{\xi}_k\|^2 - \zeta_k < \|\mathbf{x} - \boldsymbol{\xi}_m\|^2 - \zeta_m \text{ for all } m \neq k\}.$$

(i) *We have* $A_k = \{\mathbf{x} : \phi_k(\mathbf{x}) > \phi_m(\mathbf{x}) \text{ for all } m \neq k\}$. *Each* A_k *is a convex set, possibly empty. The boundary between two neighbouring sets* A_k *and* A_m *is a subset of a hyperplane orthogonal to* $\boldsymbol{\xi}_k - \boldsymbol{\xi}_m$.

(ii) *If the* ζ_k *are all equal, then* $\{A_k\}$ *forms a Dirichlet tessellation of* $\Re^d$; *i.e.,* A_k *consists of all* $\mathbf{x}$ *nearest to* $\boldsymbol{\xi}_k$. *If the* ζ_k *differ substantially, then the spatial interpretation of the* $\boldsymbol{\xi}_k$ *is less clear; e.g., it is possible that* $\boldsymbol{\xi}_k \notin A_k$.

(iii) *We have* $\partial\phi_m/\partial\gamma_k = \phi_k(1-\phi_k)$ *for* $m = k$, *and* $-\phi_k\phi_m$ *for* $m \neq k$, *so increasing* γ_k *increases the influence of* ϕ_k *and diminishes that of other* ϕ_m.

(iv) *Fix* $\zeta_1, \boldsymbol{\xi}_1, \ldots, \zeta_K, \boldsymbol{\xi}_K$. *As* τ *approaches* 0, $\phi_k(\mathbf{x})$ *converges to the indicator function of* A_k, *for all* $\mathbf{x}$ *not on the boundary of* A_k. *As* τ *approaches* ∞, $\phi_k(\mathbf{x})$ *converges to* $1/K$.

The covariates are first centered and scaled to have zero mean and unit standard deviation. This standardization is helpful when initializing parameter values. We note that for the Normal errors assumption, the response was also centered and scaled to have zero mean and unit standard deviation. However, this standardization no longer makes sense for Poisson errors assumption, creating difficulties in the stochastic approximation algorithm. We address this issue later in this section. The initial γ_k are set to zero. The initial $\boldsymbol{\xi}_k$ are obtained as a spatially representative set of points in the covariate space (see below). The initial δ_k are then defined as the logarithm of the average

of the y_i values for $\mathbf{x}_i$ in the region nearest to $\boldsymbol{\xi}_k$. The parameter τ is set to the average distance between nearest neighbors among the K initial points $\boldsymbol{\xi}_k$. This choice for τ yields a reasonable amount of overlap among neighboring basis functions.

A representative set of K points $\boldsymbol{\xi}_k$ can be obtained by minimizing

$$\sum_{i=1}^{n} \min\{\|\mathbf{x}_i - \boldsymbol{\xi}_k\|^2 : k = 1, \ldots, K\}. \tag{2.5}$$

The resulting points have been called K-means cluster centroids (MacQueen 1967) and principal points (Flury 1990). The latter term is more appropriate here, as we are not searching for clusters, but for a representative set of points. The initial $\boldsymbol{\xi}_k$ can be calculated using a K-means clustering algorithm (Hartigan & Wong 1979). However, the above minimization need not be exact. We will simultaneously initialize both $\boldsymbol{\xi}_k$ and δ_k using a vector quantization algorithm (Kohonen 1995). Begin by generating $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$ randomly from $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and set all δ_k to one. Let $\hat{P}$ denote the empirical distribution; i.e., the distribution conditioning on the data, of $(\mathbf{x}_u, y_u)$, with u distributed uniformly on $\{1, \ldots, n\}$. Then repeat the following steps for $3000\sqrt{K}$ iterations. At the mth iteration, sample $(\mathbf{x}, y)$ from $\hat{P}$, determine the point $\boldsymbol{\xi}_k$ nearest $\mathbf{x}$, replace $\boldsymbol{\xi}_k$ by $(1 - a_m)\boldsymbol{\xi}_k + a_m\mathbf{x}$, and replace δ_k by $(1 - a_m)\delta_k + a_m y$. The gain is defined as $a_m = 100\sqrt{K}/(m + 100\sqrt{K})$, gradually lowered to zero. After the last iteration, each δ_k is log-transformed. This algorithm produces approximate principal points and logarithms of y-averages, which serve as useful starting values.

Once the initial values are selected, the negative log-likelihood (2.2) is

minimized using stochastic approximation. At each iteration, an observation $(\mathbf{x}, y)$ is randomly sampled with replacement from $\hat{P}$ and the parameter vector $\boldsymbol{\theta}$ is updated as in expression (2.3). According to equation (1.4), the gradient of the log-likelihood can be written as

$$\frac{\partial l(\boldsymbol{\theta}|\mathbf{x}, y)}{\partial \boldsymbol{\theta}} = (y - \mu)\frac{\partial f_K}{\partial \boldsymbol{\theta}}.$$

The components of the gradient of f_K have already been derived in H01, equation (10). Therefore, the components of the gradient of the log-likelihood (2.1) follow easily:

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\theta}|\mathbf{x}, y)}{\partial \delta_k} &= (y - \mu)\phi_k(\mathbf{x}), \\
\frac{\partial l(\boldsymbol{\theta}|\mathbf{x}, y)}{\partial \gamma_k} &= (y - \mu)\phi_k(\mathbf{x})(\delta_k - f_K(\mathbf{x})), \\
\frac{\partial l(\boldsymbol{\theta}|\mathbf{x}, y)}{\partial \boldsymbol{\xi}_k} &= (y - \mu)\phi_k(\mathbf{x})(\delta_k - f_K(\mathbf{x}))(\mathbf{x} - \boldsymbol{\xi}_k),
\end{aligned} \tag{2.6}$$

for $k = 1, \ldots, K$. Consequently, the updating formulae for the parameters at the mth iteration are given by:

$$\begin{aligned}
\delta_k &\leftarrow \delta_k + a_m^{\delta}(y - \mu)\phi_k(\mathbf{x}) , \\
\gamma_k &\leftarrow \gamma_k + a_m^{\gamma}(y - \mu)\phi_k(\mathbf{x})(\delta_k - f_K(\mathbf{x})) , \\
\boldsymbol{\xi}_k &\leftarrow \boldsymbol{\xi}_k + a_m^{\xi}(y - \mu)\phi_k(\mathbf{x})(\delta_k - f_K(\mathbf{x}))(\mathbf{x} - \boldsymbol{\xi}_k).
\end{aligned}$$

We now explain why these formulae need to be modified. These formulae work in some situations, but present problems for large counts. The variation of the quantity $(y-\mu)$ is larger for Poisson counts than for the constant variance case, since $\mathrm{Var}(y|\mathbf{x}) = \mu$. If we use the above updating formulae, perturbations in parameter values can be large when μ is large, sometimes resulting

in convergence to a poor local optimum of the log-likelihood. Recall that, for constant variance models, the response variable was standardized prior to using stochastic approximation. However, standardization of Poisson counts does not make sense. Instead of transforming the response variable prior to using stochastic approximation, we stabilize the magnitude of perturbations by scaling the updating functions and by setting an upper bound on the perturbations. In the following, we present the modified updating formulae for the parameters at the mth iteration, and then justify the choice of the scaling constant and the upper bound on the perturbations. Set

$$h_k(\mathbf{x}, y, \boldsymbol{\theta}) = \frac{(y-\mu)}{\sqrt{\mu}}\phi_k(\mathbf{x}), \tag{2.7}$$

where $\mu = \exp(f_K(\mathbf{x}))$. We use the following updating formulae for the parameters at the mth iteration:

$$\begin{aligned}
\delta_k &\leftarrow \delta_k + a_m^{\delta} h_k(\mathbf{x}, y, \boldsymbol{\theta}) \min\{\sqrt{\mu}/c, a_0^{\delta}/a_m^{\delta}\}\,, \\
\gamma_k &\leftarrow \gamma_k + a_m^{\gamma} h_k(\mathbf{x}, y, \boldsymbol{\theta}) \min\{\sqrt{\mu}/c, a_0^{\gamma}/a_m^{\gamma}\}\{\delta_k - f_K(\mathbf{x})\}\,, \\
\boldsymbol{\xi}_k &\leftarrow \boldsymbol{\xi}_k + a_m^{\xi} h_k(\mathbf{x}, y, \boldsymbol{\theta}) \min\{\sqrt{\mu}/c, a_0^{\xi}/a_m^{\xi}\}\{\delta_k - f_K(\mathbf{x})\}(\mathbf{x} - \boldsymbol{\xi}_k),
\end{aligned} \tag{2.8}$$

where

$$c = s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{(n-1)}},$$

is the unconditional standard deviation. We note that:

$$a_m^{\delta} h_k(\mathbf{x}, y, \boldsymbol{\theta}) \min\{\sqrt{\mu}/c, a_0^{\delta}/a_m^{\delta}\} \rightarrow \begin{cases} a_m^{\delta} \frac{y-\mu}{c}\phi_k(\mathbf{x}) & \text{as } \mu \rightarrow 0, \text{ or } m \rightarrow \infty \\ a_0^{\delta} \frac{y-\mu}{\sqrt{\mu}}\phi_k(\mathbf{x}) & \text{as } \mu \rightarrow \infty, \ m \text{ fixed.} \end{cases}$$

Now we can see that the above choice for the scaling constant makes equations (2.7) and (2.8) look more similar to the constant variance case, where the response was standardized.

Two more scaling constants were tried: $c = \sum \sqrt{y_i}/n$ and $c = \sqrt{\sum y_i/n}$. These last two choices do not solve the problem for large counts. A closer look at the last two choices indicates that they do not reflect variability in μ_i. The first choice, the unconditional standard deviation, reflects variability in μ_i. This explains why the choice of the unconditional standard deviation solves the problem of large counts, while the other two choices fail. We now justify the upper bounds on the perturbations. We note that all three ratios under the min expression, a_0^δ/a_m^δ, a_0^γ/a_m^γ and a_0^ξ/a_m^ξ are equal. The stochastic approximation algorithm aims to gradually decrease the perturbations, and by imposing this upper bound we ensure that the magnitude of perturbations for all m, is no larger than in the first iteration.

A natural question is whether the modifications to the updating functions for the Poisson version of ALB are still resulting in convergence of the log-likelihood toward a local optimum. To prove this, we return to the more general setting used to review the stochastic approximation algorithm, described earlier in this section. Recall that we considered minimizing a function $Q(\boldsymbol{\theta})$ using an iterative algorithm driven by a sequence of independent and identically distributed vectors z_m, labeled earlier as equation (2.3):

$$\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m-1} + a_m \mathrm{H}(\boldsymbol{\theta}_{m-1}, z_m).$$

When the updating function H is defined so that $-\mathrm{E}\{\mathrm{H}(\boldsymbol{\theta}, z)\}$ is proportional to the gradient of $Q(\boldsymbol{\theta})$, conditions (2.4) are sufficient for the convergence of the above sequence toward a local minimum for $Q(\boldsymbol{\theta})$, provided the sequence $\{\boldsymbol{\theta}_m\}$ is bounded, (Kushner and Clark 1978). In the light of the above modifications

to the updating functions, we need to prove the following in order to ensure convergence of the log-likelihood toward a local optimum:

Theorem 2.2 *Consider minimizing a function $Q(\boldsymbol{\theta})$ using an iterative algorithm driven by a sequence of independent and identically distributed vectors z_m,*

$$\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m-1} + a_m \mathrm{H}_1(\boldsymbol{\theta}_{m-1}, z_m) \times \min\{\mathrm{H}_2(\boldsymbol{\theta}_{m-1}, z_m), a_0/a_m\}, \tag{2.9}$$

where $\mathrm{H} = \mathrm{H}_1 \times \mathrm{H}_2$ is defined so that $-E\{\mathrm{H}(\boldsymbol{\theta}, z)\}$ is proportional to the gradient of $Q(\boldsymbol{\theta})$, and $\sup_{\boldsymbol{\theta}, z} \mathrm{H}_2(\boldsymbol{\theta}, z) < \infty$. If the sequence $\{\boldsymbol{\theta}_m\}$ is bounded, then conditions (2.4) are sufficient for the convergence of the above sequence toward a local minimum for $Q(\boldsymbol{\theta})$.

PROOF. Let $\mathrm{M} = \sup_{\boldsymbol{\theta}, z} \mathrm{H}_2(\boldsymbol{\theta}, z)$. From the conditions (2.4), it follows that there exists an integer m_{M} such that

$$a_0/a_m > \mathrm{M}, \text{ for any } m > m_{\mathrm{M}}.$$

Therefore, for any $m > m_{\mathrm{M}}$ the sequence (2.9) becomes:

$$\boldsymbol{\theta}_m = \boldsymbol{\theta}_{m-1} + a_m \mathrm{H}_1(\boldsymbol{\theta}_{m-1}, z_m) \times \mathrm{H}_2(\boldsymbol{\theta}_{m-1}, z_m).$$

Since, $\{\boldsymbol{\theta}_m\}$ is bounded, and $\mathrm{H} = \mathrm{H}_1 \times \mathrm{H}_2$ is defined so that $-\mathrm{E}\{\mathrm{H}(\boldsymbol{\theta}, z)\}$ is proportional to the gradient of $Q(\boldsymbol{\theta})$, we can conclude that for a large enough m, the conditions (2.4) are sufficient for the convergence of the above sequence toward a local minimum for $Q(\boldsymbol{\theta})$, (Kushner and Clark 1978). The result therefore holds even with the above modifications on the updating functions. □

Back to our application, where Q is the negative log-likelihood, we note that the jth component of H_1 is

$$\mathrm{H}_1^j = c_j \frac{\partial Q(\boldsymbol{\theta})/\partial \boldsymbol{\theta}}{\sqrt{\mu}},$$

where $c_j = 1$ for the derivatives taken with respect to δ and γ, and $c_j = \tau^2/2$ for the derivatives taken with respect to $\boldsymbol{\xi}$. We have $\mathrm{H}_2 = \sqrt{\mu}/c$. $\mathrm{H} = \mathrm{H}_1 \times \mathrm{H}_2$ is then proportional to the gradient of $Q(\boldsymbol{\theta})$. Now we justify that $\sup_{\boldsymbol{\theta},z} \mathrm{H}_2(\boldsymbol{\theta}, z) < \infty$. Since we are sampling from a finite dataset, we can assume that z is in a compact set. Although in our applications we did not set any bounds on $\{\boldsymbol{\theta}_m\}$, it is safe to assume that $\boldsymbol{\theta}$ is in a compact set; for example, in our applications, we noticed that δ_m was always between 0 and the logarithm of the largest response count. In addition to $(\boldsymbol{\theta}, z)$ being in a compact set, $\mathrm{H}_2(\boldsymbol{\theta}, z)$ is a continuous function. Thus, $\sup_{\boldsymbol{\theta},z} \mathrm{H}_2(\boldsymbol{\theta}, z) < \infty$.

The implementation of stochastic approximation involves choosing the number of iterations and the form of the gain function. The choices described below were made in H01. The number of iterations is set at $M = 50000\sqrt{K}$. The gains are positive numbers approaching zero; more precisely,

$$a_m^{\xi} = \begin{cases} a_0^{\xi} \frac{c_g M}{m + c_g M}, & 1 \le m \le M/2, \\ a_{M/2}^{\xi} \frac{2(M-m)}{M}, & M/2 < m \le M, \end{cases}$$

with $a_0^{\xi} = 0.25$, and $c_g = 0.01$. The constants were chosen empirically based on the following considerations. The constant a_0^{ξ} controls the initial gain while c_g controls how rapidly the gains approach zero. If initial gains are too large, then large initial perturbations of the reference points are more likely to result in convergence to a poor local optimum of the log-likelihood. If initial gains

are too small or the gains decrease too quickly, then perturbations may be too small for the process to reach an optimum. If the gains decrease too slowly, then the variance of the process may remain too high. The gain function for δ is $a_m^\delta = a_m^\xi$, and the gain function for γ is $a_m^\gamma = a_m^\xi/2$. The specification of M and the gain functions is somewhat ad hoc but has been found to work well. We tried to find a theoretical justification for the choice of the gain function for δ, $a_m^\delta = a_m^\xi$, but could not. This choice is made on empirical grounds. In the following we justify the choice of the gain function for γ, $a_m^\gamma = a_m^\xi/2$. Fixing δ_k, we choose the gain function for γ such that the perturbations of γ_k and ξ_k have effects of similar magnitude on $\phi_k(\mathbf{x})$.

Set

$$w_k = \exp(\gamma_k - \tau^{-2}\|\mathbf{x} - \boldsymbol{\xi}_k\|^2).$$

Write $h_k = h_k(\mathbf{x}, y, \boldsymbol{\theta}) \min\{\sqrt{\mu}/c, a_0^\delta/a_m^\delta\}$. After some manipulations, we obtain that the perturbations of γ_k and ξ_k have the following combined effect on w_k:

$$\begin{aligned} w_k \leftarrow\ & w_k \exp\{a_m^\gamma h_k(\delta_k - f_K) + \\ & \{2a_m^\xi h_k(\delta_k - f_K) - (a_m^\xi h_k(\delta_k - f_K))^2\}\tau^{-2}\|\mathbf{x} - \boldsymbol{\xi}_k\|^2\}. \end{aligned}$$

The approximations below follow Hooper (1999). Recall that the parameter τ is set to the average distance between nearest neighbors among the K initial points $\boldsymbol{\xi}_k$. If $\mathbf{x}$ is midway between two reference points that are a distance of τ apart, then we could replace $\|\mathbf{x} - \boldsymbol{\xi}_k\|$ by $\tau/2$. If a_m^ξ is small, then $(a_m^\xi h_k(\delta_k - f_K))^2 \approx 0$ compared to other terms. With these crude approximations, the

exponent in the above equation becomes $a_m^{\gamma} h_k(\delta_k - f_K) + a_m^{\xi} h_k(\delta_k - f_K)/2$. The choice $a_m^{\gamma} = a_m^{\xi}/2$, makes the two terms equal and therefore the perturbations in γ_k and ξ_k will have effects of similar magnitude on $\phi_k(\mathbf{x})$.

The theory of stochastic approximation indicates that, after an initial transient phase, the training process typically converges toward a local optimum (Benveniste et al. 1990). There is no guarantee that a global optimum will be found, and replication of the process can produce varying results, but the algorithm typically yields reasonable results. As in H01, the quality of the estimator is improved and variation under replication is reduced by restarting the process; i.e., replicate the first 10% of the process ten times, minimizing the negative log-likelihood each time, then continue the process with the most promising vector of parameter values.

Finally, we note that, although the sequence converges in theory, we need to force the algorithm gains to approach zero at a faster rate after $M/2$ iterations. When the minimum of the two quantities in the updating functions is a_0/a_m, this may produce bias in the estimate $\hat{f}_K$. In such situations, we increase the number of iterations by one to correct eventual bias in the estimate $\hat{f}_K$.

2.2.2 Selection of K using Akaike Information Criterion

Once f_K is estimated for fixed K, we need a criterion to select an optimal value for K. We will use the Akaike Information Criterion (Akaike, 1973). The basic idea of Akaike Information Criterion (AIC) is to correct the log-likelihood of a fitted model for the effective number of parameters. Using AIC, we select K

to minimize:

$$\mathrm{AIC}(K) = -l(\hat{\boldsymbol{\theta}}_K) + p,$$

where

$$-l(\hat{\boldsymbol{\theta}}_K) = \sum_{i=1}^{n} (\exp(\hat{f}_K(\mathbf{x}_i)) - y_i \hat{f}_K(\mathbf{x}_i)) + \text{const}$$

is the negative log-likelihood evaluated at the maximum likelihood estimator, and $p = 1 + (K-1)(d+2)$ is the effective number of parameters. Therefore, AIC is equivalent to:

$$\mathrm{AIC}(K) = \sum_{i=1}^{n} (\exp(\hat{f}_K(\mathbf{x}_i)) - y_i \hat{f}_K(\mathbf{x}_i)) + (K-1)(d+2). \tag{2.10}$$

A straightforward search is used to minimize (2.10). The $\mathrm{AIC}(K)$ is evaluated for successive values of K, starting with $K = 1$. The search stops when the minimum AIC remains unchanged for m consecutive values of K. The estimate $\hat{K}$ therefore involves the calculation of $\hat{K} + m$ estimates $\hat{f}_K$. The stopping value $m = 3$ is adequate in most situations.

Many regression methods select from a large set of potential basis functions using forward selection and/or backward elimination strategies. ALB regression adopts a different approach. While K is selected by sequentially calculating $\hat{f}_K$, parameters are optimized separately for each K. The parameters and basis functions determining $\hat{f}_K$ play no role in the calculation of $\hat{f}_{K+1}$.

Computational speed is an important aspect when dealing with regression methods for high-dimensional data. Computation is reasonably fast for ALB models. particularly for large datasets. The computation time increases

Table 2.1: Time (in seconds) to calculate $\hat{f}$, including selection of $\hat{K}$. The first time corresponds to the Poisson version of ALB, while the second time, listed in parentheses, corresponds to the original H01 version.

	d			
$\hat{K}$	1	5	10	20
1	3(2)	5(3)	6(5)	9(8)
2	4(4)	6(6)	9(8)	14(13)
5	9(9)	15(15)	22(22)	43(38)
10	24(23)	40(39)	64(56)	94(88)

with $\hat{K}$ and d, but increases very slowly with n. Table 2.1 lists ALB model estimation times for a 850 MHz PC. In this table, the first time corresponds to the Poisson version of ALB, while the second time, listed in parentheses, corresponds to the original H01 version. Each value includes the total time needed to obtain $\hat{f}_K$ for $K = 1, \ldots, \hat{K} + 3$. The sample size was $n = 1000$. The time is roughly linear in d, with intercept and slope depending on $\hat{K}$, and roughly linear in $\hat{K}^2$, with slope depending on d. The sample size n has relatively little effect on time because of the sampling technique used in the training algorithm. The number of iterations M is proportional to $\sqrt{K}$, $M = 50000\sqrt{K}$, each iteration requires the evaluation of K distances, and each distance calculation time is proportional to the number of predictors d. We note that times for the Poisson version of ALB are slightly larger than for the original version, due to more complex updating functions and increase in the number of iterations. Times increase slowly with n because of increased overhead (data input and transformation, centering and scaling of the predictors, calculation of the AIC criterion for the Poisson version of ALB, or GCV

criterion for the original H01 version) and a tendency to select larger $\hat{K}$.

2.2.3 Illustrations with simulated and real data

In this section, we illustrate the method using simulated and real data sets.

Example 2.2.1 The first example is one dimensional ($d = 1$) and target function is an ALB function. A sample of size 1000 was generated with each predictor x generated from a Uniform$(-3, 3)$ distribution. The reference point parameterization was used to specify f_K: $\xi_1 = 1$, $\xi_2 = 0$, $\xi_3 = -1$, $\gamma_1 = \gamma_2 = \gamma_3 = 0$, $\delta_1 = 1$, $\delta_2 = 5$, $\delta_3 = 1$ and $\tau = 1$. For each x, a response y was generated from a Poisson distribution with mean $\exp(f(x))$. The estimate $\hat{f}$ is a linear combination of the $\hat{K} = 3$ basis functions plotted in Figure 2.2. A superimposed plot of the fitted function, $\exp(\hat{f})$ and the true function, $\exp(f)$ versus the predictor variable x displayed in Figure 2.1(a) shows an almost perfect fit. A plot of the difference between the fitted values $\exp(\hat{f})$ and the true mean function $\exp(f)$ versus the predictor x is also displayed in Figure 2.1(b). The values of the difference between the fitted values and the true mean are in the range -0.4 to 0.4.

Example 2.2.2 The second example is 4-dimensional and the target function is an ALB function with three basis functions. The reference point parameterization was used to specify f_K: $\xi_1 = (1, 0, 0, 0)$, $\xi_2 = (0, 1, 1, 2)$, $\xi_3 = (1, 0, 1, 0)$, $\gamma_1 = 1, \gamma_2 = 0, \gamma_3 = 0.5$, $\delta_1 = 4$, $\delta_2 = 0.5$, $\delta_3 = 2.5$ and $\tau = 1$. A sample of size 1000 was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-3, 3)^4$, and for each $\mathbf{x}$ a response y was generated from a Poisson distribution with mean $\exp(f(\mathbf{x}))$. The estimate $\hat{f}$ is a linear combination of

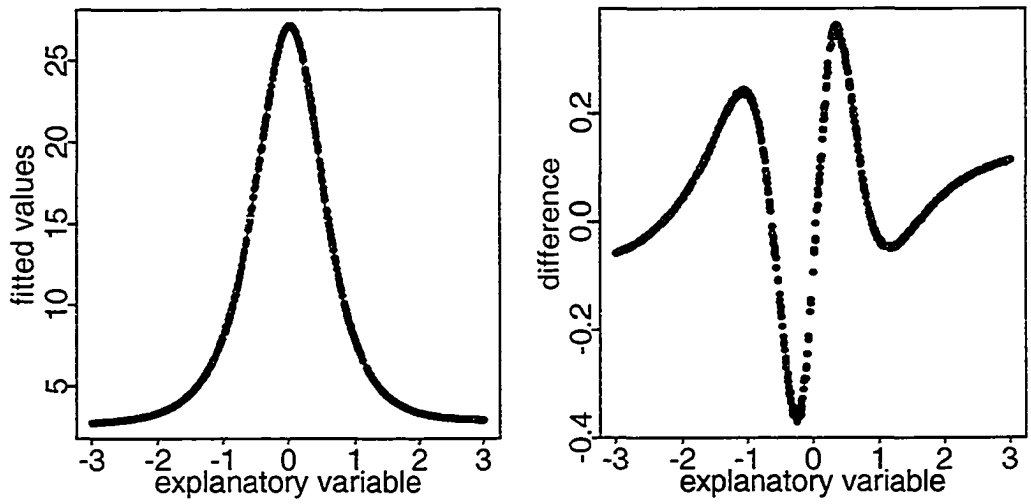

Figure 2.1: (a) Superimposed plot of the fitted function, $\exp(\hat{f})$ and the true function, $\exp(f)$ versus the predictor variable. (b) Difference $\exp(\hat{f}) - \exp(f)$ versus the predictor variable.

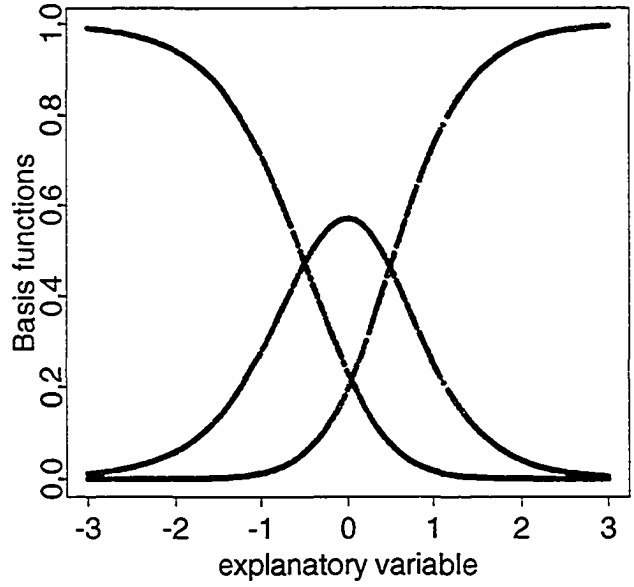

Figure 2.2: Basis functions for the ALB estimate, $\hat{K} = 3$.

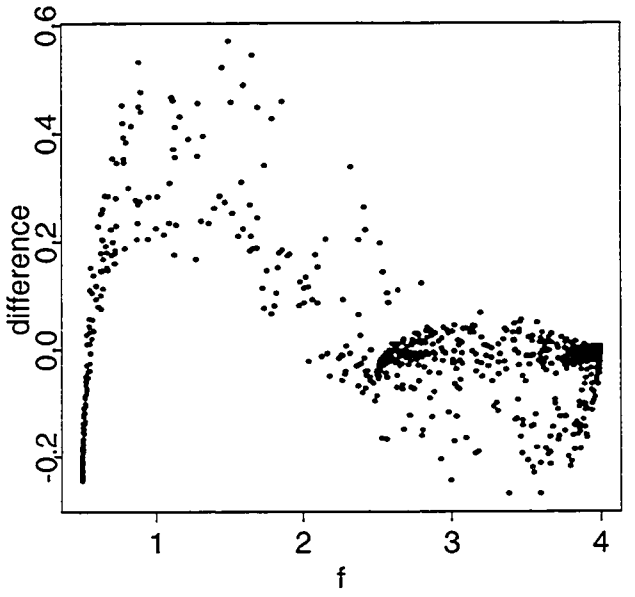

Figure 2.3: Difference $\hat{f} - f$ versus f.

the $\hat{K} = 3$ basis functions.

A plot of the difference $\hat{f} - f$ versus f is shown in Figure 2.3. The differences range from -.3 to .6. Smaller differences are observed for larger values of f.

Functions in three or more dimensions are difficult to visualize. H01 proposed a technique to visualize an ALB model in higher dimensions. Using the linear parameterization

$$\hat{\phi}_k(\mathbf{x}) = \exp(\hat{\alpha}_k + \hat{\beta}_k'\mathbf{x}) / \sum_{m=1}^{K} \exp(\hat{\alpha}_m + \hat{\beta}_m'\mathbf{x})$$

and dividing the numerator and denominator by $\exp(\hat{\alpha}_K + \hat{\beta}_K)$, the estimate $\hat{f}_K$ can be expressed as:

$$\hat{f}_K = \sum_{1}^{K} \hat{\delta}_k \exp(\hat{\alpha}_k - \hat{\alpha}_K + (\hat{\beta}_k - \hat{\beta}_K)'\mathbf{x}) / \sum_{m=1}^{K} \exp(\hat{\alpha}_m - \hat{\alpha}_K + (\hat{\beta}_m - \hat{\beta}_K)'\mathbf{x}).$$

There exist $r = \min(d, K-1)$ vectors $\mathbf{b}_1, \ldots, \mathbf{b}_r$ in $\Re^d$, spanning the subspace that is spanned by the contrasts $\hat{\beta}_1 - \hat{\beta}_K, \ldots, \hat{\beta}_{K-1} - \hat{\beta}_K$. The estimate $\hat{f}_K$ can thus be expressed as a function of r linear combinations $\mathbf{b}_1'\mathbf{x}, \ldots, \mathbf{b}_r'\mathbf{x}$. To visualize $\hat{f}_K$, directions in the covariate space that best represent variation in $\hat{f}_K$ are identified by carrying out a principal components analysis of the gradient sum-of-products matrix $\mathbf{G} = \sum \hat{g}(\mathbf{x}_i)\hat{g}(\mathbf{x}_i)'$, where $\hat{g}(\mathbf{x})$ represents the gradient of $\hat{f}_K(\mathbf{x})$. The gradient vectors $\hat{g}(\mathbf{x}_i)$ lie in the contrast subspace, so the rank of $\mathbf{G}$ is at most r. Let $e_1, \ldots, e_r$ be the eigenvalues of $\mathbf{G}$ in decreasing order and $\mathbf{b}_1, \ldots, \mathbf{b}_r$ a set of corresponding eigenvectors of length one. Then, for $j = 1, \ldots, r$, $\mathbf{G}\mathbf{b}_j = e_j\mathbf{b}_j$, which gives $\mathbf{b}_j'\mathbf{G}\mathbf{b}_j = e_j\mathbf{b}_j'\mathbf{b}_j$, or equivalently, $\sum \mathbf{b}_j'\hat{g}(\mathbf{x}_i)\hat{g}'(\mathbf{x}_i)\mathbf{b}_j = e_j$. We conclude that $\sum\{\mathbf{b}_j'\hat{g}(\mathbf{x}_i)\}^2 = e_j$, and therefore the first eigenvector $\mathbf{b}_1$ maximizes the sum of squared gradients $\sum\{\mathbf{b}'\hat{g}(\mathbf{x})\}^2$. If the first two eigenvalues are such that $(e_1 + e_2)/(e_1 + \ldots + e_r) \approx 1$, then a

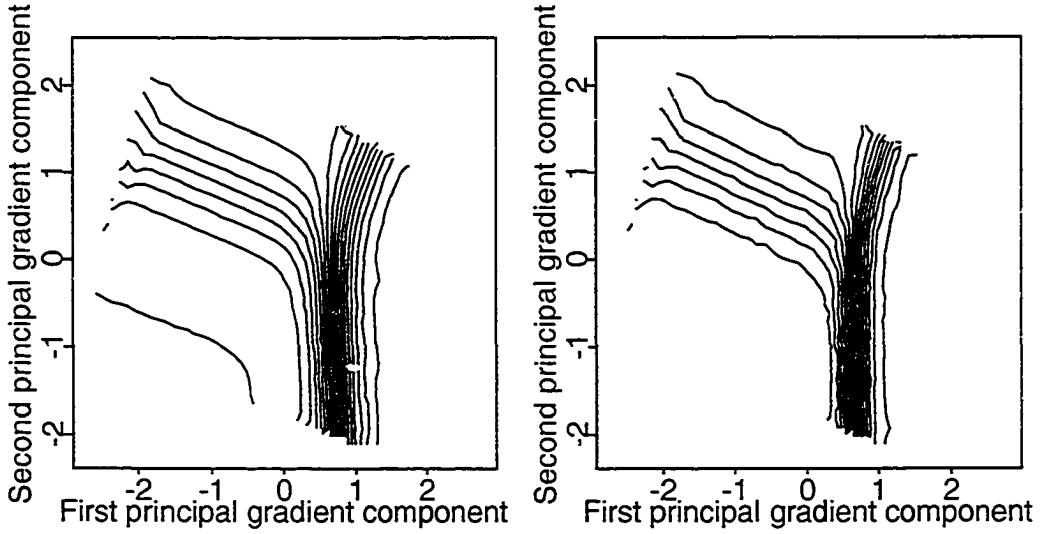

Figure 2.4: Contour plots of (a) $\hat{f}$ and (b) f as functions of two linear combinations (principal gradient components) of the original predictors.

plot of $\hat{f}_K$ versus $(\mathbf{b}_1'\mathbf{x}, \mathbf{b}_2'\mathbf{x})$ accounts for nearly all of the variation in $\hat{f}_K$, and therefore $\hat{f}_K$ can be visualized in a three-dimensional plot.

Back to the second example, the two contour plots in Figure 2.4 show f and $\hat{f}$ as functions of the two principal gradient components. We can see that $\hat{f}_K$ captures most of the curvature of the function f. We note that we did not use the mathematical functions when generating the contour plots. We used the values of the function evaluated at the 1000 data points in the sample, and interpolated them onto an evenly spaced grid of the two predictors. The contour plots of the mathematical functions are smoother than what appears in Figure 2.4. Although the plots are not smooth, they still give indication that the estimate captures most of the curvature of the true regression function.

Example 2.2.3 The third example is 2-dimensional and the target function

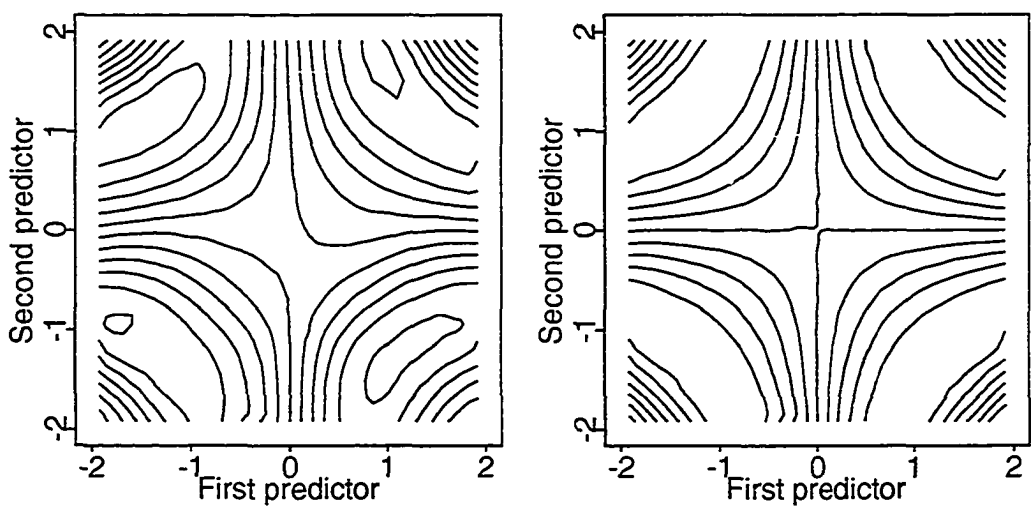

Figure 2.5: Contour plots of (a) $\hat{f}$ and (b) f as functions of the two original predictors.

is not an ALB function:

$$f(\mathbf{x}) = \sin(x_1 x_2) + 2.$$

A sample of size 1000 was generated with each predictor vector $\mathbf{x}$ generated uniformly on a hypercube $(-2, 2)^2$ and for each $\mathbf{x}$ a response y was generated from a Poisson distribution with mean $\exp(f(\mathbf{x}))$. The estimate $\hat{f}$ is a linear combination of the $\hat{K} = 10$ basis functions. Contour plots displayed in Figure 2.5 show f and $\hat{f}$ as functions of the two original predictors. The estimate $\hat{f}_K$ captured most of the curvature of the function f.

Example 2.2.4 The real data used in the fourth example are from the Butterfly Monitoring Scheme, which provides important information on butterfly population ecology and is based on transect counts at sites throughout Britain. At each site, an observer records all butterflies seen within prescribed limits

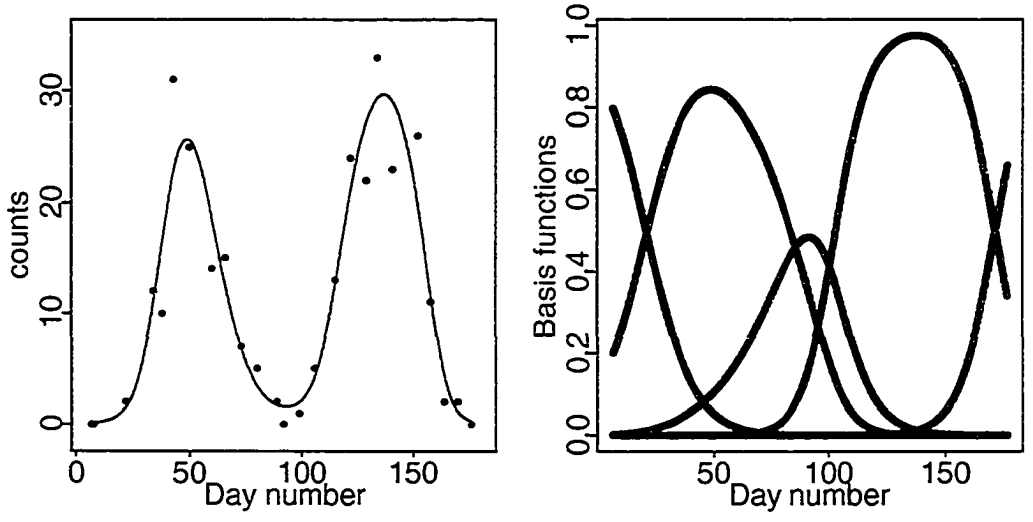

Figure 2.6: Butterfly transect count data. (a) ALB fit with counts superimposed versus Day number. (b) Basis functions for the ALB estimate.

along a fixed route (Pollard and Yates, 1993). The counts and Day number are recorded in 26 weeks from the beginning of April until the end of September, provided that weather conditions meet specified criteria. Figure 2.6(a) presents the ALB fit with the counts superimposed, for the Green-veined white species in site number 89, year 1998. The estimate $\hat{f}$ is a linear combination of the $\hat{K} = 5$ basis functions plotted in Figure 2.6(b).

A plot of the deviance residuals versus the fitted values and a normal probability plot of the deviance residuals are displayed in Figure 2.7. Several definitions for residuals are possible. Pierce and Schafer (1986) discuss the behavior of different types of residuals and argue that deviance residuals are most useful, being nearly normally distributed and a natural choice for likelihood

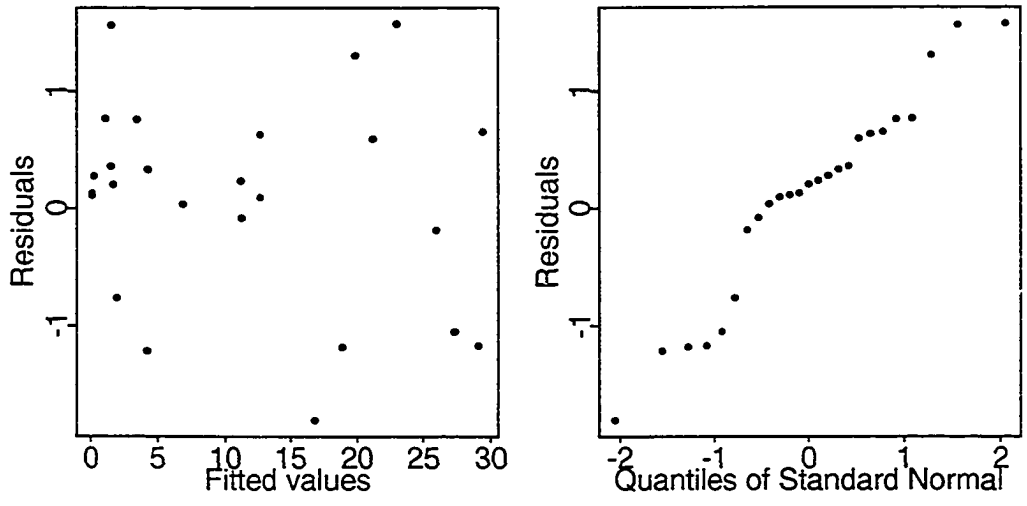

Figure 2.7: Butterfly transect count data. (a) Deviance residuals versus fitted values. (b) Normal probability plot of deviance residuals.

based methods. Deviance residuals are calculated based on the formula:

$$R_D(y,\mu) = \text{sign}(\text{y} - \hat{\mu})\{2[\text{y}\log\text{y} - \text{y}\log\hat{\mu} - \text{y} + \hat{\mu}]\}^{1/2},$$

where $y \log y = 0$. for $y = 0$, a correction motivated by

$$\lim_{y \to 0} y \log y = 0.$$

A plot of the deviance residuals versus the fitted values in Figure 2.7(a) indicates that the residuals are uniformly spread within a band around zero, in a range from −2 to 2. In Figure 2.7(b), a normal probability plot of the residuals indicates no serious departures from normality.

2.3 Predictive Accuracy

In this section we investigate the predictive accuracy of ALB models for Poisson data. A measure of prediction error based on the deviance residuals is used. The prediction error is useful for comparing predictive performance between models. Simulation studies were performed to compare predictive performance of ALB models and Generalized Additive models(GAM). We discuss examples where ALB models perform better than GAM and also examples where GAM performs better than ALB.

2.3.1 A measure of prediction error

In order to assess predictive accuracy, we fit the model on a data set $(\mathbf{x}_i, y_i), i = 1 \ldots, n$ and then evaluate the fit using an independent sample of responses $y_i^*, i = 1, \ldots, n$ generated under the same distribution. In other words, $(\mathbf{x}_i, y_i), i =$

$1, \ldots, n$ is the training set used to obtain the fit $\hat{\mu}$, and $(\mathbf{x}_i, y_i^*), i = 1, \ldots, n$ is the test set used for assessing predictive accuracy. The conditional distributions of y_i and y_i^* given $\mathbf{x}_i$ are the same. The closeness of y^* and $\hat{\mu}$ is a measure of goodness of fit.

For estimation based on maximum likelihood, it is natural to consider the deviance as a measure of closeness between the observed data and the fit:

$$\mathrm{D}(\mathbf{y}, \hat{\mu}) = 2\sum_{i=1}^{n}(y_i \log y_i - y_i \log \hat{\mu}_i - y_i + \hat{\mu}_i),$$

where $\mathbf{y} = (y_1, \ldots, y_n)$.

Let $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ be random with $y_i^*|\mathbf{x}_i \sim \text{Poisson}(\mu(\mathbf{x}_i)), i = 1, \ldots, n$. In the light of the above ideas, we consider the following quantity of interest as a measure for the prediction error:

$$\mathrm{PE} = \mathrm{E}\{\mathrm{D}(\mathbf{y}^*, \hat{\mu})\}. \tag{2.11}$$

The expectation in the expression above refers to conditional expectation, given the values of $\mathbf{x}_i, i = 1, \ldots, n$, taken with respect to the sampling distribution of $\hat{\mu}$ and $\mathbf{y}^*$. The fit $\hat{\mu}$ is calculated based on the training data $(\mathbf{x}_i, y_i), i = 1, \ldots, n$. The prediction error estimated in this way is similar to a mean predictive squared error, and under many replications of the data, it is a measure of how good the fit is.

We note that with simulation studies, the true mean μ is known and therefore a lower bound on the estimated prediction error can be obtained by using the true mean μ instead of the estimated mean $\hat{\mu}$ when evaluating the prediction error.

2.3.2 Simulation studies

In the following simulation studies we investigate the predictive performance of ALB models for Poisson data. A comparison between ALB models and GAM in terms of predictive performance is conducted.

In each example, we generated a set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of predictor vectors from the uniform distribution on a hypercube $(a, b)^d$. With this set of predictors fixed, we then generated 100 independent sets $\{y_1, \ldots, y_n\}$ of responses, where y_i has a Poisson distribution with mean $\mu(\mathbf{x}_i) = \exp(f(\mathbf{x}_i))$. The ALB estimate $\hat{\mu}$ was calculated for each sample and a measure of prediction error was evaluated over the independent sample of the remaining $99n$ observations. Averages of the prediction error over the 100 samples are reported in Table 2.2. The same measure was evaluated for GAM estimates. Details on smoothers used to fit GAM are given at the end of this section. Since the true mean μ is known, a lower bound on the prediction error is also provided. The lower bound is obtained by replacing $\hat{\mu}$ with μ in equation (2.12), and it is different for different sample sizes. Averages of $\hat{K}$ are also reported to check whether over-estimation of K occurs with increased sample size. Various examples were chosen to investigate how comparative performance depends on the function f, the dimensionality d and the sample size n. Some examples have several values of d to demonstrate the adverse effects of nuisance variables.

Example 2.3.1 Adaptive estimators should detect real structure where it exists and ignore spurious structure caused by random variation. The first example focuses on this goal by examining performance when the underlying

regression function f is constant. We would hope that, in most samples, ALB selects $\hat{K} = 1$. Indeed, ALB selected $\hat{K} = 1$ with frequencies 96, 98, 98, 92, 99, 90 and 96. Table 2.2 shows that accuracy improves with increased sample size. The accuracy deteriorates as d increases because of effects of nuisance variables.

Example 2.3.2 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-3,3)^d$. The function is a linear function of the first covariate:

$$f(\mathbf{x}) = 1 + x_1/4.$$

ALB selects $\hat{K} = 2$ in most samples, as expected. ALB and GAM prediction error measures are very close. Accuracy improves with increased sample size. The accuracy deteriorates as d increases because of effects of nuisance variables.

Example 2.3.3 In the third simulated study the function is an ALB function of the first covariate. A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-3,3)^d$. The reference point parameterization was used to specify f_K: $\xi_1 = 1$, $\xi_2 = 0$, $\xi_3 = -1$, $\gamma_1 = \gamma_2 = \gamma_3 = 0$, $\delta_1 = 1$, $\delta_2 = 5$, $\delta_3 = 1$ and $\tau = 1$. The underlying regression function f is bell-shaped, and therefore three basis functions would be needed to estimate it. ALB selects $\hat{K} = 3$ in most samples, as expected. A comparison of the average prediction error for ALB and GAM indicates that ALB performs better, as expected, since the underlying regression function is an ALB function. When adding four nuisance variables, $(d = 5)$ the accuracy deteriorates affecting ALB more than GAM, especially for the smaller sam-

ple size, $n = 50$. Methods that fit each axis separately, such as GAM, are expected to perform better with this example. However, the affine invariance property of ALB indicates that a rotation of the coordinate axes would create higher-order interactions, adversely affecting GAM but not ALB.

Example 2.3.4 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-1, 1)^d$. The function is not an ALB function:

$$f(\mathbf{x}) = \frac{2\sin(\pi(x_1 + 1)/2)}{(x_1/2 + 1)}$$

The prediction errors for the two methods are very close, especially with larger sample sizes, $n = 100$ and $n = 200$. For a smaller sample size, $n = 50$, GAM performs slightly better. When adding four nuisance variables, $(d = 5)$ the accuracy deteriorates affecting ALB more than GAM, especially for the smaller sample size, $n = 50$. Methods that fit each axis separately, such as GAM, are expected to perform better with this example. However, the affine invariance property of ALB indicates that a rotation of the coordinate axes would create higher-order interactions, adversely affecting GAM but not ALB.

Example 2.3.5 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-1, 1)^d$. The target function is an additive function of the first two covariates:

$$f(\mathbf{x}) = 1.5x_1^2 + 1.5\frac{\sin(\pi(x_2 + 1)/2)}{(x_2/2 + 1)}.$$

GAM performs better than ALB, as expected since the underlying regression function is an additive function. The difference in the prediction errors is smaller as n increases: .11 for $n = 50$, .05 for $n = 100$ and .03 for $n =$

200. When adding three nuisance variables, ($d = 5$) the accuracy deteriorates affecting ALB more than GAM, especially for the smaller sample size, $n = 50$. Methods that fit each axis separately, such as GAM, are expected to perform better with this example. However, the affine invariance property of ALB indicates that a rotation of the coordinate axes would create higher-order interactions, adversely affecting GAM but not ALB.

Example 2.3.6 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-3, 3)^4$. The target function is an ALB function defined on a 3-dimensional projection of $\Re^4$; ie., $f(\mathbf{x}) = f_K(\mathbf{z})$ where $K = 5$, $\mathbf{z} = (z_1, z_2, z_3)'$,

$$z_1 = \sqrt{3}(x_1 + x_2 + x_3 + x_4 - 2)$$

$$z_2 = \sqrt{3}(x_1 + x_2 - x_3 - x_4)$$

$$z_3 = \sqrt{3}(x_1 - x_2 + x_3 - x_4)$$

The reference point parameterization is used to specify f_K: $\xi_1 = (1, 0, 0)'$, $\xi_2 = (-1, 0, 0)'$, $\xi_3 = (0, 1, 0)'$, $\xi_4 = (0, 0, 1)'$, $\xi_5 = (0, 0, 0)'$, $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 0$, $\delta_1 = \delta_2 = .5$, $\delta_3 = \delta_4 = 3.5$, $\delta_5 = 0$ and $\tau = 1$. The function f can be expressed as an ALB function of $\mathbf{x}$ and has interactions of all orders among the four covariates. The performance of ALB is substantially better than that of GAM. GAM has difficulty modeling higher order interactions, even when n is large. With $n = 50$, there are not enough degrees of freedom for GAM to model three-way interactions. With $n = 200$, there are not enough degrees of freedom for GAM to get all three-way interactions. In further simulations, the average prediction error for GAM remains roughly constant as n increases

Table 2.2: Performance measures: prediction error averages from 100 replicated samples of size n for Poisson version of ALB models, Generalized Additive model together with the corresponding lower bound of the prediction error.

No.	d	n	$\hat{K}$	lower bound	ALB	GAM
1	1	50	1.04	1.2124	1.2424	1.2829
	1	100	1.02	1.1990	1.2080	1.2271
	1	200	1.02	1.2061	1.2113	1.2209
	5	50	1.08	1.2161	1.2634	1.4159
	5	200	1.01	1.2083	1.2139	1.2649
	10	100	1.10	1.2049	1.2439	1.4049
	10	200	1.04	1.2187	1.2304	1.3271
2	1	50	2.03	1.2840	1.3545	1.3505
	1	100	2.02	1.2864	1.3238	1.3190
	1	200	2.03	1.2641	1.2838	1.2817
	5	50	2.06	1.2936	1.5179	1.5145
	5	100	2.13	1.2751	1.3854	1.3761
3	1	50	3.01	1.2611	1.3820	1.4047
	1	100	3.00	1.2467	1.3028	1.3234
	1	200	3.00	1.2511	1.2802	1.2910
	5	50	3.26	1.2685	1.7258	1.5593
	5	200	3.02	1.2391	1.3952	1.3211
4	1	50	3.02	1.2710	1.3862	1.3797
	1	100	3.05	1.2741	1.3354	1.3321
	1	200	3.01	1.2578	1.2884	1.2901
	5	50	3.24	1.2385	1.5795	1.4855
	5	200	3.21	1.2554	1.3422	1.3329
5	2	50	4.08	1.2608	1.5430	1.4312
	2	100	4.17	1.2534	1.4063	1.3522
	2	200	4.54	1.2437	1.3258	1.2936
	5	50	4.08	1.2362	1.7628	1.5051
	5	200	4.26	1.2525	1.4042	1.3336
6	4	50	4.03	1.2619	1.7703	4.8466
	4	100	4.23	1.2369	1.7172	3.0891
	4	200	4.28	1.2548	1.6868	2.5479

from 200 to 2000.

Now we provide details regarding smoothers that have been used when fitting the above GAM's. We used the 'gam' function in R, package 'mgcv'. For the first five examples, no attempts to model interactions were made, since they were not present in the true regression function. Smooth terms are represented using penalized regression splines with smoothing parameters selected by either a Generalized Cross-Validation criterion(GCV), or an Un-Biased Risk Estimator criterion(UBRE) which is in practice an approximation to AIC, (Wood, 2003). Smoothing parameters are chosen to minimize the GCV or UBRE score for the model. To model interactions, multi-dimensional smooths are available using penalized thin plate regression splines. For the last example, with $n = 50$, we used a two-dimensional smooth surface based on thin plate regression splines to model the interaction between two of the predictors. When trying to add one more two-dimensional surfaces, there were not enough degrees of freedom to model it. With $n = 100$, we were able to include a three-dimensional smooth surface to model the interaction between three of the predictors and a two-dimensional interaction surface, but there were not enough degrees of freedom to add other second or third order interactions. With $n = 200$, we used three-dimensional smooth surfaces based on thin plate regression splines to model the interaction between three of the predictors. We were able to include two third-order interactions, but there were not enough degrees of freedom to add a third one, or a second order interaction. We also modeled three-way interactions using locally weighted

surface smoothers in S-Plus. For a sample size of $n = 200$, we were able to model only two such interactions, there were not enough degrees of freedom to add another surface. The results were worse than when using penalized thin plate regression splines in R.

2.4 Standard Errors

This section presents approximate standard errors for the ALB estimator. Given the adaptive nature of the ALB models, difficulties in deriving standard errors are expected. We derive approximate standard errors for the fit assuming that the number of basis functions is fixed and employing a standard asymptotic technique in nonlinear regression analysis. Our derivation extends that of H01 from Normal errors to the more general quasi-likelihood context. Coverage probabilities will be estimated in several simulation studies.

We tried to develop some intuition on a different technique for estimating standard errors for the fit. This was an attempt that failed to yield useful standard errors. I mention it because it was instructive to see why the method fails. The technique is derived and discussed in the Appendix.

The bootstrap is another technique to obtain standard errors. When thinking of implementing bootstrap, we need to keep in mind that the computation complexity when using stochastic approximation is high. The algorithm is reasonably fast when used to fit the model once, but the computation is too complex for the algorithm to be iterated 10000 times, as would be required by the bootstrap technique. Alternative non-stochastic optimization techniques,

such as Newton-Raphson, might be used to increase computational speed in smaller problems, allowing the use of bootstrap errors. Although we have not implemented this method, we mention here its advantages and disadvantages. First, we note that to implement Newton methods in the ALB context, adjustments have to be made to correct for the Hessian singularities caused by redundancies in the ALB parameterization. Although the convergence properties of Newton methods are unsurpassed, they are not necessarily so well behaved away from the global optimum. Good starting values are needed to ensure convergence towards a global optimum. The computation time increases with sample size and sometimes convergence is not achieved because of the large number of parameters.

2.4.1 Approximate standard errors

In this section, approximate standard errors for the ALB estimator are derived. Our derivation assumes that $K = \hat{K}$ is fixed and $f(\mathbf{x}) = f_K(\mathbf{x})$ for some parameter vector $\boldsymbol{\theta} \in \Re^{(2+d)K}$. Approximate standard errors are derived in a general quasi-likelihood context. In the following, we assume that the components of the n-dimensional response vector $\mathbf{y} = (y_1, \ldots, y_n)$ are independent with conditional mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and covariance matrix $\sigma^2 V(\mu)$, where σ^2 may be unknown and $V(\boldsymbol{\mu})$ is a diagonal matrix of known functions, $V(\boldsymbol{\mu}) = \text{diag}(V_1(\mu_1), ..., V_n(\mu_n))$, $\mu_i = \mu(\mathbf{x}_i)$. There are no other assumptions about the conditional distribution of y given $\mathbf{x}$. For the case where the conditional distribution of y_i given $\mathbf{x}_i$ is Poisson with mean μ_i, we have $\sigma^2 = 1$ and $V(\boldsymbol{\mu}) = \text{diag}(\mu_1, \ldots, \mu_n)$.

The quasi-likelihood function, for one observation, is given by:

$$Q(\boldsymbol{\theta}|\mathbf{x}, y) = \int_y^{\mu} \frac{y-t}{\sigma^2 t} dt.$$

An estimate of the p-dimensional parameter vector $\boldsymbol{\theta}$ is obtained by solving the quasi-likelihood equations:

$$U(\boldsymbol{\theta}) = \frac{dQ(\boldsymbol{\mu};\mathbf{y})}{d\boldsymbol{\theta}} = D'V^{-1}\frac{(\mathbf{y}-\boldsymbol{\mu})}{\sigma^2} = 0,$$

where D is an $n \times p$ matrix of partial derivatives:

$$\begin{aligned} D &= (D(\mathbf{x}_1, \boldsymbol{\theta}), ..., D(\mathbf{x}_n, \boldsymbol{\theta}))', \\ D(\mathbf{x}_i, \boldsymbol{\theta}) &= \frac{\partial \mu(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = (\frac{\partial \mu(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}, ..., \frac{\partial \mu(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_p})', \text{ for } i = 1, \ldots, n \end{aligned}$$

Standard errors of the fit $\hat{\mu}(\mathbf{x}) = \exp(\hat{f}_K(\mathbf{x}))$ are obtained using the Delta method, following H01. The regression function is approximated locally near $\hat{\boldsymbol{\theta}}$ by a linear function of $\boldsymbol{\theta}$,

$$\mu(\mathbf{x}, \boldsymbol{\theta}) \approx \hat{\mu}(\mathbf{x}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' \frac{\partial \mu(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \hat{\mu}(\mathbf{x}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})' D(\mathbf{x}, \hat{\boldsymbol{\theta}}).$$

Hence, given $\mathbf{x}$, the standard deviation of $\hat{\mu}(\mathbf{x})$ is estimated by

$$\text{se}\{\hat{\mu}(\mathbf{x})\} = [D'(\mathbf{x}, \hat{\boldsymbol{\theta}})\text{cov}(\hat{\boldsymbol{\theta}})D(\mathbf{x}, \hat{\boldsymbol{\theta}})]^{1/2}. \tag{2.12}$$

All we need now is an estimate of $\text{cov}(\hat{\boldsymbol{\theta}})$. We discuss here three approaches to estimate $\text{cov}(\hat{\boldsymbol{\theta}})$.

The first approach extends H01 in the quasi-likelihood context. The information matrix for $\boldsymbol{\theta}$ is given by

$$i(\boldsymbol{\theta}) = -E(\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}) = \frac{1}{\sigma^2} D'V^{-1}D = \sum_{i=1}^{n} \frac{D(\mathbf{x}_i, \boldsymbol{\theta})V_i^{-1}(\mu_i)D'(\mathbf{x}_i, \boldsymbol{\theta})}{\sigma^2}$$

Under regularity conditions, (Lehmann, 1998, page 469) the asymptotic covariance matrix of the maximum likelihood estimator is given by the inverse of the information matrix. Redundancies in the ALB parameterization present a problem here, since the information matrix is not invertible. As in H01, this problem is resolved by multiplying the diagonal elements of the information matrix by a constant close to 1, and leaving the off-diagonal elements unchanged. An estimate of $\mathrm{cov}(\hat{\boldsymbol{\theta}})$ is then given by:

$$\mathrm{cov}(\hat{\boldsymbol{\theta}}) = \sigma^2(D'V^{-1}D + k\Delta)^{-1}, \tag{2.13}$$

where $\Delta = \mathrm{diag}(D'V^{-1}D)$. A small constant $k = .01$ is used in the above formula. Notice that the operation $D'V^{-1}D + .01\Delta$ has the effect of multiplying the diagonal elements of $D'V^{-1}D$ by 1.01 and leaving the off-diagonal elements unchanged.

We now present the second approach to estimate $\mathrm{cov}(\hat{\boldsymbol{\theta}})$. An alternative solution to the redundancies in the parameterization is to fix some of the parameters, get rid of the redundancies in the parameterization, and apply the usual asymptotic techniques to calculate $\mathrm{cov}(\hat{\boldsymbol{\theta}})$. An interesting question is whether the two approaches yield essentially the same results. We start by proving this result for the usual linear regression setting, with identity link and constant variance assumption.

Theorem 2.3 *Assume* $\mathrm{E}(\mathbf{y}) = X_1\boldsymbol{\beta}$, $\mathrm{cov}(\mathbf{y}) = \sigma^2 I_n$, *where* X_1 *is* $n \times p$ *matrix with full rank* p. *Let* $\boldsymbol{\eta}$ *be a* q*-dimensional vector and* X_2 *be a* $n \times q$ *matrix, such that* $q > p$, $\mathrm{span}(X_1) = \mathrm{span}(X_2)$ *and* $X_1\boldsymbol{\beta} = X_2\boldsymbol{\eta}$. *Then,*

$$\boldsymbol{\beta} = A\boldsymbol{\eta}, \text{ where } A = (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\eta} \text{ is a } p \times q \text{ matrix}.$$

Let $\Sigma_{\hat{\beta}} = \text{cov}(\hat{\beta}) = \sigma^2(X_1'X_1)^{-1}$. *We approximate* $\text{cov}(\hat{\eta})$ *by* $\sigma^2(X_2'X_2+k\Delta)^{-1}$, *where* $\Delta = \text{diag}(X_2'X_2)$. *Then,* $\Sigma_{\hat{\beta}}$ *can be approximated by*

$$\tilde{\Sigma}_{\hat{\beta}} = \sigma^2 A(X_2'X_2 + k\Delta)^{-1}A',$$

and

$$\lim_{k \to 0} \tilde{\Sigma}_{\hat{\beta}} = \Sigma_{\hat{\beta}}.$$

PROOF. Using the Spectral Decomposition Theorem, $X_2'X_2 = PDP'$, where P is a $q \times q$ orthogonal matrix and D is a $q \times q$ diagonal matrix. Moreover, since X_2 has rank p,

$$D = \text{ diag } (d_1, \ldots, d_p, 0, \ldots, 0) = \begin{pmatrix} D_{11} & 0_{p\times(q-p)} \\ 0_{(q-p)\times p} & 0_{(q-p)\times(q-p)} \end{pmatrix},$$

where $D_{11} = \text{diag}(d_1, \ldots, d_p)$. It follows that the last $q-p$ columns of X_2P are all zeros; i.e., $X_2P = (E_{q\times p} : 0_{q\times(q-p)})$. We have,

$$\begin{aligned} X_2'X_2 + k\Delta &= PDP' + k\Delta, \\ &= P(D + kP'\Delta P)P', \\ &= P(D + kC)P', \text{ where } C = P'\Delta P. \end{aligned}$$

Then, $\tilde{\Sigma}_{\hat{\beta}} = \sigma^2 AP(D + kC)^{-1}P'A'$. Since $AP = (X_1'X_1)^{-1}X_1'X_2P$ and the last $(q-p)$ columns of X_2P are all zeros, it follows that the last $(q-p)$ columns of AP are all zeros as well; i.e., $AP = (F_{p\times p} : 0_{p\times(q-p)})$.

Now let us look more closely at $D + kC$, and its inverse. Set $H_{q\times q} = D + kC$. Both matrices $H_{q\times q}$ and $C_{q\times q}$ can be decomposed as follows:

$$H = \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \text{ and } C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

where H_{11} and C_{11} are $p \times p$ matrices, H_{12} and C_{12} are $p \times (q-p)$ matrices, H_{21} and C_{21} are $(q-p) \times p$ matrices, and H_{22} and C_{22} are $q \times q$ matrices. With these notations:

$$\begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} = \begin{pmatrix} D_{11} + kC_{11} & kC_{12} \\ kC_{21} & kC_{22} \end{pmatrix}.$$

Let us now look more closely at H^{-1}. Set

$$H^{-1} = \begin{pmatrix} H^{11} & H^{12} \\ H^{21} & H^{22} \end{pmatrix}.$$

A matrix computation result gives the following expression for H^{11}:

$$H^{11} = (H_{11} - H_{12}H_{22}^{-1}H_{21})^{-1}.$$

Using the standard notation $H_{11\cdot 2} = H_{11} - H_{12}H_{22}^{-1}H_{21}$, we have:

$$\begin{aligned} H^{11} &= (D_{11} + kC_{11} - kC_{12}(kC_{22})^{-1}kC_{21})^{-1} \\ &= (D_{11} + kC_{11\cdot 2})^{-1}, \end{aligned}$$

and therefore,

$$\lim_{k \to 0} H^{11} = D_{11}^{-1}.$$

Now let us get back to deriving $\tilde{\Sigma}_{\hat{\beta}}$. We have:

$$\begin{aligned} \tilde{\Sigma}_{\hat{\beta}} &= \sigma^2 (F_{p\times p} : 0_{p\times(q-p)}) H^{-1} (F_{p\times p} : 0_{p\times(q-p)})' \\ &= \sigma^2 F H^{11} F', \end{aligned}$$

and therefore,

$$\lim_{k \to 0} \tilde{\Sigma}_{\hat{\beta}} = \sigma^2 F D_{11}^{-1} F'.$$

On the other hand, we have:

$$\Sigma_{\hat{\beta}} = \sigma^2 (X_1' X_1)^{-1}.$$

Since, span(X_1) = span(X_2), there exists a $q \times p$ matrix G, such that $X_1 = X_2 G$. It follows that:

$$\begin{aligned} (X_1' X_1)^{-1} &= G' X_2' X_2 G \\ &= G' P D P' G \\ &= G' P_1 D_{11} P_1' G, \end{aligned}$$

where P_1 is the $q \times p$ matrix given by the first p columns of P. We note that since $X_1' X_1$ is invertible, $G' P_1$ is also invertible, and therefore:

$$(X_1' X_1)^{-1} = (P_1' G)^{-1} D_{11}^{-1} (G' P_1)^{-1}.$$

Now it remains to prove that $F = (P_1' G)^{-1}$. We have:

$$\begin{aligned} F &= (X_1' X_1)^{-1} X_1' X_2 P_1 \\ &= (P_1' G)^{-1} D_{11}^{-1} (G' P_1)^{-1} G' X_2' X_2 P_1 \\ &= (P_1' G)^{-1} D_{11}^{-1} (G' P_1)^{-1} G' P D P' P_1 \\ &= (P_1' G)^{-1} D_{11}^{-1} (G' P_1)^{-1} G' P_1 D_{11} P_1' P_1 \\ &= (P_1' G)^{-1}. \end{aligned}$$

□

We now justify this result for the ALB model in the quasi-likelihood context. Fix ξ_K and γ_K and shift $\gamma_k = \gamma_k - \gamma_K$, $\xi_k = \xi_k - \xi_K$ for $k =$

$1, \ldots, K-1$, to get rid of the redundancies in the ALB parameterization. Set

$$\begin{aligned}
\boldsymbol{\theta_1}' &= (\delta_1, \gamma_1 - \gamma_K, \boldsymbol{\xi}_1' - \boldsymbol{\xi}_K', \ldots, \delta_{K-1}, \gamma_{K-1} - \gamma_K, \boldsymbol{\xi}_{K-1}' - \boldsymbol{\xi}_K', \delta_K) \\
\boldsymbol{\beta} &= \boldsymbol{\theta_1} \\
\boldsymbol{\eta} &= \boldsymbol{\theta}
\end{aligned}$$

and

$$\begin{aligned}
X_1 &= \left(\frac{\partial \mu(\mathbf{x}_i, \boldsymbol{\theta}_1, 0, \mathbf{0}_d)}{\partial \boldsymbol{\theta}_{1,j}}\right)_{ij} \text{an } n \times (1 + (K-1)(d+2)) \text{ matrix} \\
X_2 &= \left(\frac{\partial \mu(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right)_{ij} \text{an } n \times K(d+2) \text{ matrix.}
\end{aligned}$$

Note that X_1 is obtained from X_2 suppressing the components corresponding to derivatives of $\mu(\mathbf{x}_i, \boldsymbol{\theta})$ with respect to γ_K and $\boldsymbol{\xi}_K$. With these notations, we obtain the following expression:

$$\Sigma_{\hat{\boldsymbol{\beta}}} = \text{cov}(\hat{\boldsymbol{\theta}}_1) = \sigma^2 (X_1' V^{-1} X_1)^{-1}. \tag{2.14}$$

We have $\text{span}(X_1) = \text{span}(X_2)$ and $X_1 \boldsymbol{\beta} = X_2 \boldsymbol{\eta}$, proof given in the next paragraph. It follows that:

$$\boldsymbol{\beta} = (X_1' V^{-1} X_1)^{-1} X_1' V^{-1} X_2 \boldsymbol{\eta} = A \boldsymbol{\eta},$$

where $A = (X_1' V^{-1} X_1)^{-1} X_1' V^{-1} X_2$. Now, we follow the same steps as in Theorem 2.3 and adjust for the quasi-likelihood context. We approximate $\text{cov}(\hat{\boldsymbol{\theta}}) = \text{cov}(\hat{\boldsymbol{\eta}})$ by $\sigma^2 (X_2' V^{-1} X_2 + k\Delta)^{-1}$, where $\Delta = \text{diag}(X_2' V^{-1} X_2)$. Then, $\Sigma_{\hat{\boldsymbol{\beta}}}$ can be approximated by

$$\tilde{\Sigma}_{\hat{\boldsymbol{\beta}}} = \sigma^2 A (X_2' V^{-1} X_2 + k\Delta)^{-1} A',$$

and

$$\lim_{k \to 0} \tilde{\Sigma}_{\hat{\beta}} = \Sigma_{\hat{\beta}}.$$

The proof of this result and the matrix derivations follow analogous to the proof of Theorem 2.3.

We now prove that $\text{span}(X_1) = \text{span}(X_2)$ and $X_1\boldsymbol{\beta} = X_2\boldsymbol{\eta}$. To simplify the derivations, we prove these results for a two dimensional vector $\boldsymbol{\theta} = \boldsymbol{\eta} = (\eta_1, \eta_2)'$ and $\beta = \eta_1 - \eta_2$. The proof for vectors of several dimensions follow easily analogous to the proof for two-dimensions. We have:

$$\begin{aligned} X_1 &= (g_1(\mathbf{x}_i))_{ij} \text{ an } n \times 1 \text{ matrix} \\ X_2 &= (g_j(\mathbf{x}_i))_{ij} \text{ an } n \times 2 \text{ matrix} , \end{aligned}$$

where

$$g_j(\mathbf{x}) = \frac{\partial \mu(\mathbf{x}, \boldsymbol{\theta}_j)}{\partial \theta_j}.$$

Since $\mu(\mathbf{x}, \eta_1, \eta_2) = \mu(\mathbf{x}, \eta_1 - \eta_2, 0)$, it follows that $g_1 + g_2 \equiv 0$. Hence, $\text{span}(X_1) = \text{span}(X_2)$ and $X_1\boldsymbol{\beta} = X_2\boldsymbol{\eta}$.

Here is the third approach to estimate $\text{cov}(\hat{\boldsymbol{\theta}})$. It is well known that the maximum likelihood estimator is asymptotically equivalent to the estimator obtained by a single Newton-Raphson step starting at the true $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta} + (D'V^{-1}D)^{-1}D'V^{-1}(\mathbf{y} - \boldsymbol{\mu}). \tag{2.15}$$

Here $\hat{\boldsymbol{\theta}}_1$ is different than in the previous paragraph. We derive $\text{cov}(\hat{\boldsymbol{\theta}}_1)$ and use it to approximate $\text{cov}(\hat{\boldsymbol{\theta}})$. Under regularity conditions, (Lehmann, 1998, page 469) the asymptotic covariance matrix of the maximum likelihood estimator is

given by the inverse of the information matrix. The redundant ALB parameterization of $\boldsymbol{\theta}$ presents a serious problem here, since the information matrix is singular. Even if some parameters were fixed to remove the redundancies, further difficulties might arise from multicollinearity among the estimated basis functions. These problems are addressed using a ridge regression technique.

Using the notation, $F = \sigma^{-1}V^{-1/2}D$ and $\mathbf{z} = \sigma^{-1}V^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$, equation (2.15) becomes:

$$\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta} + (F'F)^{-1}F'\mathbf{z}. \tag{2.16}$$

To correct for the singularities in $F'F$, we use the Levenberg-Marquardt compromise, (Bates and Watts 1988, Section 3.5.2). The approach is summarized below. Set $\Delta = \text{diag}(D'V^{-1}D)$ and

$$\tilde{F} = \begin{pmatrix} F \\ \sigma^{-1}k^{1/2}\Delta^{1/2} \end{pmatrix} \quad \text{and} \quad \tilde{\mathbf{z}} = \begin{pmatrix} \mathbf{z} \\ 0_{p\times 1} \end{pmatrix},$$

where k is a small positive constant. With this notation, equation (2.16) is replaced by

$$\tilde{\boldsymbol{\theta}}_1 = \boldsymbol{\theta} + (\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{\mathbf{z}}.$$

Further calculations give

$$\tilde{\boldsymbol{\theta}}_1 = \boldsymbol{\theta} + (D'V^{-1}D + k\Delta)^{-1}D'V^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

and therefore the covariance of $\tilde{\boldsymbol{\theta}}_1$ becomes

$$\text{cov}(\tilde{\boldsymbol{\theta}}_1) = \sigma^2(D'V^{-1}D + k\Delta)^{-1}D'V^{-1}D(D'V^{-1}D + k\Delta)^{-1}. \tag{2.17}$$

We now look more closely at equations (2.13) and (2.17). Let $A = D'V^{-1}D$. $B = D'V^{-1}D + k\Delta$ and $C = k\Delta$. Then $B = A + C$ and $B^{-1}AB^{-1} =$

$B^{-1}(A+C-C)B^{-1} = B^{-1}-B^{-1}CB^{-1}$. Using the ordering of positive definite matrices, it follows that $B^{-1}AB^{-1} < B^{-1}$, and therefore the approximate standard errors based on equation (2.17) are smaller than the ones based on equation (2.13), leading to more liberal confidence intervals. Based on these considerations, we decided to use the approximate standard errors based on equation (2.13).

If σ^2 is unknown, its conventional estimator is a moment estimator based on the residual vector $y - \hat{\mu}$, namely

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/V_i(\hat{\mu}_i). \tag{2.18}$$

The interval $\hat{\mu}(\mathbf{x}) \pm 2\text{se}\{\hat{\mu}(\mathbf{x})\}$ provides an approximate confidence interval for $\mu(\mathbf{x})$ with nominal 95% coverage probability. One may note four potential problems with this simple confidence interval. First, the ridge modification to $D'V^{-1}D$ produces a slight downward bias in the standard error. Second, the quality of the linear approximation of $\mu(\mathbf{x}) = \exp(f_K)$ may be poor. Third, standard errors account for variance but not bias. If μ is poorly approximated by $\exp(f_K)$ then $\exp(\hat{f}_K(\mathbf{x}))$ may have substantial bias relative to its standard deviation. Fourth, and perhaps most important, the derivation assumes that K is fixed. The effect of adaptive selection is unknown. It seems likely that variation in $\hat{K}$ will increase variation in $\hat{\sigma}$, and hence in the standard error.

The simulation studies in Section 2.4.2 suggest that the confidence intervals tend to be liberal. Given a nominal 95% confidence level, the coverage probabilities in our examples (averaged over $\mathbf{x}_1, \ldots, \mathbf{x}_n$) are between 86% and

95%.

2.4.2 Simulation studies on standard errors

In each example, a set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of predictor vectors was generated from the uniform distribution on a hypercube $(a, b)^d$. With this set of predictors fixed, we then generated 100 independent sets $\{y_1, \ldots, y_n\}$ of responses, where y_i has a Poisson distribution with mean $\mu(\mathbf{x}_i) = \exp(f(\mathbf{x}_i))$. The ALB estimate $\hat{\mu}$ and the standard errors of the fit were calculated for each sample, according to the formula on the last paragraph. Averages over the sample of the observed coverage probability of a nominal 95% confidence interval for $\mu(\mathbf{x}) = \exp(f(\mathbf{x}))$ are reported in Table 2.3. Also, a measure of variability for each coverage probability is reported. Various examples were chosen to investigate how the coverage probability depends on the function f, the dimensionality d and the sample size n. Some examples have several values of d to demonstrate the adverse effects of nuisance variables.

We denote by $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ the covariate vectors representing a sample of size n. For $i = 1, ... n$, we denote by π_i the population coverage probability of a nominal 95% confidence interval for $\mu(\mathbf{x}_i)$, and by p_i the sample proportion, or observed coverage probability of a nominal 95% confidence interval for $\mu(\mathbf{x}_i)$ obtained from the $m = 100$ replicates, i.e.:

$$p_i = \frac{1}{m} \sum_{l=1}^{m} I[|\mu(\mathbf{x}_i) - \hat{\mu}_l(\mathbf{x}_i)| \leq 2\text{se}\{\hat{\mu}_l(\mathbf{x}_i)\}],$$

where $\hat{\mu}_l$ is the fit based on the l-th sample. Let $\bar{\pi} = \sum \pi_i/n$ and $\bar{p} = \sum p_i/n$. We note that $\bar{p}$ represents the observed coverage probability of a nominal 95%

confidence interval for $\mu(\mathbf{x})$, averaged over the sample. We will also refer to it as CPμ. Then, $\sigma_\pi^2 \equiv \frac{1}{n}\sum(\pi_i - \bar{\pi})^2$ is a measure of variability in the coverage probabilities. In our simulation studies, we report the following estimate of an upper bound on σ_π^2:

$$\hat{\sigma}_\pi^2 \approx \frac{1}{n}(\sum_{i=1}^{n} p_i(1-p_i)/m + \sum_{i=1}^{n}(p_i - \bar{p})^2). \tag{2.19}$$

We give the derivation of the above estimate at the end of this section.

Example 2.4.1 In the first example the underlying regression function f is constant.

Example 2.4.2 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-3,3)^d$. The target function is a linear function of the first covariate:

$$f(\mathbf{x}) = 1 + x_1/4.$$

Example 2.4.3 In the third simulated study the target function is an ALB function of the first covariate. A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-3,3)^d$. The reference point parameterization was used to specify f_K: $\xi_1 = 1$, $\xi_2 = 0$, $\xi_3 = -1$, $\gamma_1 = \gamma_2 = \gamma_3 = 0$, $\delta_1 = 1$, $\delta_2 = 5$, $\delta_3 = 1$ and $\tau = 1$.

Example 2.4.4 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-1,1)^d$. The target function is not an ALB function:

$$f(\mathbf{x}) = \frac{2\sin(\pi(x_1+1)/2)}{(x_1/2+1)}$$

Example 2.4.5 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-1,1)^d$. The target function is an additive function of the first two covariates:

$$f(\mathbf{x}) = 1.5x_1^2 + 1.5\frac{\sin(\pi(x_2+1)/2)}{(x_2/2+1)}.$$

Table 2.3: Coverage probabilities averaged over 100 replicated samples together with corresponding measure of variability $\hat{\sigma}_\pi$.

No.	d	n	CPμ
1	1	100	.95(.02)
	1	200	.97(.02)
	1	400	.92(.03)
	5	100	.98(.01)
	5	200	.95(.02)
	5	400	.95(.02)
	10	100	.95(.02)
	10	200	.95(.02)
	10	400	.96(.02)
2	1	100	.96(.02)
	1	200	.95(.03)
	1	400	.95(.03)
	5	100	.91(.04)
	5	200	.91(.04)
	5	400	.90(.05)
3	1	100	.95(.03)
	1	200	.93(.03)
	1	400	.94(.03)
4	1	100	.94(.04)
	1	200	.94(.03)
	1	400	.91(.04)
5	2	100	.89(.06)
	2	200	.90(.07)
	2	400	.86(.08)
6	4	100	.91(.04)
	4	200	.93(.03)
	4	400	.93(.04)

Example 2.4.6 In the sixth simulation study the target function is an ALB function with $K = 2$. A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-3, 3)^d$. The reference point parameterization was used to specify f_K: $\xi_1 = (1, 2, 2, -1)'$, $\xi_2 = (2, 1, 1, 2)'$, $\gamma_1 = \gamma_2 = 0$, $\delta_1 = 0.2$, $\delta_2 = 3$ and $\tau = 2$.

The simulation studies suggest that the confidence interval are typically liberal, with average coverage probabilities between 86% and 98%. The first, second and third examples are well approximated by an ALB function, for $d = 1$. The target function in the third example is in fact an ALB function. For these examples the coverage probabilities are closer to .95. When increasing the dimensionality $d = 5$, or $d = 10$ in the first two examples, we notice a drop in the coverage probabilities. The fourth example is again one-dimensional, but the target function is not an ALB function, the coverage probability drops to .91 for $n = 400$. The coverage probabilities drop even more to .86, for the two-dimensional fifth example, where the target function is an additive function. We remind the reader that our approximate standard errors account for variance but not for bias. With an additive function, the true mean may be poorly approximated by an ALB function, and therefore $\exp(\hat{f}_K(\mathbf{x}))$ may have substantial bias relative to its standard deviation, resulting in lower coverage probabilities. In the sixth example, the target is an ALB function in four dimensions, the coverage probabilities are .93 and .91, a bit lower than the nominal 95% confidence level.

Now we give the derivation of expression (2.19). In order to get an

estimate of σ_π^2, we use the decomposition:

$$\begin{aligned}
\sum_{i=1}^{n}(\pi_i-\bar{\pi})^2 &= \sum_{i=1}^{n}(\pi_i-p_i+p_i-\bar{p}+\bar{p}-\bar{\pi})^2 \\
&= \sum_{i=1}^{n}(\pi_i-p_i)^2+\sum_{i=1}^{n}(p_i-\bar{p})^2+n(\bar{p}-\bar{\pi})^2+ \\
&+ 2\sum_{i=1}^{n}(\pi_i-p_i)(p_i-\bar{p})+2\sum_{i=1}^{n}(\pi_i-p_i)(\bar{p}-\bar{\pi})+ \\
&+ 2\sum_{i=1}^{n}(p_i-\bar{p})(\bar{p}-\bar{\pi}) \\
&= \sum_{i=1}^{n}(\pi_i-p_i)^2+\sum_{i=1}^{n}(p_i-\bar{p})^2+n(\bar{p}-\bar{\pi})^2+ \\
&+ 2\sum_{i=1}^{n}(\pi_i-p_i)(p_i-\bar{p})+2n(\bar{\pi}-\bar{p})(\bar{p}-\bar{\pi})+0 \\
&= \sum_{i=1}^{n}(\pi_i-p_i)^2+\sum_{i=1}^{n}(p_i-\bar{p})^2-n(\bar{p}-\bar{\pi})^2+ \\
&+ 2\sum_{i=1}^{n}(\pi_i-p_i)(p_i-\bar{p}).
\end{aligned}$$

Thus,

$$\sigma_\pi^2=\frac{1}{n}\sum_{i=1}^{n}(\pi_i-p_i)^2+\frac{1}{n}\sum_{i=1}^{n}(p_i-\bar{p})^2-(\bar{p}-\bar{\pi})^2+\frac{2}{n}\sum_{i=1}^{n}(\pi_i-p_i)(p_i-\bar{p}). \quad (2.20)$$

We have $mp_i \sim \text{Bin}(m,\pi_i)$, p_i's not necessarily independent. We expect $\text{Corr}(p_i,p_j)$ to be higher when $\|\mathbf{x}_i-\mathbf{x}_j\|$ is small. We note that $p_i=p_j$ if $\mathbf{x}_i=\mathbf{x}_j$. The second last term in the above equation is negative. If we assume

independence of the p_i's, then:

$$\begin{aligned}
\mathrm{E}(\bar{p}-\bar{\pi})^2 &= \mathrm{E}\left\{(\frac{1}{n}\sum_{i=1}^{n}(p_i-\pi_i))^2\right\} \\
&= \mathrm{E}\left\{\frac{1}{n^2}\sum_{i=1}^{n}(p_i-\pi_i)\sum_{j=1}^{n}(p_j-\pi_j)\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{E}(p_i-\pi_i)^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(p_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\pi_i(1-\pi_i)/m.
\end{aligned}$$

On the other hand, if all π_i are equal and all p_i are equal, then:

$$\mathrm{E}(\bar{p}-\bar{\pi})^2 = \mathrm{E}(p_i-\pi_i)^2 = \pi_i(1-\pi_i)/m.$$

We note that the last term in equation (2.20) is likely negative, because if p_i is smaller than the average $\bar{p}$, than p_i is more likely to be smaller than π_i. Conversely, if p_i is smaller than π_i, than p_i is more likely to be smaller than $\bar{p}$. Since $mp_i \sim \mathrm{Bin}(m, \pi_i)$ and $\mathrm{E}(\pi_i - p_i)^2 = \mathrm{Var}(p_i) = \pi_i(1-\pi_i)/m$, which can be estimated by $p_i(1-p_i)/m$, we obtain a conservative measure of variability in the coverage probability:

$$\hat{\sigma}^2_{\bar{\pi}} \approx \frac{1}{n}(\sum_{i=1}^{n} p_i(1-p_i)/m + \sum_{i=1}^{n}(p_i-\bar{p})^2).$$

2.5 ALB models for Poisson counts observed over time

2.5.1 Estimation

In the following, we consider the situation where the counts Y_i are recorded over time periods of different lengths t_i. We will show how ALB models can be used to model such data sets. The conditional distribution of y given $(t, \mathbf{x})$ is assumed to be Poisson with mean μ. Extending ALB in the GLM context, we end up with the following model:

$$E(y_i|t_i, \mathbf{x}_i) = \mu_i = \mu(t_i, \mathbf{x}_i),$$

$$g(\mu_i/t_i) = f_K(\mathbf{x}_i),$$

where g is known function linking μ/t (i.e., mean count per unit time) to $f_K(\mathbf{x})$. Here we use:

$$g(\mu_i/t_i) = \log(\mu_i/t_i).$$

The conditional log-likelihood for one observation, $(t, \mathbf{x}, y)$ is given by

$$\begin{aligned} l(\boldsymbol{\theta}|t, \mathbf{x}, y) &= -\mu + y\log\mu - \log y! \\ &= -(\mu/t)t + y\log(\mu/t) + y\log(t) - \log y!. \end{aligned}$$

The last term in the log-likelihood is treated as a constant. It is reasonable to work with an optimization criterion that is invariant with respect to changes in time scale (eg., second to minutes), and therefore we will keep the term $y\log(t)$ in our log-likelihood function, even though it does not depend on any of the parameters. Indeed, $l(\boldsymbol{\theta}|t, \mathbf{x}, y) = -\mu + y\log\mu$ and μ does not change when the time scale is changed.

According to equation (1.4), the gradient of the log-likelihood can be written as

$$\frac{\partial l(\boldsymbol{\theta}|t,\mathbf{x},y)}{\partial \boldsymbol{\theta}} = (y-\mu)\frac{\partial f_K}{\partial \boldsymbol{\theta}}.$$

The components of the gradient of f_K were derived in H01, equation (10). Therefore the updating formulae of the parameters at the m-th iteration stay the same as in (2.8), but with $\log(\mu/t) = f_K$.

2.5.2 Illustrations with simulated data

Two of the simulated examples from Section 2.2.3 are used here to illustrate the efficacy of the algorithm described above.

Example 2.5.1 The first example is one dimensional ($d = 1$) and the target function is an ALB function. A sample of size 1000 was generated with

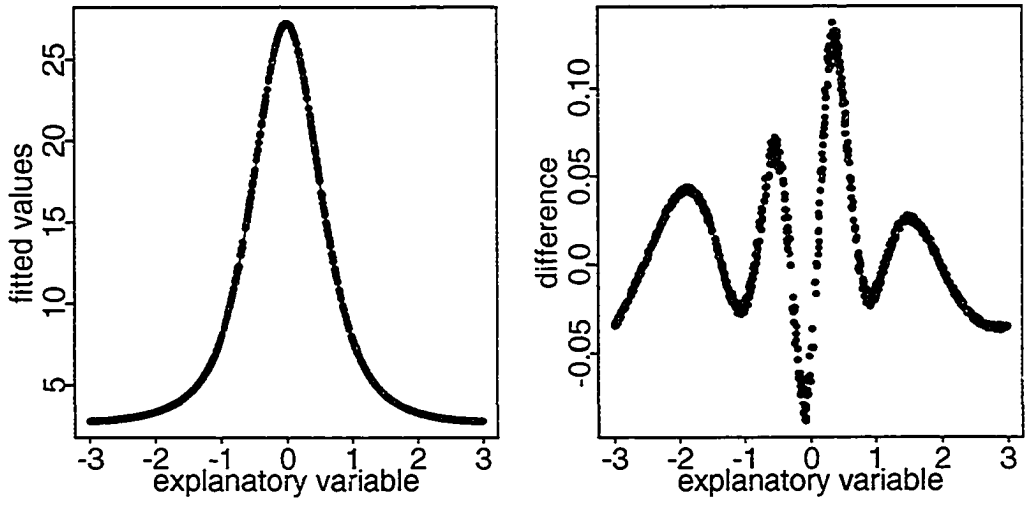

Figure 2.8: (a) Superimposed plot of the fitted function, $\exp(\hat{f})$ and the true function, $\exp(f)$ versus the predictor variable. (b) Difference $\exp(\hat{f}) - \exp(f)$ versus the predictor variable.

each predictor x generated from a Uniform$(-3, 3)$ distribution. The reference point parameterization was used to specify f_K: $\xi_1 = 1$, $\xi_2 = 0$, $\xi_3 = -1$, $\gamma_1 = \gamma_2 = \gamma_3 = 0$, $\delta_1 = 1$, $\delta_2 = 5$, $\delta_3 = 1$ and $\tau = 1$. A sample of 1000 time points t was generated with replacement ranging from 1 second to 180 seconds. For each (t, x), a response y was generated from a Poisson distribution with mean $t\exp(f_K(x))$. The estimate $\hat{f}$ is a linear combination of the $\hat{K} = 3$ basis functions. A superimposed plot of the fitted function, $\exp(\hat{f})$ and the true function, $\exp(f)$ versus the predictor variable x displayed in Figure 2.8(a) shows an almost perfect fit. A plot of the difference between the fitted values $\exp(\hat{f})$ and the true mean function $\exp(f)$ versus the predictor x is also displayed in Figure 2.8(b). The difference between the fitted values and the true mean are in the range -0.10 to 0.15. In effect, the sample size when using the time variable is between 1 and 180 times that in the similar example of Section 2.2.3. Therefore, we should obtain a better fit, when using a time variable. Indeed, the difference between the fitted values and the true mean for the similar example of Section 2.2.3 are in the range -0.4 to 0.4.

Example 2.5.2 The second example is 2-dimensional and the target function is not an ALB function:

$$f(\mathbf{x}) = \sin(x_1 x_2) + 2.$$

A sample of size 1000 was generated with each predictor vector $\mathbf{x}$ generated uniformly on a hypercube $(-2, 2)^2$ A sample of 1000 time points t was generated with replacement ranging from 1 second to 180 seconds. For each $(t, \mathbf{x})$ a response y was generated from a Poisson distribution with mean $t\exp(f(\mathbf{x}))$.

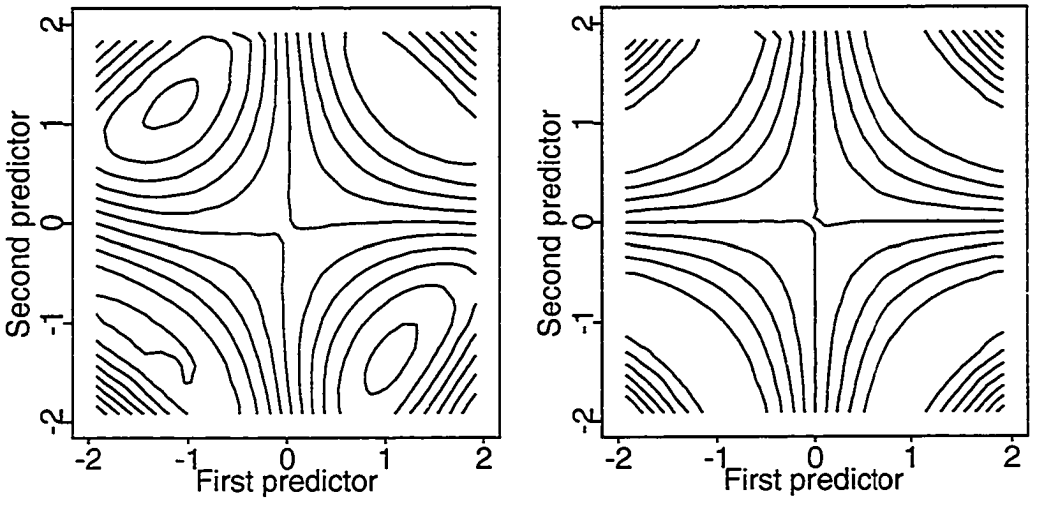

Figure 2.9: Contour plots of (a) $\hat{f}$ and (b) f as functions of the two original predictors.

The estimate $\hat{f}$ is a linear combination of the $\hat{K} = 10$ basis functions. Contour plots displayed in Figure 2.9 show f and $\hat{f}$ as functions of the two original predictors. The estimate $\hat{f}_K$ captures most of the curvature of the function f.

2.5.3 Illustration with Rongelap data

Our ALB model for Poisson counts observed over time was motivated by an example studied by Diggle, Tawn, and Moyeed (1998) and by Holmes and Mallick (2003). The example concerns radionuclide concentration on Rongelap Island, which was contaminated due to fall-out from the Bikini Atoll nuclear weapons testing programme during the 1950's. The former inhabitants of the Rongelap island have been living in self-imposed exile on the much smaller island of Mejatto since 1985. The Marshall Islands National Radiological Survey has examined the current levels of ^{137}Cs contamination by *in situ* γ-ray counting

at a set of $n = 157$ locations over the island. We denote by y_i the count at the ith location, and by t_i the length of time over which the count was recorded. According to the well-established theory of radioactive emissions, the counts y_i can be treated as realizations of Poisson random variables with expectations $\mu_i = \mu(t_i, \mathbf{x}_i)$, where $\mu(t_i, \mathbf{x}_i)$ measures the ^{137}Cs radioactivity at location $\mathbf{x}_i$ over time t_i.

We note that the ALB model assumes independence of the responses y_i given the predictors, which is unrealistic for the above data set. When the predictor $\mathbf{x} = (x_1, x_2)$ represents spatial locations of the response, we should model for spatial correlation among the responses. The above model does not take into account such correlations. Cressie, 1993 discusses methods for spatial models that estimate mean and covariance structure simultaneously. Although we have not developed theory for simultaneous estimation of mean and covariance for ALB models, we briefly consider the use of ALB models in this context in Chapter 6. We do not deal with modeling the correlations, but discuss effects of spatial correlations on the ALB fit and suggest directions for further research on how to deal with these effects.

Chapter 3

ALB models for over-dispersed count data

3.1 Introduction

The Poisson distribution is often a good model for count data, especially for processes that generate events over time and space. Sometimes, however, unmeasured effects, clustering of events, or other contaminating influences combine to produce more variation in the responses than is predicted by the Poisson model. The ALB models for Poisson counts do not take into account such variation. While this omission does not have a big impact on the fit, it can be crucial for estimating standard errors and assessing confidence.

In this chapter, we discuss ALB models for the case where there is more variation in the responses than that expected from Poisson sampling theory. The key idea is to model the conditional variance of the response given the predictors as a function of the mean. Several models for the variance, parametric and non-parametric, are proposed in the literature. We focus on the

model:

$$\text{Var}(Y|\mathbf{x}) = \mu(1 + \beta(\mu)),$$

where $\beta(.)$ is a positive-valued function. This model of the variance is motivated by a Poisson-Gamma mixture, the details of which are explained in the next section. Based on scientific reasons and model tractability, several choices of $\beta(.)$ can be made. We focus on the following choices:

(i) $\beta(\mu) = \beta_0$, $\beta_0 > 0$. This choice is appropriate to model Poisson overdispersed data with constant coefficient of variation.

(ii) $\beta(\mu) = \beta\mu$, $\beta > 0$. This choice is employed by Breslow (1984) to model extra-Poisson variation in log-linear models.

(iii) $\beta(\mu) = \exp(g(\mu))$, modeled as the exponentiate of an ALB function of the mean.

The constraints on the parameters and the exponentiate are motivated by the fact that the variance is positive. Algorithms to fit these models may be summarized in three steps:

Step1: Initialize values for weights.

Step2: Estimate $\hat{\mu}_i$ using Weighted Least Squares (WLS) or Weighted Quasi-Likelihood criterion (WQL).

Step3: Estimate 'optimal' weights $\hat{w}_i$.

After following the three steps, one can either go back to the second step and stop, or iterate steps two and three until convergence is achieved.

The choice of the weights is different for the two criteria WLS and WQL. Let us denote the conditional variance by $\text{Var}(Y_i|\mathbf{x}_i) = \sigma_i^2$, which may vary with μ_i. Then, for WLS, we minimize:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\hat{\sigma}_i^2},$$

and the weights are therefore given by $\hat{w}_i = 1/\hat{\sigma}_i^2$. For WQL, we minimize:

$$\sum_{i=1}^{n} \frac{(\mu_i - y_i) - y_i \log(\mu_i) + y_i \log(y_i)}{\hat{\sigma}_i^2/\hat{\mu}_i},$$

and the weights are therefore given by $\hat{w}_i = \hat{\mu}_i/\hat{\sigma}_i^2$.

WQL was used by Breslow (1984) to model extra-Poisson variation in Log-Linear Models. He also discusses WLS criterion that he applied to the log-transformed responses, using an additional normality assumption of the log-transformed responses. He recommends the use of WQL instead of a WLS applied to the log-transformed responses, when the counts are small and normal approximation of the log-transformed responses appears in doubt. We have not applied the WLS to the log-transformed responses, simply because we would have had to apply the ALB models for constant variance, rather than for count data, and we wanted to illustrate the latter. Therefore, we worked with the original responses. In our simulation studies and applications, the results from the two criteria were very similar. A result of the same flavor states that the WLS estimators and the Maximum Likelihood estimators are asymptotically equivalent in the class of Generalized Linear Models, (Carroll and Ruppert, 1988, Section 2.4). We have not found further discussions in the literature regarding when one of the criterion is preferred to the other.

The organization of this chapter is as follows. In section 2, we discuss the constant coefficient of variation case in Poisson over-dispersed data and illustrate the ALB model with simulated data and a real example. In section 3, we discuss the case where the variance is a function of the mean. We will summarize the ideas behind the heteroscedastic linear model, combine ALB and parametric variance function estimation to model heteroscedasticity, and use ALB to model the variance function. The accuracy of the estimators from the above models is tested in simulation studies. Also, a comparison between a competitive model and ALB model is conducted on a real example.

3.2 Modeling extra-Poisson variation using ALB

3.2.1 Estimation

In the following we will look at the case where extra-Poisson variation is present, in particular the case where the conditional variance is proportional to the conditional mean. The quasi-likelihood is a method for extending the model to allow for this possibility. The ALB model takes the form:

$$\mathrm{E}(Y|\mathbf{x}) = \mu$$

$$\mathrm{Var}(Y|\mathbf{x}) = \sigma^2\mu, \sigma^2 \geq 1$$

$$\log(\mu) = f_K(\mathbf{x}) = \sum_{k=1}^{K} \delta_k \phi_k(\mathbf{x}).$$

The extra parameter σ^2, called the dispersion parameter, captures the extra-variation and does not depend on $\boldsymbol{\theta}' = (\delta_1, \gamma_1, \boldsymbol{\xi}_1', \ldots, \delta_K, \gamma_K, \boldsymbol{\xi}_K')$. This model of the variance is motivated by a Poisson-Gamma mixture. Details of the

derivation follow. For each $\mathbf{x}$, a variable z is generated from a Gamma distribution with parameters $(\alpha(\mathbf{x}), \beta)$, such that $\alpha(\mathbf{x})\beta = \exp(f(\mathbf{x}))$. Then, for each z, a response y is generated from a Poisson distribution with mean z. We can derive the conditional mean and variance of y given x:

$$\mathrm{E}[y|\mathbf{x}] = \mathrm{E}[\mathrm{E}[y|z,\mathbf{x}]] = \mathrm{E}[z|\mathbf{x}] = \alpha(\mathbf{x})\beta = \exp(f(\mathbf{x})),$$

$$\begin{aligned}
\mathrm{Var}[y|\mathbf{x}] &= \mathrm{Var}[\mathrm{E}[y|z,\mathbf{x}]] + \mathrm{E}[\mathrm{Var}[y|z,\mathbf{x}]] \\
&= \mathrm{Var}[z|\mathbf{x}] + \mathrm{E}[z|\mathbf{x}] \\
&= \alpha(\mathbf{x})\beta^2 + \alpha(\mathbf{x})\beta \\
&= (1+\beta)\exp(f(\mathbf{x})).
\end{aligned}$$

Writing the above equations in terms of μ and σ, where $\mu = \exp(f(\mathbf{x}))$ and $\sigma^2 = 1 + \beta$, we end up with:

$$\mathrm{E}[y|\mathbf{x}] = \mu$$

$$\mathrm{Var}[y|\mathbf{x}] = \sigma^2\mu, \sigma^2 \geq 1.$$

and therefore extra-Poisson variation is present in the response. Notice that when β approaches zero while $\mu = \alpha\beta$ is fixed, then $\mathrm{Var}(z|\mathbf{x}) = \alpha\beta^2$ approaches 0 and $z \xrightarrow{P} \mu$.

The quasi-likelihood function is given by:

$$Q(\boldsymbol{\theta}, \sigma^2|\mathbf{x}, y) = \int_y^{\mu} \frac{y-t}{\sigma^2 t} dt = \frac{1}{\sigma^2}(-(\mu - y) + y\log(\mu) - y\log(y))$$

The kernel of this quasi-likelihood function is proportional to the kernel of the log-likelihood function of a Poisson distribution and therefore the vector

of parameter estimates $\hat{\theta}$ may be obtained as if $\sigma^2 = 1$; i.e., the conditional distribution of the response given the predictors is Poisson. To obtain approximate standard errors for the fit, we refer to Section 2.4.1, where we derived approximate standard errors for the ALB estimator in the quasi-likelihood context. We note that the standard errors for the fit in this case are given by multiplying the standard errors from the Poisson case by the square root of an estimate of the dispersion parameter. A moment estimator based on residuals can be used:

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

where p is the effective number of parameters in the ALB function.

3.2.2 Illustration using data on epileptic seizures

In the following example, we illustrate how ALB models work for over-dispersed Poisson data when the constant coefficient of variation assumption seems reasonable. Here is a brief description of the example, from Leppik et al, (1985). To study the anti-epileptic drug progabide, researchers randomly assigned 59 patients suffering from epileptic seizures to receive either progabide or a placebo. The data consists of the baseline number of epileptic seizures in the 8 weeks prior to administration of the treatment, the number of seizures in the 8 weeks after start of treatment, and the age of each patient. The goal is to see whether the mean number of counts is smaller for patients who received the progabide treatment than for the control patients, after accounting for age and the baseline number of seizures. In Figure 3.1, the number of post-treatment seizures is plotted versus the number of pre-treatment seizures with different

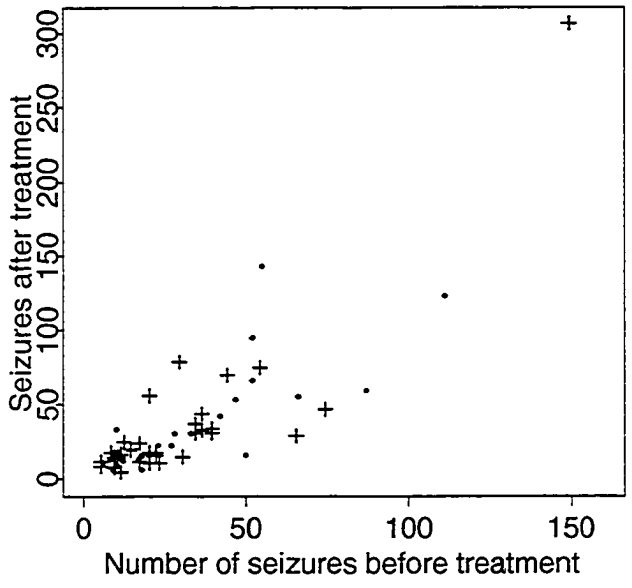

Figure 3.1: Post-treatment seizures versus Pre-treatment seizures. A case in the Control group is marked by a 'dot', and in the Progabide group by a 'plus'.

plotting symbols to distinguish the progabide patients from the controls. An initial ALB fit found a large residual corresponding to the unusual case with 302 post-treatment seizures, as seen on the upper right hand corner of the plot. This case was excluded prior to further analysis.

An ALB model was fit with number of post-treatment seizures as the response variable and logarithm of baseline number of epileptic seizures, age and treatment group as predictor variables. ALB selected two basis functions. To show the over-dispersion in the responses, a plot of the deviance residuals, assuming $\sigma^2 = 1$ is displayed in Figure 3.2. The magnitude of the residuals indicates that the responses show more variability than that explained by the Poisson distribution. The ALB fit versus the baseline number of epileptic

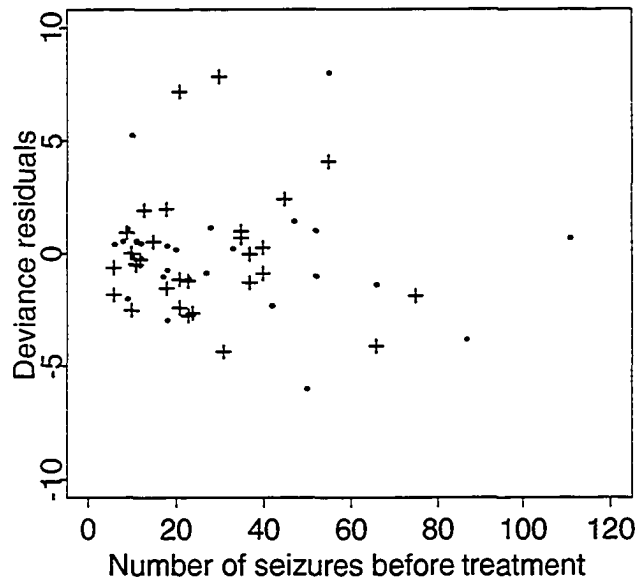

Figure 3.2: Deviance Residuals versus baseline number of seizures. A case in the Control group is marked by a 'dot', and in the Progabide group by a 'plus'.

seizures conditioning on a value of 30 for age is displayed in Figure 3.3. As we will see later, the fit does not depend on Age, so the fit is still representative when Age is fixed to a value of 30, for example. The fit is displayed separately for the progabide and control groups and has the corresponding observed data points superimposed. The two plots suggest that there is no difference between the two groups. The moment estimator for the dispersion parameter σ was evaluated at 3.0801.

We now describe a technique to investigate whether the group variable with two levels, progabide and control is statistically significant. This 'random shuffling' technique follows an idea proposed by Breiman (2001). To measure the importance of a variable, we randomly permute the values of that variable,

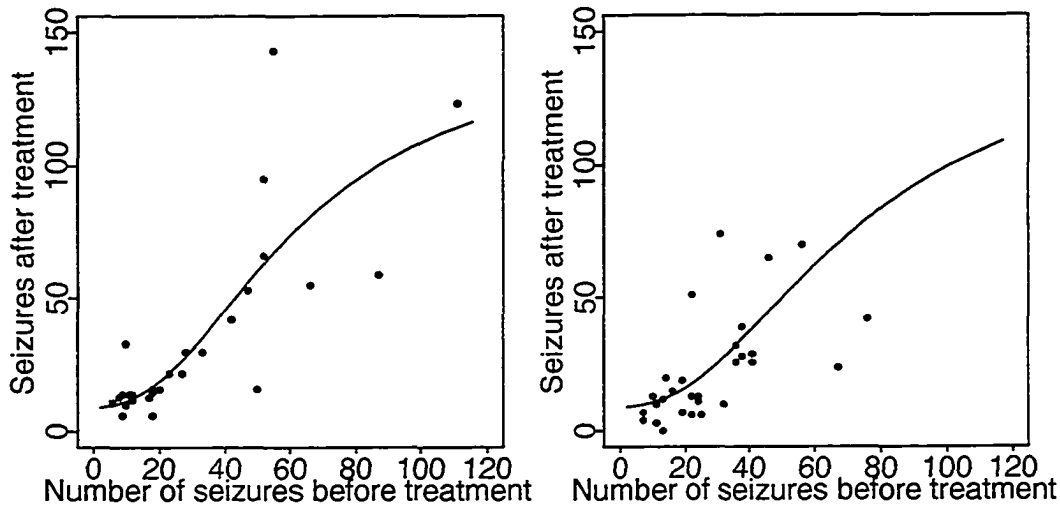

Figure 3.3: ALB fit from 3 predictors for (a)Control and (b)Progabide groups at Age 30.

leaving the rest of the variables fixed. We refit the model and evaluate the deviance on each new sample. We note that K is recomputed for each reshuffle. If the deviance evaluated on the original sample is an extreme value among the deviances calculated on the randomly shuffled samples, then this suggests that the variable is important. Figure 3.4 displays a histogram of the 200 deviances adjusted for the number of degrees of freedom, $n - (1 + (K-1)(d+2))$. The median of the distribution values is at 10, and a value of 10.85 was evaluated on the original sample. The group variable appears to be unimportant and there is no evidence of a difference in the mean number of epileptic seizures counts for the progabide and control groups.

The same approach was used to measure whether the age variable was important. Figure 3.5 displays a histogram of the 200 deviances adjusted for the number of degrees of freedom, evaluated on samples with randomly

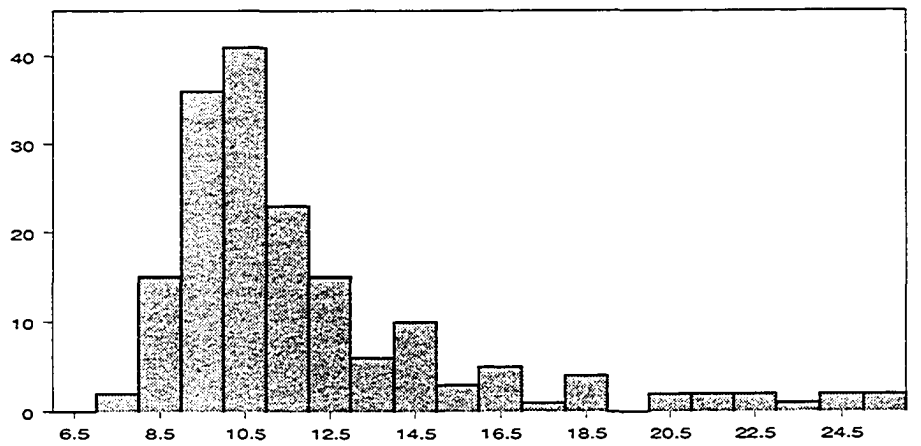

Figure 3.4: Histogram of deviances adjusted for number of degrees of freedom from shuffled groups with 10.85 being the adjusted deviance for the original data.

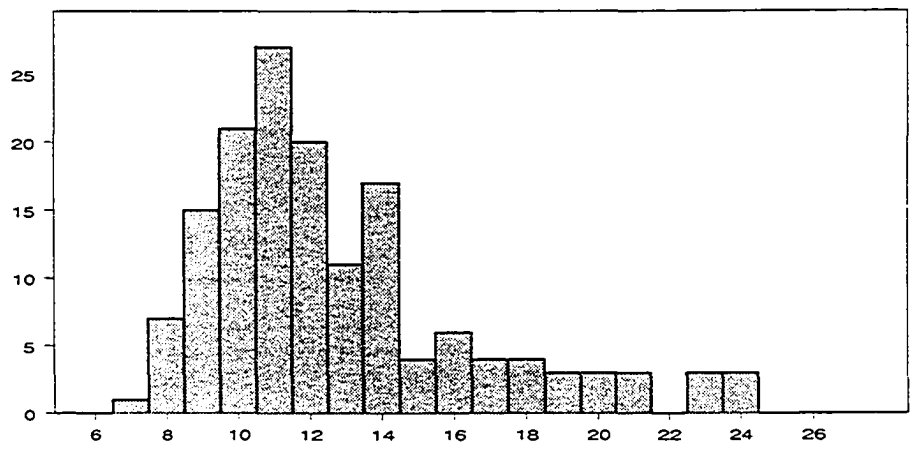

Figure 3.5: Histogram of deviances adjusted for number of degrees of freedom from shuffled age with 10.85 being the adjusted deviance for the original data.

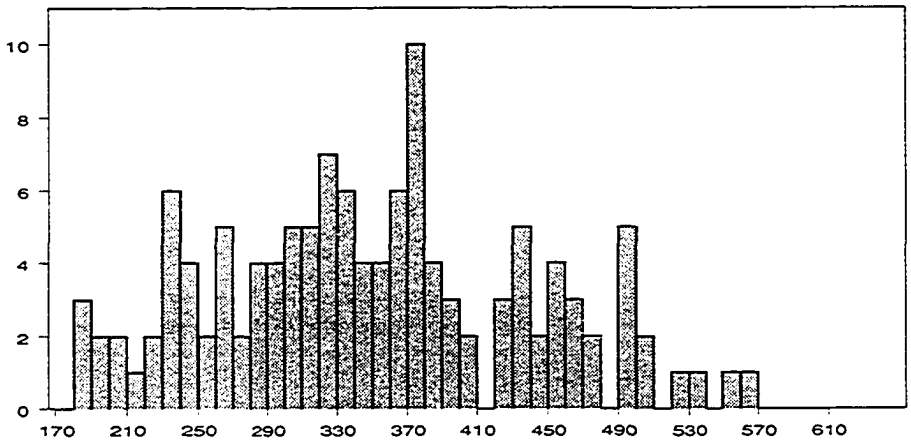

Figure 3.6: Histogram of deviances adjusted for number of degrees of freedom from shuffled baseline number of seizures with 10.85 being the adjusted deviance for the original data.

permuted values of age. The median of the distribution values is at 10.5, and

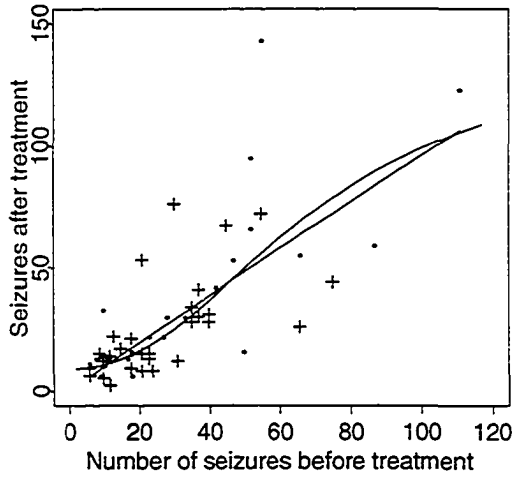

Figure 3.7: ALB fit and GLM fit superimposed using baseline number of seizures as the only predictor.

a value of 10.85 was evaluated on the original sample. The age variable appears to be unimportant. The same conclusion was obtained when both group and age were randomly permuted, leaving the baseline number of seizure variable fixed. The histogram corresponding to the baseline number of seizures variable is displayed in Figure 3.6. The deviance value for the original data falls in the left hand tail of the distribution, suggesting the variable is important.

We fit both ALB and GLM using baseline number of seizures as the only predictor. ALB uses two basis functions, suggesting a log-linear model. The ALB fit and GLM fit are displayed in Figure 3.7.

H01 discusses the use of standardized gradients as an exploratory tool to determine possible nuisance variables. The gradient components are defined as:

$$g_j(\mathbf{x}) = \frac{\partial}{\partial x_j} f(\mathbf{x}), \hat{g}_j(\mathbf{x}) = \frac{\partial}{\partial x_j} \hat{f}(\mathbf{x}),$$

where $\mathbf{x}' = (x_1, ..., x_d)$. Approximate standard errors for the gradient components, $\text{se}\{\hat{g}_j(\mathbf{x})\}$, can be defined in a manner similar to (2.16). Boxplots of the standardized gradients $\hat{g}_j(\mathbf{x})/\text{se}\{\hat{g}_j(\mathbf{x})\}$ may suggest possible nuisance variables. If f does not involve the covariate x_j, then $g_j(\mathbf{x}) = 0$ for all $\mathbf{x}$. Figure 3.8(a) indicates that most of the standardized gradients values of the variable age are in the interval $(-2, 2)$, suggesting age is a nuisance variable. Similarly, in Figure 3.8(b), most of the standardized gradient values for logarithm of baseline number of seizures are outside the interval $(-2, 2)$, suggesting the importance of the variable. We note that standardized gradient plots, like t statistics for linear regression coefficients, could be misleading given depen-

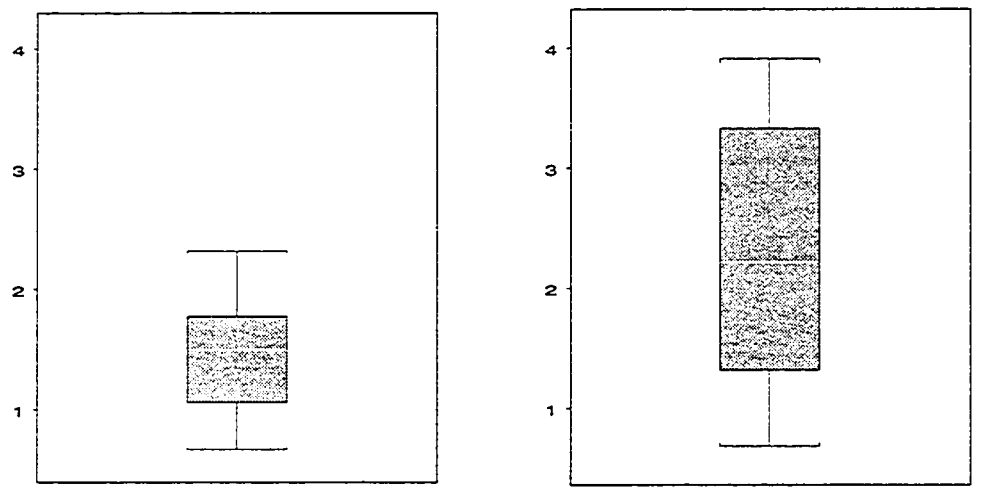

Figure 3.8: Boxplot of gradients, (a) for age and (b) for logarithm of baseline number of seizures.

dencies among covariates.

3.3 Variance function estimation and ALB models

3.3.1 Approaches to heteroscedasticity in linear models

In their book, 'Transformation and Weighting in Regression', Carroll and Ruppert (1988) discuss several approaches to heteroscedasticity. They view heteroscedasticity of variance as a regression problem in the sense that systematic smooth change of variability occurs as predictors are perturbed. From this point of view, there are similarities with modeling the mean. The basic assumption for modeling the mean is that the mean vector varies smoothly as we perturb continuous predictors. In simple linear regression the scatterplot of the predictors against the response helps us determine an appropriate model.

In modeling variability, the basic assumption is that there is a smooth change of variability as continuous predictors are perturbed. The corresponding residual plot replaces the usual scatterplot used to model the mean. The residual plot is used to suggest models for the variability. The most widely used diagnostic for heterogeneity is the unweighted least squares residual plot. The plot consists of residuals from an unweighted fit versus the predicted values. A fan-shaped pattern indicates that the residual variability depends on the mean response. In their book, 'Residuals and Influence in Regression', Cook and Weisberg (1982) argue for plotting squared residuals versus predicted value, mainly because plots of raw residuals are often sparse and difficult to interpret. A problem with squared residuals is that moderately large residuals can cause problems. Transformations of absolute residuals can help, for example, absolute residuals themselves, their logarithms, or their 2/3 power. Carroll and Rupert (1982) treat absolute residuals as the basic building block in the analysis of heteroscedasticity.

Carroll (1982) discusses regression parameter estimation in a heteroscedastic linear model given by:

$$\begin{aligned} \mathrm{E}\{y_i|\mathbf{x}_i\} &= \mu_i, \\ \mathrm{Var}\{y_i|\mathbf{x}_i\} &= \sigma_i^2. \end{aligned}$$

The variance σ_i^2 is usually modeled as a function of the mean μ_i, or the covariates. Depending on the problem, either a parametric or a nonparametric model can be used to estimate variability. In a parametric model, the variance is modeled as a known function of the mean. In a nonparametric model, the

variance is modeled as an unknown, but smooth function of the mean. Let us denote by $\hat{\mu}_L$ the usual unweighted least squares fit, and by r_i, the raw residuals, $r_i = y_i - \hat{\mu}_L(\mathbf{x}_i)$. For both parametric and nonparametric models,

$$\mathrm{E}(r_i^2) = \mathrm{E}(Y_i - \hat{\mu}_L(\mathbf{x}_i))^2 \approx \sigma_i^2.$$

Either a parametric or a nonparametric regression model, more precisely a kernel-type estimate, can be used to regress r_i^2 on $\hat{\mu}_L(\mathbf{x}_i)$ and obtain an estimate of σ_i^2. The estimates $\hat{\sigma}_i^2$ can then be used in a weighted least squares regression to get a weighted fit, $\hat{\mu}_E$. If we denote by $\hat{\mu}_T$ the weighted estimate using the optimal weights $1/\sigma_i^2$ then, under regularity conditions, there is asymptotically no cost due to estimating σ_i, ie. $\hat{\mu}_T$ and $\hat{\mu}_E$ have the same normal limit distribution, (Carroll, 1982).

3.3.2 Combining ALB models and parametric variance function estimation. An example

In the following we summarize an approach proposed by Breslow (1984) to model heteroscedasticity in generalized linear models using parametric variance function estimation. Then we combine ALB models and parametric variance estimation to model heteroscedasticity and compare the results of the two approaches on an example used in Breslow (1984). We note that because of the small sample size, $n = 18$, small dimensionality and not very complex data structure, this example is not ideal to reveal advantages of the ALB fit over other competing models. However, we found Breslow's approach to model heteroscedasticity interesting and considered incorporating ALB models. First we describe Breslow's approach and later in this section we introduce

the example.

Breslow (1984) proposes the following heteroscedastic model using a parametric variance function:

$$\mathrm{E}(Y|\mathbf{x}) = \mu$$

$$\mathrm{Var}(Y|\mathbf{x}) = \mu(1 + \beta\mu)$$

$$\log(\mu) = \boldsymbol{\gamma}'\mathbf{x}.$$

This is a quasi-likelihood model since no assumptions are made on the distribution function, other than specifying the mean and the variance. If β were known, the maximum quasi-likelihood solution to the above model would be easily obtained using the Poisson error function, the log-link, and defining prior weights, $w_i = (1 + \beta\hat{\mu}_i)^{-1}$. Even if the wrong weights are used, a new value of β may again be estimated by setting the chi-square criterion to its degrees of freedom, i.e. $\sum(y_i - \hat{\mu}_i)^2/\{\hat{\mu}_i(1 + \beta\hat{\mu}_i)\} = n - p$, where p is the number of parameters used to estimate the mean and n is the sample size. Multiplying both sides by β, this equation may be solved recursively using:

$$\beta \leftarrow (n-p)^{-1} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(\hat{\mu}_i + \beta^{-1})}, \tag{3.1}$$

and substituting from left to right-hand side.

The algorithm proposed by Breslow may be summarized as follows.

Algorithm 3.1:

Step1: Fit the quasi-likelihood log-linear model using weights w_i. At the first iteration use $w_i = 1$. If the chi-square criterion is close to its degrees

of freedom, stop and conclude that the residual variation is adequately explained.

Step2: Update $\hat{\beta}$ using (3.1). In the first iteration substitute following expression for $\hat{\beta}$ in the right hand side of (3.1):

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/\hat{\mu}_i - (n-p)}{\sum_{i=1}^{n}\hat{\mu}_i(n-p)/n}. \tag{3.2}$$

In subsequent iterations, use the value $\hat{\beta}$ from previous iteration in the right hand side of (3.1).

Step3: Define new weights $w_i = (1 + \beta\hat{\mu}_i)^{-1}$ and return to Step 1.

In the light of the above ideas, we propose the following heteroscedastic ALB model with a parametric variance function:

$$\mathrm{E}(Y|\mathbf{x}) = \mu,$$

$$\mathrm{Var}(Y|\mathbf{x}) = \mu(1 + \beta\mu),$$

$$\log(\mu) = f_K(\mathbf{x}) = \sum_{k=1}^{K} \delta_k \phi_k(\mathbf{x}).$$

Algorithm 3.1 is applied as before, but using a quasi-likelihood ALB model to fit the conditional mean.

We also investigated the WLS criterion, log-link, and weights $w_i = (\hat{\mu}_i(1 + \beta\hat{\mu}_i))^{-1}$. Let us look more closely at the two criteria. The WQL criterion minimizes:

$$\sum_{i=1}^{n}((\mu_i - y_i) - y_i \log(\mu_i) + y_i \log(y_i))w_i,$$

where $w_i = (1 + \beta\hat{\mu}_i)^{-1}$. Our WLS criterion minimizes:

$$\sum_{i=1}^{n}(y_i - \mu_i)^2 w_i,$$

where $w_i = (\hat{\mu}_i(1 + \beta\hat{\mu}_i))^{-1}$. Modifications to the updating functions are obtained as in Section 2.2.1, equations (2.7) and (2.8). We note that the results from the two criteria, WQL and WLS, were very similar on various simulation studies and real examples. Results of the simulation studies are displayed in the last section of this chapter.

Another approach for estimating the parametric variance function in Step 2, would be to use a regression through origin of $(y_i - \hat{\mu}_i)^2/\hat{\mu}_i - 1$ vs $\hat{\mu}_i$. The weights for Step 3, would then be given by $w_{LM} = (\hat{\mu}_i(1+\beta\hat{\mu}_i))^{-1}$, where β is given by the slope in Step 2.

Now let us apply these heteroscedastic models with a parametric variance function to the following example used by Breslow (1984) to illustrate the use of Algorithm 3.1. The example concerns an Ames Salmonella reverse mutagenicity assay. The Salmonella bacterium carries a defective (mutant) gene. If in reaction with a certain chemical this type of mutation can be reversed, causing a back mutation with the gene regaining its function and ability to form many revertant colonies, then this chemical is called mutagenic for the Salmonella bacterium. The purpose of the Ames Salmonella reverse mutagenicity assay is to determine whether the chemical quinoline is mutagenic for the Salmonella bacterium. This is investigated by counting the number of revertant colonies corresponding to various doses of quinoline. Table 3.1 shows the number of revertant colonies observed on each of three replicate plates

Table 3.1: Number of revertant colonies of TA98 Salmonella

Dose of quinoline (μg per plate)					
0	10	33	100	333	1000
15	16	16	27	33	20
21	18	26	41	38	27
29	21	33	60	41	42

tested at each of six dose levels of quinoline.

Breslow noted that scientific reasons suggest the following log-linear model:

$$\mathrm{E}(Y|x) = \mu(x)$$

$$\log(\mu(x)) = \gamma_1 + \gamma_2 \log(x + c) - \gamma_3 x,$$

where y denotes the number of revertant colonies, x the dose level and c is a constant that was set to 10, the smallest non-zero dose level. Margolin, Kaplan and Zeiger (1981) questioned the use of the Poisson model to analyze the data from this example. They argued that there is substantial evidence that the Poisson model, nearly universal in its adoption as the sampling distribution of revertants per plate from a Salmonella test, lacks the flexibility to adequately describe the variability in a set of plate counts. This can lead to false conclusions concerning the outcome of the assay. Scientific purpose of the study is to determine the mutagenic effect, and to do so it is necessary to accommodate excess variation typically observed among replicates.

We applied the heteroscedastic GLM model using the two covariates, the dose level and the logarithm of the dose level, as suggested by scientific reasons. We also applied the heteroscedastic ALB model using only one covari-

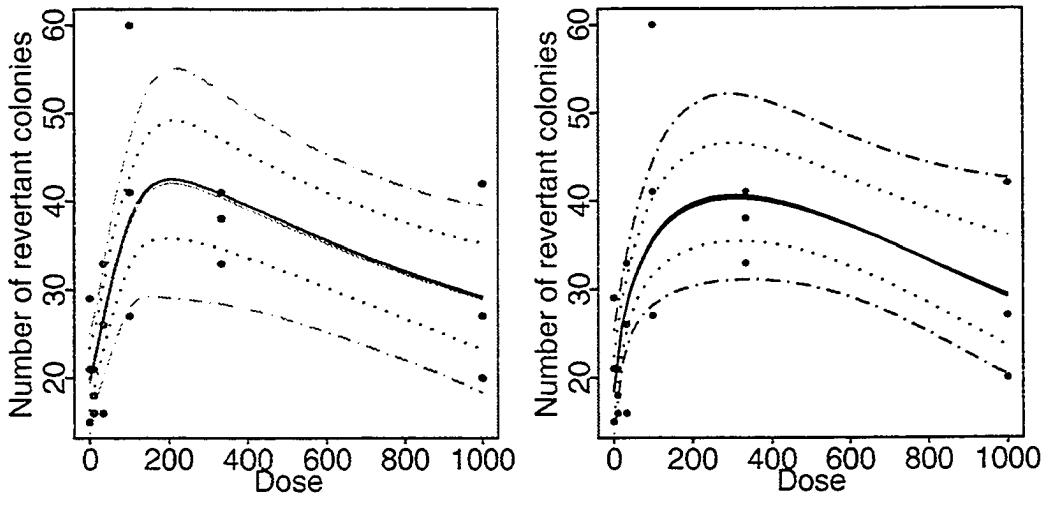

Figure 3.9: (a) The ALB Poisson fit and heteroscedastic ALB fit versus Dose level together with the 95% confidence bands (the narrower intervals are for the ALB Poisson fit) and the counts superimposed. (b)The GLM Poisson fit and heteroscedastic GLM fit versus Dose level together with the 95% confidence bands (narrower intervals are for the GLM Poisson fit) and the counts superimposed.

ate, the dose level. ALB selected three basis functions. We obtained similar estimates for the parameter in the variance function, $\hat{\beta} = .0718$ and .0717 respectively. The ALB Poisson fit and heteroscedastic ALB fit versus Dose level together with the 95% confidence bands and the counts superimposed are displayed in Figure 3.9(a). The GLM Poisson fit and heteroscedastic GLM fit versus Dose level together with the 95% confidence bands and the counts superimposed are displayed in Figure 3.9(b). The confidence bands are larger for the heteroscedastic models, as expected. The ALB confidence intervals are symmetric, given by the $\hat{\mu} \pm 2\text{se}$, while the GLM confidence intervals are asymmetric, given by the exponentiate of the confidence interval for the lin-

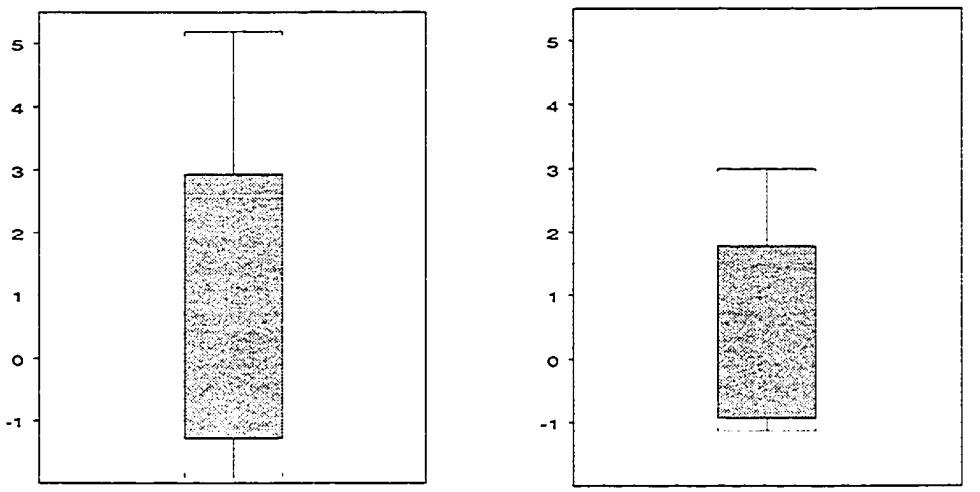

Figure 3.10: Boxplot of standardized gradients for logarithm of Dose level from (a) ALB model and (b) heteroscedastic ALB model.

ear component, that is $\exp(\log \hat{\mu} \pm 2\text{se})$. We remind the reader that results from simulations studies were very similar for ALB standard errors calculated in these two ways. We note that, compared to GLM where we needed two predictors, ALB adaptively estimates the mean function based on a single predictor, the Dose. The flexibility of ALB models, comes with a price of more parameters. In this example, ALB uses 8 parameters, including the parameter in the variance function, compared to GLM that uses 4 parameters. We also tried to fit ALB using two covariates, same as GLM, and ALB selected two basis functions, and therefore 6 parameters, including the parameter in the variance function. The fit based on the two covariates is very similar to the GLM fit.

Boxplots of the standardized gradients for Dose level displayed in figure 3.10, suggest the importance of the Dose level in predicting the number of

revertant colonies. We note that the magnitude of the standardized gradients is reduced from the ALB Poisson model to the heteroscedastic ALB model as expected. Although the Poisson fit and Heteroscedastic fit are very close, modeling for heteroscedasticity is crucial for estimating standard errors and assessing confidence.

3.3.3 Modeling the variance function using ALB

So far we assumed parametric models for the variance function, such as:

$$\text{Var}(Y|\mathbf{x}) = \sigma^2\mu, \sigma^2 \geq 1.$$

A natural generalization of the above model would be to assume the conditional variance to be a function of the mean:

$$\text{Var}(Y|\mathbf{x}) = H(\mu), H \text{ unknown.}$$

A key feature of many heteroscedastic regression problems is that the variances appear to be smooth functions of the mean response.

Our ALB model to accommodate heteroscedasticity takes the form:

$$\text{Var}(Y|\mathbf{x}) = \mu(1 + \beta(\mu)), \tag{3.3}$$

where $\beta(\mu)$ is modeled as the exponentiate of an ALB function, $\beta(\mu) = \exp(g(\mu))$. This model of the variance can be motivated by a Poisson-Gamma mixture. The derivation follows closely the one in Section 3.2.1, except that here β is no longer fixed, but is a positive smooth function of μ.

We suggest the following algorithm to fit model (3.3):

Algorithm 3.2:

Step1: Fit an ALB Poisson model of $\mu_i = \mu(\mathbf{x}_i)$, get fitted values $\hat{\mu}_i$ and residuals $r_i = y_i - \hat{\mu}_i$.

Step2: Estimate $g(\cdot)$ using an ALB L_q regression with a log-link function of transformed residuals $r_i^2/(\hat{\mu}_i + .1) - 1$ vs $\hat{\mu}_i$. That is,

$$\min \sum_{i=1}^{n} \left(\frac{n}{n-p}\right)^q \left| \frac{r_i^2}{\hat{\mu}_i + .1} - 1 - \exp(g(\hat{\mu}_i)) \right|^q .$$

Calculate variance estimates $\hat{\sigma}_i^2 = \hat{\mu}_i(1 + \exp(\hat{g}(\hat{\mu}_i)))$.

Step3: Fit an ALB weighted least squares model of $\mu_i = \mu(\mathbf{x}_i)$ with log-link, using the weights from the second step, $w_i = 1/\hat{\sigma}_i^2$.

We now discuss this algorithm and give details on its implementation. As a first step in our algorithm, we run an ALB model assuming that the conditional distribution of Y given $\mathbf{x}$ is Poisson with mean μ. The second step uses an ALB model to estimate the variance function. We explain here the choice of the transformed residuals. We investigated several transformations of the residuals. Given a Poisson random variable with mean μ, we have:

$$\mathrm{Var}((Y-\mu)^2) = \mu(1+3\mu).$$

This suggests that standardized residuals are more stable. Indeed,

$$\mathrm{Var}\left\{\frac{(Y-\mu)^2}{\mu}\right\} = \frac{\mu(1+3\mu)}{\mu^2} = 3 + \frac{1}{\mu}.$$

Here is the behavior of the square root of the above variance function as μ varies:

μ	.1	1	10	100
$\sqrt{\frac{\mu(1+3\mu)}{\mu^2}}$	3.6	2	1.76	1.73

To attenuate the effect of outliers and obtain a more stable behavior, we add a small constant in the denominator:

μ	.1	1	10	100
$\sqrt{\frac{\mu(1+3\mu)}{(\mu+.1)^2}}$	1.80	1.81	1.74	1.73

To accommodate over-dispersion, we model $\mathrm{Var}(Y|\mathbf{x}) = \mu(1+\exp(g(\mu)))$. Let $\hat{\mu}_i$ denote the ALB estimate from the first step. Then,

$$\mathrm{E}\left\{\frac{(Y_i-\hat{\mu}_i)^2}{\hat{\mu}_i+.1}\right\} \approx 1+\exp(g(\hat{\mu}_i)).$$

As a second step in fitting the heteroscedastic model, we estimate $g(\cdot)$ by an ALB L_q regression of $r_i^2/(\hat{\mu}_i+.1)-1$ vs $\hat{\mu}_i$, using a log-link function. Modifications to the updating functions are obtained as in Section 2.2.1, equations (2.7) and (2.8).

There are a few considerations in choosing the power q. ALB L_q regression was introduced in H01. Conditional mean and median are obtained by choosing $q=2$ and $q=1$, respectively. Residual plots tend to be scattered and to have outliers. An ALB L_1 fit corresponding to the median is resistant to outliers. On the other hand, too many residuals close to zero would produce undesirable near zero variances. An ALB L_2 estimate, corresponding to the conditional mean, would not be resistant to outliers. A value of $q=1.5$ provides a compromise and seems to work well. For a discussion on robustness and L_q estimators in linear regression, we refer to Forsythe (1972).

We note that, for the case where the conditional variance is proportional

to the conditional mean, the ALB fit in the second step should select a constant model.

The last step consists of running an ALB weighted least squares fit with a log-link, using the weights from the second step:

$$w_i = \frac{1}{\hat{\mu}_i(1 + \exp(\hat{g}(\hat{\mu}_i)))}.$$

As it was mentioned earlier in this section, we can also use a WQL criterion, log-link, and weights

$$w_i = \frac{1}{(1 + \exp(\hat{g}(\hat{\mu}_i)))}.$$

Modifications to the updating functions in this last step are obtained as in Section 2.2.1, equations (2.7) and (2.8).

The estimator obtained from this algorithm will be referred to as the 3-step estimator. We denote by $\hat{\mu}_A$ the estimate obtained in Step 3, based on the estimated weights and by $\hat{\mu}_T$ the estimate obtained using an ALB weighted least squares fit with a log-link and optimal weights, $1/\sigma_i^2$. Simulation studies at the end of this chapter show that mean predictive squared errors associated with the two estimators are similar.

3.3.4 Simulation Studies

In these simulation studies we evaluate estimates of the mean predictive squared errors associated with the heteroscedastic ALB estimators, and investigate coverage probabilities corresponding to the different choices of weights. In each example, $m = 100$ samples of n independent observations of $(\mathbf{x}, y)$ were generated as follows. A sample of size n was generated, with each predictor vector

$\mathbf{x}$ generated uniformly on a hypercube $(a,b)^d$. For each $\mathbf{x}$, the mean vector $\mu(\mathbf{x}) = \exp(f(\mathbf{x}))$ was calculated. A variable z was generated from a Gamma distribution with parameters $\alpha(\mu) = \mu/\beta(\mu)$ and $\beta(\mu)$. For each z, a response y was generated from a Poisson distribution with mean z. The variable z and the response y differs from sample to sample, while the predictor vectors $\mathbf{x}$ remain the same over the $m = 100$ samples. The conditional variance is thus:

$$\text{Var}(Y|\mathbf{x}) = H(\mu) = \mu(1 + \beta(\mu))$$

Four estimates were calculated for each sample: $\hat{\mu}_T$ using true weights, $\hat{\mu}_P$ using Poisson weights, $\hat{\mu}_{LM}$ using linear model weights, (see (ii) at the beginning of Section 3.1), and $\hat{\mu}_A$ using an ALB variance function to estimate the weights, (see (iii) at the beginning of Section 3.1). Reverse cross-validation is used to obtain an estimate of the mean predictive squared error for each of the four estimates. The technique was derived in Chapter 2. For each of the $m = 100$ samples, the mean estimate is obtained. Then, a measure of mean predictive squared error corresponding to the mean estimate is evaluated over each of the remaining 99 samples, as follows. For a sample $(\mathbf{x}_i, y_i^*)$, $i = 1, n$,

$$\text{MPSE} = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i^* - \hat{\mu}(\mathbf{x}_i))^2}{H(\mu(\mathbf{x}_i))},$$

where $\hat{\mu}$ can be any of the four estimators: $\hat{\mu}_T$, $\hat{\mu}_P$, $\hat{\mu}_{LM}$ and $\hat{\mu}_A$. The average over the 99 remaining samples is taken. This is repeated for each of the $m = 100$ samples and an estimate of the mean predictive squared error is obtained by averaging the estimates over the $m = 100$ samples. The mean

predictive squared error in the above equation estimates:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{E}\{(Y_i^*-\hat{\mu}(\mathbf{x}))^2|\mathbf{x}_i\}}{\mathrm{E}\{(Y_i^*-\mu(\mathbf{x}))^2|\mathbf{x}_i\}}.$$

Averages over the 100 samples of the mean predictive squared errors for the four estimators are reported in Table 3.2. For each of the five examples, several values of the sample size n are used to demonstrate how accuracy improves with increased sample size. With simulation studies, the true mean μ is known and therefore a lower bound on the mean predictive squared error can be obtained by using the true mean μ instead of the estimates when evaluating the mean predictive squared error:

$$\text{lbound} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i^*-\mu(\mathbf{x}_i))^2}{H(\mu(\mathbf{x}_i))}.$$

Table 3.2: MPSE for the four estimates: averages over 100 replicated samples of size n. Lower bound is displayed on the third column.

Eg	n	lbound	$\text{MPSE}_{\hat{\mu}_T}$	$\text{MPSE}_{\hat{\mu}_P}$	$\text{MPSE}_{\hat{\mu}_{LM}}$	$\text{MPSE}_{\hat{\mu}_A}$
1	50	1.0065	1.1465	1.1306	1.1406	1.1942
	100	1.0053	1.0665	1.0616	1.0672	1.0939
	200	.9705	1.0132	1.0061	1.0125	1.0152
2	50	1.0099	1.1302	1.1273	1.1142	1.1594
	100	.9873	1.0203	1.0316	1.0235	1.0437
	200	1.0092	1.0293	1.0331	1.0274	1.0362
3	50	.9924	1.1164	1.1236	1.1173	1.1721
	100	1.0031	1.0605	1.0759	1.0644	1.0885
	200	1.0022	1.0316	1.0362	1.0327	1.0380
4	50	.9693	1.2542	1.2734	1.2077	1.3539
	100	.9987	1.1508	1.1652	1.1169	1.1655
	200	.9943	1.0631	1.0793	1.0494	1.0718
5	50	1.0050	1.4678	1.5693	1.5612	1.6611
	100	1.0024	1.2408	1.3680	1.2188	1.4298
	200	1.0172	1.0984	1.1896	1.0992	1.1506

We note that the lower bound estimates should then be close to one, but not precisely one, since the numerator in the above expression is calculated using a test set and therefore different from the denominator, which is simply the conditional variance at each predictor vector $\mathbf{x}$ in the sample.

Example 3.3.1 The first simulated example is one-dimensional and the target function is an ALB function. A sample of size n was generated with each predictor vector x generated uniformly on $(-3, 3)$. For each x, the mean vector $\mu = \exp(f_K)$ was calculated, where f_K was specified by the reference point parameterization: $\xi_1 = 1$, $\xi_2 = 0$, $\xi_3 = -1$, $\gamma_1 = \gamma_2 = \gamma_3 = 0$, $\delta_1 = .5$, $\delta_2 = 5$, $\delta_3 = .5$ and $\tau = 1$. The parameters for the Gamma distribution are $\alpha(\mu) = \mu/\beta(\mu)$ and $\beta(\mu) = .72 + .16\mu$.

Example 3.3.2 The second simulated example is one-dimensional. A sample of size n was generated with each predictor vector x generated uniformly on $(-3, 3)$. The target function is linear:

$$f(x) = 1 + x/4.$$

The parameters for the Gamma distribution are $\alpha(\mu) = \mu/\beta(\mu)$ and $\beta(\mu) = .35 + .37\mu$.

Example 3.3.3 For the third simulated example, a sample of size n was generated with each predictor vector x generated uniformly on $(-1, 1)$. The target function is not an ALB function:

$$f(x) = \frac{2\sin(\pi(x+1)/2)}{(x/2+1)}$$

The parameters for the Gamma distribution are $\alpha(\mu) = \mu/\beta(\mu)$ and $\beta(\mu) = .47 + .31\mu$.

Example 3.3.4 The fourth example is two-dimensional. A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on $(-1,1)^2$ and the target function is an additive function:

$$f(\mathbf{x}) = 1.5x_1^2 + 1.5\frac{\sin(\pi(x_2+1)/2)}{(x_2/2+1)}.$$

The parameters for the Gamma distribution are $\alpha(\mu) = \mu/\beta(\mu)$ and $\beta(\mu) = .34 + .66\mu$.

Example 3.3.5 The last simulated example is four-dimensional and the target function is an ALB function with $K = 2$. A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-1,1)^4$. The reference point parameterization is used to specify f_K: $\xi_1 = (1,2,2,-1)'$, $\xi_2 = (2,1,1,2)'$, $\gamma_1 = \gamma_2 = 0$, $\delta_1 = 0.2$, $\delta_2 = 2$ and $\tau = 2$. The parameters for the Gamma distribution are $\alpha(\mu) = \mu/\beta(\mu)$ and $\beta(\mu) = .34 + .5\mu$.

The mean predictive squared errors associated with the four estimates are very close to 1, showing that the four estimates are approximating very well the true mean function. On one hand this shows that the estimates obtained using the true weights and estimated weights are very close. The fact that the estimate using Poisson weights is doing also very well in approximating the true mean function reminds of the result stating that using Poisson log-linear regression when heteroscedasticity is present still yields a roughly unbiased fit. The results in Table 3.2 indicate that the fitted values of the heteroscedastic models show substantial agreement in fitting the data. However, the mean predictive squared errors in Table 3.2 are based on averages across the predictors. The accuracy could vary for different values of the predictor

vectors depending on how the variance function is modeled to obtain weights. A closer look at Table 3.2 indicates that the mean predictive squared errors corresponding to $\hat{\mu}_A$ are consistently slightly larger than the rest of the estimators. We noted earlier that the choice of weights does not have a huge effect on the estimate of the mean, but can lead to important differences in the standard errors. As a consequence, the confidence intervals corresponding to the Poisson fit are narrower than the ones corresponding to the weighted fit using the correct weights. A simulated study on the coverage probabilities, similar to the one in Section 2.4.2, is summarized in Table 3.3. Table 3.3 indicates that the coverage probabilities corresponding to $\hat{\mu}_T$ and $\hat{\mu}_L$ are closer to .95. The confidence intervals corresponding to $\hat{\mu}_P$ and $\hat{\mu}_A$ are narrower resulting

Table 3.3: Coverage probabilities averaged over 100 replicated samples together with corresponding measure of variability $\hat{\sigma}_\pi$.

Eg	n	CP$\hat{\mu}_T$	CP$\hat{\mu}_P$	CP$\hat{\mu}_{LM}$	CP$\hat{\mu}_A$
1	50	.92(.03)	.76(.05)	.90(.03)	.78(.04)
	100	.94(.03)	.72(.03)	.94(.03)	.80(.03)
	200	.94(.02)	.73(.03)	.94(.02)	.81(.04)
2	50	.95(.03)	70(.04)	.94(.03)	.80(.05)
	100	.96(.02)	.74(.06)	.95(.02)	.85(.03)
	200	.96(.03)	.77(.05)	.94(.02)	.83(.03)
3	50	.95(.04)	.78(.06)	.89(.03)	.72(.05)
	100	.95(.03)	.80(.05)	.90(.03)	.72(.04)
	200	.95(.03)	.75(.05)	.94(.03)	.85(.05)
4	50	.92(.08)	.69(.07)	.90(.07)	.70(.06)
	100	.90(.07)	.72(.09)	.89(.08)	.73(.05)
	200	.93(.06)	.73(.06)	.92(.07)	.71(.07)
5	50	.91(.04)	.72(.05)	.89(.04)	.75(.06)
	100	.93(.03)	.74(.06)	.91(.04)	.74(.06)
	200	.94(.04)	.72(.04)	.91(.03)	.75(.05)

in smaller coverage probabilities ranging in our examples from .71 to .81. Our simulation studies indicated that the estimate $\hat{\mu}_A$ is not very efficient. Modeling the variance function using ALB does not present an advantage over using a linear model to estimate the weights. In fact, in our examples, the coverage probabilities corresponding to an ALB model for the variance function are much smaller than the ones corresponding to a linear model for the variance function. We suspect and give an argument here that modeling the residuals as an ALB function of the predictors from the Poisson fit is not appropriate. The residuals are very scattered even after they are transformed, and from our experience with the simulation studies, when fitting an ALB model to the variance function, ALB tends to select a larger number of basis functions than would be needed.

Similar tables, not shown, corresponding to the quasi-likelihood approach, using a log-link and defining prior weights, display the same pattern. As we have stated in Section 3.3.2, we did not notice any changes in the results between using the quasi-likelihood or the weighted sum of squares methods.

Chapter 4

Examples

In this chapter we illustrate ALB models using two examples. We indicate ALB strengths and weaknesses in comparison to other models. The first example presented in Section 4.1 concerns dependence of ozone on three meteorological variables at sites in the New York metropolitan region. The second example presented in Section 4.2 concerns dependence of ozone on eight meteorological measurements made in the Los Angeles basin. Ozone can be beneficial or harmful, depending on where it is found in the atmosphere. In the stratosphere, ozone protects us from ultraviolet radiation, so it is beneficial to human health. On the other hand, ground-level ozone is a pollutant that can cause breathing difficulty, permanent lung damage, and eye irritation, and may trigger asthma attacks and reduce resistance to infection. It can also be harmful to vegetation and contribute to smog formation. In the studies considered, we are concerned with ground-level ozone as a pollutant. Daily weather conditions affect whether and how much we are exposed to pollutants in the air. In the two studies considered, meteorological measurements are taken. The response variable is represented by ozone level counts, i.e., number

of ozone molecules in a fixed volume of air.

4.1 Dependence of ozone on three meteorological variables at sites in the New York region

The data for this example come from a study by Bruntz, Cleveland, Kleiner and Warner (1974) of the dependence of ozone on three meteorological variables over 111 days from May to September 1973 at sites in the New York metropolitan region. We will refer to this data as the NY Ozone data. The response variable is represented by ozone level counts and the three predictors are measurements of solar radiation, temperature and wind. A matrix scatterplot of the four variables is displayed in Figure 4.1. Cleveland, W.S., Devlin,

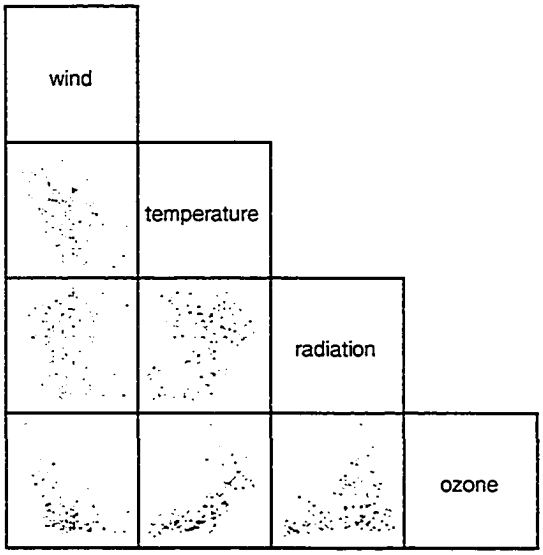

Figure 4.1: Matrix Scatterplot of the four variables in NY Ozone data.

S.J. and Grosse (1988) and Hastie and Tibshirani (1990) analyzed this data set using the cube root of ozone and Normal errors assumptions. Although Hastie and Tibshirani (1990) have suggested an analysis based on Poisson errors assumption, we have not found such an analysis in the literature. We assume the conditional distribution of the ozone counts given the three predictors is Poisson and apply ALB, GLM and GAM.

The ALB fit uses $\hat{K} = 5$ basis functions. A plot of the deviance residuals, assuming $\sigma^2 = 1$ is displayed in Figure 4.2. The magnitude of the residuals indicates that the responses show more variability than that explained by the Poisson distribution. The deviance statistic is 434.1979 based on 90 degrees of freedom. A chi-squared goodness of fit test indicates substantial evidence that

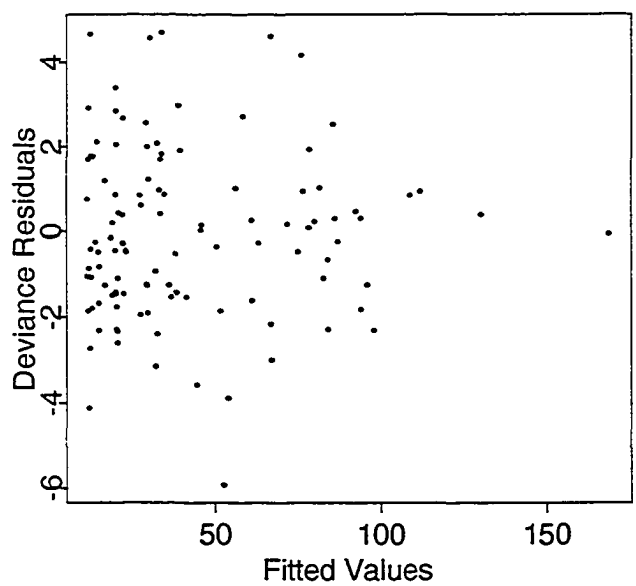

Figure 4.2: Deviance Residuals versus fitted values in NY Ozone data.

the Poisson model does not fit, p-value = 0. To obtain approximate standard errors, the quasi-likelihood approach was used. The deviance based estimator for the dispersion parameter σ was evaluated at 2.1964.

Since the measurements are taken over time on consecutive days, it is necessary to investigate whether serial correlation occurs. Serial correlation occurs when residuals from adjacent measurements are not independent of one another. Although in the presence of serial correlation the mean estimates are still unbiased, the standard errors are underestimated. Serial correlation can be detected either using residual plots or calculating an estimator of a first-order serial correlation. A residual plot displayed in Figure 4.3 does not indicate patterns in residuals over time. Cressie (1993) derives a robust estimator

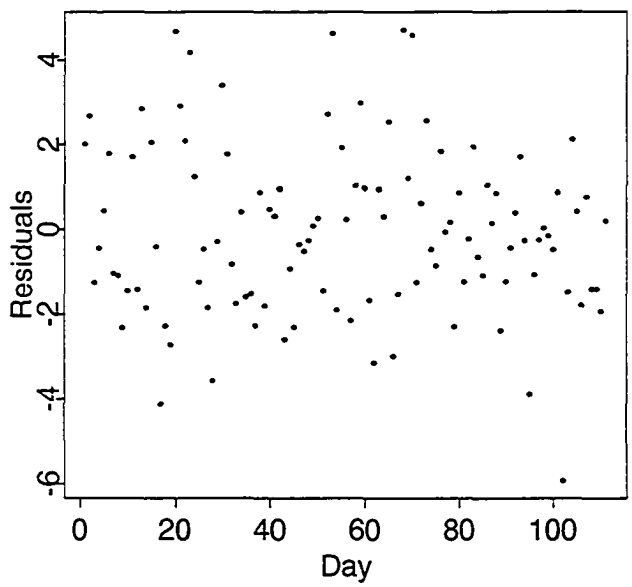

Figure 4.3: Residuals versus day in NY Ozone data.

of the first-order serial correlation:

$$\tilde{\rho} = 1 - \frac{1}{2}\left\{\frac{\sum_{i=1}^{n} |z_{i+1} - z_i|^{1/2}/(n-1)}{\sum_{i=1}^{n} |z_i - \tilde{z}|^{1/2}/n}\right\}^4,$$

where $\tilde{z} = \text{median}\{z_1, \ldots, z_n\}$ and z_i represents the residual at observation i. A value of -.0369 was obtained, indicating that serial correlation is not a problem for this data.

Boxplots of the standardized gradients of f displayed in Figure 4.4, indicate that all three variables are important in predicting ozone levels. We note that such plots could be misleading given dependencies among predictors, as suggested by the matrix scatterplot in Figure 4.1. An alternative method to measure importance of each predictor, derived in Chapter 3, is applied to this

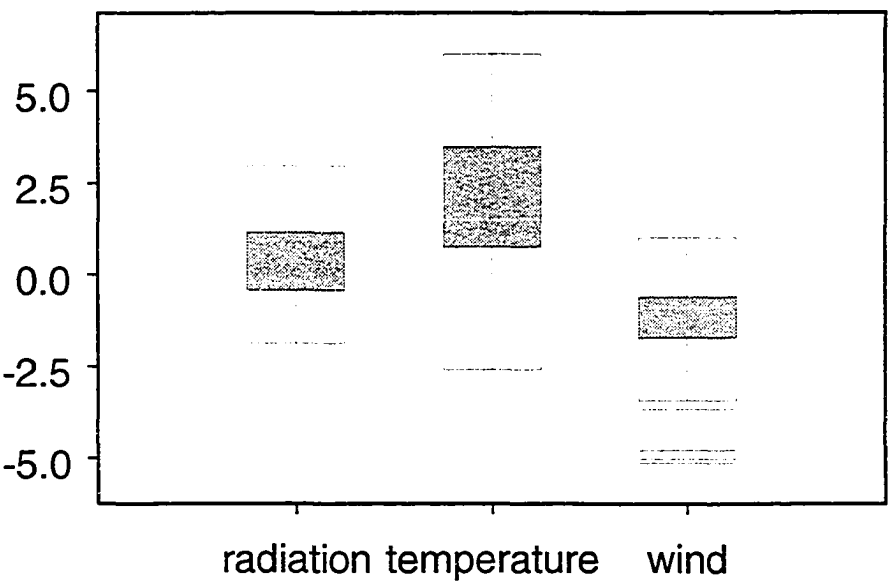

Figure 4.4: Boxplots of the three standardized gradients in NY Ozone data.

data. To measure the importance of a predictor, the values of that predictor are randomly permuted and the deviance from the ALB model is evaluated on each sample. The proportion of deviances as extreme as the one from the original sample, together with the standard error are listed for each predictor: .01(.0099) for radiation, .00(.00) for temperature and .01(.0099) for wind. Each proportion was calculated from 100 randomly permuted samples. The results suggest that the three variables are important in predicting ozone levels.

We use conditional plots to represent the ALB surface in one dimensional plots. In Figure 4.5, we condition the surface on first, second and third quantiles of radiation and temperature, respectively. In Figure 4.6, we condition the surface on first, second and third quantiles of radiation and wind, respectively. In Figure 4.7, we condition the surface on first, second and third quantiles of temperature and wind, respectively. The plots suggest that the ozone level decreases with increasing the wind speed, increases with increasing temperature, and increases with radiation up to a maximum and than decreases. As temperature increases, the curvature in the effect of the wind gets milder, closer to a roughly linear function, as seen in Figure 4.5. As radiation decreases, the curvature in the effect of the wind gets milder, closer to a roughly linear function, as seen in Figure 4.5. As temperature decreases, the curvature in the effect of the radiation gets milder, closer to a roughly linear function, as seen in Figure 4.7. As wind increases and temperature is in the first or second quantile, the curvature in the effect of the radiation gets milder, closer to a roughly linear function, as seen in Figure 4.7. As radiation decreases and wind

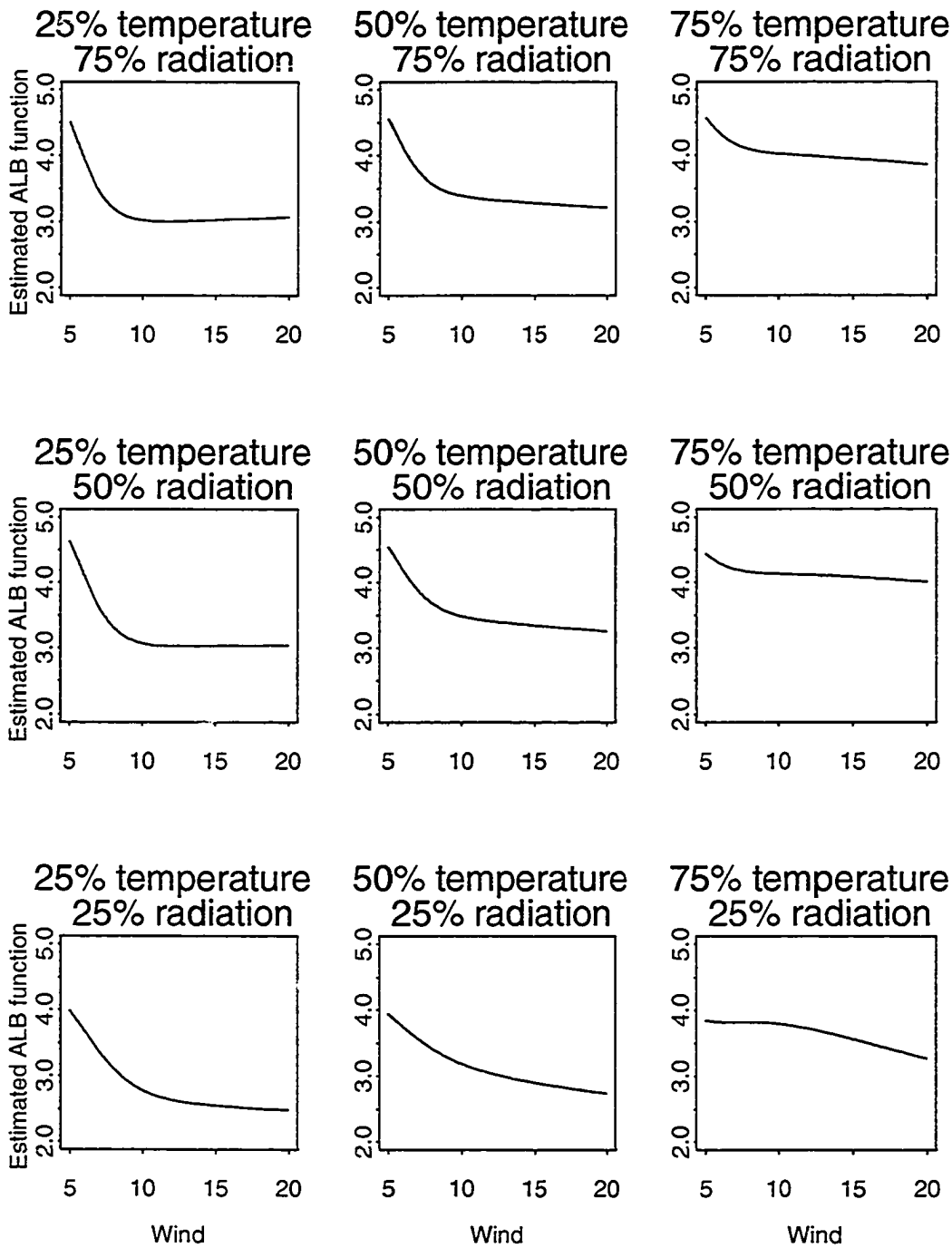

Figure 4.5: Conditional plots of estimated ALB function, $\hat{f}$, at 25%, 50% and 75% quantiles of radiation(113.5, 207, 255.5), and quantiles of temperature(71, 79, 84.5).

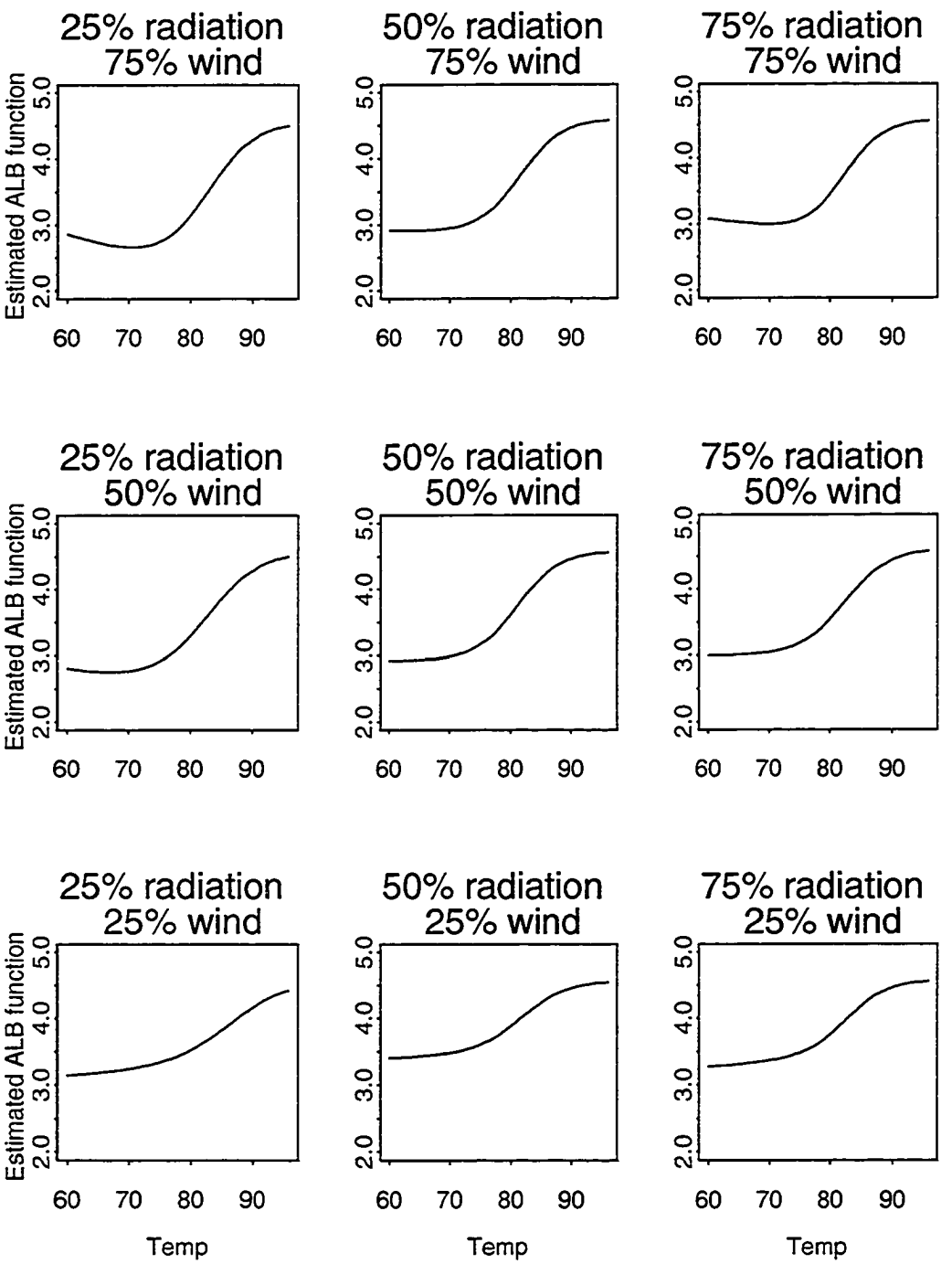

Figure 4.6: Conditional plots of estimated ALB function, $\hat{f}$, at 25%, 50% and 75% quantiles of radiation(113.5, 207, 255.5), and quantiles of wind(7.4, 9.7, 11.5).

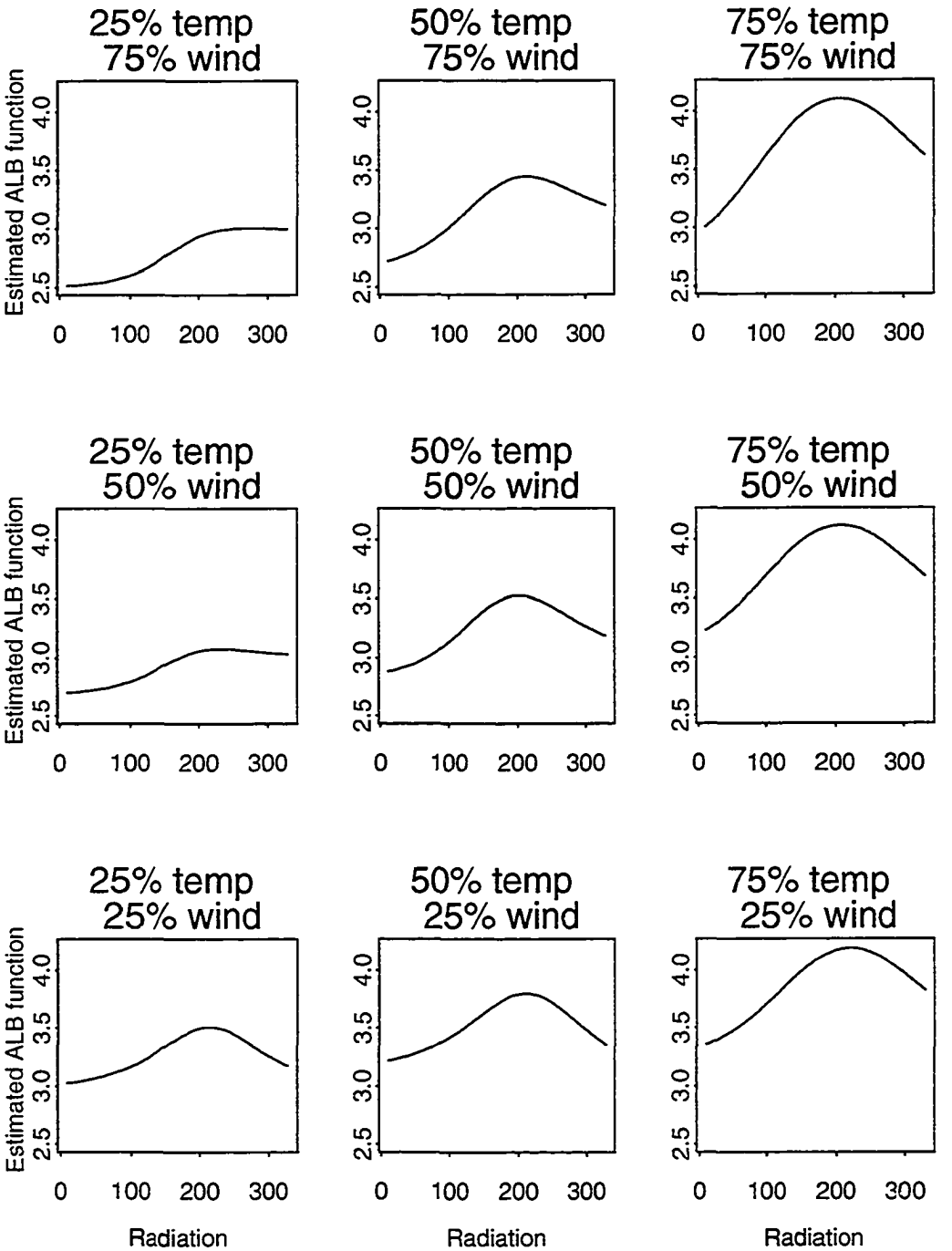

Figure 4.7: Conditional plots of estimated ALB function, $\hat{f}$, at 25%, 50% and 75% quantiles of temperature(71, 79, 84.5), and quantiles of wind(7.4, 9.7, 11.5).

is in the first quantile, the effect of temperature gets milder, closer to a roughly linear function, as seen in Figure 4.6. There appears to be a strong interaction between effect of radiation and temperature, Figure 4.7. Also, there appears to be a mild interaction between wind and temperature, Figure 4.5. We give a more through discussion on detecting interactions later in this section.

We applied GAM to the NY Ozone data. We started with the main effects and then fitted separately all three pairwise interactions using locally weighted surface smoothers. Based on a crude F-test, only the interaction between wind and temperature was found significant. The full three-dimensional surface was not found significant either. The same findings were obtained by Hastie and Tibshirani (1990), using the cube root of ozone and additive models with normal errors assumption. They have also applied MARS to the cube root of ozone. In addition to the interaction between temperature and wind, MARS also found a significant interaction between radiation and temperature. Later in this section we compare the predictive performance of ALB and GAM, based on the log-likelihood. We note that Hastie and Tibshirani (1990) conduct a comparison of predictive performance of GAM and other models, based on cube root of ozone and normal error assumptions. A comparison of predictive performance of ALB and GAM with Poisson errors assumption versus GAM and other models using cube root of ozone and normal errors assumption is therefore inappropriate.

It is interesting to investigate whether ALB detects possible interactions for the NY Ozone data. As it was noted in H01, plots of the ALB gradient

components

$$\hat{g}_j(\mathbf{x}) = \frac{\partial \hat{f}}{\partial x_j}$$

can be used to detect additive structure. If the effect of x_1 is additive, then the partial derivative $g_1(\mathbf{x})$ is a function of x_1 and therefore a plot of the estimate $\hat{g}_1(\mathbf{x})$ versus x_1 reveals little scatter about the curve. We note that, with a non-identical link, we should focus on the gradients of the ALB fit $\hat{f}(\mathbf{x})$, rather than $\hat{\mu}(\mathbf{x})$. The gradient of $\hat{\mu}(\mathbf{x})$ would indicate interactions between predictors generated by the inverse-link transformation, even if there are no interactions in the function f. For example, using a log-link, even if the effect of x_1 is additive, the partial derivative

$$\frac{\partial}{\partial x_1}\mu(\mathbf{x}) = \exp(f)\frac{\partial f}{\partial x_1},$$

would depend not only on x_1, but also on the rest of the predictors through $\exp(f)$, therefore suggesting interactions even if they are not present in the function f. Back to our example, plots of the gradient ALB estimates displayed in Figure 4.8 suggest possible interactions between the predictors, rather than

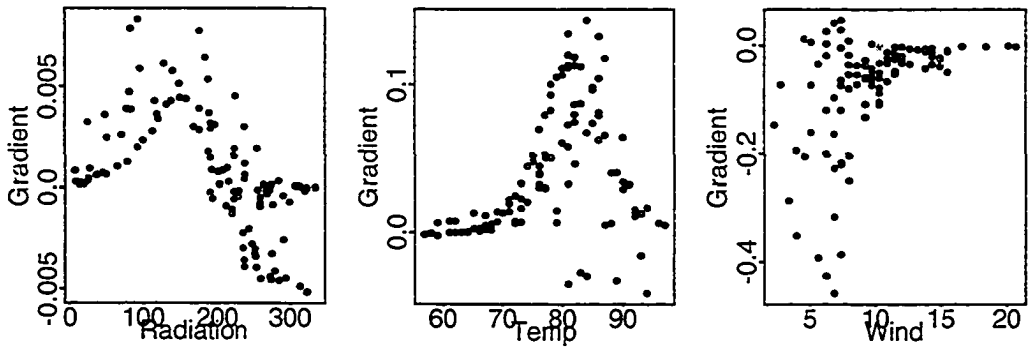

Figure 4.8: Matrix scatterplot of the three gradients of the log of the mean ALB estimate versus predictors in NY Ozone data.

just an additive structure. A GLM with all two-way interactions, indicate the interaction between wind and temperature as being the strongest among the two-way interactions.

Partial additivity can be investigated with rotating scatterplots. For example, if x_1 and x_2 interact but their joint effect on f is additive, then g_1 and g_2 are both functions of (x_1, x_2). Following this idea, we can investigate whether the combined interaction effect between wind and temperature is additive with respect to radiation. The rotating scatterplots of the gradient with respect to wind versus wind and temperature and the gradient with respect to temperature versus wind and temperature suggest a wind and temperature interaction and an additive effect of radiation.

The conclusions based on these scatterplots of the gradients are quite subjective. GAM has the advantage of providing conclusions based on tests and p-values, although it should be noted that their inference techniques are approximate.

Table 4.1: Ten-fold cross validated prediction error based on Kullback-Leibler distance together with the standard errors, for the different methods.

Method	Prediction error
GLM with main effects only	8.0658(1.0109)
GLM with temperature by wind interaction	8.9162(1.2169)
GLM with two-way interactions	9.1273(1.2549)
GLM with three-way interaction	9.6281(1.6438)
GAM with additive effects only, one df for each	8.0658(1.0109)
GAM with additive effects only, two df for each	9.2611(1.8341)
GAM with temperature by wind interaction	7.8040(0.9967)
ALB	6.4437(0.9262)

In order to compare the predictive performance of ALB and GAM, we have evaluated the 10-fold cross validated prediction error based on the Kullback-Leibler divergence, derived in section 2.3.1. The same groups were used for all methods, and the cross-validation partition is based on serial order of data. For a baseline comparison, we applied GLM with all three predictors. The prediction errors together with the standard errors are displayed in Table 4.1. Several GLM and GAM models were considered. GAM with one nominal degree of freedom corresponds to a GLM, main effects only. Details on how GAM were fitted are given in the next paragraph. Both ALB and GAM with an interaction surface between temperature and wind improve upon GLM and ALB performs better than GAM. There is a question whether these differences among GLM, GAM and ALB, are statistically significant. Also, is there evidence that some of the cross-validation groups are harder to predict than others, i.e., the prediction errors are significantly higher for some cross-validation groups than others. A two-way ANOVA assuming additive effects for methods and cross-validation groups, indicates there are significant differences among the ten cross-validation groups, p-value = .002, and that there are no significant differences among the three methods considered, GLM with main effects only, GAM with interaction surface between wind and temperature, and ALB, p-value = .189. The third and seventh cross-validation groups were harder to predict than the rest.

Now we give details regarding smoothers that have been used when fitting the above GAM's. As mentioned before, Hastie and Tibshirani (1990)

modeled pairwise interactions using locally weighted surface smoothers and found that the only significant interaction was the one between wind and temperature. A three-way interaction surface was not found significant. They used the cube root of ozone and normal errors assumption. We used the same approach in fitting GAM, assuming Poisson errors and obtained the same conclusions as they did regarding interaction terms. The same conclusions regarding interaction terms were obtained using a quasi-likelihood model with constant coefficient of variation. The above models were fitted using the whole data set. We encountered difficulty when trying to calculate the 10-fold cross-validated prediction error. The software S-Plus returned error messages regarding the predicted values on the test set when using GAM. The error messages were coming from Fortran, stating that the predictions cannot be calculated when using locally weighted surface smoothers and extrapolating. We considered replacing the locally weighted surface smoothers by smoothing splines, but S-Plus does not allow modeling interactions using smoothing splines. Switching to the software R and using the 'gam' package did not solve the predictions problem. However, the predictions were working when using GAM within a different package in R, called 'mgcv'. There are major differences between what this package provides and S-Plus. Locally weighted surface smoothers are no longer available. Instead, multi-dimensional smooths are available. They are based on penalized thin plate regression splines, providing a sensible way of modeling interaction terms in GAM, (Wood, 2003). The GAM models in Table 4.1 are based either on main effects only, or in-

teraction terms. The first two GAM models correspond to main effects only using smoothing splines in S-Plus, with a nominal one degree of freedom per term for the first model, and two degrees of freedom per term for the second model. To fit interaction terms we used R. The last GAM model in Table 4.1 corresponds to a four degrees of freedom thin-plate regression spline term for radiation and a two-dimensional smooth based on thin plate regression splines to model the interaction between wind and temperature.

4.2 Dependence of ozone on eight meteorological measurements made in the Los Angeles basin

Data for this example come from a case study regarding the dependence of ozone level on eight daily meteorological measurements made in the Los Angeles basin in 1976. Although measurements were made every day that year, some measurements were missing; we use the 330 complete cases. We will refer to this data as the LA Ozone data. The response variable is represented by ozone level counts and there are nine predictors, including day of the year and eight meteorological measurements: 500 millibar pressure height, measured at the Vandenberg air force base (vdht), wind speed at Los Angeles airport (wind), humidity at Los Angeles airport (hmdt), Sandburg Air Force Base temperature (temp), inversion temperature base height (ibht), pressure gradient from Los Angeles airport to Daggert (dgpg), inversion base temperature at Los Angeles airport (ibtp), visibility (vis). A scatterplot matrix of the ten variables is displayed in Figure 4.9. Hastie and Tibshirani (1990) analyzed

this data set using the log of ozone level counts and Normal errors assumptions. We assume the conditional distribution of the ozone counts given the nine predictors is Poisson and apply ALB, GLM and GAM. We mention that previous analysis of this data set indicate additive structure, so an additive model is likely to provide a more parsimonious fit compared with models allowing interactions.

The ALB fit uses $\hat{K} = 5$ basis functions. A plot of the deviance residuals, assuming $\sigma^2 = 1$ is displayed in Figure 4.10. The magnitude of the residuals indicates that the Poisson distribution is appropriate. The deviance statistic is 246.1218 based on 285 degrees of freedom. A chi-squared goodness of fit test indicates that the data are consistent with a Poisson model, (p-value = 0.9535).

Since the measurements are taken over time on approximately consecutive days, it is necessary to investigate whether serial correlation occurs. A residual plot displayed in Figure 4.11 does not indicate patterns in residuals over time. The same formula for an estimator of the first-order serial correlation as in the previous example was used. A value of -0.0664 was obtained, indicating that serial correlation is not a problem for this data.

Boxplots of the standardized gradients of f displayed in Figure 4.12, suggest the following variables as being important in predicting ozone levels: humidity, dgpg , ibtp, wind, visibility and day of the year. The rest of the variables: vdht, temperature, and ibht may not be important as suggested by the boxplots of the standardized gradients. Wind and visibility are less clear,

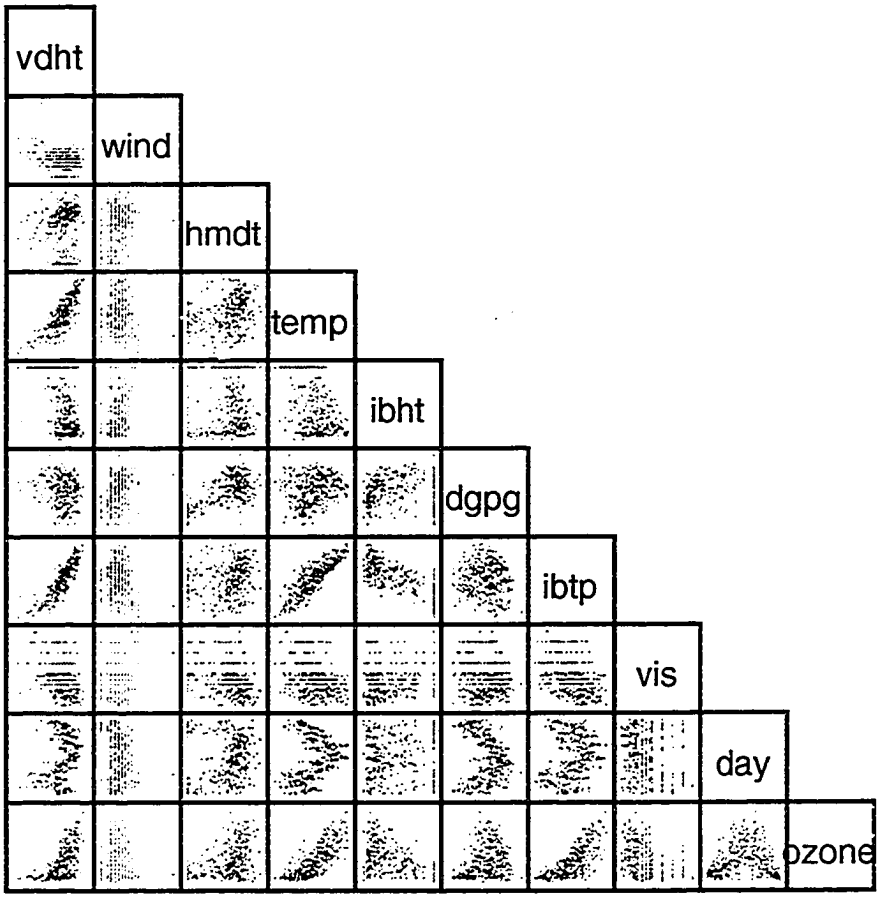

Figure 4.9: Scatterplot matrix of the nine predictors in LA Ozone data.

they may be important or not. We note that such plots could be misleading given dependencies among predictors, as suggested by the scatterplot matrix in Figure 4.9. The shuffling technique to measure importance of each predictor, derived in Chapter 3, is applied to this data. The proportion of deviances as extreme as the one from the original sample together with the margin of er-

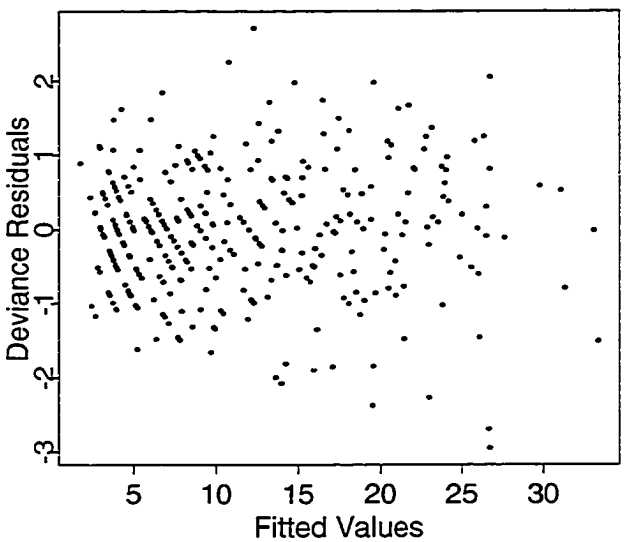

Figure 4.10: Deviance Residuals versus fitted values in LA Ozone data.

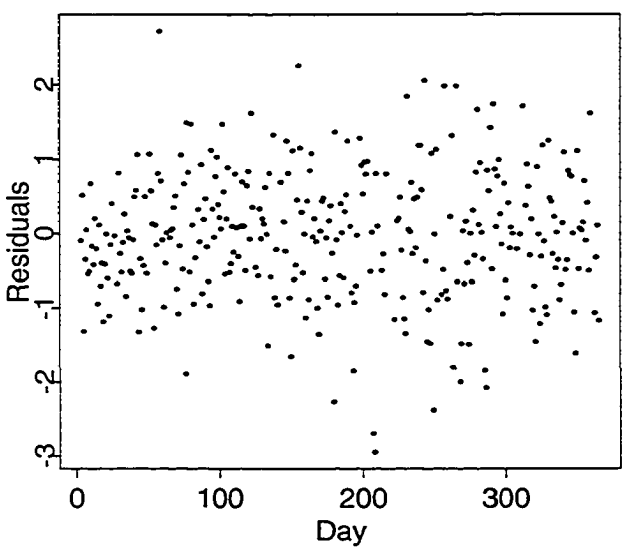

Figure 4.11: Residuals versus day in LA Ozone data.

ror, calculated using the Agresti-Coull(AC) interval for a binomial proportion, are listed for each predictor: .36(.09) for vdht, .11(.06) for wind, .00(.03) for humidity, .28(.09) for temperature, .18(.08) for ibht, .00(.03) for dgpg, .01(.03) for ibtp, .01(.03) for visibility and .00(.03) for day of the year. The AC interval for a binomial proportion is discussed in Brown et al., 2001. Brown et al., 2001 remind the reader that the actual coverage probability of a standard

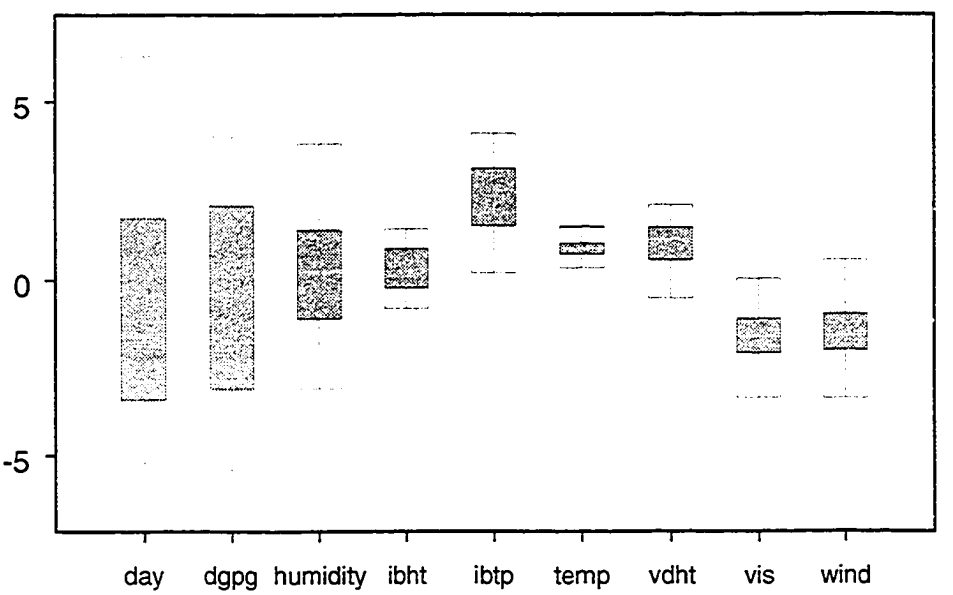

Figure 4.12: Boxplots of the nine standardized gradients in LA Ozone data.

Table 4.2: Suggested important variables for the different methods.

Method	Suggested important variables
ALB, using boxplots	hmdt, dgpg , ibtp, wind, vis, day
ALB, using shuffling technique	hmdt, dgpg , ibtp, vis, day
GAM	hmdt, dgpg , ibtp, vis, day

interval is poor for p near 0 or 1. They discuss and test several alternatives for the standard interval and recommend the AC interval for larger n ($n > 40$). The AC interval(Agresti and Coull, 1998) is also called the " add 2 successes and 2 failures" interval. To compute the AC interval, we add 2 successes and 2 failures and then use the same formula as for the standard interval. Table 4.2 displays the suggested important variables for ALB using boxplots and the shuffling technique and for GAM. We note that except for the wind variable, the same sets of important variables were suggested by both techniques. We note that GAM returns the same sets of important variables as the numeric shuffling technique. We give more details about how GAM was fitted to the LA Ozone data later in this section.

We have mentioned before that Hastie and Tibshirani (1990) analyzed this data set using GAM, the log of ozone level counts and Normal errors assumptions. They used smoothing regression splines with a nominal four df for each predictor, no interaction terms. We fitted GAM using the ozone level counts and Poisson errors assumption following their approach of using smoothing regression splines with a nominal four df each. In order to compare

Table 4.3: Ten-fold cross validated prediction error based on Kullback-Leibler distance together with the standard errors, for the different methods.

Method	Prediction error
GLM with main effects only	1.9319(.3199)
GAM with additive effects only, four df for each	1.3125(.1476)
GAM with interaction term	1.2228(.1329)
ALB	1.7114(.3363)

the predictive performance of ALB and GAM, we have evaluated the 10-fold cross validated prediction error based on the Kullback-Leibler divergence, derived in Section 2.3.1. The same groups were used for all methods, and the cross-validation partition was based on serial order of data. For a baseline comparison, we applied GLM to all nine predictors, main effects only. The prediction errors together with the standard errors are displayed in Table 4.3. As mentioned before, we fitted GAM using the ozone level counts, Poisson errors assumption, and smoothing regression splines with a nominal four df each. Details on how GAM with an interaction surface were fitted are given in the next paragraph. Both ALB and GAM improve upon GLM and GAM performs better than ALB. There is a question whether these differences among GLM, GAM and ALB, are statistically significant. Also, is there evidence that some of the cross-validation groups are harder to predict than others, i.e., the prediction errors are significantly higher for some cross-validation groups than others. A two-way ANOVA assuming additive effects for methods and cross-validation groups, indicates there are significant differences among the ten cross-validation groups, p-value = .000, and that there are significant differences among the four methods considered, GLM with main effects only, GAM with additive effects, GAM with interaction surface, and ALB, p-value = .025. A post-hoc comparison indicates that GAM with interaction surface did significantly better than GLM, there are no significant differences among GLM, GAM with additive effects, and ALB, and there are no significant differences among GAM with additive effects, GAM with interaction surface, and

ALB. The first cross-validation group was harder to predict than the rest.

It is interesting to investigate whether ALB detects possible interactions for the LA Ozone data. Plots of the gradient ALB estimates displayed in Figure 4.13 suggest possible interactions between the predictors, rather than just an additive structure. We investigated the effect of modeling the interaction between humidity and inversion base temperature (ibtp) using a locally-weighted surface smoother within GAM. We remind the reader that S-Plus does not allow modeling interactions using smoothing splines. Based on a crude F-test, the interaction between humidity and inversion base temperature was found significant. It is interesting to see whether including this significant interaction term between humidity and inversion base temperature would have an effect on the 10-fold cross-validated prediction error. Since there are problems with the use of locally-weighted surface smoothers in S-Plus when trying to predict on a test set, we switched to R and calculated the 10-fold cross-validated prediction error using a two-dimensional smooth based on thin plate regression splines to model the interaction between humidity and inversion base temperature, and one-dimensional thin-plate regression spline terms for the rest of seven predictors. As displayed in Table 4.3, GAM with the interaction term improves upon GAM with main effects, but the difference between the two methods was not found to be statistically significant. However, GAM with interaction surface performs significantly better than GLM.

This example illustrates how ALB models can be used in high dimensional data sets to explore possible interactions among predictors. Although

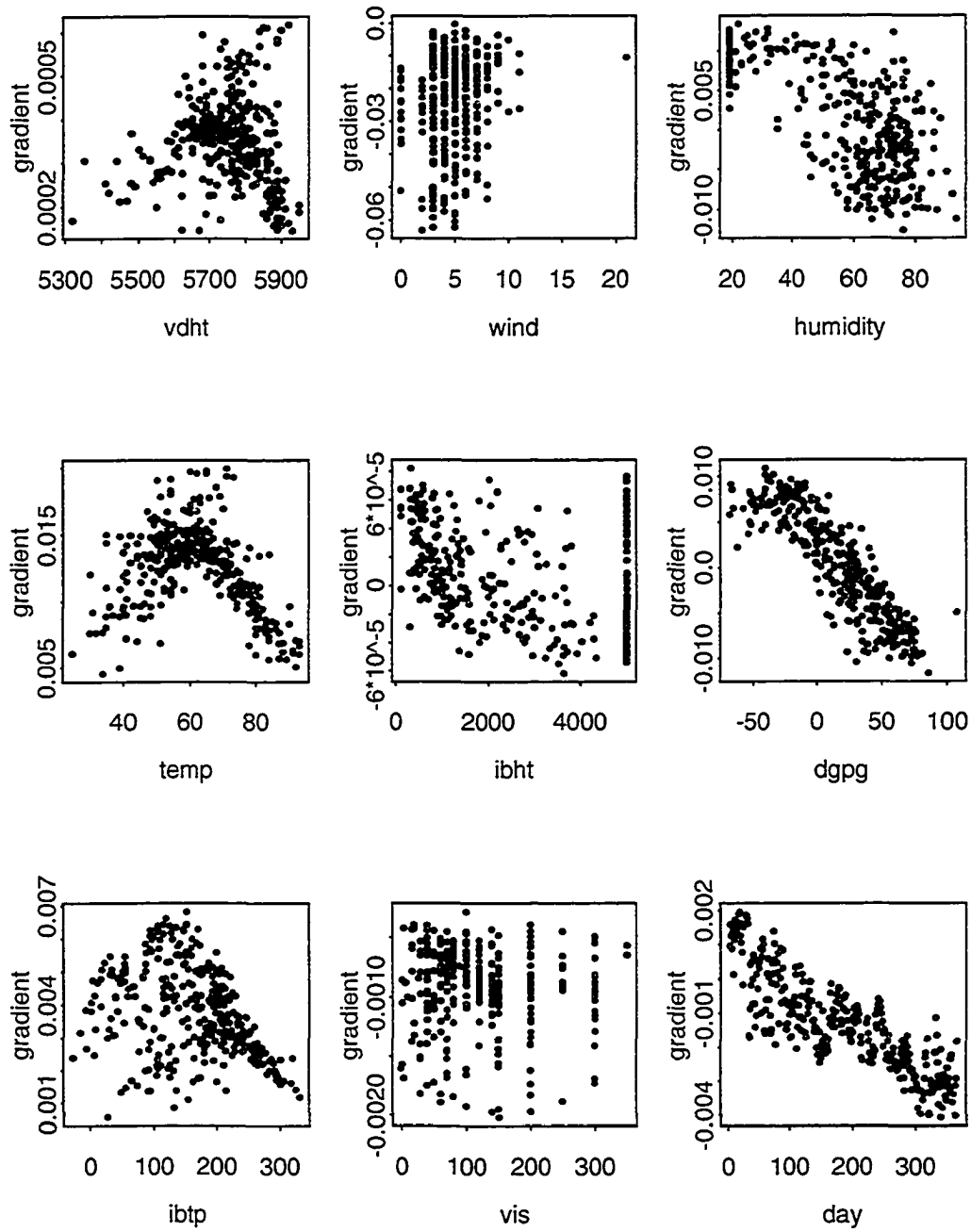

Figure 4.13: Matrix scatterplot of the nine gradients of the log of the mean estimate versus predictors in LA Ozone data.

GAM performs better than ALB in this example, we can still use ALB to identify possible interactions and then use GAM for further exploration, based on the guidelines offered by ALB regarding interactions. We showed how the cross-validated prediction error was improved when modeling an interaction suggested by ALB, leading to a GAM model with an improved prediction power. ALB proves to be an automatic method that provides interesting insights into multidimensional data sets.

Chapter 5

A comparison of ALB models, GPP and GAM

5.1 Introduction

In Chapter 2 of this thesis we investigate and compare predictive accuracy of ALB models and GAM. We remind the reader that we used GAM based on thin plate regression splines with smoothing parameters selected by either a GCV criterion or an Un-Biased Risk Estimator criterion (UBRE) which is an approximation to AIC, (Wood 2003). GAM based on thin plate regression splines was implemented in R, package 'mgcv'. To model interactions, multidimensional smooths are available using penalized thin plate regression splines.

In this chapter, we investigate predictive accuracy of ALB compared with GAM restricted to component-wise additive models, that is:

$$g(E(Y|x_1,\ldots,x_d)) = \alpha + f_1(x_1) + \ldots + f_d(x_d),$$

as well as Generalized Projection Pursuit(GPP) models, that is:

$$g(E(Y|\mathbf{x})) = \beta_0 + \sum_{j=1}^{M} \beta_j f_j(\alpha_j'\mathbf{x}).$$

We note that we actually consider the Poisson distribution, so g is the log-link function.

Additive models are motivated by the failure of linear models in situations where the effects of the predictors are not linear. Additive models identify and characterize nonlinear effects retaining an important feature of linear models: they are additive in the predictor effects, making interpretation easier. Although the additive models approach has a number of attractive properties, not all possible underlying regression functions can be modeled as a sum of smooth functions. The additive model is a special case of a more general model that goes beyond component-wise additivity, the Projection Pursuit model:

$$E(Y|\mathbf{x}) = \beta_0 + \sum_{j=1}^{M} \beta_j f_j(\alpha_j'\mathbf{x}).$$

The scope of this chapter is to compare predictive accuracy of ALB, GPP and GAM (component-wise additivity models), for the case where the conditional distribution of the response given the predictors is Poisson. The organization of this chapter is as follows. In Section 2, we give background on Projection Pursuit and the extension to the exponential family (GPP). In Section 3, we investigate and compare predictive accuracy of ALB, GPP and GAM (component-wise additivity models).

5.2 Background on Projection Pursuit models

The central idea behind Projection Pursuit models is to extract linear combinations of the predictors as derived features, and then model the target as a nonlinear function of these features. The Projection Pursuit regression (PPR) model has the form:

$$E(Y|\mathbf{x}) = \beta_0 + \sum_{j=1}^{M} \beta_j f_j(\alpha_j'\mathbf{x}).$$

As in other smoothing problems, we need to impose complexity constraints for identifiability of the model components:

$$\begin{aligned} \sum_{i=1}^{d} \alpha_{ij}^2 &= 1 \\ \frac{1}{n}\sum_{i=1}^{n} f_j(\alpha_j'\mathbf{x}_i) &= 0 \\ \frac{1}{n}\sum_{i=1}^{n} f_j^2(\alpha_j'\mathbf{x}_i) &= 1. \end{aligned}$$

This is an additive model, but in the derived features $\alpha_j'\mathbf{x}$, rather than the predictors themselves. The functions f_j are unspecified and are estimated along with the directions α_j using some flexible smoothing method.

The function $f_j(\alpha_j'\mathbf{x})$ is called a ridge function in $\Re^p$ and it varies in the direction defined by the vector α_j. The PPR model is very general, since the operation of forming nonlinear functions of linear combinations generates a large class of models. For example, the product x_1x_2 can be written as $((x_1 + x_2)^2 - (x_1 - x_2)^2)/4$ and higher-order products can be represented similarly.

PPR is a generalization of additive models, addressing two of their limitations: additive models can at best find the additive function closest to the

true regression function, and additive models are not invariant to rotations of the predictor space. The PPR has the universal approximation property; i.e., asymptotically in the number of linear combinations M, PPR can approximate any continuous function (Diaconis and Shahshahani, 1984). However, this generality comes with a price. Interpretation of the fitted model is usually difficult, because each predictor enters the model in a complex way. As a result, PPR is most useful for prediction, and not as useful for producing an understandable model for the data.

The standard PPR algorithm of Friedman and Stuetzle (1981) estimates the smooth functions f_j using the supersmoother nonparametric scatterplot smoother. Friedman's algorithm constructs a model with $M_{\max}$ linear combinations, then prunes back to a simpler model of size $M \leq M_{\max}$, where M and $M_{\max}$ are specified by the user.

Friedman et al. (1983) propose a regression spline PPR algorithm which they call the Multidimensional Additive Spline Approximation (MASA). In this algorithm, the user specifies the number of terms M to add, and the number of knots to use in each regression spline. Another PPR algorithm is that of Hwang et al. (1993). They estimate the smooth functions using high degree polynomials (in the examples they present, they use polynomials of degree 7 or 9). The same degree polynomial is used for each term in the fit and the user specifies M and $M_{\max}$.

Roosen and Hastie (1993) present a smoothing spline PPR algorithm based on the PPR algorithm described by Hwang et al. (1993) and Friedman

(1985). The general approach is to fit the terms in a stepwise manner using backfitting between the addition of terms. A total of $M_{\max} \geq M$ terms are fitted, and a backwards selection procedure is used to prune the fit down to M terms.

Hwang et al. (1993) note that two desirable qualities of polynomials are that they provide a fast and accurate derivative calculation, and they provide a smooth interpolation. Roosen and Hastie (1994) note that smoothing splines also have these properties, and in addition can be used to choose the smoothing parameters and the number of terms automatically. They also note that the local smoothness of the fit provided by smoothing splines will, in many cases, make them more reliable than the supersmoother. Roosen and Hastie (1994) propose an Automatic Smoothing Spline Projection Pursuit (ASP) algorithm with the smoothing parameters and number of terms selected automatically using a GCV criterion. The ASP algorithm is a direct descendant of the non-adaptive smoothing spline PPR algorithm described in the previous paragraph.

Roosen and Hastie (1993) develop a Generalized Projection Pursuit (GPP) framework for the exponential family models using the canonical link:

$$g(E(Y|\mathbf{x})) = \beta_0 + \sum_{j=1}^{M} \beta_j f_j(\alpha_j'\mathbf{x}).$$

The GPP algorithm is a modification of the ASP algorithm, using the local scoring loops around the backfitting loops, analogous to the way GAM algorithm generalizes the Additive Models algorithm. We note that Roosen and Hastie (1993) only provide simulated results for a binary response using a logit

link:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{M} \beta_j f_j(\alpha_j' \mathbf{x}),$$

where $p = P(Y = 1|\mathbf{x}) = E(Y|\mathbf{x})$. No simulations or applications are provided for a Poisson distributed response. We contacted the authors asking for the code and they confirmed that the code was never adjusted to fit a Poisson distributed response. Only logistic regression with a binary response was performed. We have modified the code to account for the Poisson distributed response, i.e., we modified the adjusted dependent variable and weights in the local scoring algorithm, and the corresponding starting values.

In the next section, we present results of simulation studies regarding predictive accuracy of ALB, GPP, and GAM (component-wise additivity, as well as interactions).

5.3 Comparing predictive accuracy of ALB, GPP and GAM

The affine invariance of ALB suggests comparison with PPR. Both methods employ a linear combination of simpler functions and neither is affected by scaling or rotation of the covariates. PPR approximates the regression function by a sum of one-dimensional ridge functions. The ridge function are estimated using one-dimensional smoothers and can incorporate several bumps. The logistic basis functions employed by ALB are more complex than ridge functions in one respect, being multi-dimensional, but are simpler in other respects, with quasi-concave shape constrained by a parametric family. Regarding starting

values. Roosen and Hastie (1994) note that due to the highly nonlinear nature of the PPR problem, the particular initial direction coefficients $\{\alpha_{kj}\}$ used can have a dramatic effect on the results of the algorithm.

In the following simulation studies we compare predictive performance of ALB models, GPP and GAM for Poisson distributed response. For a description of the prediction error and simulation protocols we refer the reader to Sections 2.3.1 and 2.3.2. We note that in Section 2.3.1 and 2.3.2 we used GAM with thin-plate splines to model interaction surfaces, (Wood, 2003). For the rest of this section, we refer to this method as GAM(W). Here, we also look at GAM as component-wise additive models, (Hastie and Tibshirani, 1986), and refer to it as GAM(HT). GAM(W) using thin-plate splines to model interaction surfaces was fit using package 'mgcv' in R, the code being written by the author (Wood, 2003). The smooth terms are represented using penalized thin-plate regression splines, allowing for interaction surfaces, with smoothing parameters selected by a GCV criterion. We note that GAM(W) allows a more flexible additive model, but the interaction surfaces need to be specified by the user. When there is a priori knowledge about the form of the target function, for example, we know that there is an interaction between the first two predictors and an additive effect of the third predictor, GAM(W) is able to take advantage of it employing a model of the form $s(x_1, x_2) + s(x_3)$, leading to a better performance. GAM(HT) as a component-wise additive model was fit in S-Plus using smoothing regression splines with a fixed number of degrees of freedom for each of the predictor variables. As Roosen and Hastie (1993)

have already used GPP on a number of functions, their work provides a valuable test bed in which to compare ALB with GPP. In fact, Roosen and Hastie (1993) used functions from Hwang et al. (1993) to examine the supersmoother and polynomial based PPR. They only use two-dimensional functions, and also investigate performance of their methods using additional nuisance predictors. In the last three simulation studies, we use functions from Hwang et al. (1993). In the first three simulations studies we use functions from Section 2.3.2. The simulation studies results are displayed in Table 5.1.

Example 5.3.1 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-1,1)^d$. The target function is not an ALB function:

$$f(\mathbf{x}) = \frac{2\sin(\pi(x_1+1)/2)}{(x_1/2+1)}$$

We note that it does not make sense to use GPP to model data with only one predictor. However, it is interesting to compare ALB with GPP and GAM using additional nuisance predictors. In this example, we used four nuisance variables, for a total of five predictors. The prediction errors for the methods we investigate are very close, especially with larger sample sizes, $n = 100$ and $n = 200$. For a smaller sample size, $n = 50$, GAM(W) performs slightly better. GAM(W) is here represented by $s(x_1) + \ldots + s(x_5)$, no interaction surfaces. A closer look indicates that, especially for the smaller sample size, $n = 50$, ALB and GPP are more affected by the nuisance variables than GAM(W), or GAM(HT). Methods that fit each axis separately, such as GAM, are expected to perform better with this example.

Table 5.1: Performance measures: prediction error averages from 100 replicated samples of size n for Poisson version of ALB models, Generalized Additive model (Wood, 2003), Component-wise GAM (Hastie and Tibshirani, 1986), and GPP together with the corresponding lower bound of the prediction error.

No.	d	n	$\hat{K}$	lower bound	ALB	GAM(W)	GAM(HT)	GPP
1	5	50	3.24	1.2385	1.5795	1.4855	1.6025	1.5825
	5	100	3.20	1.2214	1.3911	1.3821	1.4219	1.3901
	5	200	3.21	1.2554	1.3422	1.3329	1.3585	1.3392
2	2	50	4.08	1.2608	1.5430	1.4312	1.5219	1.6015
	2	100	4.17	1.2534	1.4063	1.3522	1.4031	1.4408
	2	200	4.54	1.2437	1.3258	1.2936	1.3204	1.3419
	5	50	4.08	1.2362	1.7628	1.5051	1.5512	1.9235
	5	100	4.12	1.2418	1.5021	1.4052	1.4132	1.7124
	5	200	4.26	1.2525	1.4042	1.3336	1.3521	1.5735
3	4	50	4.03	1.2619	1.7703	4.8466	9.2317	3.9205
	4	100	4.23	1.2369	1.7172	3.0891	8.5076	2.7104
	4	200	4.28	1.2548	1.6868	2.5479	8.1224	2.3220
4	2	50	4.57	1.2364	1.5529	1.4809	7.1719	1.6019
	2	100	4.76	1.2015	1.3973	1.3905	6.3372	1.4157
	2	200	5.05	1.2286	1.3233	1.3160	5.7549	1.3367
	5	50	4.21	1.2313	1.7591	1.6355	8.6984	1.7621
	5	100	4.55	1.2299	1.5541	1.4341	7.6760	1.5528
	5	200	5.14	1.2150	1.3577	1.3271	7.3507	1.3614
5	2	50	5.18	1.2419	1.6584	1.5741	2.1094	1.7753
	2	100	5.79	1.2567	1.5166	1.4800	2.0591	1.6054
	2	200	6.51	1.2408	1.4051	1.3987	1.9404	1.4135
	5	50	4.81	1.2345	1.9480	1.7104	2.2901	2.0142
	5	100	6.29	1.2409	1.7102	1.5034	2.2617	1.9341
	5	200	6.93	1.2348	1.5157	1.3863	2.1439	1.7912
6	2	50	8.89	1.2200	1.9356	1.9663	4.2602	1.9241
	2	100	11.53	1.2101	1.6218	1.8831	3.0441	1.6042
	2	200	15.62	1.2181	1.7612	1.7942	2.4082	1.6839
	5	50	7.59	1.2279	3.3134	1.8623	9.9684	2.7361
	5	100	10.64	1.2537	3.0522	1.7294	4.6386	2.5116
	5	200	14.42	1.2215	2.3334	1.7823	4.2287	2.2813

Example 5.3.2 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-1,1)^d$. The target function is an additive function of the first two covariates:

$$f(\mathbf{x}) = 1.5x_1^2 + 1.5\frac{\sin(\pi(x_2+1)/2)}{(x_2/2+1)}.$$

GAM performs better than ALB and GPP, as expected since the underlying regression function is additive. The difference in the prediction errors is smaller as n increases. When adding three nuisance variables, $(d = 5)$ the accuracy deteriorates affecting ALB and GPP more than GAM, especially for the smaller sample size, $n = 50$. Methods that fit each axis separately, such as GAM, are expected to perform better with this example. We note that ALB performs better than GPP on this additive example and seems to be less affected by the nuisance variables. We also note that GAM(W) performs better than GAM(HT). GAM(W) is here represented by $s(x_1) + s(x_2)$ for $d = 2$, and by $s(x_1) + \ldots + s(x_5)$ for $d = 5$, no interaction surfaces.

Example 5.3.3 A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(-3,3)^4$. The target function is an ALB function defined on a 3-dimensional projection of $\Re^4$; ie., $f(\mathbf{x}) = f_K(\mathbf{z})$ where $K = 5$, $\mathbf{z} = (z_1, z_2, z_3)'$,

$$z_1 = \sqrt{3}(x_1 + x_2 + x_3 + x_4 - 2)$$

$$z_2 = \sqrt{3}(x_1 + x_2 - x_3 - x_4)$$

$$z_3 = \sqrt{3}(x_1 - x_2 + x_3 - x_4)$$

The reference point parameterization is used to specify f_K: $\xi_1 = (1,0,0)'$,

$\xi_2 = (-1, 0, 0)'$, $\xi_3 = (0, 1, 0)'$, $\xi_4 = (0, 0, 1)'$, $\xi_5 = (0, 0, 0)'$, $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 0$, $\delta_1 = \delta_2 = .5$, $\delta_3 = \delta_4 = 3.5$, $\delta_5 = 0$ and $\tau = 1$. The target f can be expressed as an ALB function of $\mathbf{x}$ and has interactions of all orders among the four covariates. The performance of ALB is substantially better than that of GPP and GAM(W). GAM(W) has difficulty modeling higher order interactions, even when n is large. With $n = 50$, there are not enough degrees of freedom for GAM to model three-way interactions. With $n = 200$, there are not enough degrees of freedom for GAM to get all three-way interactions. On this example, GAM(W) performs much better than component-wise GAM(HT), as expected. GPP performs better than GAM(W), but worse than ALB. The logistic basis functions employed by ALB are more complex than ridge functions, being multi-dimensional, and therefore leading to a better performance of ALB on this example. In further simulations, the average prediction error for GAM and GPP remains roughly constant as n increases from 200 to 2000.

Example 5.3.4 This example was used by Hwang et al. (1993) and Roosen and Hastie (1993, 1994) to test performance of the supersmoother and polynomial PPR algorithms, ASP and GPP. A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(0, 1)^d$. The target function is a simple interaction of the first two covariates:

$$f(\mathbf{x}) = 10.391((x_1 - .4)(x_2 - .6) + .36).$$

The prediction errors for ALB, GPP and GAM(W) are very close, especially with larger sample sizes, $n = 100$ and $n = 200$. For a smaller sample size,

$n = 50$, GAM(W) performs slightly better. GAM(W) is here based on a two-dimensional smooth interaction surface $s(x_1, x_2)$. GAM(HT) performs much worse than the other models we consider, since the target function is an interaction of two covariates. ALB performs slightly better than GPP for $d = 2$. When adding three nuisance predictor variables, the prediction errors of ALB and GPP are very close. For $d = 5$, we note that GAM(W) performs better than ALB and GPP, but we remind the reader that in this case GAM(W) has an advantage; in this example GAM(W) is based on a two-dimensional smooth interaction surface and additive effects for the rest of the three predictors, i.e.,

$$g(E(y|\mathbf{x})) = s(x_1, x_2) + s(x_3) + s(x_4) + s(x_5).$$

So, we take advantage of the known form of the target function when fitting GAM(W), leading to improved performance over ALB and GPP.

Example 5.3.5 This example was used by Hwang et al. (1993) and Roosen and Hastie (1993, 1994) to test performance of the supersmoother and polynomial PPR algorithms, ASP and GPP. A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(0, 1)^d$. The target function is a radial function in the first two covariates:

$$f(\mathbf{x}) = 24.234(r^2(0.75 - r^2)), \text{ where } r^2 = (x_1 - .5)^2 + (x_2 - .5)^2.$$

On this example, ALB performs better than GPP. Component-wise GAM(HT) performs much worse than the other models we consider, since the target function is an interaction of two covariates. GAM(W) is here based on a

two-dimensional smooth interaction surface $s(x_1, x_2)$. GPP performs slightly better than ALB for $d = 2$. When adding three nuisance predictor variables, the prediction errors of ALB and GPP are very close. As in the previous example, for $d = 5$, we note that GAM(W) performs better than ALB and GPP, but we remind the reader that in this case GAM(W) has an advantage; in this example GAM(W) is based on a two-dimensional smooth interaction surface and additive effects for the rest of the three predictors, i.e.,

$$g(E(y|\mathbf{x})) = s(x_1, x_2) + s(x_3) + s(x_4) + s(x_5).$$

So, we take advantage of the known form of the target function when fitting GAM(W), leading to improved performance over ALB and GPP.

Example 5.3.6 This example was used by Hwang et al. (1993) and Roosen and Hastie (1993, 1994) to test performance of the supersmoother and polynomial PPR algorithms, ASP and GPP. A sample of size n was generated with each predictor vector $\mathbf{x}$ generated uniformly on the hypercube $(0, 1)^d$. The target function is a complicated interaction function in the first two covariates:

$$f(\mathbf{x}) = 1.9(1.35 + \exp(x_1)\sin(13(x_1 - .6)^2)\exp(-x_2)\sin(7x_2)).$$

On this example, GPP performs better than ALB. We note that this complicated interaction function produces a surface with many bumps and ripples occurring in multiple directions. This is a good example to illustrate limitations on the complexity of the ALB function. A large number of basis functions is required to approximate functions with many bumps and ripples occurring in multiple directions. The averages over the 100 samples of the number of

basis functions selected by ALB is displayed in Table 5.1. The ridge regression functions estimated using one-dimensional smoothers can incorporate several bumps, and they are more successful in this example, than the quasi-concave logistic basis functions employed by ALB. We note that, for $d = 2$, ALB and GPP perform better than GAM(W). Component-wise GAM(HT) performs much worse than the other models we consider, since the target function is an interaction of two covariates. GPP performs slightly better than ALB for $d = 2$. As in the previous two examples, when adding three nuisance predictors ($d = 5$), we note that GAM(W) performs better than ALB and GPP, but we remind the reader that in this case GAM(W) has an advantage; the same discussions as in the previous two examples apply here.

To summarize, ALB performs better than GPP on some of these examples, and worse than GPP on others. The performance of these methods depends on the target function, and is affected by the different basis functions employed by the two methods. PPR approximates the regression function by a sum of one-dimensional ridge functions. The ridge functions are estimated using one-dimensional smoothers and can incorporate several bumps. The logistic basis functions employed by ALB are more complex than ridge functions in one respect, being multi-dimensional, but are simpler in other respects, with quasi-concave shape constrained by a parametric family. When we have previous knowledge of the additive or partially additive structure of the regression function, GAM(W) has a greater advantage over ALB and GPP, using models that exploit additivity. Especially in the last example, we note that when

nuisance variables are added to the predictors, ALB tends to compensate by reducing the number of basis functions. This results in smoother estimates and reduced predictive performance. In this situation, it is useful to investigate whether the target function can be represented on a lower dimensional projection of the predictor space. Dimension reduction techniques (Li 1991, 1992; Cook, 1998; Ferre 1998) can be used to estimate a lower dimension projection before using ALB.

Chapter 6

Concluding Remarks

A problem common to many disciplines is that of adequately approximating a function of several to many variables, given only the value of the function, often perturbed by noise, at various points in the dependent variable space. In statistics, flexible regression models for multi-dimensional data have been devised in response to this problem. Researchers started by fitting data for which the constant variance of the response given the predictors is a reasonable assumption. Some of the flexible regression models have been extended to the exponential family of distributions. This thesis extends a flexible regression model for multidimensional data, called Adaptive Logistic Basis (ALB) models to accommodate some of the exponential family distributions. Comparisons with competitive flexible regression models are presented on both simulation studies and real data.

We have attempted to show that various extensions of ALB to the exponential distribution family provides a useful addition to regression methodology. Its strengths, such as affine invariance, complement those of other techniques. Extensions of ALB models to the exponential family appear well

suited for large multidimensional data sets, where the target function contains higher-order interactions. Some limitations may be addressed by combining this method with other techniques.

6.1 Main contributions

1. Extensions of ALB models for the case where the conditional distribution of the response given the predictors is Poisson. The original ALB methodology employs a squared error and absolute error loss functions. Generalizations for count data are achieved by introducing a log-link function and an appropriate likelihood or quasi-likelihood function. While the idea is straightforward, several technical complications arise in the implementation. Stochastic approximation was used to solve the minimization problem. For constant variance models, the response variable was standardized prior to using stochastic approximation. However, standardization of Poisson counts does not make sense. Instead of transforming the response variable prior to using stochastic approximation, we stabilize the magnitude of perturbations by scaling the updating functions and by setting an upper bound on the predictions. We proved that these modifications to the updating functions for the Poisson version of ALB still result in convergence of the log-likelihood toward a local optimum (Theorem 2.2).

We derived approximate standard errors for the ALB estimator. Given the adaptive nature of the ALB estimator, difficulties in deriving standard errors are expected. We derived approximate standard errors for the fit assuming

that the number of basis functions is fixed and employing a standard asymptotic technique in nonlinear regression analysis. Our derivation extends that of H01 from Normal errors to the more general quasi-likelihood context. Coverage probabilities were estimated in several simulation studies. We discuss three approaches to estimate the covariance matrix of the parameter vector $\boldsymbol{\theta}$. The first approach extends H01 in the quasi-likelihood context. The second approach gives an alternative solution to the redundancies in the parameterization by fixing some of the parameters, and applying the usual asymptotic techniques to calculate $\text{cov}(\hat{\boldsymbol{\theta}})$. We proved that these two approaches yield essentially the same results (Theorem 2.3). The third approach to estimate $\text{cov}(\hat{\boldsymbol{\theta}})$ yields a sandwich estimator based on a first Newton-Raphson step approximation of $\hat{\boldsymbol{\theta}}$. We found that the approximate standard errors based on the third approach are lower than the ones based on the first approach, leading to more liberal confidence intervals. Based on these considerations, we recommended the approximate standard errors based on the first approach.

2. ALB models for over-dispersed count data. Unmeasured effects, clustering of events, or other contaminating influences combine to produce more variation in the responses than is predicted by the Poisson model. We discussed ALB models for the case where there is more variation in the responses than that predicted by the Poisson sampling theory. The model of the variance is motivated by a Poisson-Gamma mixture. Based on scientific reasons and model tractability, we focused on the following choices:

(i) Poisson over-dispersed data with constant coefficient of variation.

(ii) Breslow (1984) to model extra-Poisson variation in log-linear models.

(iii) A flexible variance model that employs an ALB function.

Algorithms based on a Weighted Least Square(WLS) or a Weighted Quasi-likelihood(WQL) criterion are employed to fit these models (Algorithms 3.1 and 3.2). We found that the choice of weights does not have a huge effect on the estimate of the mean, but can lead to important differences in the standard errors. The confidence intervals corresponding to the Poisson fit and the flexible ALB variance model are narrower, resulting in lower coverage probabilities. Our simulation studies indicated that modeling the variance function using ALB does not present an advantage over using a linear model to estimate the weights. Modeling residuals as an ALB function of the predictor from the Poisson fit yields relatively poor results. The residuals are very scattered, even after transformations. From our experience with simulation studies, when fitting ALB models to the variance function, ALB tends to select a larger number of basis functions than would be needed.

In our simulation studies and examples, the results from the two criteria WLS and WQL, were very similar. This finding is consistent with earlier results showing that WLS and ML estimators are asymptotically equivalent in the class of Generalized Linear Models (Carroll and Rupert, 1988, Section 2.4).

3. A comparison of ALB models, GPP and GAM. We compared predictive accuracy of ALB models, GPP and GAM using several simulation studies. We used several target functions, some of them previously used by Hwang et

al.(1993) to examine the supersmoother and polynomial based PPR, and by Roosen and Hastie (1993) to test ASP and Logistic PPR. Although Roosen and Hastie (1993) extended PPR to the exponential family, they never investigated the performance for a Poisson distributed response. Their code and simulations were restricted to a binary response. We modified the code to account for the Poisson distributed response, i.e., we modified the adjusted dependent variable and weights in the local scoring algorithm, and the corresponding starting values. The performance of these methods was tested on several examples. We found that ALB performs better than GPP on some of these examples, and worse than GPP on others. The performance of these methods is affected by the different basis functions employed by the two methods. When we have previous knowledge of the structure of the regression function, GAM(W) has a greater advantage over ALB and GPP, using models that exploit this previous knowledge. When complicated interactions are present, producing a surface with many bumps and ripples occurring in multiple directions, GPP performs better than ALB, and GAM(HT) performs much worse than both GPP and ALB. ALB works better than GPP when the covariate space can be covered with a small number of overlapping regions where f is well approximated by simple low-dimensional functions (Example 5.3.3 and Example 5.3.5).

6.2 Future directions

In the last section of Chapter 2, we presented an example (Rongelap Island data) that motivated work on generalizing ALB models to accommodate Pois-

son counts over time. We remind the reader that the independence assumption was unrealistic for the Rongelap Island example because of spatial correlations at close locations.

Spatial correlations may appear in the context of modeling data as a (partial) realization of a random process $\{Z(\mathrm{s}) : \mathrm{s} \in D\}$, where the index D allows $\mathbf{s}$ to vary continuously through a region of d-dimensional Euclidian space. The following two model assumptions are often made:

$$E(Z(\mathrm{s})) = \mu,$$

$$\mathrm{Cov}(Z(s_1), Z(s_2)) = C(s_1 - s_2),$$

where $C(.)$ is called a covariogram or a stationary covariance function. If $C(s_1 - s_2)$ is a function only of $\|s_1 - s_2\|$, then C(.) is called isotropic. In order to model the spatial correlations, one approach is to assume multivariate normality, $Z \sim N_n(\mu, \Sigma(\eta))$, and model the mean as a linear combination of the predictor variables:

$$\mu(\mathbf{s}) = \beta' \mathbf{x}(\mathrm{s}), \tag{6.1}$$

where $\mathbf{x}(\mathrm{s})$ are the predictor variables at location $\mathbf{s}$. Assuming a parametric model for the covariogram, one can obtain Maximum Likelihood estimates of the parameters. An example of an isotropic covariogram model is:

$$\eta = (\sigma, \nu_1, \nu_2),$$

$$\Sigma(\eta) = \mathrm{Cov}(Z(s_1), Z(s_2)) = \sigma^2 \exp\{-(\nu_1 \|s_1 - s_2\|)^{\nu_2}\},$$

for some $\nu_1 > 0$, $\sigma > 0$ and $0 < \nu_2 < 2$. We note that the above expression generates a strictly positive definite covariance matrix only if $0 < \nu_2 < 2$. One

might apply ALB models in this context, by replacing the linear predictor in (6.1) with the more flexible ALB predictor.

We mentioned in the previous chapter that, when we have previous knowledge of the additive or partially additive structure of the regression function, additive models have a greater advantage over ALB, using models that exploit additivity. In this situation, it is useful to investigate whether the target function can be represented on a lower dimensional projection of the predictor space. Dimension reduction techniques (Li 1991, 1992; Cook, 1998; Ferre 1998) can be used to estimate a lower dimension projection before using ALB.

Bibliography

[1] Agresti, A. and Coull, B.A. (1998) Approximate is better than "exact" for an interval estimation of binomial proportions. *The American Statistician*, **52**, 119–126.

[2] Akaike, H. (1973) Information Theory and an Extension of the Entropy Maximization Principle. *Proceedings of the 2nd International Symposium on Information Theory*, 267–281.

[3] Altman, N.S. (1992) An introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, **46**, 175–185.

[4] Bates, D.M. and Watts, D.G. (1988) *Nonlinear Regression Analysis and its Applications*. New-York: John Willey.

[5] Bellman,R.E. (1961) *Adaptive Control Processes*. Princeton Univ. Press.

[6] Benveniste, A., Metivier, M., and Priouret, P. (1990) *Adaptive Algorithms and Stochastic Approximations*. New York: Springer-Verlag.

[7] Breiman, L. (2001) Statistical Modeling: The Two Cultures. *Statistical Science*, **16**, 199–231.

[8] Breiman, L., Friedman, J.H., Olshen R.A. and Stone, C.J. (1984) *Classification and Regression Trees.* Wadsworth, Belmont, Calif.

[9] Breiman, L. and Meisel, W.S. (1976) General estimates of the intrinsic variability of data in nonlinear regression models. *Journal of the American Statistical Association*, **71**, 301–307.

[10] Breslow, N.E. (1984) Extra-Poisson Variation in Log-Linear Models. *Applied Statistics*, **33**, 38–44.

[11] Bridle, J.S. (1990) Probabilistic Interpretation of Feedforward Classification Network Outputs, with relationships to Statistical Pattern Recognition. *Neuro-computing: Algorithms, Architectures and Applications*, eds F. Fogleman Soulie and J. Herault, 227–236, Berlin: Springer.

[12] Brown, L.D., Cai. T.T. and DasGupta, A. (2001) Interval Estimation for a binomial proportion (with discussion). *Statistical Science*, **16**, 101–133.

[13] Bruntz, S.M., Cleveland, W.S., Kleiner, B. and Warner, J.L. (1974) The Dependence of Ambient Ozone on Solar Radiation, Temperature and Mixing height . *Symposium on Athmospheric Diffusion and Air Pollution*, Boston, American Meteorological Society, 125–128.

[14] Carroll, R.J (1982) Adapting for Heteroscedasticity in Linear Models. *Annals of Statistics*, **10**, 1224–1233.

[15] Carroll, R.J. and Ruppert, D. (1988) *Transformation and Weighting in Regression.* Chapman Hall, New York.

[16] Cheng, B. and Titterington, D.M. (1994) Neural Networks: A review from a statistical perspective. *Statistical Science*, **9**, 2–30.

[17] Cleveland, W.S., Devlin, S.J. and Grosse, E.H. (1988) Regression by local fitting: methods, properties and computational algorithms. *Journal of Econometrics*, **37**, 87–114.

[18] Cook, R.D. (1998) Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association*, **93**, 84–100.

[19] Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. Chapman Hall, New York.

[20] Cressie, N. (1993) *Statistics for Spatial Data*. Wiley, New York.

[21] Diaconis, P. and Shahshahani, M. (1984) On nonlinear functions of linear combinations. *J. Sci. Statist. Comput.*, **5**, 175–191.

[22] Diggle, P.J., Tawn, J.A. and Moyeed, M. (1998) Model-Based Geostatistics (with discussion). *Applied Statistics*, **47**, 299–350.

[23] Donoho, D.L. and Johnstone, I. (1989) Projection based approximation and a duality with kernel methods. *Annals of Statistics*, **17**, 58–106

[24] Efron, B., Hastie, T.J., Johnstone, I., and Tibshirani, R.J. (2004) Least Angle Regression. *Annals of Statistics*, **32**, 407–499

[25] Epanenchinikov, V.A. (1969) Nonparametric Estimation of a multivariate probability desity. *Theory of Probability and its Applications*, **14**, 153–158.

[26] Ferre, L. (1998) Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, **93**, 132–140.

[27] Flury, B.D. (1990) Principal Points. *Biometrika*, **77**, 33–41.

[28] Forsythe, A.B. (1972) Robust estimation of straight line regression coefficients by minimizing p-th power deviations. *Technomwtrics*, **14**, 159–166.

[29] Friedman, J.H. (1979) *A tree-structured approach to nonparametric multiple regression. In smoothing techniques for curve estimation.* (T.H. Gasser and M. Rosenblatt, eds.) 5–22. Springer, New York.

[30] Friedman, J.H. (1985) *Classification and multiple response regression through projection pursuit.* Technical report LCS012, Department of Statistics, Standford University.

[31] Friedman, J.H. (1991) Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–114.

[32] Friedman, J.H., Grosse, E. and Stuetzle, W. (1983) Multidimensional additive spline approximation. *SIAM J.Sci.Statist. Comput.*, **4**, 291–301.

[33] Friedman, J.H. and Stuetzle, W. (1981) Projection pursuit regression. *Journal of the American Statistical Association*, **76**, 817–823.

[34] Gu, C., Bates, D.M., Chen, Z. and Wahba, G. (1989) The computation of GCV functions through householder tridiagonalization with application to

the fitting of interaction spline models. *SIAM Journal of Matrix Analysis*, **10**, 457–480.

[35] Hansen, M., Kooperberg, C., Sardy, S. (1998) Triogram Models. *Journal of the American Statistical Association*, **93**, 101–119.

[36] Hartigan, J.A. and Wong, M.A. (1979) Algorithm AS136. A K-means Clustering Algorithm. *Applied Statistics*, **28**, 100–108.

[37] Hastie, T. and Tibshirani, R. (1986) Generalized Additive Models (with discussion). *Statistical Science*, **1**, 297–318.

[38] Hastie, T. and Tibshirani, R. (1987) Generalized Additive Models: Some Applications. *Journal of the American Statistical Association*, **82**, 371–386.

[39] Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Chapman Hall, New York.

[40] Holmes, C.C. and Mallick, B.K. (2003) Generalized Nonlinear Modeling with Multivariate Free-Knot Regression Splines. *Journal of the American Statistical Association*, **98**, 352–368.

[41] Hooper, P.M. (1999) Reference Point Logistic Classification. *Journal of Classification*. **16**. 91–116.

[42] Hooper, P.M. (2001) Flexible regression modelling with adaptive logistic basis functions. *The Canadian Journal of Statistics*, **29**, 1–34.

[43] Huber, P.J. (1985) Projection pursuit. *Annals of Statistics*, **13**, 435–475.

[44] Hwang, J-N, Lay, S-R, Maechler, M., Martin, D. and Schimert, J. (1993) Regression Modeling in Back Propagation and Projection Pursuit Learning. *IEEE Transactions on Neural Networks. In press.*

[45] Kohonen, T. (1995) *Self-Organizing Maps.* New-York: Springer.

[46] Kullback, T. (1959) *Information Theory and Statistics.* New-York: John Willey.

[47] Kushner, H.J. and Clark, D.S. (1978) *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer-Verlag, New York.

[48] Lehmann, E.L. (1998) *Elements of Large Sample Theory.* New-York: Springer.

[49] Leppik, I.E., Dreifuss, F.E., Porter, R., Bowman, T., Santilli, N., Jacobs, M., Crosby, C., Cloyd, J., Stackman, J. and Graves, N. (1987) A controlled study of progabide in partial seizures: methodology and results. *Neurology*, **37**, 963–968.

[50] Li, K.C. (1991) Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316–342.

[51] Li, K.C. (1992) On principal Hessian directions for data visualization and dimension reduction: another application of Stein's Lemma. *Journal of the American Statistical Association*, **87**, 1025–1039.

[52] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281–297.

[53] Margolin, B.H., Kaplan N. and Zeiger, E. (1981) Statistical Analysis of the Ames Salmonella/microsome test. *Proceedings of the National Academy of Sciences, USA*, **76**, 3779–3783.

[54] Moody, J.E. and Darken, C. (1989) Fast Learning in networks of locally-tuned processing units. *Neural Computation*, **1**, 281–294.

[55] Morgan, J.N. and Sonquist, J.A. (1963) Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, **58**, 415–434.

[56] Pierce, D.A. and Schafer, D.W. (1986) Residuals in generalized linear models. *Journal of the American Statistical Association*, **81**, 977–986.

[57] Pollard, E. and Yates, T.J. (1993) *Monitoring Butterflies for Ecology and Conservation*. London, Chapman Hall.

[58] Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

[59] Robbins, H. and Monro, S. (1951) A stochastic approximation method. *The Annals of Mathematical Statistics*, **22**, 400–407.

[60] Roosen, C.B. and Hastie, T.J. (1993) Logistic Response Projection Pursuit Regression. *Statistics and Data Analysis Research Department. ATT Bell Laboratories. Document No. BL011214-930806-09TM.*

[61] Roosen, C.B. and Hastie, T.J. (1994) Automatic Smoothing Spline Projection Pursuit. *Journal of Computational and Graphical Statistics*, **3**, 235–248.

[62] Silverman, B.W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society*, Series B. **47**, 1–52.

[63] Wahba, M. (1990) *Spline Models for Observational Data.* SIAM, Philadelphia, PA.

[64] Wood, S.N. (2003) Thin-plate regression splines. *Journal of the Royal Statistical Society*, Series B, **65**, 95–114.

[65] Xu, L., Kryzak, A. and Yuille, A. (1994) On radial basis function nets and kernel regression: statistical consistency, convergence rates, and receptive field sizes. *Neural Networks*, **7**, 609–628.

Appendix A

An attempt to compute standard errors for the ALB estimator using a splitting technique

I derived approximate standard errors for the fit in the quasi-likelihood context, assuming that the number of basis functions is fixed and using a linear approximation of the regression function around the parameter vector. In the following I will derive an attempt to compute standard errors for the ALB estimator using a splitting technique. A simpler starting point was to look at the case where the errors are normal, more precisely, $y_i = f(\mathbf{x}_i) + \epsilon_i$. The method may be summarized in a few steps. To keep things simple, assume the sample size n is an even number, $n = 2m$. We start by splitting the data into two subsets of m cases:

$$(\mathbf{x}_i^a, y_i^a), i = 1, \ldots, m$$

$$(\mathbf{x}_i^b, y_i^b), i = 1, \ldots, m$$

The split is chosen in matched pairs, that means $\mathbf{x}_i^a \approx \mathbf{x}_i^b$, for $i = 1, \ldots, m$.

An ALB estimator is obtained on the first half,

$$\hat{f}^a(\mathbf{x}) = \sum_{k=1}^{K_a} \hat{\delta}_k^a \phi_k^a(\mathbf{x})$$

Then we carry out least squares linear regression on the second half using the basis functions $\phi_k^a(\mathbf{x})$ evaluated at the data points in the second half. The new estimator can be written as:

$$\hat{f}^{ab}(\mathbf{x}) = \sum_{k=1}^{K_a} \hat{\delta}_k^{ab} \phi_k^a(\mathbf{x})$$

The estimated standard errors, $se\{\hat{f}^{ab}(\mathbf{x})\}$ follow from the linear regression formula. The last step is to repeat the same algorithm described above switching the two halves. Another estimator, $\hat{f}^{ba}(\mathbf{x})$ is obtained together with standard errors, $se\{\hat{f}^{ba}(\mathbf{x})\}$. The two estimators $\hat{f}^{ab}(\mathbf{x})$ and $\hat{f}^{ba}(\mathbf{x})$ can than be averaged to give:

$$\tilde{f}(\mathbf{x}) = \frac{\hat{f}^{ab}(\mathbf{x}) + \hat{f}^{ba}(\mathbf{x})}{2}$$

In order to obtain standard errors for this estimator, a natural question would be whether the two estimators $\hat{f}^{ab}(\mathbf{x})$ and $\hat{f}^{ba}(\mathbf{x})$ are uncorrelated. If they are uncorrelated, a formula for the estimated standard errors would be:

$$se\{\tilde{f}(\mathbf{x})\} = \frac{1}{2}\sqrt{se^2\{\hat{f}^{ab}(\mathbf{x})\} + se^2\{\hat{f}^{ba}(\mathbf{x})\}}$$

In order to develop some intuition on this splitting approach, we looked at a simulated one-dimensional example. A sample of size $m = 1000$ was generated, with each predictor x generated from a Uniform$(-3, 3)$ distribution. The target function is an ALB function and the reference point parameterization was used to specify f_K: $\xi_1 = 1$, $\xi_2 = 0$, $\xi_3 = -1$, $\gamma_1 = \gamma_2 = \gamma_3 = 0$,

$\delta_1 = 1$, $\delta_2 = 5$, $\delta_3 = 1$ and $\tau = 1$. To keep things simple, the same predictor values are used for both halves. The response vectors corresponding to the two halves are generated from the models: $y^a = f(x) + \epsilon^a$ and $y^b = f(x) + \epsilon^b$ where x and ϵ^a are independent, x and ϵ^b are independent and ϵ^a and ϵ^b are $N(0, \sigma_\epsilon^2)$ random variables, $\sigma_\epsilon = .2$. The following estimates are then obtained: $\hat{f^{ab}}, \hat{f}^{ba}, se\{\hat{f}^{ab}\}, se\{\hat{f}^{ba}\}$. A super-imposed plot of the target function f and the estimates, $\hat{f^{ab}}, \hat{f}^{ba}, \tilde{f}$ versus predictor variable x. displayed in Figure A.1, shows an almost perfect fit.

In order to get some intuition on the correlation between the two estimators $\hat{f^{ab}}$ and $\hat{f}^{ba}$, a sample of size $m = 400$ was generated, with each predictor x generated from a Uniform$(-3, 3)$ distribution. The target function stays the

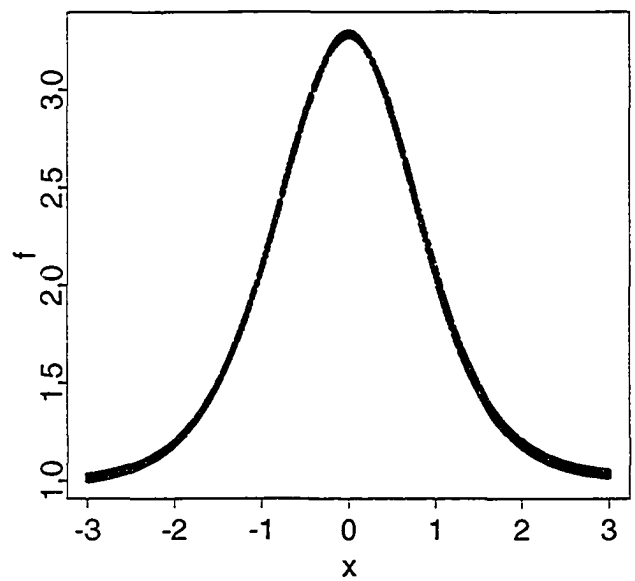

Figure A.1: Superimposed plot of the target function, f and the estimates $\hat{f^{ab}}$, $\hat{f}^{ba}$ and $\tilde{f}$

same as in the previous paragraph and to keep things simple, the same predictor values are used for both halves described as part of the splitting approach. Then, we generated 100 replicated samples of independent response vectors y^a and y^b, corresponding to the two halves, in the same manner described in the previous paragraph. 100 values for the estimates $\hat{f}^{ab}$ and $\hat{f}^{ba}$ were obtained at each fixed x. The correlations between $\hat{f}^{ab}$ and $\hat{f}^{ba}$ are then obtained at each fixed x. A plot of these correlations against x, displayed in Figure A.2, indicates small correlations between the two estimates. A natural estimator for the standard error of $\tilde{f}$ would be:

$$se\{\tilde{f}(x)\} = \frac{1}{2}\sqrt{se^2\{\hat{f}^{ab}(x)\} + se^2\{\hat{f}^{ba}(x)\}}$$

Coverage probabilities at a 95% nominal level are obtained for $\hat{f}^{ab}$, $\hat{f}^{ba}$

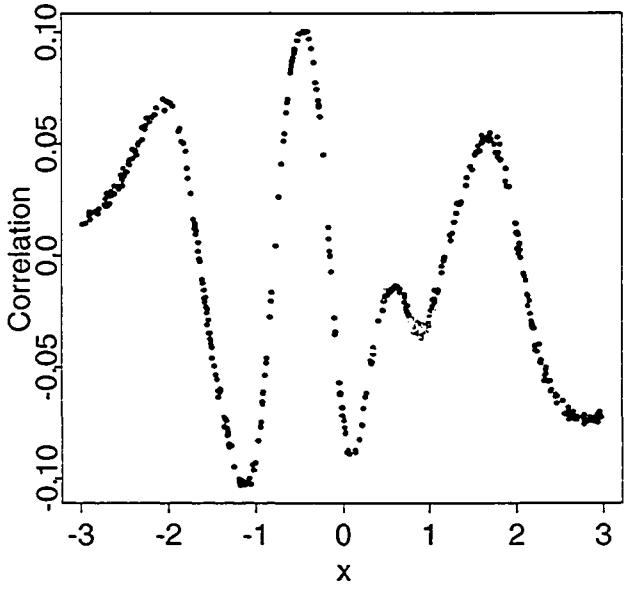

Figure A.2: Plot of the correlations between $\hat{f}^{ab}$ and $\hat{f}^{ba}$ versus the predictor variable x

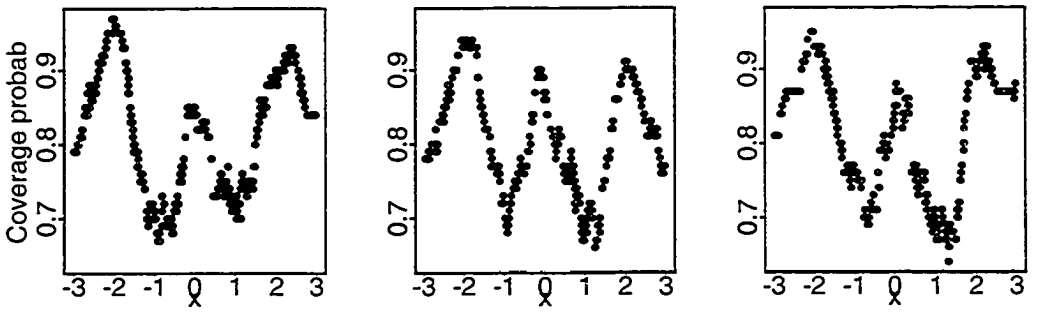

Figure A.3: (a) Plot of coverage probabilities for $\hat{f}^{ab}$ versus predictor $\mathbf{x}$ (b) Plot of coverage probabilities for $\hat{f}^{ba}$ versus predictor $\mathbf{x}$ (c) Plot of coverage probabilities for $\tilde{f}$ versus predictor $\mathbf{x}$

and $\tilde{f}$, at each fixed x. Plots of these coverage probabilities against x, displayed in Figure A.3, indicated problems with the splitting approach. In some regions, the coverage probabilities appear to be poor, with minimum values as low as .66. A closer look at normal probability plots of the z-scores corresponding to $\hat{f}^{ab}$ and $\hat{f}^{ba}$ at a fixed x, displayed in Figure A.4, indicated departures from the standard normal distribution. The bias in the linear regression model may be an explanation for the too liberal standard errors $se\{\hat{f}^{ab}\}$ and $se\{\hat{f}^{ba}\}$. Note that we carried out least squares linear regression on the second half using the estimated basis functions $\phi_k^a(\mathbf{x})$ obtained from the data points in the first half, assuming that they are correct. Taking into account the bias, the linear regression model becomes:

$$y^b = \Phi^a \delta + \eta + \epsilon,$$

where $\Phi^a = (\Phi^a(\mathbf{x}_1), ..., \Phi^a(\mathbf{x}_m))'$, $\Phi^a(\mathbf{x}) = (\phi_1^a(\mathbf{x}), ..., \phi_{k_a}^a(\mathbf{x}))'$, and k_a is the

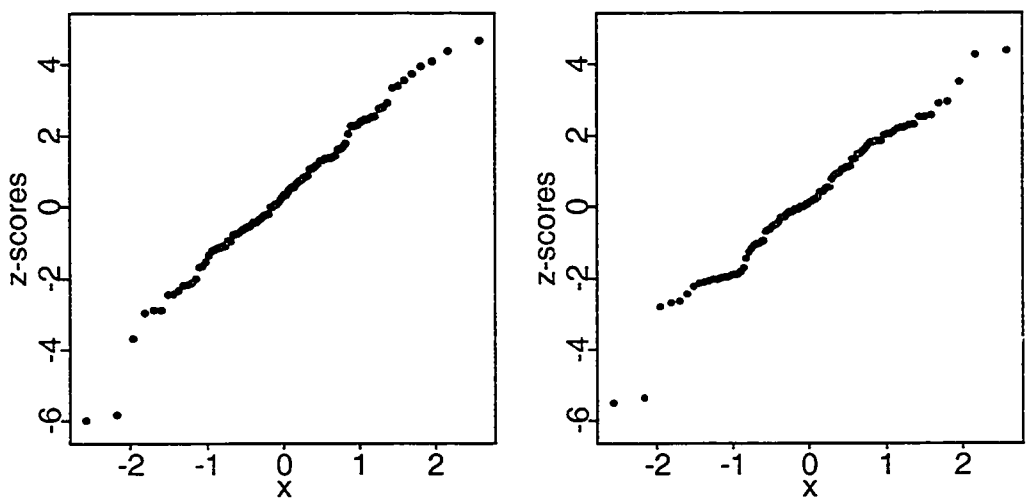

Figure A.4: (a) Normal probability plot of z-scores for $\hat{f}^{ab}$ for a case with the lowest cpf. (b) Normal probability plot of z-scores for $\hat{f}^{ba}$ for the same case as in (a)

number of basis functions selected from the first half. Note that,

$$\delta = (\Phi^{a\prime}\Phi^a)^{-1}\Phi^{a\prime}\mathrm{E}\{y^b|\mathbf{x}\}\,,$$

$$\eta = (I_m - \Phi^a(\Phi^{a\prime}\Phi^a)^{-1}\Phi^{a\prime})\mathrm{E}\{y^b|\mathbf{x}\}.$$

The linear regression formulae give us $\delta^{ab} = (\Phi^{a\prime}\Phi^a)^{-1}\Phi^{a\prime}y^b$ and $\hat{f}^{ab}(\mathbf{x}) = \Phi^a(\mathbf{x})'\delta^{ab}$, and therefore:

$$\mathrm{E}\{\delta^{ab}|\Phi^a\} = (\Phi^{a\prime}\Phi^a)^{-1}\Phi^{a\prime}(\Phi^a\delta + \eta)\,,$$

$$\mathrm{E}\{\hat{f}^{ab}(\mathbf{x})|\Phi^a\} = \Phi^a(\mathbf{x})'\delta = \Phi^a(\mathbf{x})'(\Phi^{a\prime}\Phi^a)^{-1}\Phi^{a\prime}\mathrm{E}\{y^b|\mathbf{x}\}.$$

We obtain the conditional bias and variance of $\hat{f}^{ab}$:

$$\mathrm{Bias}\{\hat{f}^{ab}(\mathbf{x})|\Phi^a\} = \mathrm{E}\{\hat{f}^{ab}(\mathbf{x})|\Phi^a\} - f(\mathbf{x})\,,$$

$$\mathrm{Var}\{\hat{f}^{ab}(\mathbf{x})|\Phi^a\} = \sigma^2\Phi^a(\mathbf{x})'(\Phi^{a\prime}\Phi^a)^{-1}\Phi^a(\mathbf{x}).$$

We estimate σ^2 by

$$s^2_{ab} = \frac{1}{m - k_a}\|y^b - \Phi^a \delta^{ab}\|^2$$

so,

$$\mathrm{E}\{s^2_{ab}|\Phi^a\} = \sigma^2 + \frac{1}{m - k_a}\|\eta\|^2.$$

The standard errors for $\hat{f}^{ab}$ follow:

$$se\{\hat{f}^{ab}\} = s_{ab}\{\Phi^a(\mathbf{x})'(\Phi^{a\prime}\Phi^a)^{-1}\Phi^a(\mathbf{x})\}^{1/2}$$

It seems that the conditional bias in s^2_{ab} did not account enough for the bias in $\hat{f}^{ab}$, and therefore the standard errors $se\{\hat{f}^{ab}\}$ are biased and the coverage probabilities are poor.

These findings suggest that we should not draw any conclusions based upon individual estimated basis functions, they are useful to estimate the fit, but can be misleading when used for further interpretation.