OVERDISPERSED GENERALIZED LINEAR MODELS

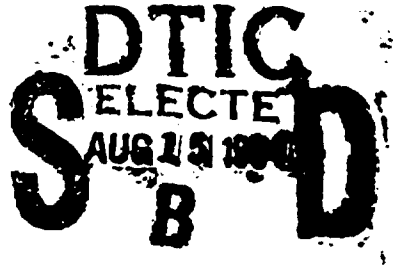
D.K. Dey
A.E. Gelfand
F. Peng

TECHNICAL REPORT No. 481
MAY 31, 1994

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Professor David Siegmund, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065



Overdispersed Generalized Linear Models

Dipak K. Dey, Alan E. Gelfand and Fengchun Peng

ABSTRACT

Generalized linear models have become a standard class of models for data analysts. However in some applications, heterogeneity in samples is too great to be explained by the simple variance function implicit in such models. Utilizing a two parameter exponential family which is overdispersed relative to a specified one parameter exponential family enables the creation of classes of overdispersed generalized linear models (OGLM's) which are analytically attractive. We propose fitting such models within a Bayesian framework employing noninformative priors in order to let the data drive the inference. Hence our analysis approximates likelihood-based inference but with possibly more reliable estimates of variability for small sample sizes. Bayesian calculations are carried out using a Metropolis-within-Gibbs sampling algorithm. An illustrative example using a data set involving damage incidents to cargo ships is presented. Details of the data analysis are provided including comparison with the standard generalized linear models analysis. Several diagnostic tools reveal the improved performance of the OGLM.

KEY WORDS: Exponential families; Exponential dispersion models; Jeffreys's prior; Mixture models; Metropolis-within-Gibbs algorithm; Orthogonal parameters.

1 INTRODUCTION

Generalized linear models (GLM) have by now become a standard class of models in the data analyst's tool box. The evolution of these models as well as details regarding inference, fitting, model checking, etc, is documented in the book by McCullagh and Nelder (1989). The GLIM software package is widely available and utilized.

Recently, such models have been criticized in certain applications as being too restrictive due to the fact that the variance is a specified function of the mean. In practice samples are often found to be too heterogeneous to be explained by such a simple functional relationship; variability tends to be larger than that captured through this function. A natural remedy is to consider a larger class of models.

Historically, the most frequently used approach for creating a larger class has been through mixture models. For instance the one parameter exponential family defining the GLM is mixed with a two parameter exponential family for the canonical parameter θ (equivalently the mean parameter μ) resulting in a two parameter marginal mixture family for the data. Shaked (1980) showed that such mixing necessarily inflates the model variance. Such overdispersion is of a certain type. Since the likelihood depends upon sample size while the mixture distribution does not, the relative overdispersion of the resulting mixture family to the original exponential family tends to infinity as sample size does. In other words, taking additional observations within a population does not increase our knowledge regarding heterogeneity across populations. (See Gelfand and Dalal, 1990 in this regard.) We also note that the resulting overdispersed family of mixture models will generally be awkward to work with since it will no longer be an exponential family (e.g. Beta-binomial, Poisson-gamma).

Efron (1986) presents an alternative approach through so-called double exponential families. Such families are derived as the saddle point approximation to the density of an average of n^* random variables from a one parameter exponential family for large n^* . The parameter n^*; written suggestively by Efron as $n\rho$, $0 < \rho < 1$ for actual sample size n, introduces ρ as a second parameter in the model along with the canonical parameter θ. Customary

inclusion of a dispersion parameter, say ϕ, to the usual one parameter exponential family results in an exponential dispersion model, EDM (Jorgensen, 1987). Because ϕ enters as a sample size or shape parameter, associated inference is usually handled differently from that for θ or μ. A one parameter exponential family in θ arises for each given ϕ but, as a two parameter model in (θ, ϕ), we no longer have an exponential family . Recently, Ganio and Schafer (1992) circumvent this problem in an approximate fashion by viewing the EDM as embedded within Efron's double exponential family . The asymptotics associated with Efron's family reveal overdispersion relative to the original exponential family which tends to a constant as $n \to \infty$, unlike the mixture case (Efron, 1986; Gelfand & Dalal, 1990).

Gelfand & Dalal (1990) argue that an appeal to asymptotics is not necessary to justify such models. More generally, for a given one parameter exponential family they introduce a two parameter exponential family of models which is overdispersed. This family includes Efron's model as a special case and also includes a family discussed in Lindsay (1986). Retaining the exponential family structure simplifies inference (as we shall detail in the subsequent sections). Relative overdispersion behaves as in Efron's models. Gelfand and Dalal suggested that, with the specification of link functions, these parameters could each be given the usual GLM structure but take the matter no further. Our objective here is to fully examine such models which we refer to as overdispersed generalized linear models(OGLM's). Again the selling points of our approach include the familiarity of exponential families, the ready interpretation of model parameters, the unification of modeling by absorbing earlier cases and exact inference rather than that from possibly inappropriate asymptotic approximations.

Another approach for handling heterogeneity in GLM's is through the use of random effects. See, e.g., Breslow and Clayton (1993) for a discussion and review of the literature. In the simplest version a standard GLM form is employed, but for each individual or population in the sample, a random effect is added to the fixed covariate effect term in the definition of the mean structure resulting in a generalized linear mixed model(GLMM). These random effects, sometimes called frailties, are introduced to "soak up variability" and hence clarify the explanation provided by the covariates. The likelihood retains all of the random effects.

Often these individual effects are of interest but even if they are not, marginalization over them is generally intractable. Such a marginalized density or likelihood is an example of the aforementioned mixing. With the random effects, the likelihood is nonregular in that, as sample size tends to infinity so does the number of model parameters. Customary approaches for model adequacy and model choice fail since the usual asymptotic distribution theory is no longer valid. OGLM's are regular and parsimonious compared to a GLMM but carry parameters in addition to those of the standard GLM which permit the possibility of capturing overdispersion within an exponential family framework.

We adopt a Bayesian perspective in fitting these models since we are drawn to the unifying use of inference summaries based upon the posterior implicit therein. However we assume that primary concern lies with the modeling incorporated in the likelihood and thus adopt an objective Bayesian stance employing noninformative prior specifications. For large sample sizes our inference will be close to that arising from maximum likelihood; for smaller samples our estimates of variability should be more appropriate than asymptotic ones associated with likelihood methods. Required Bayesian computation is handled through a Metropolis-within-Gibbs Markov chain Monte Carlo approach resulting in samples essentially from the joint posterior distribution which may be summarized to provide any desired inference. Such samples may also be used as the starting point for sampling from predictive distributions to investigate questions of model adequacy and model choice. Bayesian fitting of GLM's is discussed in Dellaportas and Smith (1993).

The format of this paper is then the following. In section 2 we formalize notation to develop required likelihoods. We also review properties of the overdispersed exponential family of models. In section 3 we develop Fisher's information matrix for these models which is important for both likelihood-based and Bayesian inference. For the Bayesian approach there is concern regarding the propriety of posteriors arising under Jeffreys's prior, which is obtained from Fisher's information matrix, as well as under a flat prior. In an appendix we address this question extending the results of Ibrahim and Laud (1991). Finally in section 4 we undertake an analysis of a data set involving wave damage to cargo vessels. We offer a reasonably complete posterior analysis of an OGLM fitted to the data as well as comparison

with the likelihood analysis of the standard GLM. We also use several diagnostic tools to reveal the improved performance of the OGLM.

2 OVERDISPERSED GENERALIZED LINEAR MODELS

Gelfand and Dalal (1990) discuss the class of two-parameter exponential family models of the form

$$f(y \mid \theta, \tau) = b(y)e^{\theta y + \tau T(y) - \rho(\theta,\tau)} \tag{1}$$

where if y is continuous, f is assumed to be a density with respect to Lebesgue measure while if y is discrete, f is assumed to be a density with respect to counting measure. Assuming that (1) is integrable over $y \epsilon\, \mathcal{Y}$, they show that, if T(y) is convex then, for a common mean, var(y) increases in τ.

It is presumed that the natural parameter space contains a two dimensional rectangle which, by translation, can be taken to contain $\tau = 0$. Then the associated one parameter exponential family arises at $\tau = 0$ and takes the form

$$f(y \mid \theta) = b(y)e^{\theta y - \chi(\theta)} \tag{2}$$

with $\chi(\theta) = \rho(\theta, 0)$. Thus, as τ increases from 0, var(y) increases relative to that under the associated one parameter exponential family, capturing the notion of overdispersion.

Expression (2) is the customary one parameter exponential family from which a GLM is developed. In particular $\mu \equiv E(y) = \chi'(\theta)$, $var(y) = \chi''(\theta) = V(\mu)$. Here $\chi'(\theta)$ is strictly increasing in θ so that μ and θ are one-to-one ($\theta = (\chi')^{-1}(\mu)$). $V(\mu)$ is called the variance function. A GLM is defined through a link function g, a strictly increasing differentiable transformation from μ to $\eta \epsilon R^1$, i.e., $g(\mu) = \eta = x^T\beta$, where x and β are, respectively, a known vector of explanatory variables and an unknown vector of model parameters. A more general version of (2) is

$$f(y \mid \theta, \phi) = b(y, \phi)e^{(\theta y - \chi(\theta))/a(\phi)}. \tag{3}$$

Now $var(y) = (a(\phi))^{-1}V(\mu)$. If y is viewed as an average of say n random variables then a usual form for $a(\phi)$ is $(n\phi)^{-1}$. Setting $\phi^* = n\phi$, ϕ^* is referred to as the dispersion parameter and (3) is called an exponential dispersion model(EDM) (Jorgensen, 1987). Ganio and Schafer (1992) extend the GLM based on (2) to an EDM, where $\phi = h(z^T\alpha)$ with z a known vector and α an unknown parameter vector. They assume $z^T\alpha$ includes an intercept. The two parameter family in (3) differs from that in (1) in the sense that (1) is a customary two parameter exponential family whereas (3) is a customary one parameter family for each fixed ϕ.

Efron (1986) defined the double exponential family of models through the density

$$\bar{f}(y \mid \theta, \rho,\ n) = c(\theta, \rho, n)\rho^{\frac{1}{2}}e^{n\rho(\theta y - \chi(\theta)) + n(1-\rho)(\theta(y)y - \chi(\theta(y)))} \tag{4}$$

where $\theta(y) = (\chi')^{-1}(y)$. Here y is viewed as an average of n i.i.d. random variables, θ is the canonical parameter and ρ is a dispersion parameter. In regression problems he assumes a GLM in θ (canonical link), i.e., $\theta = x^T\beta$ and that $\rho = h(z^T\alpha)$ for a suitable h. Using various expansions, Efron shows that (4) permits attractive approximation as n grows large. Most notably, $\bar{f}$ behaves like (3) with $a(\phi) = (n\rho)^{-1}$. Hence Ganio and Schafer (1992) treat their extended GLM, based upon (3), as an example of Efron's double exponential family and carry out their model fitting following his examples.

We note that, regardless of n, (4) is of the form (1) with $T(y) = \theta(y)y - \chi(\theta(y))$, $\tau = n(1-\rho)$ and $\theta = n\rho\theta$. Straightforward calculation shows this $T(y)$ is convex so that, in fact, (4) is a special case of (1). As Gelfand and Dalal (1990) show, other choices of $T(y)$ may be more appropriate and in any event it seems preferable to work with the exact form (1) rather than with approximation to (4).

Returning to (1), under usual regularity conditions, we have the following properties. If $\rho^{(r,s)} \equiv \frac{\partial \rho^{r+s}}{\partial \theta^r \partial \tau^s}$ then $\rho^{(1,0)} = E(y \mid \theta, \tau) \equiv \mu$, $\rho^{(2,0)} = var(y \mid \theta, \tau)$, $\rho^{(0,1)} = E(T(y) \mid \theta, \tau)$, etc. It is sometimes convenient to consider (1) through a mean parametrization

$$f(y \mid \mu, \tau) = b(y)e^{(y-\mu)\psi^{(1,0)}(\mu,\tau) + \tau T(y) + \psi(\mu,\tau)}. \tag{5}$$

In (5), we employ the same sort of notation for ψ as for ρ. By comparison with (1) we

have $\theta = \psi^{(1,0)}(\mu,\tau)$ and $\rho(\theta,\tau) = -\psi(\mu,\tau) + \mu\psi^{(1,0)}(\mu,\tau)$. Using (5) and straightforward calculation, we can show that $E\left(\frac{\partial^2 \log f}{\partial\tau\partial\mu}\right) = 0$, i.e, that μ and τ are orthogonal parameters in the sense of Barndorff-Nielsen (1978, p184) and Cox and Reid (1987).

The only practical drawback to working with (1) is that $\rho(\theta,\tau)$ is not available explicitly. While $\chi(\theta)$ in (2) is usually an explicit function of θ, $\rho(\theta,\tau) = \log \int b(y)e^{\theta y+\tau T(y)}dy$ usually requires a univariate numerical integration or summation. In the examples we have investigated thus far this has not presented a problem.

To create an overdispersed generalized linear model (OGLM) from (1) suppose we have independent responses y_i with associated covariates $x_i, p \times 1$ and $z_i, q \times 1$, i=1,2,..,n. The components of x and z need not be exclusive. Let $y = (y_1, y_2, \cdots, y_n)$ and define $\theta_i = g(x_i^T\beta)$ and $\tau_i = h(z_i^T\alpha)$ where g and h are strictly increasing. The resulting likelihood is,

$$L(\beta,\alpha;y) = \prod_{i=1}^{n} e^{\theta_i y_i+\tau_i T(y_i)-\rho(\theta_i,\tau_i)}. \tag{6}$$

The monotonicity of g and h is natural and insures that θ_i is monotonic in x_{il} and that τ_i is monotonic in z_{il} facilitating interpretation. Of course such monotonicity does not imply that, e.g., θ_i is monotone in each covariate. The form $x_i^T\beta$ allows a covariate to enter as a polynomial. If an explanatory variable appears in x_i but not in z_i, say x_{il}, then $\frac{\partial\mu_i}{\partial x_{il}} = \rho^{(2,0)}(\theta_i,\tau_i)g'(x_i^T\beta)\beta_l$. But since $\rho^{(2,0)}$ and g' are strictly positive, μ_i is strictly monotone in x_{il} with the sign of β_l determining the direction. If the variable appears in z_i as well, its influence on μ_i is less clear since now $\frac{\partial\mu_i}{\partial x_{il}}$ involves $\rho^{(1,1)}$, the covariance between y and $T(y)$. In any event, using a Bayesian framework with a sampling-based implementation enables straightforward computation of $E(\mu_i|y)$, the posterior mean of μ_i at any x_i and z_i allowing us to study its change as covariate levels change. In the example of section 4 we have taken g and h to be the identity functions, in the spirit of canonical links. If we wanted to *force* overdispersion ($\tau_i > 0$) we could, for example, take $\tau_i = \exp(z_i^T\alpha)$.

3 INFORMATION CALCULATIONS AND PRIOR SPECIFICATIONS FOR AN OGLM

The Fisher information matrix associated with (6) is of interest for both likelihood-based inference as well as Bayesian inference. Evaluated at the MLE, it provides an estimate of the large sample covariance structure for the MLE's of β and α. In the Bayesian setting, again evaluated at the MLE and again when sample sizes are large, it provides an approximation to the posterior covariance matrix for β and α. Also, the square root of the determinant of this matrix, as a function of β and α, is known as Jeffreys's prior and is commonly used as a "noninformative" specification. Ibrahim and Laud (1991) obtain this matrix in the case of a generalized linear model developed from (3), assuming ϕ is known, and discuss its use as a prior. We extended their calculation to (6).

Straightforwardly we may show that

$$
\begin{aligned}
E\left(\frac{\partial^2 log L(\beta,\alpha,y)}{\partial\beta_i\partial\beta_j}\right) &= -\sum_i \rho^{(2,0)}(\theta_i,\tau_i)x_{ij}x_{ik}(g'(x_i^T\beta))^2 \\
E\left(\frac{\partial^2 log L(\beta,\alpha,y)}{\partial\alpha_i\partial\alpha_j}\right) &= -\sum_i \rho^{(0,2)}(\theta_i,\tau_i)z_{ij}z_{ik}(h'(z_i^T\alpha))^2 \\
E\left(\frac{\partial^2 log L(\beta,\alpha,y)}{\partial\beta_j\partial\alpha_k}\right) &= -\sum_i \rho^{(1,1)}(\theta_i,\tau_i)x_{ij}z_{ik}(g'(x_i^T\beta))(h'(z_i^T\alpha)).
\end{aligned}
$$

Let X denote the nxp design matrix arising from the $x_i's$, Z the nxq design matrix arising from the $z_i's$, M_θ an nxn diagonal matrix with $(M_\theta)_{ii} = \rho^{(2,0)}(\theta_i,\tau_i)(g'(x_i^T\beta))^2$, M_τ an nxn diagonal matrix with $(M_\tau)_{ii} = \rho^{(0,2)}(\theta_i,\tau_i)(h'(z_i^T\alpha))^2$ and $M_{\theta,\tau}$ an nxn diagonal matrix with $(M_{\theta,\tau})_{ii} = \rho^{(1,1)}(\theta_i,\tau_i)(g'(x_i^T\beta))(h'(z_i^T\alpha))$. Then

$$
I(\beta,\alpha) = \begin{pmatrix} X^T M_\theta X & X^T M_{\theta,\tau} Z \\ Z^T M_{\theta,\tau} X & Z^T M_\tau Z \end{pmatrix} \tag{7}
$$

and Jeffreys's prior is $|I(\beta,\alpha)|^{\frac{1}{2}}$.

To work with (7) requires calculation of ρ, $\rho^{(2,0)}$, $\rho^{(0,2)}$ and $\rho^{(1,1)}$. This in turn, requires

calculation of six integrals of the form $\int y^c T^d(y) b(y) e^{\theta y + \tau T(y)} dy$ for the set of (c,d) from {(0,0) (1,0) (2,0) (0,1) (0,2) (1,1)}. Numerical integration or summation for such univariate integrals is generally routine.

Suppose instead we define an extension of the GLM using the mean parametrization (5), setting $\mu = g(x^T\beta)$ with again $\tau = h(z^T\alpha)$. In the case of (2), i.e. $\tau = 0$, this is, in fact, the more usual way of formulating a GLM. Paralleling (7) let us now define M_μ as an nxn diagonal matrix with $(M_\mu)_{ii} = \psi^{(2,0)}(\mu_i, \tau_i)(g'(x_i{}^T\beta))^2$ and M_τ an nxn diagonal matrix with $(M_\tau)_{ii} = \psi^{(0,2)}(\mu_i, \tau_i)(h'(z_i{}^T\alpha))^2$. The orthogonality of μ and τ results in

$$I(\beta, \alpha) = \begin{pmatrix} X^T M_\mu X & 0 \\ 0 & Z^T M_\tau Z \end{pmatrix}. \tag{8}$$

Hence Jeffreys's prior is $|I(\beta,\alpha)|^{\frac{1}{2}} = |X^T M_\mu X|^{\frac{1}{2}} |Z^T M_\tau Z|^{\frac{1}{2}}$.

It is worth observing that obtaining the MLE for β and α under (6) is challenging. Customary iteratively reweighted least squares algorithms such as Fisher scoring will require the calculation of ρ and several of its partial derivatives at each (θ_i, τ_i). A grid search algorithm would only require calculation of ρ but will be very inefficient in higher dimensions. Hence, fitting the Bayesian model, using a sampling-based approach, is no harder than performing a likelihood analysis. But then, an important question to ask is whether either a flat prior for (β, α) or Jeffreys's prior using (7), both of which are improper, in combination with the likelihood in (6), results in a proper posterior for (β, α). We provide an answer in the appendix, extending work of Ibrahim and Laud (1991) who address this question in the case of a GLM developed from (3).

4 AN OVERDISPERSED POISSON MODEL

McCullagh and Nelder (1989, p204) discuss a data set where the response variable is the number of damage incidents by waves to cargo ships. For each of 34 ships the aggregate months in service were recorded as well as the number of damage incidents over that period.

Explanatory factors are ship type having 5 levels (A, B, C, D, E), year of construction having 4 levels (CP1, CP2, CP3, CP4) and period of operation having two levels (SP1, SP2). Since the response is a count, McCullagh and Nelder propose a Poisson regression presuming that the expected number of damage incidents is directly proportional to the aggregate months in service, i.e., the total period of risk. Using a canonical link the GLM sets

$$\begin{aligned}\theta &= \log(\text{aggregate months service}) + \beta_0 + \text{effect due to ship type} \\ &+ \text{effect due to year of construction} + \text{effect due to service period} \qquad (9)\end{aligned}$$

The log(aggregate months service) term is called an offset, its coefficient fixed at 1 as a result of the proportionality assumption.

The parameter estimates and associated confidence intervals for (9) under the standard GLM appears in the first column of Table 1. For each factor, the first level is taken as a baseline and its effect is set to zero. McCullagh and Nelder incorporate a dispersion parameter ϕ, obtaining an estimate $\hat{\phi} = 1.69$, indicating overdispersion relative to the standard Poisson density which, intrinsically, has $\phi = 1$. Of course, in fitting such an EDM, the estimates of the effects, of the θ_i and of the μ_i are unaffected by the presence of ϕ. Our model in (1) incorporates overdispersion in a much different way; under (1), $\mu_i = \rho^{(0,0)}(\theta_i, \tau_i)$, a function of the dispersion parameter. McCullagh and Nelder note that some points are not well fit using their EDM.

We consider three Bayesian models, all fit using a flat prior for the effects. Model 1 fits the GLM in (9) as a reduced case of (1) with $\tau = 0$ (and, of course, $\phi = 1$) resulting in a 9 parameter model. With regard to inference about the mean structure, model 1 is essentially equivalent to the model of McCullagh and Nelder. Model 2 incorporates a constant dispersion parameter $\tau = \alpha_0$ with the convex function, $T(y) = (y+1)\log(y+1)$. Finally, anticipating that overdispersion might increase with exposure, model 3 sets $\tau = \alpha_0 + \alpha_1 \log(\text{aggregate months service})$, using the same $T(y)$, an 11 parameter model. The Bayesian fits for models 1, 2 and 3 are produced in Table 1 as well. Comparing the likelihood analysis with the Bayesian analysis for model 1 there is an indication that the likelihood based confidence intervals, arising under asymptotic theory, may be too long and

too symmetric.

The deviations between the observed y_r and the posterior mean $E(\mu_r \mid Y)$ are given in Table 2. The abundance of small counts suggest, that the exact analysis under our model will be more appropriate that the asymptotic models of Efron (1986) and of Ganio and Schafer (1992). The posterior densities under model 3 for α_0 and α_1 are presented in Figures 1a and 1b respectively. These figures, along with the last column of Table 1 provide evidence of overdispersion ($\tau > 0$) and, in addition, evidence that $\alpha_1 > 0$, supporting the hypothesis that overdispersion increases with exposure to risk. Also from Table 2, model 3 seems to better fit the most of larger y_r, which are associated with greater exposure.

Fitting models 2 and 3 requires calculation of $\rho(\theta, \tau)$ which is the sum of the form

$$\rho(\theta, \tau) = \log \sum_{y=0}^{\infty} (y!)^{-1} e^{\theta y + \tau T(y)}. \tag{10}$$

Calculation of Jeffreys' prior requires of computation of sums similar to (10) as described after (7). Under a flat prior, a proper posterior results following the argument of the appendix since, with canonical links, log concavity of (6) holds. We suspect a proper posterior arises using Jeffreys' prior but the analytical approaches suggested in the appendix are not easily applied.

Lastly, with regard to models 1, 2 and 3 we ask two questions. Are these models adequate? Amongst those that are, which one would we choose? The Bayesian approach to answering these questions is based upon predictive distributions. Under an improper flat prior the marginal density of the data (the prior predictive density) is improper hence impossible to calibrate. As an alternative we adopt a cross-validation approach, paralleling widely used classical regression strategy. In particular, we consider the proper densities $f(y_r \mid y_{(r)})$, r = 1, 2, ..., 34, where $y_{(r)}$ denotes y with y_r removed. We, in fact, condition on the actual observations $y_{(r),obs}$ creating the predictive distribution for y_r under the model and all the data except y_r. For model determination we would then compare, in some fashion, $f(y_r \mid y_{(r),obs})$ with the r_{th} observation, $y_{r,obs}$. Such cross validation is discussed in Gelfand, Dey and Chang (1992) and in further references provided there.

A natural approach for model adequacy is to draw, for each r, a sample from $f(y_r \mid y_{(r),obs})$, and compare this sample with $y_{r,obs}$. In particular using this sample we might obtain the .025 and .975 quantiles of the $f(y_r \mid y_{(r),obs})$ say $\underline{y}_r$ and $\overline{y}_r$ and see how many of the 34 $y_{r,obs} \in [\underline{y}_r, \overline{y}_r]$. Under each model at least 28 of the 34 intervals contained the corresponding $y_{r,obs}$. Suppose instead we obtain the lower and upper quartiles of $f(y_r \mid y_{(r),obs})$ and see how many $y_{r,obs}$ fall in their interquartile ranges. We find 11 for model 1, 18 for model 2 and 16 for model 3. Under the true model we would expect half, i.e., 17. Hence both model 2 and model 3 perform close to expectation though all three models seem adequate.

A well established tool for model choice is the conditional predictive ordinate (CPO), $f(y_{r,obs} \mid y_{(r),obs})$. A large value implies agreement between the observation and the model. For comparing models, the ratio $d_r = \frac{f(y_{r,obs}|y_{(r),obs},M_i)}{f(y_{r,obs}|y_{(r),obs},M_{i'})}$ (or perhaps $\log d_r$) indicates support by point r for one model versus the other (see Pettit and Young, 1990). Figure 2 provides a plot of logCPO ratios for model 3 vs model 1 and for model 3 vs model 2. Model 3 emerges as best. A simple diagnostic with a frequentist flavor is $\sum_{r=1}^{34} (y_{r,obs} - E(\mu_r|y))^2/34$. For model 1 this statistic is 25.37, for model 2 it is 12.09 and for model 3 it is 4.60, again supporting model 3.

We recognize that the above model determination diagnostics are informal. However they are in the spirit of widely used classical EDA approaches and do permit examination of model performance at the level of the individual observation.

5 BRIEF REMARKS ON THE SAMPLING BASED ANALYSIS OF THE MODELS

We conclude with a brief discussion regarding the Bayesian fitting of the models, computation of the $f(y_r \mid y_{(r)})$ and the sampling from them in the present context. For each model we used a Metropolis-within-Gibbs Markov chain Monte Carlo algorithm (Müller, 1994) to develop samples from the posterior, beginning with multiple starts in the vicinity of the maximum likelihood estimate. Evaluation of the likelihood required repeated calculation of the function

$\rho(\theta_j, \tau_j)$. These samples provide the Bayesian inference associated with Tables 1 and 2 and Figure 1.

As for model determination, let (β_j^*, α_j^*), j = 1, ..., m denote a sample of size m from a particular posterior. Since

$$\begin{aligned} f(y_r \mid y_{(r)}) &= \frac{\int f(y \mid \beta, \alpha) f(\beta, \alpha) d\beta d\alpha}{\int f(y_{(r)} \mid \beta, \alpha) f(\beta, \alpha) d\beta d\alpha} \\ &= \frac{\int \frac{f(y \mid \beta, \alpha) f(\beta, \alpha)}{f(\beta, \alpha \mid y)} f(\beta, \alpha \mid y) d\beta d\alpha}{\int \frac{f(y_{(r)} \mid \beta, \alpha) f(\beta, \alpha)}{f(\beta, \alpha \mid y)} f(\beta, \alpha \mid y) d\beta d\alpha} \\ &= \frac{1}{\int f(y_r \mid \beta, \alpha)^{-1} f(\beta, \alpha \mid y) d\beta d\alpha}, \end{aligned}$$

a Monte Carlo estimate of $(f(y_r \mid y_{(r),obs})$ based upon (β_j^*, α_j^*) is

$$\hat{f}(y_r \mid y_{(r),obs}) = (m^{-1} \sum_{j=1}^{m} (f(y_r \mid \beta_j^*, \alpha_j^*))^{-1})^{-1}. \tag{11}$$

(11) is used for CPO calculations. See Gelfand and Dey (1994) for more general discussion.

Samples from $f(y_r \mid y_{(r)})$ are drawn in two stages. Given samples (β_j^*, α_j^*), j = 1, 2, ..., m, from $f(\beta, \alpha \mid y)$ we convert these to samples from $f(\beta, \alpha \mid y_{(r)})$ by resampling with weights $q_j = \frac{(f(y_{r,obs} \mid \beta_j^*, \alpha_j^*)^{-1}}{\sum_{j=1}^{m} (f(y_{r,obs} \mid \beta_j^*, \alpha_j^*))^{-1}}$. See Smith and Gelfand (1992) in this regard. Since $f(y_r \mid y_{(r)}) = \int f(y_r \mid \beta, \alpha) f(\beta, \alpha \mid y_{(r)}) d\beta d\alpha$, if β', α' is a draw from $f(\beta, \alpha \mid y_r)$ then if $y_r' \sim f(y_r \mid \beta', \alpha')$, the marginal distribution of y_r' is $f(y_r \mid y_{(r)})$.

APPENDIX: INTEGRABILITY OF POSTERIORS FOR OGLM'S UNDER OBJECTIVE PRIOR SPECIFICATIONS

Ibrahim and Laud (1991) investigate the propriety of a Bayesian GLM, i.e., the special case when $\tau = 0$ in (6), under Jeffreys' prior which becomes $|X^T M_\theta X|^{\frac{1}{2}}$. They show that if X is full column rank and the likelihood is bounded above then a sufficient condition for the posterior of β to be proper is that for each y_i

$$\int_\Theta e^{\theta y_i - \chi(\theta)} (\chi''(\theta))^{\frac{1}{2}} d\theta < \infty \tag{A1}$$

posterior of β to be proper is that for each y_i

$$\int_{\Theta} e^{\theta y_i - \chi(\theta)} (\chi''(\theta))^{\frac{1}{2}} d\theta < \infty$$

where Θ denotes the natural parameter space for the canonical parameter θ.

Since $I(\beta, \alpha)$ is positive definite, $|I(\beta,\alpha)|^{\frac{1}{2}} \leq |X^T M_\theta X|^{\frac{1}{2}} |Z^T M_\tau Z|^{\frac{1}{2}}$. Hence, if, in (6) $L(\beta, 0; Y)$ is bounded above, (A1) implies that $f(\beta \mid \alpha, Y)$ is proper. Similarly if $L(0, \alpha; Y)$ is bounded above, if $\gamma(\tau) \equiv \rho(0, \tau)$ for $\tau \epsilon \Gamma$ and if for each y_i

$$\int_{\Gamma} e^{\tau T(y_i) - \gamma(\tau)} (\gamma''(\tau))^{\frac{1}{2}} d\tau < \infty \qquad \text{(A2)}$$

then $f(\alpha, \mid \beta, y)$ is proper. Of course the fact that these conditional distributions are proper does not imply that the joint distribution is. Combining (6) and (7), we need to establish when

$$\int \int L(\beta, \alpha; Y) |I(\beta, \alpha)|^{\frac{1}{2}} d\beta d\alpha < \infty. \qquad \text{(A3)}$$

Suppose we assume that X and Z are of full column rank and that, $L(\beta, \alpha; Y)$ is bounded above. The fact that

$$\begin{aligned} |I(\beta,\alpha)|^{\frac{1}{2}} &\leq (\prod_{l=1}^{p} (X^T M_\theta X)_{ll})^{\frac{1}{2}} (\prod_{m=1}^{q} (Z^T M_\tau Z)_{mm})^{\frac{1}{2}} \qquad \text{(A4)} \\ &\leq \prod_{l=1}^{p} (\sum_{j=1}^{n} |x_{ij}| (M_\theta)_{jj}^{\frac{1}{2}}) \prod_{m=1}^{q} (\sum_{j=1}^{n} |z_{mj}| (M_\tau)_{jj}^{\frac{1}{2}}) \end{aligned}$$

enables us to imitate the argument of Ibrahim and Laud to conclude that, if for each y_i,

$$\int_{\Gamma} \int_{\Theta} e^{\theta y_i + \tau T(y_i) - \rho(\theta,\tau)} (\rho^{(2,0)}(\theta,\tau) \rho^{(0,2)}(\theta,\tau))^{\frac{1}{2}} d\theta d\tau < \infty \qquad \text{(A5)}$$

then (A3) holds. We omit details. Barndorff-Nielsen (1978) provides conditions for a bounded likelihood and that these will usually hold for densities of the form (1). The condition (A5) is easier to work with than (A3) since it is a two, rather than $p + q$, dimensional integral. However since $\rho^{(2,0)}$ and $\rho^{(0,2)}$ are not available explicitly they must be approximated within (A5) making analytic investigation difficult.

Integrability of the posterior may also be investigated using a very different approach. Suppose $L(\beta, \alpha; Y)$ is log concave. Then, if the prior $f(\beta, \alpha)$ is bounded, the posterior is

the "lower" one has all positive slope coefficients, the "upper" one all negative. The lower one bounds $\log L$ on a set say S_L, the upper one on the complement, S_U. But then upon exponentiating, in each case we obtain a product of exponential curves. The exponentiated lower curve is immediately integrable over S_L, the upper curve over S_U whence L is integrable and hence the posterior, if $f(\beta, \alpha)$ is bounded.

Log concavity of $L(\beta, 0; Y)$ is discussed in Wedderburn (1976) and in Dellaportas and Smith (1993). The former is concerned with the behavior of MLE's, the latter with simplifying Monte Carlo sampling of β. What follows has implications for OGLM's in either context. In particular, for the OGLM defined in (6), log concavity can be established by verifying a simple nonnegative definiteness condition. Letting $\theta = g(\eta)$, $\tau = h(\gamma)$ consider the function $a(\eta, \gamma) = \theta(\eta)y + \tau(\gamma)T(y) - \rho(\theta(\eta), \tau(\gamma))$. If $-\frac{\partial^2 a}{\partial \eta^2} \geq 0$, $-\frac{\partial^2 a}{\partial \gamma^2} \geq 0$ and $\begin{vmatrix} -\frac{\partial^2 a}{\partial \eta^2}, & - \\ -\frac{\partial^2 a}{\partial \gamma \partial \eta}, & -\frac{\partial^2 a}{\partial \gamma^2} \end{vmatrix} \geq 0$, (6) is log concave. For the canonical case $\theta = \eta$, $\tau = \gamma$, $-\frac{\partial^2 a}{\partial \eta^2} = var(y), -\frac{\partial^2 a}{\partial \gamma^2} = var(T(y))$ and the determinant becomes $var(y)var(T(y)) - cov^2(y, T(y))$ which is nonnegative whence log concavity holds.

While the flat prior is obviously bounded, Jeffreys's prior need not be (consider the case of the univariate normal standard deviation). Using the inequality in (A4), $|I(\beta, \alpha)|^{\frac{1}{2}}$ will be bounded if the $(M_\theta)_{jj}$ and $(M_\tau)_{jj}$ are. In the canonical case this reduces to boundedness of $var(y)$ and $var(T(y))$.

References

[1] Barndorff-Nielsen, O.E. (1978), *Information and Exponential Families in Statistical Theory*, John Wiley & Sons, New York.

[2] Breslow, N. and Clayton, D. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9-25.

[3] Cox, D.R. and Reid, N. (1987), "Parameter Orthogonality and Approximate Conditional Inference (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, 49, 1-39.

[4] Dellaportas, P. and Smith, A.F.M. (1993), "Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling," *Applied Statistics.*, 42, 443-460.

[5] Efron, B. (1986), "Double Exponential Families and Their Use in Generalized Linear Regression," *Journal of the American Statistical Association*, 81, 709-21.

[6] Ganio, L.M. and Schafer, D.W. (1992). "Diagnostics for Overdispersion," *Journal of the American Statistical Association*, 87, 795-804.

[7] Gelfand, A.E. and Dalal, S.R. (1990), "A Note on Overdispersed Exponential Families," *Biometrika*, 77, 55-64.

[8] Gelfand, A.E. and Dey, D.K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society*, Ser. B, 56, 501-514.

[9] Gelfand, A.E., Dey, D.K. and Chang H. (1992), "Model Determination Using Predictive Distributions with Implementation Via Sampling-Based Methods," In *Bayesian Statistics* 4, (J. Bernardo et al. eds.), Oxford University Press, Oxford, 147-167.

[10] Ibrahim, J.G., and Laud, P.W. (1991), "On Bayesian Analysis of General Linear Models Using Jeffreys's Prior", *Journal of the American Statistical Association*, 86, 981-986.

[11] Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press, London.

[12] Jorgensen, B. (1987), "Exponential Dispersion Models (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, 49, 127-162.

[13] Lindsay, B. (1986), "Exponential Family Mixture Models (with least squares estimators)," *The Annals of Statistics*, 14,124-37.

[14] McCullagh, P., and Nelder, J.A. (1989), *Generalized Linear Models*, Chapman & Hall, London.

[15] Müller P. (1994), "Metropolis Posterior Integration Schemes" *Journal of the American Statistical Association*, To appear.

[16] Pettit, L.I. and Young, K.D.S. (1990), "Measuring the Effect of Observations on Bayes Factors," *Biometrika*, 77, 455-466.

[17] Shaked, M. (1980), "On Mixtures from Exponential Families," *Journal of the Royal Statistical Society*, Ser. B, 42, 192-198.

[18] Smith, A.F.M. and Gelfand, A.E. (1992), "Bayesian Statistics Without Tears: a Sampling-Resampling Perspective, " *The American Statistician*, 46, 2, 84-88.

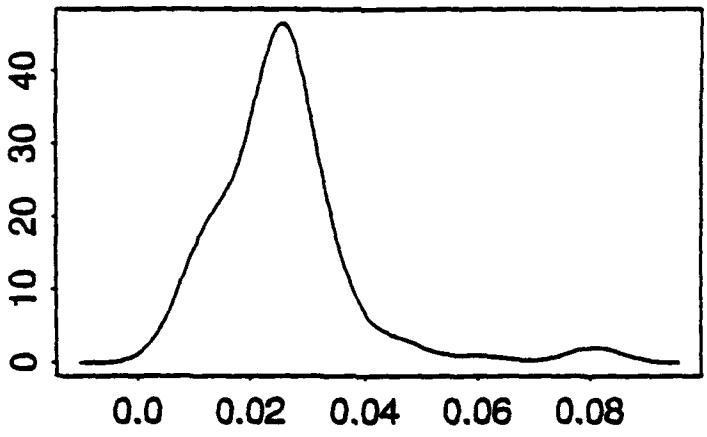
[19] Wedderburn, R. (1976), "On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models," *Biometrika*, 63, 27-32.

Table 1: Inference Summaries for Models 1, 2 and 3.

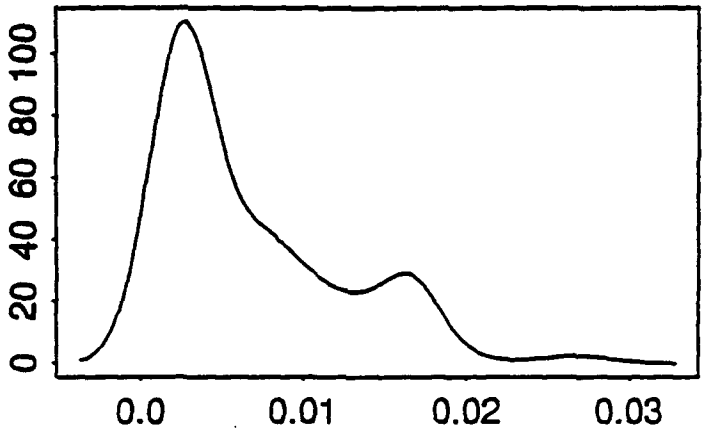
Parameters	EDM MLE (.95 CI)	Model 1 Posterior Mean (.95 Credible Interval)	Model 2 Posterior Mean (.95 Credible Interval)	Model 3 Posterior Mean (.95 Credible Interval)
INTERCEPT	-6.410 (-6.836, -5.984)	-6.444 (-6.863, -6.055)	-6.337 (-6.414, -6.226)	-6.351 (-6.467, -6.171)
B	-0.540 (-0.991, -0.089)	-0.517 (-0.813, -0.179)	-0.571 (-0.625, -0.525)	-0.567 (-0.738, -0.339)
C	-0.690 (-1.533, 0.153)	-0.608 (-1.156, 0.094)	-0.611 (-1.049, -0.275)	-1.132 (-2.163, -0.508)
D	-0.080 (-0.813, 0.653)	-0.089 (-0.616, 0.368)	0.290 (0.276, 0.303)	0.444 (0.194, 1.131)
E	0.330 (-0.277, 0.937)	0.317 (-0.030, 0.731)	0.248 (0.170, 0.309)	0.521 (0.108, 0.734)
CP2	0.700 (0.328, 1.072)	0.719 (0.380, 1.081)	0.534 (0.420, 0.631)	0.356 (0.182, 0.480)
CP3	0.820 (0.391, 1.249)	0.822 (0.531, 1.103)	0.640 (0.635, 0.645)	0.568 (0.177, 0.737)
CP4	0.450 (-0.138, 1.038)	0.476 (0.055, 0.889)	0.695 (0.532, 0.810)	-0.764 (-1.575, -0.498)
SP2	0.380 (0.086, 0.674)	0.383 (0.151, 0.600)	0.248 (0.172, 0.343)	0.515 (0.219, 0.794)
α_0			0.034 (0.031, 0.040)	0.038 (0.024, 0.072)
α_1				0.009 (0.001, 0.014)

Table 2: Comparison of Fits for Models 1, 2 and 3. $(\text{Dev})_r \equiv y_r - E(\mu_r|\mathbf{y})$

		Model 1		Model 2		Model 3	
r	y_r	$E(\mu_r\|\mathbf{y})$	$(\text{Dev})_r$	$E(\mu_r\|\mathbf{y})$	$(\text{Dev})_r$	$E(\mu_r\|\mathbf{y})$	$(\text{Dev})_r$
1	0.0	0.24	-0.24	0.21	-0.21	0.23	-0.23
2	0.0	0.15	-0.15	0.15	-0.15	0.20	-0.20
3	3.0	3.61	-0.61	3.64	-0.64	2.97	0.03
4	4.0	4.69	-0.69	5.33	-1.33	5.03	-1.03
5	6.0	5.61	0.39	5.58	0.42	5.20	0.80
6	18.0	16.51	1.49	18.07	- 0.07	20.04	-2.04
7	11.0	3.47	7.53	8.55	2.45	11.53	-0.53
8	39.0	53.52	-14.52	53.26	-14.26	42.95	-3.95
9	29.0	25.51	3.49	24.17	4.83	34.14	-5.14
10	58.0	48.33	9.67	56.04	1.96	58.04	-0.04
11	53.0	52.97	0.03	58.40	-5.40	57.94	-4.94
12	12.0	15.21	-3.21	15.34	-3.34	14.14	-2.14
13	44.0	37.36	6.64	41.71	2.29	45.63	-1.63
14	18.0	6.32	11.68	21.05	-3.05	16.16	1.84
15	1.0	1.24	-0.24	1.07	-0.07	0.74	0.26
16	1.0	0.73	0.27	0.73	0.27	0.59	0.41
17	0.0	1.39	-1.39	1.45	-1.45	0.70	-0.70
18	1.0	1.53	-0.53	1.84	-0.84	1.03	-0.03
19	6.0	1.56	4.44	1.61	4.39	0.89	5.11
20	2.0	5.08	-3.08	5.84	-3.84	3.83	-1.83
21	1.0	0.74	0.26	0.58	0.42	0.14	0.86
22	0.0	0.63	-0.63	0.38	-0.38	0.77	-0.77
23	0.0	0.34	-0.34	0.23	-0.23	0.55	-0.55
24	0.0	1.24	-1.24	0.90	-0.90	1.27	-1.27
25	0.0	1.06	-1.06	0.87	-0.87	1.42	-1.42
26	2.0	1.68	0.32	1.20	0.80	1.88	0.12
27	11.0	7.77	3.23	6.06	4.94	11.39	-0.39
28	4.0	14.18	-10.18	7.34	-3.34	5.04	-1.04
29	0.0	0.11	-0.11	0.10	-0.10	0.14	-0.14
30	7.0	3.34	3.66	3.57	3.43	3.79	3.21
31	7.0	2.37	4.63	2.91	4.09	3.40	3.60
32	5.0	5.50	-0.50	5.80	-0.80	7.10	-2.10
33	12.0	22.26	-10.26	15.94	-3.94	13.58	-1.58
34	1.0	3.44	-2.44	2.84	-1.84	1.42	-0.42



(1a)



(1b)

Figure 1: Posterior Densities for α_0 and α_1 under Model 3.

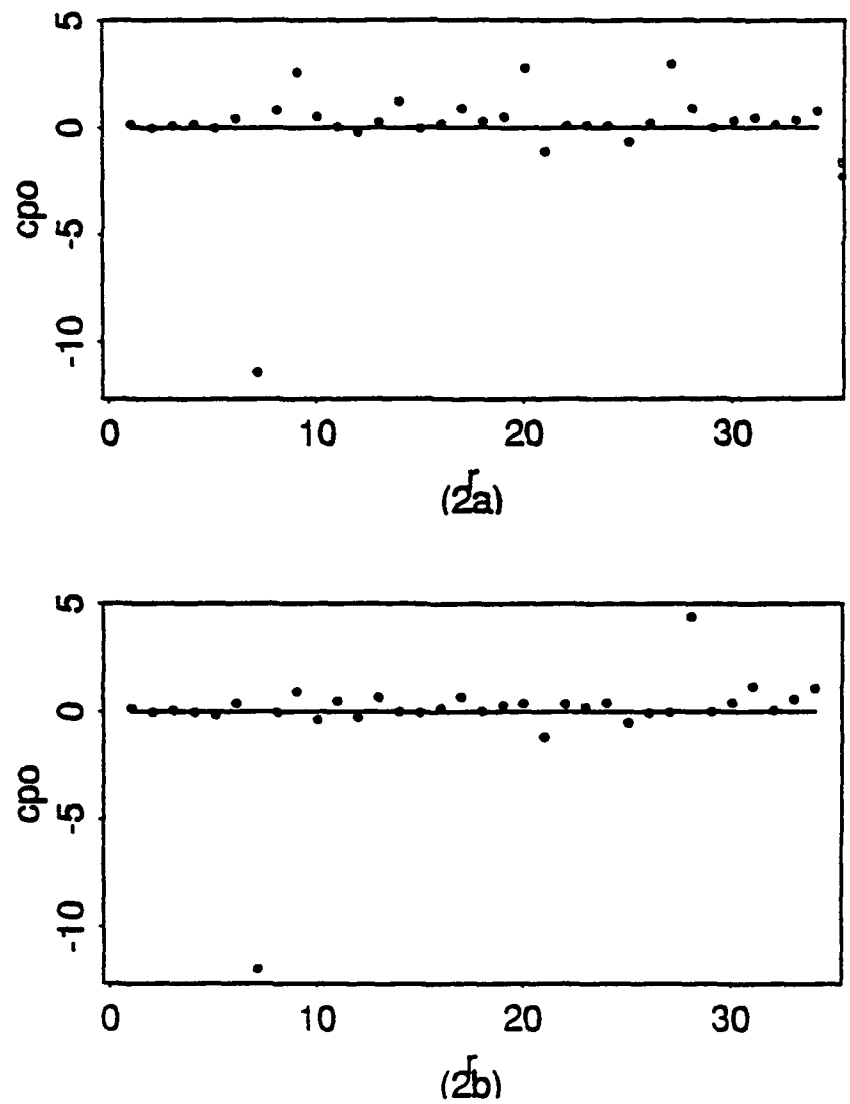


Figure 2: Log CPO Ratios for Model 1, 2 and 3 as Described in Section 4.
(2a): Model 3 vs Model 1; (2b); Model 3 vs Model 2.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Overdispersed Generalized Linear Models		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER 481
7. AUTHOR(s) Dipak K. Dey, Alan E. Gelfand and Fengchun Peng		8. CONTRACT OR GRANT NUMBER(s) N00014-92-J-1264
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305-4065		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE 31 May 1994 13. NUMBER OF PAGES 21
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Sponsored in part by NSF grant DMS 9301316 and National Cancer Institute Grant CA09667.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Exponential families; Exponential dispersion models; Jeffreys's prior; Mixture models; Metropolis-within-Gibbs algorithm; Orthogonal parameters.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See reverse side

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT

Generalized linear models have become a standard class of models for data analysts. However in some applications, heterogeneity in samples is too great to be explained by the simple variance function implicit in such models. Utilizing a two parameter exponential family which is overdispersed relative to a specified one parameter exponential family enables the creation of classes of overdispersed generalized linear models (OGLM's) which are analytically attractive. We propose fitting such models within a Bayesian framework employing noninformative priors in order to let the data drive the inference. Hence our analysis approximates likelihood-based inference but with possibly more reliable estimates of variability for small sample sizes. Bayesian calculations are carried out using a Metropolis-within-Gibbs sampling algorithm. An illustrative example using a data set involving damage incidents to cargo ships is presented. Details of the data analysis are provided including comparison with the standard generalized linear models analysis. Several diagnostic tools reveal the improved performance of the OGLM.