GENERALIZED LINEAR MODELS WITH UNKNOWN LINK FUNCTIONS

B.K. Mallick
A.E. Gelfand

TECHNICAL REPORT No. 482
JULY 18, 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

Generalized linear models with unknown link functions

Bani K. Mallick and Alan E. Gelfand

Abstract

Generalized linear models are widely used by data analysts. However, the choice of the link function, i.e., the scale on which the mean is linear in the explanatory variables is often made arbitrarily . Here we permit the data to estimate the link function by incorporating it as an unknown in the model. Since the link function is usually taken to be strictly increasing, by a strictly increasing transformation of its range to the unit interval we can model it as a strictly increasing cumulative distribution function. The transformation results in a domain which is [0,1] as well. We model the cumulative distribution function as a mixture of Beta cumulative distribution functions, noting that the latter family is dense within the collection of all continuous densities on [0,1]. For the fitting of the model we take a Bayesian approach, encouraging vague priors, to focus upon the likelihood. We discuss choices of such priors as well as the integrability of the resultant posteriors. Implementation of the Bayesian approach is carried out using sampling based methods, in particular, a tailored Metropolis-within-Gibbs algorithm. An illustrative example utilising data involving wave damage to cargo ships is provided.

Key Words: Bayesian model determination; Jeffreys's prior; Metropolis-within-Gibbs algorithm; Mixture-of-Betas distributions;

1 Introduction

Generalized linear models have by now become a standard class of models for exploration within the data analyst's tool kit. The evolution of these models along with details on fitting them is provided in McCullagh and Nelder (1989). The GLIM software for carrying out the model fitting is widely available.

Generalized linear models have been advocated as an advance over standard linear models in that they allow for (i) nonnormal sampling mechanisms, (ii) heterogeneous variances which are captured through the mean-variance relationship of the sampling model , and (iii) a mean for the observations which need only be linear on a transformed scale. This transformation, referred to as the link function, is the focus of the present paper. That is, often the stochastic mechanism for the observations arises naturally as, for example, a binomial or Poisson in the case of count data. However choice of the scale upon which the transformed mean is presumed linear is often made arbitrarily. Informal classical diagonostic tools for selecting a link and for assessing the adequacy of a link are discussed in the McCullagh and Nelder (1989) drawing upon work of Pregibon (1980) and Hinkley (1985). In particular, employing a family of power link functions, an approach in the spirit of the Box-Tidwell transformation can suggest an appropriate power. However, this family insists upon a positive mean and in addition may be too small within the class of strictly monotone functions.

We treat the link function as another unknown in the generalized linear model specification and estimate it jointly with the mean structure. Our approach is thus semiparametric; we assume a linear parametric form for the mean on a transformed scale where the transformation is expressed nonparametrically. Moreover, for a given linear form our fitted link function may be compared with say the canonical link to identify shortcomings of the latter.

Our fitting is within the Bayesian framework treating all model unknowns as random with inference proceeding from the posterior distribution of these unknowns. If prior information is available say about coefficient parameters we would be happy to use it. However, interest usually focusses more upon the likelihood whence we would tend to use noninformative priors. Recent advances in Bayesian computation using sampling based methods (Gelfand and Smith, 1990; Smith & Gelfand, 1993) enable reasonably straight forward fitting of such models. In fact Gibbs sampling was utilised by Dellaportas and Smith (1993) for implementing the Bayesian analysis of standard generalized linear models.

The Dirichlet process prior (Ferguson, 1973) has been the usual modeling tool for nonparametric Bayesian inference. In an unpublished thesis, Escobar employs the Dirichlet process prior in investigating a nonparametric version of the simultaneous normal means problem. This idea has been extended in a series of, as yet unpublished, papers authored by Escobar, West and Erkanli amongst others to include generalized linear models. In all of this work the nonparametric aspect of the modelling is introduced at the second stage; the first stage specification is a fully parametric likelihood. No nonparametric estimate of the link function results say for comparison with the link assumed in the first stage.

Unpublished work of Czado and Newton introduces a random link into the likelihood in the binary regression problem. They assume $\Pr(y_i = 1|x_i, \beta)$ =G($x_i^T\beta$) where G is an unknown cumulative distribution function. G is assumed to be a random draw from a Dirichlet process prior independent of β. The base measure for G might be normal or logistic. Czado and Newton show that, with the inclusion of latent variables $u_i \sim$ G such that $\Pr(y_i = 1|x_i, \beta)$= $\Pr(u_i \leq x_i^T\beta)$, G can be marginalised out and a straightforward Gibbs sampler arises. Here the Dirichlet process prior is convenient since the inverse link is a distribution function. Extension to other generalized linear models is not obvious.

For an arbitrary generalized linear model our approach describes the strictly increasing link function g, suitably transformed to have range in (0,1), again as an unknown cumulative distribution function. In the process the resulting domain also becomes (0,1). We model this function as a mixture of Beta distribution functions appealing to the well known result that any continuous density on (0,1) can be arbitrarily well approximated by a discrete mixture of Beta densities. Unlike distributions arising under a Dirichlet process, which could also be used here, we have a continuous, dense class of distributions admitting an explicit form. In practice we have treated the number of mixands r, as fixed though a discrete prior could be attempted. We have experimented with a range of r's, for a number of examples, discovering, perhaps not surprisingly, that robustness occurs with quite small r. In introducing randomness to this finite mixture model it is simpler assume the mixture weights to be random rather than the parameters of the Beta densities.

Interesting related nonparametric Bayesian regression work, which also does not employ Dirichlet process priors, includes Blight and Ott (1975), O'Hagan (1978), Weerahandi and Zidek (1988) and Angers and Delampady (1992).

The outline of this paper is thus the following. In section 2 we detail our general approach. Section 3 considers non informative priors appropriate for our likelihood specification. In section 4 we describe the fitting of these models using a sampling based approach. Section 5 examines a data set taken from McCullagh and Nelder (1989) where a Poisson regression is fit assuming a canonical link. Allowing an unknown link, not surprisingly, we obtain an improved model. But also, comparison of the estimated link with the canonical link reveals the nature of the shortcomings of the latter. We conclude with a brief summary.

2 Likelihood form

Recall that a generalized linear model assumes a one-parameter exponential family form for the distribution of the response y, i.e.,

$$f(y|\beta,\phi,g,x) = d(y,\phi)exp[\{\theta y - b(\theta)\}/a(\phi)] \tag{1}$$

is a density with respect to Lebesgue measure if y is continuous, with respect to counting measure if y is discrete. Here $\mu \equiv E(y) = b'(\theta)$ and var(y)=$b''(\theta)/a(\phi)$. Furthermore, $g(\mu) = \eta = x^T\beta$ where x is a p×1 known vector of covariates, β is an unknown p×1 vector of coefficients and g is a strictly increasing differentiable function. The link function g, often taken to be the so called canonical link, $g(\mu)$=$(b')^{-1}(\mu)$, is assumed unknown as well. In some generalized linear models such as the binomial and Poisson, $a(\phi)$ is a known constant . In general we assume that $a(\phi)$=ϕ/ν, i.e., that ϕ enters as an unknown scale parameter with

ν a known "sample size". Then given a sample of responses y_i, $i = 1, 2 \cdots n$ with associated covariates vectors x_i and sample sizes ν_i the resulting likelihood becomes

$$L(\beta, \phi, g) = \prod_{i=1}^{n} d(y_i, \phi) exp[\phi^{-1} \nu_i \{\theta_i y_i - b(\theta_i)\}] \tag{2}$$

where θ_i=$(b')^{-1}(\mu_i)$ and $\mu_i = g^{-1}(x_i{}^T \beta)$. The likelihood in (2) is infinite dimensional; without further assumptions it need not be identifiable. For example, if x is a single continuous covariate, i.e., $\mu = g^{-1}(\beta_0 + \beta_1 x)$, then g, β_0 and β_1 cannot be identified.

Our inference approach is Bayesian requiring the specification of a prior $f(\beta, \phi, g)$. The identifiability question from a Bayesian point of view, becomes whether the data can inform about all of the unknown parameters in the model. If yes, provided a proper posterior results, there is no identifiability problem. If not and if f is improper then the posterior necessarily is as well and we have an ill-defined Bayesian model. If not and if f is proper then the prior drives the posterior.

The mapping g is from the space of μ, say Ω, into R^1. As will become obvious shortly, it is convenient to work with g^{-1}, a mapping from R^1 into Ω. Suppose T is a strictly increasing differentiable transformation from Ω into (0,1) with $J(\eta)$=T$(g^{-1}(\eta))$. Then J is a strictly increasing differentiable distribution function. So modeling the function g is equivalent to modeling an unknown distribution function. A rich class of models may be created as follows. Let g_0 be a baseline link function for g, perhaps the canonical link, and let $J_0(\eta)$=T$(g_0^{-1}(\eta))$ be the cumulative distribution function associated with g_0. Diaconis and Ylvisaker (1985) argue that discrete mixtures of Beta densities provide a continuous dense class of models for densities on (0,1). A general member has the form

$$h(u) = \sum_{l=1}^{r} w_l Be(u|c_l, d_l) \tag{3}$$

where r denotes the number of mixands, $w_l \geq 0$, $\sum w_l = 1$ and $Be(u|c_l, d_l)$ denotes the Beta density in standard form with parameters, c_l and d_l. If $IB(u; c_l, d_l)$ denotes the incomplete Beta function associated with $Be(u|c_l, d_l)$ then let

$$J(\eta) = \sum_{l=1}^{r} w_l IB(J_0(\eta); c_l, d_l). \tag{4}$$

Clearly $J(\eta)$ is a distribution function and $\mu=g^{-1}(\eta)= T^{-1}(J(\eta))$ is readily calculated. Thus $\theta = (b')^{-1}(\mu)$ is and so given a set of y_i, x_i ν_i and a β and ϕ, we can evaluate the likelihood (2) directly. Note that calculation of $g(\mu)$ for a given μ, requires a clumsy inversion for a corresponding quantile of $J(\eta)$ clarifying the advantage to modeling g^{-1}. Mixtures other than Beta could be used, e.g., gammas on R^+, uniforms on R^1.

We could assume r unknown but do not since in practice this gains little. In our experience, inference , e.g., estimation of μ_i, prediction of y_i, is very robust to choice of r; mixtures with r=3 or 4 are virtually indistinguishable from those with much larger r. In fact, allowing r$\geq$n does not insure perfect fit since g is restricted to be monotone. Given r, it is mathematically easier to assume that the component Beta densities are specified but that the weights are unknown. We choose the set of c_l, d_l to provide a collection of Beta densities which blanket (0,1). In particular we work with $c_l=\lambda l$, $d_l = \lambda$(r+1-l). Hence specification of g is equivalent to specification of w and we can denote (2) by $L(\beta, w, \phi)$.

The choice of T is not a modeling issue. In principle any strictly increasing differentiable function from Ω to R^1 could be used. For $\Omega = R^1$ we might use T$(\cdot)$ =e$^{(\cdot)}$/(1+e$^{(\cdot)}$), for $\Omega = R^+$ we might use T$(\cdot)$ =$(\cdot)/\{1+(\cdot)\}$, for Ω=[0,1] we might use T$(\cdot)$=$(\cdot)$. In practice computational difficulties can be ameliorated by centering and scaling these choices. For example, if $g_0(\mu) = log\mu$, $g_0^{-1}(\eta) = e^{\eta}$ and over/under flow problems will arise if $|\eta_i|$ can be large. However in (4), $J_0(\eta)$ is required, not $g_0^{-1}(\eta)$. How can we choose T so that, given η, computation of the composite function, T$(g_0^{-1}(\eta))$ avoids these problems? If we try T$(\cdot)$ =$(\cdot)/\{1 + (\cdot)\}$, for large $|\eta_i|$, within the accuracy of the computer, T will be 0 or 1. Consider instead $T_{k_1,k_2}(\cdot) = k_1(\cdot)^{k_2}/\{1 + k_1(\cdot)^{k_2}\}$ with $k_1 = e^{a/b}$, $k_2 = 1/b$. Then $J_0(\eta) = e^{(\eta-a)/b}/\{1 + e^{(\eta-a)/b}\}$. Hence if a "centers" the η_i and b "scales" them, J_0 can be computed without problem. Treating min y_i and max y_i as a range for μ_i enables a range for η_i from which simple choices for a and b hence k_1 and k_2, can be made. There is no notion of a best choice and, to keep notation simpler, we suppress $k_1$and k_2 in the sequel.

3 Prior specification and proper posteriors

Since primary interest is in the likelihood and since only occasionally will there be useful prior information we consider vague specification of the prior $f(\beta, w, \phi)$. In the case of generalized linear models, where g is specified, a multivariate normal prior for β is customary yielding a flat prior as the precision matrix tends to 0. For such models, assuming ϕ is known, Ibrahim and Laud (1991) consider Jeffreys's prior, the square root of the determinant of Fisher's information matrix. Let X be the n× p matrix whose rows are the x^T's and let M be an n×n diagonal matrix such that M_{ii}=$[\{g'(\mu_i)\}^2 V(\mu_i)]^{-1}$ where $V(\mu) \equiv b''(\theta)$. Then Jeffreys's prior is proportional to $|X^T MX|^{1/2}$. In our case w determines g, so given ϕ, Jeffreys's prior is a specification of $f(\beta|w, \phi)$, i.e.,

$$f(\beta|w, \phi)) \propto |X^T MX|^{1/2}. \tag{5}$$

In evaluating (5) we require $g'(\mu_i) = d\eta_i/d\mu_i$. But $d\mu_i/d\eta_i = J'(\eta_i)(T^{-1})'(J(\eta_i)$. From (4)

$$J'(\eta_i) = J_0'(\eta_i) \sum_{l=1}^{r} w_l Be(J_0(\eta_i); c_l, d_l). \tag{6}$$

with $J_0'(\eta_i) = (g_0^{-1})'(\eta_i)T'(g_0^{-1}(\eta_i))$. Thus $d\eta_i/d\mu_i = \{J'(x_i^T\beta)(T^{-1})'(J(x_i^T\beta))\}^{-1}$

The prior specification is completed by providing $f(w, \phi)$. If ϕ is intrinsically given we only require f(w). This is the case with our Poisson regression example in section 5. If not, we take f(w,ϕ)= f(w)$f(\phi)$ where, since ϕ is a scale parameter, customary choice for $f(\phi)$ would be an inverse gamma density or a limit thereof.

$f(w)$ is a distribution on the r-dimensional simplex. If g_0 is a baseline link for g then we might choose $f(w)$ such that, a priori, g is "centered" around g_0. The data would then revise this prior in terms of support for g_0. Centering g around g_0 corresponds to centering J around J_0, i.e., from (4), centering $\sum_{l=1}^{r} w_l IB(u; c_l, d_l)$ around u. If we center using the mean, as is typically done in the case of Dirichlet processes, we obtain

$$\sum_{l=1}^{r} E(w_l) IB(u; c_l, d_l) = u \tag{7}$$

Suppose $\mathbf{w} \sim Dir(\gamma\ 1)$. Then (7) requires $r^{-1} \sum IB(u; c_l, d_l) = u$. If we use c_l and d_l as in section 2 and take r even, expansion of the terms in this summation about u_0=1/2, yields, to a first order approximation, an average which is u. We omit details.

When does a proper posterior result under the Bayesian model $L(\beta, w, \phi) f(\beta, w, \phi)$? The prior for w, Dir(γ1) is proper if γ >0. Assuming $f(\phi)$ is proper, if ϕ is unknown, we may investigate propriety of the posterior by considering the integrability of

$$\int L(\beta, w, \phi) f(\beta | w, \phi) d\beta \tag{8}$$

If $f(\beta|w, \phi)$ is proper, then (8) exists. If $f(\beta|w, \phi)$ is flat, integrability follows if log$L$ is concave in β for each w and ϕ. If $f(\beta|w, \phi)$ is Jeffreys's prior as in (5) we may utilise Theorem 2.1 of Ibrahim and Laud (1991). In particular, if as a function of β for fixed w and ϕ, L is bounded above and if for each observed y_i

$$\int exp[\phi^{-1} \nu_i \{zy_i - b(z)\}] \{b''(z)\}^{1/2} dz < \infty \tag{9}$$

then (8) is integrable. Note that (9) does depend upon the link function g enabling (8) to be integrable for all w. Boundedness of L in β holds under very general conditions as in Barndorff-Nielsen, (1977). These conditions hold for the usual models arising under (1).

4 Implementing the Bayesian model.

The posterior distribution of (β, w, ϕ) is proportional to L$(\beta, w, \phi) \cdot f(\beta|w, \phi) \cdot f(w) \cdot f(\phi)$. Analytic investigation of this p+(r-1)+1 dimensional nonnormalised joint distribution is infeasible so we adopt a sampling based approach using a Markov chain Monte Carlo algorithm to obtain observations essentially from this posterior distribution. In particular, we use a version of the Gibbs sampler (Gelfand and Smith, 1990) updating β, then w, then ϕ to complete one iteration. The associated nonstandardised complete conditional distributions are not available explicitly. From (2) and (4), for a given β, to evaluate $L(\beta, w, \phi)$ requires nr incomplete Beta function evaluations so making draws directly from these distributions is inconvenient. Instead, we utilise a Metropolis-within-Gibbs algorithm (Müller, 1993) with a multivariate normal proposal density for β, a univariate normal proposal density for $\log\phi$ and a Dirichlet proposal density for w. Under a logit transformation of w to r-1 dimensional space this last proposal could be a multivariate normal. We run short Metropolis sub-trajectories, typically 20 to 50 iterations, for each update within each iteration of the Gibbs sampler. Such trajectories requires only one new evaluation of the likelihood to proceed from the current step of the trajectory to the next. Starting points for replications of the Gibbs sampler are taken in the vicinity of the maximum likelihood estimates for β under the baseline link g_0 and the associated moments estimator for ϕ when ϕ is present. If we update w first, no starting w values are needed. We adaptively improve the proposal densities, following Müller's suggestions, for typically 25 iterations after which we run five to ten parallel Gibbs replications using various "convergence" diagnostics to decide when to stop. The retained output of the Gibbs sampler, denoted by $(\beta_j^*, w_j^*, \phi_j^*)$ $j = 1, 2, \ldots, m$ is approximately a sample from the posterior enabling us to carry out any desired posterior inference.

Matters of model comparison are examined in the Bayesian framework using predictive distributions. Here we need to study sensitivity to the choice of number of mixands and to compare the fit of the generalized linear model with unknown link to the associated model using the baseline link. Under an improper prior specification the *prior* predictive distribution, as a function of y; $\int L(\beta, w, \phi) f(\beta, w, \phi) d\beta dw d\phi$, is improper hence difficult to compare amongst models. We adopt a *cross validation* approach, in particular, considering the proper predictive densities $f(y_i|y_{(i)})$, $i = 1, 2, \ldots, n$ where $y_{(i)}$ denotes y with y_i removed. See Gelfand, Dey and Chang (1992), for further discussion in this regard. In particular, for each i we obtain a 95% equal tail predictive interval for y_i using f$(y_i|y_{(i),obs})$ and compare it with $y_{i,obs}$. A so-called adequacy plot graphs these intervals along with $y_{i,obs}$ vs. i, perhaps after relabeling the observations to increasing order. We also calculate $f(y_{i,obs}|y_{(i),obs})$, the conditional predictive ordinate. A large value indicates agreement between the observation and the model (Pettit and Young, 1990). A conditional predictive ordinate plot graphs $f(y_{i,obs}|y_{(i),obs})$ vs i. Using these diagnostics, models can be compared at the observation level, i.e., for each i, $i = 1, 2, \ldots, n$. Monte Carlo estimation of $f(y_i|y_{(i),obs})$ and sampling from $f(y_i|y_{(i),obs})$ is taken up in Gelfand & Dey (1993). No details are given here.

5 An illustrative example

To illustrate the modeling of the previous section we study a data set from McCullagh and Nelder (1989, p.204) concerning the number of damage incidents to cargo vessels caused by waves. There are three qualitative covariates : ship type (five levels), year of construction (four levels) and period of operation (two levels). Year of construction is taken to be a continuous covariate. Because the response is a count, Poisson regression seems appropriate. The n=34 counts ranging from 0 to 58 with four bigger than 30. As in section 2, we assume an unknown link of the form $T^{-1}(J(\eta))$ with $J(\eta)$ as in (4). We take the canonical link $\eta = g_0(\mu) = log(\mu)$as the baseline. Under the Poisson model ϕ is intrinsically equal to one. The likelihood is log concave in β given w. We take $f(\beta|w) = 1$ with $f(w)$= Dir(1), i.e., a flat prior.

Figure 1 shows a conditional predictive ordinate plot, where the i's are associated with increasing $y_{i,obs}$, for the baseline model as well as the cases r=3,4. The r =3 and 4 models are quite similar both improving upon the baseline model. We tried larger r's with little gain primarily because, regardless of r, the link function is still strictly increasing. The declining pattern of the conditional predictive ordinate is expected. Since $f(y_i|y_{(i),obs})$ is discrete, when it is concentrated say at 0 and $y_{i,obs}$=0, the conditional predictive ordinate becomes a substantial point mass. We next compare the r=3 model with the baseline model, the latter fitted as in Dellaportas and Smith (1992). Figures 2a and 2b show the adequacy plots for these two models. Better fit would be expected for the r=3 model since it incorporates two additional parameters (weights). Indeed, 17 intervals fail in 2a, only 8 in 2b. Again if $f(y_i|y_{(i),obs})$ predicts few incidents, shorter predictive intervals arise.

Finally, to compare link functions between the two models it is easier to work with $g^{-1}(\eta)$. We "estimated" g^{-1} by $\hat{g}^{-1}$, the Monte Carlo posterior mean, i.e., the average of the m =2000 link functions arising from each of the w_j^*. Comparison of $\hat{g}^{-1}$ and ${g_0}^{-1}$ is on a range of η

values roughly that over which the model is fitted, $[g_0(y_{min}), g_0(y_{max})]$. Figure 3 plots the ratio $\hat{g}^{-1}/{g_0}^{-1}$ over this interval. $\hat{g}^{-1}$ is *larger* than ${g_0}^{-1}$ for small η, i.e., $\hat{g}$ is *smaller* than g_0 for small μ, vice versa for large μ. This is reflected in figure 2. Relative to the baseline model the intervals for the r=3 model are lower for small y_r, higher for large y_r.

6 Summary

Our approach to modeling an unknown link function using a mixture of beta densities is directly applicable to other statistical settings involving modeling a monotone function. These include integrated hazards in survival models and bias functions in size-biased sampling. The Bayesian inference framework with a sampling-based implementation offers a relatively straightforward fitting technique.

References

[1] Angers, J. F. and Delampady, M. (1992a), Hierarchical Bayesian curve fitting and smoothing. *Can.J.Statist.*, 20, 35-49.

[2] Barndörff-Nielsen, O. (1977), Information and exponential families in statistical theory. John Wiley, Inc., New York.

[3] Blight, B. J. N. and Ott, L. (1975), A Bayesian approach to model inadequacy for polynomial fegression. *Biometrika*, 62, 79-80.

[4] Dellaportas, P. and Smith, A. F. M. (1992), Bayesian inference for generalised linear and proportional hazard models via Gibbs sampling. *Appl. Statist.* 42, 443-460.

[5] Diaconis, P. and Ylvisaker, D. (1985), Quantifying prior opinion. In:*Bayesian Statistics* 2, eds. J.M.Bernardo et.al. 133-156. North Holland, Amsterdam.

[6] Ferguson, T. (1973), A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1, 209-230.

[7] Gelfand, A. E. and Smith, A. F. M. (1990), Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 85 , 398-409.

[8] 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, eds. J.M.Bernardo et.al. 133-156. Oxford University Press. Oxford, 147-167.

[9] Gelfand, A.E. and Dey, D.K. (1992), Bayesian model choice: asymptotics and exact calculations. To appear in *J.R.Statist. Soc.*, B. (To appear)

[10] Hinkley, D. V. (1985), Transformation diagnostics for linear models. *Biometrika*, 72, 487-96.

[11] Ibrahim. J, and Laud, P. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *J. Amer. Statist. Assoc.*, 86, 981-986.

[12] McCullagh. P. and Nelder. J. A. (1989), *Generalized linear models.* Chapman and Hall, London.

[13] Müller, P. (1991), A generic approach to posterior integration and Gibbs sampling. *J. Amer. Statist. Assoc.* (to appear).

[14] O'Hagan, A. (1978), Curve fitting and optimal design for prediction. *J.R. Statist. Soc.*, *B*, 40, 1-42.

[15] Pettit, L. I. and Young, K. D. S. (1990), Measuring the effect of observation on Bayes factors. *Biometrika*, 77, 455-466.

[16] Pregibon, D. (1980). *g*oodness of link tests for generalized linear models. *Appl. Statist.*, 29, 15-24.

[17] Smith, A. F. M. and Gelfand, A. E.(1992), Bayesian Statistics without tears : a sampling resampling perspective. *Amer. Statist.*, 46, 84-88.

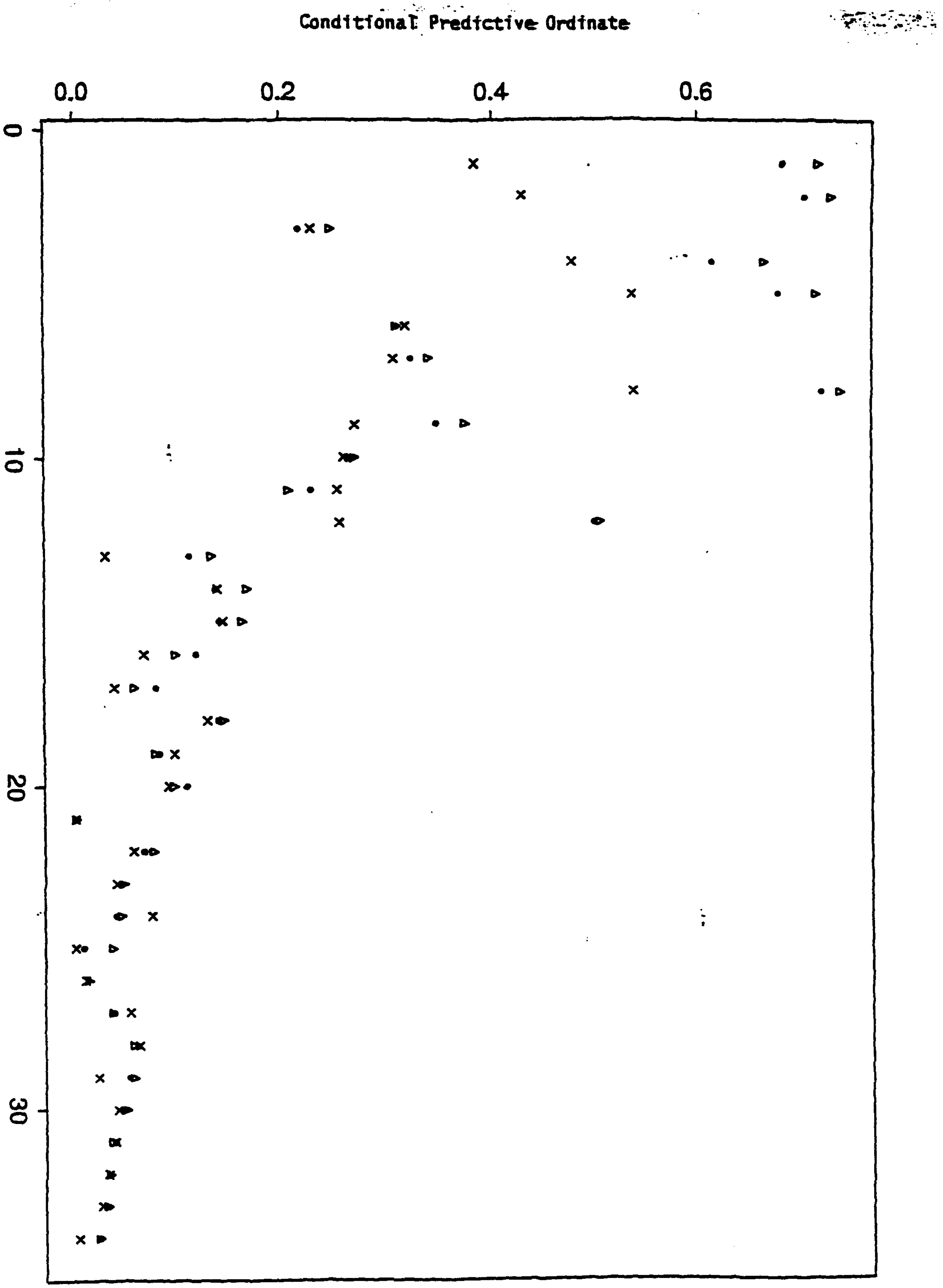
[18] Weerahandi, S. and Zidek, J. V. (1988), Bayesian nonparametric smoothers. *Can.J.Statist*, 16, 61-74.

Figures

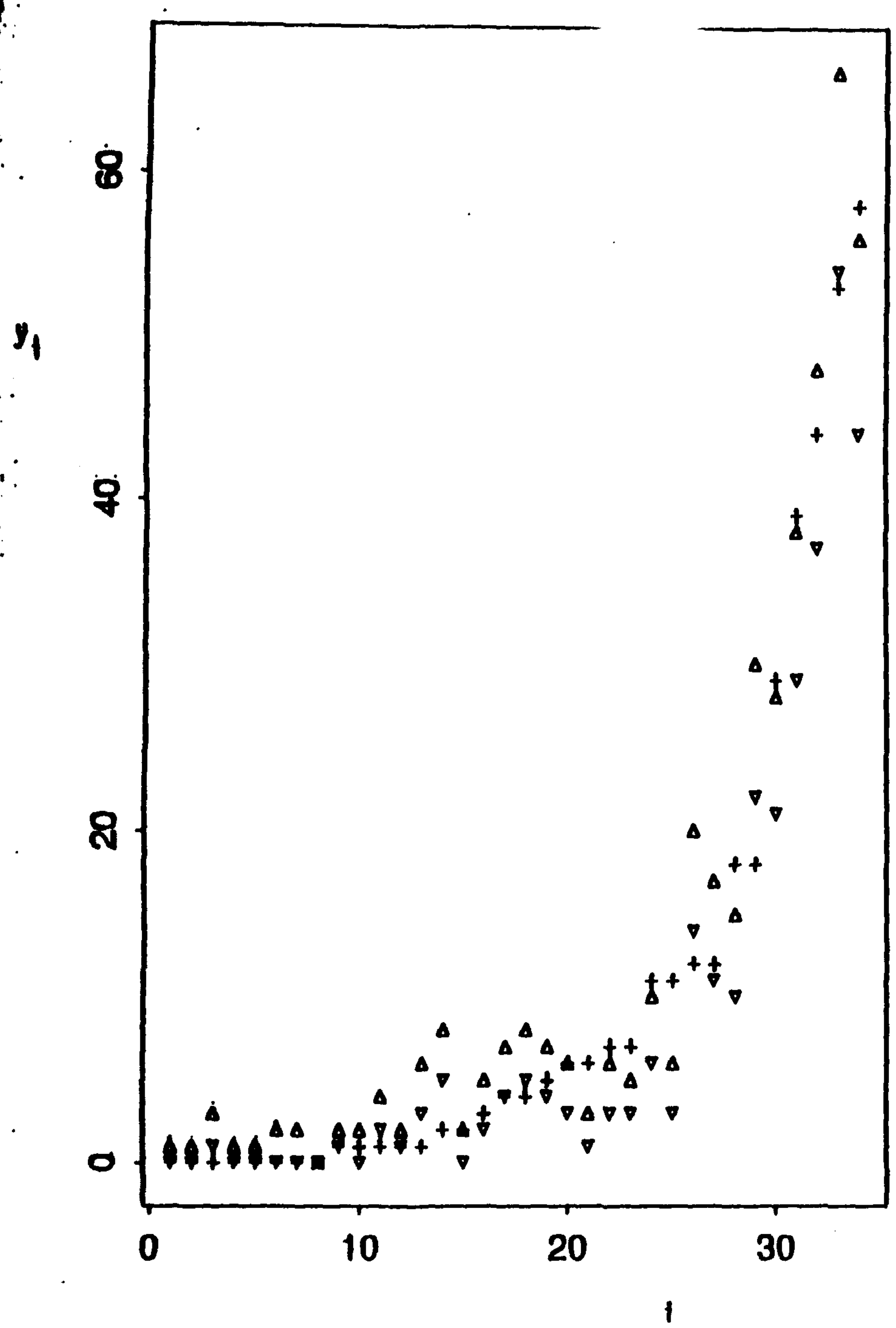
Figure 1: Conditional predictive ordinate plot for r=3 model (•), r=4 model($\triangle$) and baseline model ($\times$).

Figure 2: 2a is an adequacy plot for the baseline model, 2b for r=3 model. 95% equal tail predictive interval are indicated by ∇ and $\triangle$, observed value by $+$.

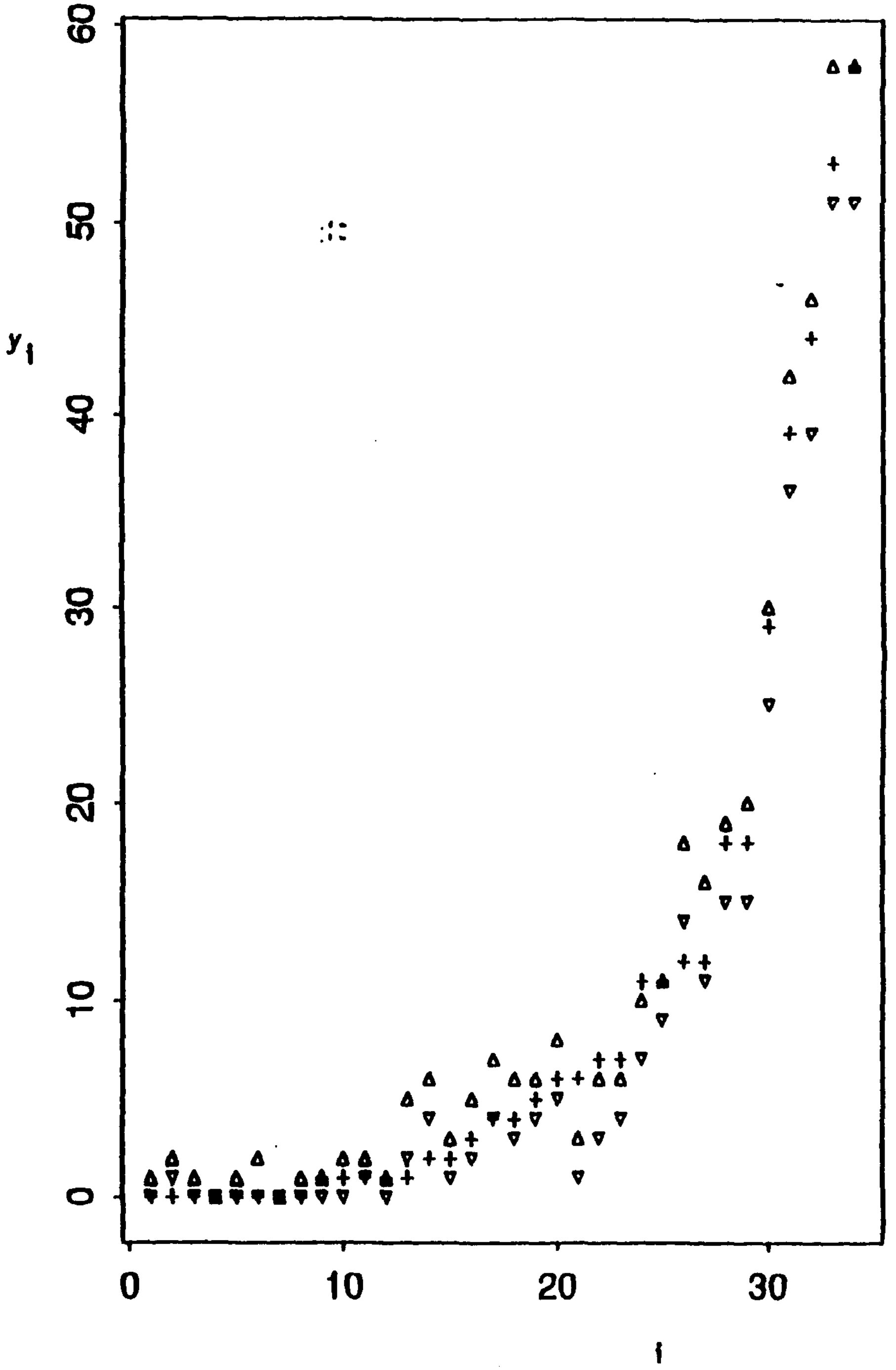
Figure 3: Plot of $\hat{g}^{-1}(\eta)/{g_0}^{-1}(\eta)$ vs η.

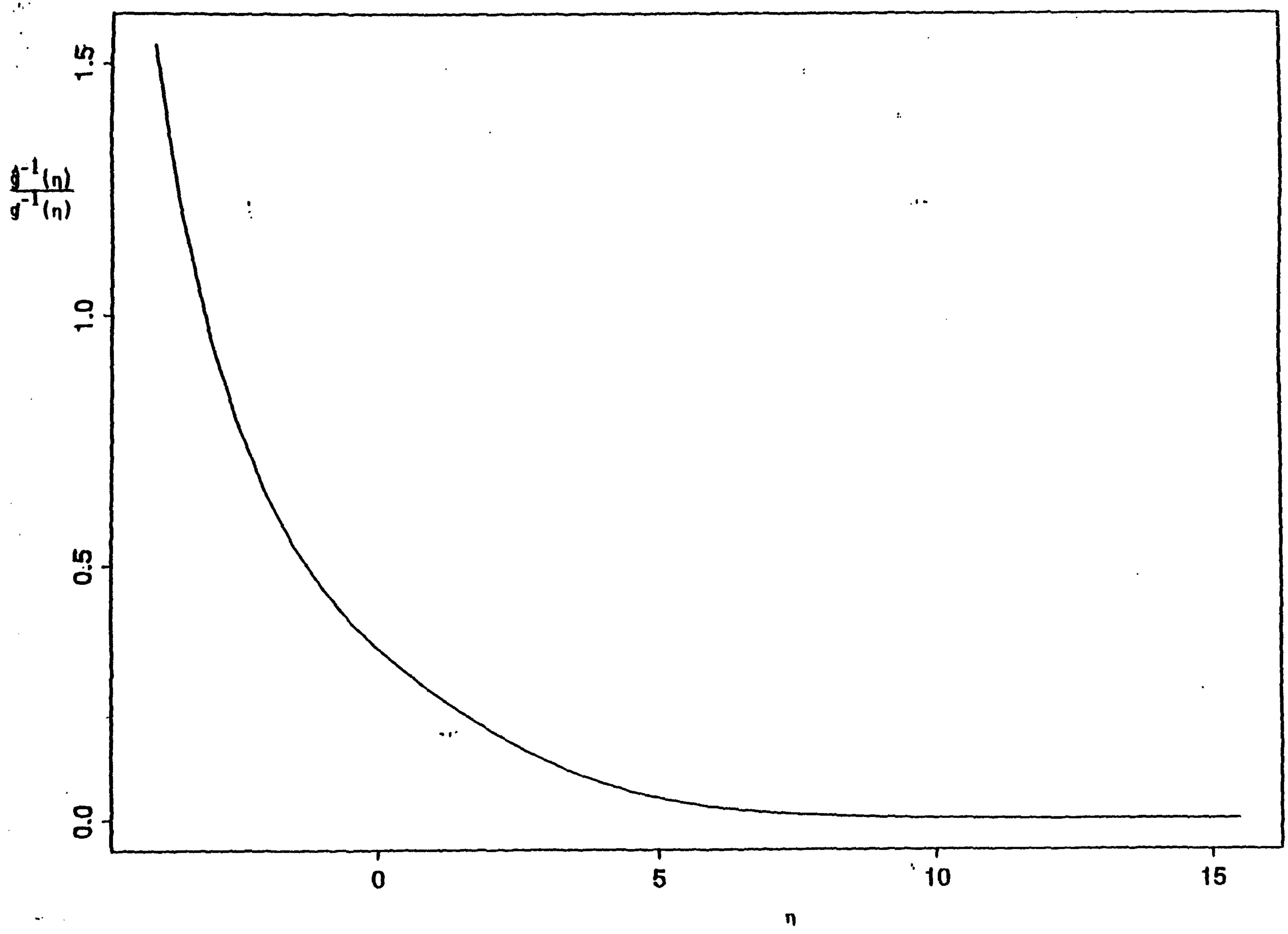

Conditional Predictive Ordinate
0.0
0.2
0.4
0.6
0
10
20
30

[illegible] 2a



2b




$\frac{\dot{g}^{-1}(\eta)}{g^{-1}(\eta)}$
1.5
1.0
0.5
0.0
0
5
10
15
η

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)

Generalized Linear Models with Unknown Link Functions

5. TYPE OF REPORT & PERIOD COVERED

Technical

6. PERFORMING ORG. REPORT NUMBER

482

7. AUTHOR(s)

Bani K. Mallick and Alan E. Gelfand

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

18 July 1994

13. NUMBER OF PAGES

16

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

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

Bayesian model determination; Jeffreys's prior; Metropolis-within-Gibbs algorithm; Mixture-of-Betas distributions

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

See reverse side

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT

Generalized linear models are widely used by data analysts. However, the choice of the link function, i.e., the scale on which the mean is linear in the explanatory variables is often made arbitrarily. Here we permit the data to estimate the link function by incorporating it as an unknown in the model. Since the link function is usually taken to be strictly increasing, by a strictly increasing transformation of its range to the unit interval we can model it as a strictly increasing cumulative distribution function. The transformation results in a domain which is [0,1] as well. We model the cumulative distribution function as a mixture of Beta cumulative distribution functions, noting that the latter family is dense within the collection of all continuous densities on [0,1]. For the fitting of the model we take a Bayesian approach, encouraging vague priors, to focus upon the likelihood. We discuss choices of such priors as well as the integrability of the resultant posteriors. Implementation of the Bayesian approach is carried out using sampling based methods, in particular, a tailored Metropolis-within-Gibbs algorithm. An illustrative example utilizing data involving wave damage to cargo ships is provided.