Bayesian Analysis (2007) **2**, Number 2, pp. 239–280

Nonparametric Estimation of Kullback-Leibler Information Illustrated by Evaluating Goodness of Fit

Kert Viele*

Abstract. We describe a method for quantifying the lack of fit in a proposed family of distributions. The method involves estimating the posterior distribution of the Kullback-Leibler information between the true distribution generating the data and the proposed family. We include an implementation for discrete data involving Dirichlet Processes, for continuous data involving Dirichlet Process Mixtures, and for regression data involving a common "perturbation" distribution also estimated by a Dirichlet Process Mixture. We examine the effectiveness of the method through simulation. We also show that, for independent, identically distributed discrete data, the posterior distribution from a Dirichlet Process provides a consistent estimate of the KL information. Because the entire posterior distribution is computed, one can readily acquire interval estimates of the distance without resorting to asymptotics.

Keywords: Goodness of Fit, Kullback-Leibler Information, Consistency

1 Introduction

One important aspect of statistical modeling is evaluating the fit of the chosen model. Let D_Θ be a family of models indexed by a parameter θ. Evaluating the fit of D_Θ to a set of data may take many forms. In some situations one may wish to determine if there exists a θ such that D_θ is an exact description of the actual process P that generated the data. Alternatively, one may only wish to evaluate if there exists a θ such that D_θ is a "reasonable" approximation of P. Gelfand, Dey, and Chang (1992) make this point, noting that typically one is looking for a "sufficient" model rather than the "correct" one. In cognitive psychology, for example, linear models attempt to quantify such variables as cognitive capacity which are probably nonlinear. In criminology an individual's criminal career is modeled through several environmental variables, not to provide an exact model but to provide an approximation useful for guiding public policy.

Our goal in this paper is to provide an estimate for $d(D_\Theta, P)$, thus quantifying the lack of fit. A model may still be useful even if it clearly incorrect, so long as $d(D_\Theta, P)$ is sufficiently small. An alternative framework is conducting a hypothesis test on the value of $d(D_\Theta, P)$. The question of determining whether the model is an exact fit can be phrased as testing $H_0 : d(D_\Theta, P) = 0$ against $H_1 : d(D_\Theta, P) > 0$. In the Bayesian framework, this hypothesis test may be done using Bayes Factors (Kass and Raftery 1995) or predictive p-values (Gelman, Meng, Stern 1996). One way of

*Department of Statistics, University of Kentucky, Lexington, KY, mailto:viele@ms.uky.edu

evaluating whether D_Θ is a "sufficient" model is to test $H_0 : d(D_\Theta, P) \leq a$ against $H_1 : d(D_\Theta, P) > a$ for some $a > 0$. This approach has been used in Verdinelli and Wasserman (1998), Goutis and Robert (1998), and Mengerson and Robert (1996). In frequentist terms, the Kolmogorov Smirnov (KS) test, the VonMises goodness of fit test, and the χ^2 goodness of fit test produce a "reject or do not reject" result based on a test statistic that is an estimate of a distance measure. For example, the "D-statistic" used in a Kolmogorov Smirnov test is an estimate of the Kolmogorov-Smirnov distance between the distributions. One could consider using the test statistic directly to estimate $d(D_\Theta, P)$ directly.

In this paper we explore theoretical and methodological issues involved in estimating $d(D_\Theta, P)$ for discrete D_Θ using Dirichlet Process priors, concentrating on $d(D_\Theta, P) = \inf_\theta KL(P, D_\theta)$, where

$$KL(P, D) = \int \ln \frac{dP(\omega)}{dD(\omega)}\, dP(\omega)$$

is the Kullback-Leibler (KL) information. KL information is used similarly to our formulation (though parametrically) in Goutis and Robert (1998) and in Mengerson and Robert (1996). An alternative approach for evaluating fit is described in Carota et. al. (1995), who focus on the KL information between the prior and posterior distributions, not on estimating distances between D_Θ and P. We show that Dirichlet Processes can produce consistent estimates of $\inf_\theta KL(D_\theta, P)$. Previous results (Diaconis and Freedman 1986, Carota and Parmigiani 1994) have illustrated that consistency must be checked carefully in nonparametric Bayesian methods, and KL information is a stronger criteria than most others commonly used in exploring consistency such as Hellinger distance, which was considered in Barron et. al. (1999) and Robert and Rousseau (2003).

In fact, KL information sits "atop" several distance/divergence measures that are part of a unified framework formulated by Csiszar (1963) and Ali and Silvey (1966). For any two distributions Q and P, an "f-divergence" has the form

$$I_f(Q, P) = \int q(x) f(p(x)/q(x))\, dx$$

The class of f-divergences includes many commonly used divergence measures, including among others Kullback-Leibler information ($f(u) = -u \ln u$), Total Variation ($f(u) = |u-1|$, note Total Variation and Hellinger produce the same open sets), and χ^2 ($f(u) = (u-1)^2$). For many common f-divergences, KL information can provide an upper bound. One famous inequality is the Csiszar-Kullback-Pinsker inequality, which bounds Hellinger distance above by the square root of KL information (Kullback 1967, note this equality depends on the constant used in defining Hellinger information). In turn, Kolmogorov-Smirnov distance is bounded by total variation, etc. Such bounds imply that any consistency result for KL information implies consistency for those measures bounded by KL, and thus theoretical results in KL information are particularly useful.

Although our central goal here is estimation of the KL information, acquiring the prior and posterior distributions of $d(D_\Theta, P)$ provides the Bayes Factor for testing $H_0 : d(D_\Theta, P) \leq a$ against $H_1 : d(D_\Theta, P) > a$ by comparing the prior and posterior probabilities of each hypothesis. This avoids some of the difficulties involved in computing a Bayes Factor in nonparametric context, for example those found in Carota and Parmigiani (1994). Other solutions to the problem of computing Bayes Factors in nonparametric contexts may be found in Berger and Guglielmi (2001) and Conigliani et. al (2000).

In Section 2 we describe our choice for $d(D_\Theta, P)$ and the method for estimating $d(D_\Theta, P)$. Section 3 provides a detailed implementation for discrete distributions with examples. The method is most developed for discrete distributions, including a results demonstrating consistency of the posterior distribution in KL information. In Section 4 we discuss possible extensions of the method to continuous and regression data, including simulation results. Section 5 provides a discussion of the results.

2 Method

2.1 Null and Alternative Families

As stated in the introduction, we are attempting to assess the fit of a proposed, or null, family of distributions D_Θ. Since we are evaluating the fit of D_Θ, we cannot assume the true distribution that generated the data, P_0, is an element of D_Θ. A fundamental component of any Bayesian analysis is defining the space that the parameter of interest (in this case the entire distribution P) resides in. If possible, we would prefer this space to contain every possible value of P, and to have a practical method of estimating P within this space. In the case of independent, identically distributed data we will make the minimal assumption that P is supported on the same set as D_Θ and estimate P nonparametrically using either a mixture of Dirichlet Processes (Ferguson 1973, Antoniak 1974) or a Dirichlet Process Mixture of Normals (Escobar and West 1995). For regression models we introduce a "perturbation distribution" in Section 4.2 similar to a nonparametric extension of overdispersion. The remainder of the analysis consists of computing the posterior distribution of P and the induced posterior distribution on $d(P, D_\Theta)$.

2.2 Distance Measures

The distance $d(P, D_\Theta)$ is intended to quantify the lack of fit between the P and the null family D_Θ. In a full decision theoretic framework, d should be a loss function between distributions. If such a loss function is available, it may be readily incorporated into what follows. However, it is often the case that a loss function is either unavailable and or is difficult to elicit from investigators because of the nonparametric space involved. Our goal in this section is to motivate a default d for use when no specific loss function is specified.

Our default distance is

$$d(P, D_\Theta) = \inf_{\theta \in \Theta} KL(P, D_\theta) = \kappa \tag{1}$$

where KL stands for the Kullback-Leibler information

$$KL(P, D) = \int \ln \frac{dP}{dD}\, dP$$

This particular distance is motivated by asymptotics and has been used similarly in Mengerson and Robert (1996) and Goutis and Robert (1998). This distance d is a special case of distance measures of the form $d(P, D_\Theta) = d_1(P, D_{\theta_P})$, where d_1 is an unspecified distance function and $\theta_P = \arg\min_\theta KL(P, D_\Theta)$. Berk (1966) shows that if we observe data from $P \notin D_\Theta$ but incorrectly model the data as being from D_Θ, then as the sample size increases the posterior distribution of Θ will, under fairly general conditions, converge to a point mass at θ_P. Thus, distance measures of this form measure the long term loss involved in using D_Θ incorrectly. The function d above chooses KL information for d. KL information has been used in a large number of contexts, such as influence measures (Carlin and Polson 1991) and model sensitivity (McCulloch 1989) in addition to the context here. It is also fundamental to hypothesis testing using Bayes Factors (Kass and Raftery 1995). This previous research indicates KL information forms a natural distance measure between distributions.

Our goal in this paper is to estimate

$$\kappa_0 = d(P_0, D_\Theta)$$

using the posterior distribution of κ as P ranges over its posterior distribution within the alternative family. The value κ_0 quantifies the distance from the true distribution P_0 to the null family of distributions.

2.3 Computing the Posterior Distribution of κ

Usually it is not practical to find the posterior distribution of κ analytically. However, since the prior on P is a mixture of Dirichlet Processes, the posterior distribution of P is a mixture of Dirichlet Processes as well, and thus it is fairly straightforward to draw a large sample $P_1, P_2, \ldots, P_M$ from the posterior distribution of P by Markov Chain Monte Carlo (MCMC) techniques (Gilks, Richardson, and Spiegelhalter 1996). We discuss the details in context in subsequent sections. For each distribution P_m from the posterior distribution of P, we evaluate κ to produce a sequence $\kappa_1, \ldots, \kappa_M$.

We may evaluate κ_m for a particular P_m in a variety of ways. Often direct minimization of the KL information may be done analytically. If necessary, the minimizing θ may also be found using Theorem 4 in Berk (1966), which states under general conditions the maximum likelihood estimate converges to the value of θ minimizing KL

information. In the examples in this paper, analytical solutions are found. The values $\kappa_1, \ldots, \kappa_M$ may be used in the following ways to evaluate fit

1. The posterior density of κ given the data, $\pi(\kappa|X)$, may be estimated using a kernel density estimate $\hat{\pi}(\kappa|X)$. Theorem 3.2 in Section 3.2 shows that $\pi(\kappa|X)$ converges in probability to a point mass at κ_0, the true KL information. Thus, for large samples, the estimated posterior mean/median and/or HPD region from $\hat{\pi}(\kappa|X)$ may be used to estimate κ_0.

2. For more moderate samples, one may see the effect of the data by comparing the prior and posterior distributions of κ. The induced prior density of κ, $\pi(\kappa)$, may be estimated by drawing a large sample from $\pi(P)$, evaluating κ for each observation in the sample, and computing a kernel density estimate $\hat{\pi}(\kappa)$ analogously to the computation of $\hat{\pi}(\kappa|X)$. The estimated prior and posterior distributions on κ may then be compared to see the effect of the data.

3. Although our goal here is not to produce Bayes Factors, the notion of "more mass" from the previous item may be made formal by computing estimated Bayes Factors. The Bayes Factor for testing $H_0 : \kappa \leq a$ against $H_A : \kappa > a$ is

$$\frac{\Pr(X|H_0)}{\Pr(X|H_A)} = \frac{\Pr(H_0|X)}{\Pr(H_A|X)}\frac{\Pr(H_A)}{\Pr(H_0)} = \frac{\Pr(H_0|X)}{\Pr(H_0)}\frac{1-\Pr(H_0)}{1-\Pr(H_0|X)} \tag{2}$$

   For any particular a, $\Pr(H_0)$ may be estimated by the empirical proportion of prior κ values that are less than a while $\Pr(H_0|X)$ may be estimated by the empirical proportion of posterior κ values that are less than a. The resulting estimates may be placed in Equation (2) to produce the estimated Bayes Factor. We demonstrate this with examples in Section 3.3.

Inevitably, one is faced with the decision whether a particular KL information is small or large. This is a complicated problem that has been explored by several authors. The simplest method, which we will follow in this paper, may be found in McCulloch (1989), who suggests calibrating the KL information in terms of Bernoulli distributions. For a particular KL information κ, find $q(\kappa)$ such that the KL information between a Bern(0.5) random variable and a Bern($q(\kappa)$) random variable is κ. Soofi, Ebrahimi, and Habibullah (1995) extend this method to other reference distributions and propose a calibration method based on a normalized transformation of the KL information given by the information distinguishability measure. Hoeffding and Wolfowitz (1958) provide inequalities relating Total Variation distance to KL information which may be useful in interpreting smaller values of the KL information on a probability scale.

Whether a "line in the sand" such as 0.05 for p-values may be placed on KL information is a complicated and controversial problem, which we do not pretend to solve here. The posterior distribution of the KL information (1) may be placed in McCulloch's proposed calibration scheme or any other.

3 Discrete Distributions

3.1 Method

In this section we assume D_Θ is a family of discrete distributions. The support of D_Θ is therefore countable and can be placed in one to one correspondence with the nonnegative integers. Since KL information is invariant to such transformations (Kullback 1968) we can without loss of generality assume that D_Θ is supported on the nonnegative integers.

Our default Mixture of Dirichlet Processes prior begins with a prior distribution on the θ within the null family and proceeds according to

$$\begin{array}{rcl} \theta & \sim & F \\ \gamma^{-1} & \sim & N^+(0,(0.25)^2) \\ P & \sim & \mathrm{Dir}(D_\theta \times \gamma) \\ Y_1,\ldots,Y_n & \sim & P. \end{array} \tag{3}$$

where the notation $\mathrm{Dir}(D_\theta \times \gamma)$ indicates a Dirichlet Process with base probability measure D_Θ and confidence parameter γ, and a $N^+(\mu,\sigma^2)$ distribution is the distribution proportional to a $N(\mu,\sigma^2)$ but supported only on positive real numbers. The prior mixes across γ to place more mass near the null family (Viele 2000). Using results from Escobar (1994), one can show

$$\Pr(\mathbf{y}|\Theta,\gamma) = \frac{\Gamma(\gamma)\Gamma(n+1)\prod_{i=1}^m \Gamma(d_{i,\theta}\gamma + n_i)}{\Gamma(\gamma+n)\prod_{i=1}^m \Gamma(n_i+1)\prod_{i=1}^m \Gamma(d_{i,\theta}\gamma)}$$

where $\mathbf{x} = (x_1,\ldots,x_m)$ are the distinct values observed in $\mathbf{y}$, n_i is the number of times x_i appears in $\mathbf{y}$, and $d_{i,\theta} = D_\theta(\{x_i\})$.

To determine the posterior distribution of the KL information measure (1), one may simulate a sequence of (Θ,γ) using Gibbs Sampling with the Metropolis Algorithm. For each observation in the sequence, also generate

$$P_i \sim \mathrm{Dir}(\gamma D_\theta + n\hat{P})$$

where $\hat{P}$ is the empirical distribution of the observations. For each P_i, compute $\kappa_i = \inf_\theta KL(P_i, D_\theta)$. The resulting κ_i are observations approximately distributed as the posterior distribution of (1).

3.2 Theoretical Results

In this section we demonstrate conditions under which the posterior distribution of κ converges to a point mass at κ_0. Lemma 3.1 is used to reduce the problem from

demonstrating an infimum is consistent to demonstrating consistency at a single point. We then demonstrate, in Theorem 3.2, that a Dirichlet Process prior produces consistent estimates of κ_0.

Lemma 3.1 *Let $Y_1, Y_2, \ldots \sim P_0$ and $\theta_0 = \text{argmin}_\theta KL(P_0, D_\theta)$. Also, for each P in the alternative family define $\theta_P = \text{argmin}_\theta KL(P, D_\theta)$. Let π be a prior over the alternative class and let π_n be the posterior distribution based on $Y_1, \ldots, Y_n$. Also define three sets of neighborhoods*

$$N_{1,\delta} = \{P : |KL(P, D_{\theta_0}) - KL(P_0, D_{\theta_0})| < \delta\}$$

$$N_{2,\delta} = \{P : |KL(P, D_{\theta_P}) - KL(P, D_{\theta_0})| < \delta\}$$

$$N_{\text{inf},\delta} = \left\{P : |\inf_\theta KL(P, D_\theta) - \inf_\theta KL(P_0, D_\theta)| < \delta\right\}$$

Suppose that, for all ϵ and δ (all probabilities are over the distribution of the data),

(a) $\lim_n \Pr(\pi_n(N_{1,\delta}) > 1 - \epsilon) = 1$

(b) $\lim_n \Pr(\pi_n(N_{2,\delta}) > 1 - \epsilon) = 1$

Then, for all ϵ and δ,

$$\Pr(\pi_n(N_{\text{inf},\delta}) > 1 - \epsilon) = 1$$

Proof

By definition, $KL(P, D_{\theta_P}) = \inf_\theta KL(P, D_\theta)$ and $KL(P_0, D_{\theta_0}) = \inf_\theta KL(P_0, D_\theta)$. Therefore, $(N_{1,\delta/2} \bigcap N_{2,\delta/2}) \subset N_{\text{inf},\delta}$ by the triangle inequality and the result follows.

The neighborhood $N_{2,\delta}$ is those P satisfying

$$|KL(P, D_{\theta_P}) - KL(P, D_{\theta_0})| = \left|\int \ln \frac{d_{\theta_0}(y)}{d_{\theta_P}(y)} dP(y)\right| < \delta$$

Thus, condition (b) of Lemma 3.1 is a smoothness condition on the proposed model. Often condition (b) is easy to verify. If, for example, D_θ is the family of Poisson distributions, then it is easy to show $\theta_P = \mu_P$ and thus

$$\int \ln \frac{d_{\theta_0}(y)}{d_{\theta_P}(y)} dP(y) = \theta_P - \theta_0 + \theta_P \ln(\theta_0/\theta_P)$$

Under most reasonable nonparametric priors (for example Dirichlet Processes), θ_P converges to θ_0, resulting in condition (b) being satisfied.

Condition (a) of Lemma 3.1 concerns the consistency of $KL(P, D_{\theta_0})$. Thus, we are only interested in consistency at θ_0 instead of worrying about the entire family D_Θ. If we had a parametric prior on P and the KL information was a continuous function of the parameters, then standard asymptotic results would often allow condition (a) to be verified simply by showing the consistency of the parameters. In nonparametric settings, however, the standard asymptotic results do not necessarily apply since the parameter is infinite dimensional. Theorem 3.2 below proves condition (a) of Lemma 3.1 for a Dirichlet Process prior. Theorem 3.3, which may be proven analogously, shows $KL(\hat{P}_n, D_{\theta_0})$ converges to $KL(P_0, D_{\theta_0})$, where $\hat{P}_n$ is the empirical distribution based on the first n observations.

In the remainder of this section we use the following notation and assumptions. Let $\mathcal{F}$ be the set of all probability measures on the nonnegative integers, and let D, P_0, and Q be elements of $\mathcal{F}$. Let d_i, p_i, and q_i refer to the respective probabilities for each nonnegative integer i. Assume we observe data $X_1, X_2, \ldots \sim P_0$ and let P_0^n be the n-fold product measure of P_0. We place a Dirichlet Process on P with base measure γQ for $\gamma \geq 0$ (the results in this section apply to the "noninformative" prior with $\gamma = 0$). The posterior distribution of P given $X_1, \ldots, X_n$, $\pi_n(P|X)$, is therefore (Ferguson 1973) a Dirichlet Process with base measure $(n\hat{P}_n + \gamma Q)$, where $\hat{P}_n$ is the empirical distribution of the first n observations. We will also define $\hat{p}_{in}$ to be the empirical proportion of $X_1, \ldots, X_n$ that are equal to i. Let $\tilde{P}_n$ be a randomly drawn distribution from $\pi_n(P|X)$, and let $\tilde{p}_{in}$ be the probability that $\tilde{P}_n$ assigns to i. In addition

assume there exists an η such that, for $b_i = \max(p_i^\eta, q_i^\eta)$, the sums $\sum p_i^{1-2\eta}$, $\sum q_i^{1-2\eta}$, $\sum b_i$, $\sum b_i \ln(p_i/d_i)$, and $\sum b_i \ln(1 + (b_i/p_i))$ are all finite.

Let c_n be a sequence such that $c_n > 1$ for all n, $\lim_n c_n = \infty$, and $\lim_n (c_n/n^{1/2}) = 0$.

Let B_{in} be the event that $\tilde{p}_{in} \leq p_i + (b_i/c_n)$, and let $A_n = \bigcap_{i=0}^{\infty} B_{in}$.

Let $N_\epsilon = \{F \in \mathcal{F} : |KL(F, D) - KL(P, D)| < \epsilon\}$

Note that the assumption concerning the five sums is satisfied if D, Q, and P_0 are all of the form $\exp\{-\psi(i)\}$, where $\psi(i)$ is a finite degree polynomial in i.

Theorem 3.2 *Under the assumptions listed above, for all $\epsilon > 0$,* $\lim_n P_0^n (\pi_n(N_\epsilon) > 1 - \epsilon) = 1$.

Proof: This follows from Propositions 5.2 and 5.3, proven in the Appendix.

Theorem 3.3 *Under the assumptions listed above, for all* $\epsilon > 0$, $\lim_n P_0^n \left(|KL(\hat{P}_n, D) - KL(P_0, D)| < \epsilon \right) = 1$.

Proof: See Appendix.

In practice, of course, one only truly estimates a parametric model. For example, if all the observations are less than 20, one practical method is to estimate probabilities for the integers between 0 and 50, making the implicit assumption that very little mass is contained past 50. Even though the actual implementation is parametric, Theorem 3.2 is useful in justifying this procedure. By demonstrating consistency in the nonparametric setting, we establish that the tails have little effect on the KL information. This establishes that the choice of upper bound (so long as there truly is little mass above the upper bound) is unimportant to the analysis.

3.3 Examples

In the following examples we assess the fit of the Poisson family to two real datasets. In each, D_Θ is the Poisson family and we use the prior from Equation (3), with the prior $\pi(\theta) = Exp(0.2)$ (mean 5).

3.3.1 Prussian Horsekicking Data

The Prussian Horsekicking data (shown in Table 1) is a well analyzed dataset which all tests confirm is Poisson. The data show the number of horsekicking deaths for 280 corps-years in the Prussian army. We show it here to demonstrate the method proposed in Section 2 agrees with the previous analyses. The upper left plot in Figure 1 shows a probability plot of the data (dark bars) together with the closest Poisson distribution (light bars). As can be seen, the Poisson distribution closely approximates the data.

Number of Deaths	0	1	2	3	4
Number of Corps-Years	144	91	32	11	2

Table 1: Number of Prussian Army Horsekicking Deaths

We ran 100,000 iterations of an MCMC sampler and provide the estimated posterior density of κ in the upper right plot of Figure 1. As with the simulated Poisson datasets, the posterior mass of κ is near 0. The posterior mean of κ is 0.0065, the difference between Bernoulli distributions with probabilities 0.500 and 0.557, indicating a small difference if any from Poisson. Bayes Factors also confirm the horsekicking data is Poisson. The estimated Bayes Factor for testing $H_0 : \kappa \leq 0.005$ against $H_A : \kappa > 0.005$ is 22.04, strong evidence in favor of H_0.

3.3.2 Epilepsy Data

Table 2 shows data on the number of epileptic seizures by an individual per day over 351 days. If the seizures followed a Poisson process, we would expect the number of seizures per day to follow a Poisson distribution. A probability plot of the data and the closest Poisson distribution are shown in the lower left plot in Figure 1. The probability plot indicates the tail of the data is longer than would be expected from a Poisson distribution.

Number of Seizures	0	1	2	3	4	5	6	7	8
Number of Days	126	80	59	42	24	8	5	4	3

Table 2: Number of Seizures for 351 Days

Again, we computed 100,000 values from the posterior distribution of κ. The estimated posterior density of κ is shown in the lower right plot in Figure 1. The results confirm the impression from the probability plot that the seizure data is not Poisson. The posterior mass is distinct from 0, while a 95% credible interval for κ is $(0.0747, 0.2038)$. The posterior mean is 0.1319, the difference between Bernoulli distributions with probabilities 0.5 and 0.74, a fairly large difference. Bayes Factors also confirm the data is not Poisson. Only 87 of the 100,000 κ values simulated from the prior were less than 0.05 compared to 16071 from the prior distribution, indicating the Bayes Factor is decisively against H_0 in testing $H_0 : \kappa \leq 0.05$ against $H_A : \kappa > 0.05$. Note that while most tests (frequentist Kolmogorov Smirnov tests, etc.) would indicate these data are not Poisson, providing the estimate and the credible region provides more information that a "not Poisson" answer, as one can then decide whether the lack of fit justifies a more complicated model.

4 Possible extensions to continuous and regression data

4.1 Continuous Independent, Identically Distributed Data

The methods in Section 3 cannot be used for continuous data directly because the Dirichlet Process is discrete with probability 1. The KL information between any discrete distribution and any continous distribution is infinity, and thus we must find a nonparametric method that produces continuous distributions. We employ a Dirichlet Process Mixture (DPM).

To evaluate the fit of a continuous family D_θ, we employ a device similar to that in Verdinelli and Wasserman (1998). To generate $Y \sim D_\theta$, we may first generate $U \sim Uni(0,1)$ and then let $Y = D_\theta^{-1}(U)$. We may then construct an alternative distribution by replacing $U \sim Uni(0,1)$ with $U \sim G$. Verdinelli and Wasserman (1998) specify an infinite dimensional exponential family for G. More recently, Robert and Rousseau (2002) use a mixture of Beta distributions and evaluate Hellinger distance rather than KL information. Finally, Kottas (2006) uses a Dirichlet Process Mixture of

Beta distributions in this way for estimation purposes rather than for evaluating fit.

While using a $Uni(0,1)$ distribution has intuitive appeal, we use a slightly different formulation. In general, one can generate $Y \sim D_\theta$ by first generating $V \sim G_0$ and then setting $Y = D_\theta^{-1}(G_0(V))$. This can be expanded into a nonparametric alternative by allowing $V \sim G$ where G is arbitrary and still setting $Y = D_\theta^{-1}(G_0(V))$. In this general setting, the density of Y is

$$f_{Y|G,\theta}(y) = \frac{g(G_0^{-1}(D_\theta(y)))}{g_0(G_0^{-1}(D_\theta(y)))}\, d_\theta(y) \tag{4}$$

Thus, when $G = G_0$, the density of Y reduces the null family D_θ.

We will use $G_0 = \Phi$, the standard normal distribution function. While this creates complexity in writing the formula for Y, it provides a distinct computational advantage. Specifically, instead of estimating a distribution on $(0,1)$ using a Dirichlet Process Mixture of Betas as in Kottas (2006), we can use the more computationally efficient Dirichlet Process Mixture of Normals (Escobar and West 1995). The efficiency arises from the fact that conjugate priors exist for the normal parameters. For the differences between sampling for conjugate priors versus nonconjugate priors, see the review by Neal (2000).

Thus, our complete formulation is

$$\begin{aligned} \theta &\sim \pi(\theta) \\ G &\sim DPM(G^*, \omega) \\ V_1, \ldots, V_n &\sim G \\ Y_i &= D_\theta^{-1}(\Phi(V)) \end{aligned}$$

where $DPM(G^*, \omega)$ is a Dirichlet Process Mixture of normals with base measure G^* and confidence parameter ω. This prior states that G is generated by first drawing a distribution G' over the normal parameter space by a Dirichlet Process (Ferguson 1973) with base measure G^* and confidence parameter ω. This G' results in sequences of means $\mu_1, \mu_2, \ldots$, variances $\sigma_1^2, \sigma_2^2, \ldots$, and probabilities $p_1, p_2, \ldots$ (see Sethuraman 1994 for details) which are then used to form the continuous distribution G through the equality

$$g(\cdot) = \sum_{k=1}^{\infty} p_k N(\mu_k, \sigma_k^2)(\cdot)$$

where $N(\mu_k, \sigma_k^2)(\cdot)$ is the normal density function. The base measure G^* is a joint distribution for the means and variances, while ω controls the probabilities p_k. A common choice of G^* is the conjugate normal-inverse gamma prior where $\sigma^2 \sim IGamma(a, b)$ and $\mu|\sigma^2 \sim N(\eta, \tau^2\sigma^2)$

This prior structure is not identifiable, in that there are multiple G and θ which produce the same distribution for the observable data Y. However, our goal is not to estimate G and θ, but to estimate the KL information between the distribution $F_{Y|G,\theta}(y)$ and the null family D_θ

$$\kappa = \inf_{\theta^* \in \Theta} KL(F_{Y|G,\theta}, D_{\theta^*}) \tag{5}$$

as G and θ range over their posterior distributions. Note we still require the infimum since, after including the perturbation distribution G, the closest D_θ in KL information may not be the same as the D_θ used to generate the Y values.

It is straightforward to implement Markov Chain Monte Carlo (MCMC) in this context. We alternatively sample from the distributions of $G|\theta$ and $\theta|G$. The distribution of $G|\theta$ may be found by first transforming each Y_i back to $V_i = \Phi^{-1}(D_\theta(Y))$ and then updating G according to a standard Dirichlet Process Mixture updating scheme. We use algorithm 2 from Neal (2000). The samples from $\theta|G$ are drawn using the Metropolis algorithm. Since G is general, there are no conjugate priors for θ.

We make straightforward choices for ω and G^*. We simply choose $\omega = 1$ and found little dependence (with respect to the distribution of κ) on this parameter. For G^* our default choices were $a = 1$, $b = 1$, $\eta = 0$ and $\tau^2 = 1e + 08$. Note that since the null value of G is a standard normal, we can focus our prior in the range $(-5, 5)$.

To examine the effectiveness of this strategy we ran simulations with D_θ being the normal family. To avoid confusion between θ in this case, which includes the mean and variance of the normal family to be evaluated, and G^*, which generates μ and σ^2 values as part of the generation of G, we refer to the mean and standard deviation within the null family as θ_μ and θ_σ. We used the prior $\pi(\theta) = 1/\theta_\sigma$.

When D_θ is the normal family, κ may be simplified dramatically. First note the quantity $\Phi^{-1}(D_\theta(y))$ appearing in the density $f_{Y|G,\theta}$ is $\Phi^{-1}(\Phi((y - \theta_\mu)/\theta_\sigma)) = (y - \theta_\mu)/\theta_\sigma$ and thus

$$f_{Y|G,\theta} = \frac{1}{\theta_\sigma}\, \phi\left(\frac{y-\theta_\mu}{\theta_\sigma}\right) \frac{g\left(\frac{y-\theta_\mu}{\theta_\sigma}\right)}{\phi\left(\frac{y-\theta_\mu}{\theta_\sigma}\right)} = \frac{1}{\theta_\sigma} g\left(\frac{y-\theta_\mu}{\theta_\sigma}\right)$$

Thus, for fixed G, θ, and θ^*

$$KL(F_{Y|G,\theta}, D_{\theta^*}) = KL\left(\frac{1}{\theta_\sigma} g\left(\frac{y-\theta_\mu}{\theta_\sigma}\right), \frac{1}{\theta^*_\sigma}\phi\left(\frac{y-\theta^*_\mu}{\theta^*_\sigma}\right)\right)$$

$$= KL\left(G, N\left(\frac{\theta^*_\mu - \theta_\mu}{\theta_\sigma}, \frac{\theta^*_\sigma}{\theta_\sigma}\right)\right)$$

Distribution	g(v)	κ	E_G	$\sqrt{V_G}$
A - Normal	$\phi(v)$	0	0	1
B - Mixture	$0.7N(0,1)(v) + 0.3N(3,0.5)(v)$	0.0959	0.9	1.6544
C - Skew	$0.5Exp(1.66)(-v)I_{v<0} + 0.5Exp(1)(v)I_{v>=0}$	0.1188	0.1985	1.1492
D - Uniform	$Uni(-3,3)(v)$	0.1763	0	1.7321
E - t_3	$t_3(v)$	0.1035	0	1.7321

Table 3: The five G distributions used for simulation. This table provides a short title used for description, the density, the value of κ, and the mean and variance of G used to find the normal distribution the minimizes KL information to G

Since for any θ, allowing θ^* to vary over the parameter space allows $(\theta^*_\mu - \theta_\mu)/(\theta_\sigma)$ and $\theta^*_\sigma/\theta_\sigma$ to vary over the parameter space, we find

$$\kappa = \inf_{\theta^* \in \Theta} KL(F_{Y|G,\theta}, D_{\theta^*}) = \inf_{\theta^* \in \Theta} KL(G, N(\theta^*_\mu, \theta^*_\sigma)) \tag{6}$$

which achieves its infimum at $\theta^*_\mu = E_G$ and $\theta^*_\sigma = V_G^{1/2}$. Thus, when evaluating the fit of a normal distribution, we can work with G directly rather than the induced distribution on Y. Note κ is not found by projecting G to a $N(0,1)$, but to the entire normal family.

The five G distributions are described in Table 3, with their corresponding value of κ and the normal parameters used to achieve the infimum of the KL information. Distribution A is a standard normal, the null value. Thus, the KL information is 0 and the family D_θ is an exact fit. Distribution B is a standard mixture of two normals. Distribution C is a skewed distribution, a "double exponential" but with differing parameters for the exponential halves. Distribution D is a uniform and distribution E is a t distribution. Figure 2 shows the five densities in black, with the closest normal density in KL information shown in red.

We do not expect the method to work well for distribution D. As noted in Kottas (2006), mixtures of normals do not estimate distributions with bounded support well. We are unaware of any distribution that simultaneously handles all possible alternatives well. It might be interesting to see if one could create a sampler that includes both mixtures of normals and mixture of betas. For each of the 5 distributions, we generated data using sample sizes of 30, 100, and 500, and ran the sampler for 1000 iterations of "burn-in" and 10000 iterations used for results.

Table 4 summarizes the results for each of the 15 simulations. We include the estimated posterior mean of the κ values, an estimated 95% highest posterior density (HPD) region, and for reference the true value of κ.

Histograms of the simulated values of κ are shown in Figures 3-7. For the normal distributions, the posterior distribution of κ quickly converges to 0, the true value. For all other distributions, the posterior distributions retain a mode at $\kappa = 0$ for the $n = 30$

Distribution	n	Mean κ	$95\%HPD$	True value
A - Normal	30	0.0259	(0.0000,0.1149)	0.0000
A - Normal	100	0.0142	(0.0000,0.0546)	0.0000
A - Normal	500	0.0040	(0.0000,0.0153)	0.0000
B - Mixture	30	0.0822	(0.0000,0.1977)	0.0959
B - Mixture	100	0.0986	(0.0296,0.1805)	0.0959
B - Mixture	500	0.0653	(0.0379,0.0992)	0.0959
C - Skew	30	0.0454	(0.0000,0.1887)	0.1188
C - Skew	100	0.1574	(0.0296,0.4697)	0.1188
C - Skew	500	0.1317	(0.0834,0.1975)	0.1188
D - Uniform	30	0.0244	(0.0000,0.0828)	0.1763
D - Uniform	100	0.0584	(0.0109,0.1225)	0.1763
D - Uniform	500	0.0786	(0.0532,0.1063)	0.1763
E - t_3	30	0.0515	(0.0000,0.1781)	0.1035
E - t_3	100	0.0931	(0.0144,0.2526)	0.1035
E - t_3	500	0.1195	(0.0540,0.2551)	0.1035

Table 4: Results for the Continuous iid data. With the exception of the uniform distribution, known to be a problem, all the HPD regions contain the true value of κ.

simulations, with the posterior distributions becoming more concentrated as the sample size increases. As expected, the method performs poorly at estimating κ when G is a uniform distribution. While the posterior distribution is separated from 0 for $n = 100$ and $n = 500$ for the uniform, the HPD regions remain consistently (and dramatically) below the true value of $\kappa = 0.1763$. All other HPD regions contain the true values of κ.

4.2 Regression Models

In this section we describe evaluating fit for a regression model where we observe known $x_1, \ldots, x_n$ and independent random variables $Y_1, \ldots, Y_n$ where each $Y_i \sim D_{\theta_i}$ for $\theta_i = \eta(x_i)$.

We first define our goals for evaluating fit. One could use the term "lack of fit" to describe either a misspecification of the function η (for example, using a simple linear regression where a quadratic regression may be more appropriate) or in the family D_θ (e.g. Y may be linearly related to x, but with nonnormal errors). We focus on the latter. The first problem, a misspecification of η, has been considered nonparametrically by Gelfand, et. al. (2005), where the function η is estimated by dependent Dirichlet Processes which build η out of a sum of stochastic processes. Alternatively, one can consider nonparametrically estimating the joint distribution of X and Y as in Mueller et. al. (1996). The difficulty with estimating the joint distribution is that all estimates of the distribution of Y given X are based primarily on local information, rather than combining the lack of fit information across all values of x.

The proposed method is a nonparametric analogue of overdispersion in generalized linear models (McCullagh and Nelder 1989). In overdispersion, one includes an extra parameter that is common across all values of x which increases the variance of the distribution of Y given x. The central point is that the overdispersion parameter is common across x.

Our nonparametric analogue is defined as follows. We assume we know the parametric form for η that depends on parameters $\beta = (\beta_0, \ldots, \beta_p)$, for example a linear regression or other generalized linear model. We also assume the x_i values are fixed and known. In the null family, the random variables $Y_1, \ldots, Y_n$ are formed by

$$\begin{array}{rcl} \beta & \sim & \pi(\beta) \\ \theta_i & = & \eta(x_i) \\ Y_i|\theta_i & \sim & D_{\theta_i} \end{array} \tag{7}$$

This can be placed in the same formulation as in Section 4.1 by

$$\begin{array}{rcl} \beta & \sim & \pi(\beta) \\ \theta_i & = & \eta(x_i) \\ G & \sim & DPM(G^*, \omega) \\ V_1, \ldots, V_n & \sim & G \\ Y_i & = & D_{\theta_i}^{-1}(\Phi(V_i)) \end{array} \tag{8}$$

where the null model (7) can be acquired by setting $G = \Phi$, the standard normal distribution function. As in Section 4.1, we estimate G nonparametrically using a Dirichlet Mixture of Normals.

Sampling in this context is similar to the method for continuous data described in Section 4.1. At each iteration of a Markov Chain, we update G using fixed β by transforming the Y_i to V_i and then using the update step in algorithm 2 of Neal (2000), and then we update β using fixed G using the Metropolis algorithm. These steps become more complicated when D_Θ is a discrete distribution because V_i is known only up to the interval where $D_{\theta_i}^{-1}(\Phi(V_i)) = Y_i$. This can be handled using the interval censoring approach taken in Hanson and Johnson (2004), which we do not discuss here.

Since we have a covariate x, we are interested in estimating the KL information function

$$\kappa(x) = \inf_{\theta^* \in \Theta} KL(F_{Y|\theta=\eta(x),G}, D_{\theta^*})$$

where $F_{Y|\theta=\eta(x),G}$ is as in Equation (4). Note that we require a separate value of KL for each x because given perturbation functions may have more of an affect for some values of x than others. For example, Let consider a Poisson regression where $D_\theta = Poi(\theta)$,

$\eta(x) = \exp\{x-2\}$, and $G = Uni(-3,3)$. In this case $\kappa(2) = 0.234$ while $\kappa(4) = 0.575$. Thus, the KL information depends on x and the quality of the fit, in terms of KL information, depends on the location of the observed x values.

Note an exception to this phenomenon (the dependence of κ on x) occurs when D_θ is the normal family. Equation (6) in Section 4.1 still holds, resulting in $\kappa(x) = \inf_{\theta^*} KL(G, D_{\theta^*})$ for all x.

To examine the fit of regression data, we ran simulations using simple linear regression. For each simulation we used evenly spaced X_i values over the range 0 to 5. We have parameters $\beta_0 = 2$, $\beta_1 = 3$, and $\sigma^2_{reg} = 1$ (the regression variance, with the subscript intended to differentiate σ^2_{reg} from the variances used in generating G) and the model is

$$\begin{array}{rcl} (\beta_0, \beta_1, \sigma^2_{reg}) & \sim & \pi(\beta_0, \beta_1, \sigma^2_{reg}) \\ \theta_i = (\theta_\mu, \theta_\sigma) & = & (\beta_0 + \beta_1 x_i, \sigma_{reg}) \\ G & \sim & DPM(G^*, \omega) \\ V_1, \ldots, V_n & \sim & G \\ Y_i & = & D^{-1}_{\theta_i}(\Phi(V_i)) \end{array} \tag{9}$$

In generating our simulations, we used the same five G distributions from Table 3 in Section 4.1 and the same samples sizes (30, 100, and 500). We also used the same simulated $V_1, \ldots, V_n$ as in Section 4.1. While this does reduce the number of independent samples used in this paper, it also allows us to see the direct effect of estimating the regression parameters. As with the continuous data, we ran 1000 iterations of burn-in and 10000 iterations were used to estimate the posterior distribution. Figure 8 shows the data used in the regression simulations with $n = 500$ (for smaller samples, of course, the pattern is similar).

The results are described in Table 5 and Figures 9-13. Note that as stated above, $\kappa(x)$ is constant across all x for normal regression data and thus we report the common κ.

The results for the regression data are similar to that for continuous iid data. There is a tendency for slightly wider intervals, but not a dramatic one. For normal G the value of κ appears to tend toward 0, and for the remaining G we notice a mode at 0 for $n = 30$ but the distribution has separated from $\kappa = 0$ for all other situations except Uniform $n = 100$ (recall we know the uniform distribution is difficult to estimate using a Dirichlet Mixture of Normals). As with the continuous iid data, all the 95% HPD regions contain the exact answer except for the uniform G intervals.

5 Discussion

The Dirichlet process method for evaluating fit described in this paper provides a method for quantifying the inaccuracy of a discrete distribution. Theorem 3.2 demonstrates

Distribution	n	Mean κ	$95\% HPD$	True value
A - Normal	30	0.0331	(0.0000,0.1382)	0.0000
A - Normal	100	0.0189	(0.0000,0.0753)	0.0000
A - Normal	500	0.0058	(0.0000,0.0224)	0.0000
B - Mixture	30	0.0383	(0.0000,0.1500)	0.0959
B - Mixture	100	0.0592	(0.0003,0.1540)	0.0959
B - Mixture	500	0.0618	(0.0315,0.1016)	0.0959
C - Skew	30	0.0447	(0.0000,0.1801)	0.1188
C - Skew	100	0.1599	(0.0276,0.4781)	0.1188
C - Skew	500	0.1293	(0.0793,0.1986)	0.1188
D - Uniform	30	0.0308	(0.0000,0.1277)	0.1763
D - Uniform	100	0.0217	(0.0000,0.0694)	0.1763
D - Uniform	500	0.0509	(0.0260,0.0807)	0.1763
E - t_3	30	0.0514	(0.0000,0.2044)	0.1035
E - t_3	100	0.0981	(0.0106,0.2982)	0.1035
E - t_3	500	0.1213	(0.0535,0.2604)	0.1035

Table 5: Results for the Simple Normal Linear Regression data. With the exception of the uniform, known to be a problem, all HPD regions contain the true answer.

that the method can provide a consistent estimate of $\inf_\theta KL(P, D_\theta)$. Note that the theoretical results are in need of at least one key improvement. The results are for single Dirichlet Processes, not mixtures of Dirichlet Processes. This would at first seem an "easy" problem, since the mixture consists of a bivariate parameter (θ,γ). Simulations indicate that the posterior distribution $P(\theta, \gamma|\mathbf{y})$ converges setwise to a nondegenerate distribution, which would seem sufficient to invoke general convergence theorems such as those in Chapter 11, Section 4 of Royden (1988) (these involve convergence of $\int f_n d\mu_n$ as opposed to the more common $\int f_n d\mu$ or $\int f d\mu_n$).

By providing the entire posterior distribution of $d(P, D_\theta)$, information on lack of fit may be communicated without drawing lines in the sand concerning what makes a "large" model error, a property that is useful in situations where defining a "large" model error is either difficult or controversial. In addition, the posterior distribution allows easy calculation of Bayes Factors for testing whether the distance is less than any particular value.

Appendix - Proofs of Theoretical Results

The notation in this section is the same as that introduced in Section 3.2.

Proof of Theorem 3.2

Lemma 5.1

$$E[\tilde{p}_{in}] = \frac{np_i + \gamma q_i}{n+\gamma}$$

$$V[\tilde{p}_{in}] = \frac{n(n-1)p_i(1-p_i) + n\gamma p_i(1-q_i) + n\gamma q_i(1-p_i) + \gamma^2 q_i(1-q_i)}{(n+\gamma)^2(n+\gamma+1)} + \frac{np_i(1-p_i)}{(n+\gamma)^2}$$

$$V[\tilde{p}_{in}] \leq \frac{(2+\gamma)p_i + (\gamma+\gamma^2)q_i}{n}$$

Proof:

$$n\hat{p}_{in} \sim Bin(n, p_i)$$

and

$$\tilde{p}_{in}|\hat{p}_{in} \sim Beta(n\hat{p}_{in} + \gamma q_i, n(1-\hat{p}_{in}) + \gamma(1-q_i))$$

Therefore

$$E[\tilde{p}_{in}] = E[E[\tilde{p}_{in}|\hat{p}_{in}]] = E\left[\frac{n\hat{p}_{in} + \gamma q_i}{n+\gamma}\right] = \frac{np_i + \gamma q_i}{n+\gamma}$$

and

$$V[\tilde{p}_{in}] = E[V[\tilde{p}_{in}|\hat{p}_{in}]] + V[E[\tilde{p}_{in}|\hat{p}_{in}]]$$

$$= E\left[\frac{(n\hat{p}_{in} + \gamma q_i)(n(1-\hat{p}_{in}) + \gamma(1-q_i))}{(n+\gamma)^2(n+\gamma+1)}\right] + V\left[\frac{n\hat{p}_{in} + \gamma q_i}{n+\gamma}\right]$$

$$= E\left[\frac{n^2\hat{p}_{in}(1-\hat{p}_{in}) + n\gamma\hat{p}_{in}(1-q_i) + n\gamma q_i(1-\hat{p}_{in}) + \gamma^2 q_i(1-q_i)}{(n+\gamma)^2(n+\gamma+1)}\right] + \frac{np_i(1-p_i)}{(n+\gamma)^2}$$

$$= \frac{n(n-1)p_i(1-p_i) + n\gamma p_i(1-q_i) + n\gamma q_i(1-p_i) + \gamma^2 q_i(1-q_i)}{(n+\gamma)^2(n+\gamma+1)} + \frac{np_i(1-p_i)}{(n+\gamma)^2}$$

Note one can place an upper bound on the variance by

$$V[\tilde{p}_{in}] \leq \frac{n(n-1)p_i + n\gamma(p_i+q_i) + \gamma^2 q_i}{(n+\gamma)^2(n+\gamma+1)} + \frac{np_i}{(n+\gamma)^2}$$

$$\leq \frac{n^2 p_i + n\gamma(p_i+q_i) + \gamma^2 q_i}{n^3} + \frac{np_i}{n^2} = \frac{2p_i}{n} + \frac{\gamma(p_i+q_i)}{n^2} + \frac{\gamma^2 q_i}{n^3}$$

$$\leq \frac{(2+\gamma)p_i + (\gamma+\gamma^2)q_i}{n}$$

Proposition 5.2 *For any $\epsilon > 0$, $\lim_n P_n(KL(\tilde{P}, D) > KL(P, D) + \epsilon) = 0$.*

Proof: Using Chebychev's inequality and lemma 5.1, for all i and n,

$$P_n(B_{in}^c) = P_n\left(\tilde{p}_{in} > p_i + \frac{b_i}{c_n}\right) = P_n\left(\tilde{p}_{in} - \frac{np_i+\gamma q_i}{n+\gamma} > p_i + \frac{b_i}{c_n} - \frac{np_i+\gamma q_i}{n+\gamma}\right)$$

$$\leq P_n\left(\left|\tilde{p}_{in} - \frac{np_i+\gamma q_i}{n+\gamma}\right| > \frac{b_i}{c_n} + \frac{\gamma(p_i-q_i)}{n+\gamma}\right) \leq \frac{V[\tilde{p}_{in}]}{\left[\frac{b_i}{c_n} + \frac{\gamma(p_i-q_i)}{n+\gamma}\right]^2} \leq \frac{(2+\gamma)p_i + (\gamma+\gamma^2)q_i}{n\left[\frac{b_i}{c_n} + \frac{\gamma(p_i-q_i)}{n+\gamma}\right]^2}$$

Let N be an integer such that $(\gamma c_n)/(n+\gamma) < 1$ for all $n > N$ and $q_i^\eta - q_i > (q_i^\eta/2)$ for all $i > N$ (such an N exists by the construction of c_n and the fact that Q is a distribution with $\lim_i q_i = 0$). Then

$$P_n(A_n^c) \leq \sum_{i=0}^{N} P_n(B_{in}^c) + \sum_{i=N+1}^{\infty} P_n(B_{in}^c)$$

$$\leq \sum_{i=0}^{N} P_n(B_{in}^c) + \frac{c_n^2}{n} \sum_{i=N+1}^{\infty} \frac{(2+\gamma)p_i + (\gamma+\gamma^2)q_i}{\left[b_i + \frac{\gamma c_n(p_i-q_i)}{n+\gamma}\right]^2}$$

$$\leq \sum_{i=0}^{N} P_n(B_{in}^c) + \frac{c_n^2}{n}\left[\sum_{i>N:p_i\geq q_i} \frac{(2+\gamma)p_i + (\gamma+\gamma^2)q_i}{\left[b_i + \frac{\gamma c_n(p_i-q_i)}{n+\gamma}\right]^2} + \sum_{i>N:q_i>p_i} \frac{(2+\gamma)p_i + (\gamma+\gamma^2)q_i}{\left[b_i + \frac{\gamma c_n(p_i-q_i)}{n+\gamma}\right]^2}\right]$$

$$\leq \sum_{i=0}^{N} P_n(B_{in}^c) + \frac{c_n^2}{n}\left[\sum_{i>N:p_i\geq q_i} \frac{(2+2\gamma+\gamma^2)p_i}{\left[p_i^\eta + \frac{\gamma c_n(p_i-q_i)}{n+\gamma}\right]^2} + \sum_{i>N:q_i>p_i} \frac{(2+2\gamma+\gamma^2)q_i}{\left[q_i^\eta + \frac{\gamma c_n(p_i-q_i)}{n+\gamma}\right]^2}\right]$$

Using the assumptions defining N, we find this quantity is

$$\leq \sum_{i=0}^{N} P_n(B_{in}^c) + \frac{(2+2\gamma+\gamma^2)c_n^2}{n}\left[\sum_{i>N:p_i\geq q_i} p_i^{1-2\eta} + 4\sum_{i>N:q_i>p_i} q_i^{1-2\eta}\right]$$

The bound on the last term in the brackets follows from

$$q_i^\eta + \frac{\gamma c_n}{n+\gamma}(p_i - q_i) = q_i^\eta + \frac{\gamma c_n}{n+\gamma}p_i - \frac{\gamma c_n}{n+\gamma}q_i > \frac{q_i^\eta}{2} > 0$$

The first sum converges to 0 as n increases because it involves a finite number of parameters $p_0, \ldots, p_N$. Returning the definition of B_{in}, each $\tilde{p}_{in}$ converges to p_i at rate $n^{-1/2}$, which is faster than $1/c_n$ by assumption. In addition, the two sums inside the bracket are finite by assumption. Therefore, the entire quantity tends to 0 as n increases, resulting in $P_n(A_n)$ tending to 1 as n increases. When A_n occurs

$$KL(\tilde{P}, D) = \sum_{i=0}^{\infty} \tilde{p}_{in} \ln \frac{\tilde{p}_{in}}{d_i} \leq \sum_{i=0}^{\infty}\left[p_i + \frac{b_i}{c_n}\right] \ln\left[\frac{p_i + \frac{b_i}{c_n}}{d_i}\right] = \sum_{i=0}^{\infty}\left[p_i + \frac{b_i}{c_n}\right] \ln\left[\frac{p_i}{d_i}\left(1 + \frac{b_i}{c_n p_i}\right)\right]$$

$$= \sum p_i \ln \frac{p_i}{d_i} + \sum p_i \ln\left(1 + \frac{b_i}{c_n p_i}\right) + \frac{1}{c_n}\sum b_i \ln \frac{p_i}{d_i} + \frac{1}{c_n}\sum b_i \ln\left(1 + \frac{b_i}{c_n p_i}\right)$$

$$\leq KL(P, D) + \frac{1}{c_n}\left[\sum b_i + \sum b_i \ln \frac{p_i}{d_i} + \sum b_i \ln\left(1 + \frac{b_i}{p_i}\right)\right]$$

The three sums in the brackets are finite by assumption, hence for all ϵ,

$$\lim_n P_n(KL(\tilde{P}, D) > KL(P, D) + \epsilon) = 0$$

Proposition 5.3 *For any* ϵ, $\lim_n P_n(KL(\tilde{P}_n, D) \geq KL(P, D) - \epsilon) = 1$.

Proof: Since $KL(P, D)$ is finite by assumption, for any ϵ we may find a k such that $\sum_{i=0}^{k} p_i \ln(p_i/d_i) > KL(P, D) - \epsilon/2$ and $\sum_{i=k+1}^{\infty} d_i < e\epsilon/2$. Let $k_n = 1 - \sum_{i=0}^{k} \tilde{p}_{in}$ and define

$$CKL_n = \sum_{i=0}^{k} \tilde{p}_{in} \ln(\tilde{p}_{in}/d_i) + \sum_{i=k+1}^{\infty} k_n d_i \ln(k_n d_i/d_i)$$

By Kullback (1968, page 13),

$$\begin{array}{rcl} KL(\tilde{P}, D) \geq CKL_n & = & \sum_{i=0}^{k} \tilde{p}_{in} \ln(\tilde{p}_{in}/d_i) + k_n \ln k_n \sum_{i=k+1}^{\infty} d_i \\ & \geq & \sum_{i=0}^{k} \tilde{p}_{in} \ln(\tilde{p}_{in}/d_i) - (1/e) \sum_{i=k+1}^{\infty} d_i \\ & \geq & \sum_{i=0}^{k} \tilde{p}_{in} \ln(\tilde{p}_{in}/d_i) - (\epsilon/2) \end{array} \tag{10}$$

Since $p_0, \ldots, p_k$ are a finite set of parameters and each $\tilde{p}_{in}$ converges almost surely to p_i,

$$\lim_n P_n \left(\sum_{i=0}^{k} \tilde{p}_{in} \ln(\tilde{p}_{in}/d_i) \geq \sum_{i=0}^{k} p_i \ln(p_i/d_i) - \epsilon/2 \right) = 1$$

Combining this with the Equation (10)

$$\lim_n P_n(KL(\tilde{P}, D) \geq KL(P, D) - \epsilon) = 1$$

Proof of Theorem 3.3

The proof proceeds analogously to the proof of Theorem 3.2. The bounds in Lemma 5.1 on the mean and variance of $\tilde{p}_{in}$ may be used for $\hat{p}_{in}$ since 1) $\hat{p}_{in}$ is unbiased for p_i and 2) $V[\hat{p}_{in}] \leq 2p_i/n$, which is less than the variance bound for $\tilde{p}_{in}$. Thus, the bounds used in Proposition 5.2 may be used for the empirical distribution $\hat{P}$. The proof of Proposition 5.3 is identical with $\tilde{p}_{in}$ replaced with $\hat{p}_{in}$.

LaTeX Warning: Label(s) may have changed. Rerun to get cross-references right.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, 267–281. Budapest, Hungary: Akademiai Kiado.

Andrews, D. W. K. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59: 817–858.

Anscombe, F. J. 1972. Contribution to the discussion of H. Hotelling's paper. *Journal of the Royal Statistical Society - Series B* 15(1): 229–230.

Arminger, G., C. C. Clogg, and M. E. Sobel, eds. 1995. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum Press.

Barndorff-Nielsen, O. 1976. Factorization of likelihood functions for full exponential families. *Journal of the Royal Statistical Society - Series B* 38(1): 37–44.

Barndorff-Nielsen, O. E. and D. R. Cox. 1984. Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *Journal of the Royal Statistical Society - Series B* 46(3): 483–495.

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Ben-Akiva, M. and S. R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA: MIT Press.

Berkson, J. 1944. Application of the logistic function to bio-assay. *Journal of the American Statistical Association* 39: 357–365.

Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman. 1974. Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* 3/4: 653–665.

Binder, D. A. 1983. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51: 279–292.

—. 1992. Fitting Cox's proportional hazards models from survey data. *Biometrika* 79(1): 139–147.

Bliss, C. I. 1934. The method of probits. *Science* 79: 38–39, 409–410.

Booth, J. G. and R. W. Butler. 1990. Randomization distributions and saddlepoint approximations in generalized linear models. *Biometrika* 77(4): 787–796.

Booth, J. G. and J. P. Hobert. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society - Series B* 61(1): 265–285.

von Bortkewitsch, L. 1898. *Das Gesetz der Kleinen Zahlen*. Leipzig: Teubner.

Breslow, N. E. 1984. Extra-Poisson variation in log-linear models. *Applied Statistics* 33(1): 38–44.

—. 1990. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association* 85(410): 565–571.

—. 1996. Generalized linear models: Checking assumptions and strenghening conclusions. *Statistica Applicata* 8: 23–41.

Breslow, N. E. and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88(421): 9–25.

Brown, D. 1992. A graphical analysis of deviance. *Applied Statistics* 41(1): 55–62.

Cameron, A. C. and F. A. G. Windmeijer. 1997. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics* 77: 329–342.

Carroll, R. J. and D. Ruppert. 1982. Robust estimation in heteroscedastic linear models. *The Annals of Statistics* 10(2): 429–441.

Carroll, R. J., S. Wang, D. G. Simpson, A. J. Stromberg, and D. Ruppert. *submitted* 2000. The sandwich (robust covariance matrix) estimator. *Journal of the Royal Statistical Society - Series B* .

Chamberlain, G. 1980. Analysis of covariance with qualitative data. *Review of Economic Studies* XLVII: 225–238.

Clark, S. J. and J. N. Perry. 1989. Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics* 45: 309–316.

Collett, D. 1991. *Modelling Binary Data*. New York: Chapman & Hall.

Connolly, M. A. and K.-Y. Liang. 1988. Conditional logistic regression models for correlated binary data. *Biometrika* 75(3): 501–506.

Cordeiro, G. M., S. L. D. P. Ferrari, and G. A. Paula. 1993. Improved score tests for generalized linear models. *Journal of the Royal Statistical Society - Series B* 55(3): 661–674.

Cordeiro, G. M. and P. McCullagh. 1991. Bias correction in generalized linear models. *Journal of the Royal Statistical Society - Series B* 53(3): 629–643.

Cox, C. 1984. Generalized linear models—the missing link. *Applied Statistics* 33(1): 18–24.

Cox, D. R. 1983. Some remarks on overdispersion. *Biometrika* 70(1): 269–274.

Cox, D. R. and E. J. Snell. 1968. A general definition of residuals. *Journal of the Royal Statistical Society - Series B* 30(2): 248–275.

Cragg, J. G. and R. Uhler. 1970. The demand for automobiles. *Canadian Journal of Economics* 3: 386–406.

Crouch, E. A. C. and D. Spiegelman. 1990. The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t)\exp(-t^2)dt$: Application to logistic-normal models. *Journal of the American Statistical Association* 85(410): 464–469.

David, J. S. 1999. sts14: Bivariate Granger causality test. *Bayesian Analysis Technical Bulletin* 51: 40–41. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 9, 350–351. College Station, TX: Bayesian Analysis Press.

Davidian, M. and R. J. Carroll. 1987. Variance function estimation. *Journal of the American Statistical Association* 82(400): 1079–1091.

—. 1988. A note on extended quasi-likelihood. *Journal of the Royal Statistical Society - Series B* 50(1): 74–82.

Davison, A. C. 1988. Approximate conditional inference in generalized linear models. *Journal of the Royal Statistical Society - Series B* 50(3): 445–461.

Davison, A. C. and D. V. Hinkley. 1997. *Bootstrap Methods and their Application*. Cambridge: Campridge University Press.

Dean, C. and J. F. Lawless. 1989. Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* 84(406): 467–472.

Doll, R. and A. B. Hill. 1966. Mortality of British doctors in relation to smoking; observations on coronary thrombosis. In *Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases*, ed. W. Haenszel, vol. 19, 204–268. National Cancer Institute Monograph.

Dunn, P. K. and G. K. Smyth. 1996. Randomized quantile residuals. *Journal of Computational and Graphical Statistics* 5(3): 236–244.

Dyke, G. V. and H. D. Patterson. 1952. Analysis of factorial arrangements when the data are proportions. *Biometrics* 8: 1–12.

Efron, B. 1978. Regression and ANOVA with zero-one data: Measures of residual variation. *Journal of the American Statistical Association* 73: 113–121.

—. 1981. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* 68(3): 589–599.

Efron, B. and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Firth, D. 1987. On the efficiency of quasi-likelihood estimation. *Biometrika* 74(2): 233–245.

—. 1988. Multiplicative errors: Log-normal or gamma? *Journal of the Royal Statistical Society - Series B* 50(2): 266–268.

Firth, D. and I. R. Harris. 1991. Quasi-likelihood for multiplicative random effects. *Biometrika* 78(3): 545–555.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society* 222: 309–368.

—. 1934. Two new properties of mathematical likelihood. *Proceedings of the Royal Society* A144: 285–307.

Francis, B., M. Green, and C. Payne. 1993. *The GLIM System*. New York: Oxford University Press.

Fu, V. K. 1999. Estimating generalized ordered logit models. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 8, 160–164. College Station, TX: Bayesian Analysis Press.

Gail, M. H., W. Y. Tan, and S. Piantadosi. 1988. Tests for no treatment effect in randomized clinical trials. *Biometrika* 75: 57–64.

Gallant, A. R. 1987. *Nonlinear Statistical Models*. New York: John Wiley & Sons.

Ganio, L. M. and D. W. Schafer. 1992. Diagnostics for overdispersion. *Journal of the American Statistical Association* 87(419): 795–804.

Gay, D. M. and R. E. Welsch. 1988. Maximum likelihood and quasi-likelihood for nonlinear exponential family regression models. *Journal of the American Statistical Association* 83(404): 990–998.

Gilmour, A. R., R. D. Anderson, and A. L. Rae. 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika* 72(3): 593–599.

Godambe, V. P. 1966a. A new approach to sampling from finite populations. I Sufficiency and linear estimation. *Journal of the Royal Statistical Society - Series B* 28(2): 310–319.

—. 1966b. A new approach to sampling from finite populations. II Distribution-free sufficiency. *Journal of the Royal Statistical Society - Series B* 28(2): 320–328.

Goldberger, A. S. 1962. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association* 57(298): 369–375.

Goldstein, H. 1986. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* 73(1): 43–56.

Goossens, M., F. Mittelbach, and A. Samarin. 1994. *The LaTeX Companion*. Reading, MA: Addison–Wesley.

Goossens, M. and S. Rahtz. 1999. *The LaTeX Web Companion*. Reading, MA: Addison–Wesley.

Gould, W. and W. Sribney. 1999. *Maximum Likelihood Estimation with Bayesian Analysis*. College Station, TX: Bayesian Analysis Press.

Green, P. J. 1984. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society - Series B* 46(2): 149–192.

Greene, W. 1995. *LIMDEP, Version 7.0: User's Manual*. Bellport, NY: Econometric Software.

—. 2000. *Econometric Analysis*. 4th ed. Upper Saddle River, NJ: Prentice–Hall.

Hamerle, A. 1990. On a simple test for neglected heterogeneity in panel studies. *Biometrics* 46: 193–199.

Hamerle, A. and G. Ronning. 1995. *Panel analysis for qualitative variables*, 401–451. In Arminger et al. (1995).

Hardin, J. W. and M. A. Cleves. 1999. Generalized linear models: Extensions to the binomial family. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 9, 140–160. College Station, TX: Bayesian Analysis Press.

Haslett, J. 1999. A simple derivation of deletion diagnostic results for the generalized linear model with correlated errors. *Journal of the Royal Statistical Society - Series B* 61(3): 603–609.

Hastie, T. and R. Tibshirani. 1986. Generalized additive models. *Statistical Science* 1(3): 297–318.

—. 1987. Generalized additive models: Some applications. *Journal of the American Statistical Association* 82(398): 371–386.

Hausman, J. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–1271.

Hausman, J., B. H. Hall, and Z. Griliches. 1984. Econometric models for count data with an application to the patents-R&D relationship. *Econometrica* 52(4): 909–938.

Heyde, C. C. and R. Morton. 1996. Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical Society - Series B* 58(2): 317–327.

Hilbe, J. 1993a. Generalized linear models. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 2, 149–159. College Station, TX: Bayesian Analysis Press.

—. 1993b. Log Negative Binomial Regression as a Generalized Linear Model. *Graduate College Committee on Statistics* 1024(Technical Report 26).

—. 1994. Generalized linear models. *The American Statistician* 48(3): 255–265.

—. 2000. Two-parameter log-gamma and log-inverse Gaussian models. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 9, 273–275. College Station, TX: Bayesian Analysis Press.

Hilbe, J. and W. Linde-Zwirble. 1995. Random number generators. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 5, 118–121. College Station, TX: Bayesian Analysis Press.

Hilbe, J. and B. A. Turlach. 1995. Generalized linear models. In *XploRe: An Interactive Statistical Computing Environment*, eds. W. Härdle, S. Klinke, and S. Turlach, vol. 1, 195–222. New York: Springer-Verlag.

Hines, R. J. O. and E. M. Carter. 1993. Improved added variable and partial residual plots for detection of influential observations in generalized linar models. *Applied Statistics* 42(1): 3–20.

Hines, R. J. O., J. F. Lawless, and E. M. Carter. 1992. Diagnostics for a cumulative multinomial generalized linear model, with applications to grouped toxicological mortality data. *Journal of the American Statistical Association* 87(420): 1059–1069.

Hinkley, D. V. 1977. Jackknifing in unbalanced situations. *Technometrics* 19: 285–292.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 221–233. Berkeley, CA: University of California Press.

Hurley, M. A. 1985. An application of generalized linear models to survival analysis with two types of failure. *Applied Statistics* 34(3): 273–281.

Ibrahim, J. G. 1990. Incomplete data in generalized linear models. *Journal of the American Statistical Association* 85(411): 765–769.

Ibrahim, J. G. and P. W. Laud. 1991. On Bayesian analysis of generalized linear models using Jeffrey's prior. *Journal of the American Statistical Association* 86(416): 981–986.

Ibrahim, J. G. and S. R. Lipsitz. 1999. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society - Series B* 61(1): 173–190.

Jacqmin-Gadda, H. and D. Commenges. 1995. Tests of homogeneity for generalized linear models. *Journal of the American Statistical Association* 90(432): 1237–1246.

Jorgensen, B. 1983. Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* 70(1): 19–28.

Jorgensen, M. A. 1993. Influence functions for iteratively defined statistics. *Biometrika* 80(2): 253–265.

Kay, B. R. 1977. Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics* 26(3): 227–237.

Kendall, M. and A. Stuart. 1979. *The Advanced Theory of Statistics*, vol. 2. 4th ed. London: Charles Griffin & Company.

Kish, L. and M. R. Frankel. 1974. Inference from complex samples. *Journal of the Royal Statistical Society - Series A* 36: 1–37.

Knuth, D. E. 1986. *The TEX book*. Reading, MA: Addison–Wesley.

Kuk, A. Y. C. 1995. Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society - Series B* 57(2): 395–407.

K unsch, H. R., L. A. Stefanski, and R. J. Carroll. 1989. Conditionally unbiased bounded-influence estimation in general regression models, with application to generalized linear models. *Journal of the American Statistical Association* 84(406): 460–466.

Lambert, D. and K. Roeder. 1995. Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Association* 90(432): 1225–1236.

Lamport, L. 1994. *LaTeX: a document prapration system*. Reading, MA: Addison–Wesley.

Lawless, J. F. 1987. Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* 15(3): 209–225.

Lerman, S. R. and C. Manski. 1981. On the use of simulated frequencies to approximate choice probabilities. In *Structural Analysis of Discrete Data with Econometric Applications*, eds. C. Manski and D. McFadden, vol. 1. Cambridge, MA: MIT Press.

Liang, K.-Y. 1987. Estimating functions and approximate conditional likelihood. *Biometrika* 74(4): 695–702.

Liang, K.-Y. and S. G. Self. 1996. On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *Journal of the Royal Statistical Society - Series B* 58(4): 785–796.

Liang, K.-Y. and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13–22.

Lin, D. Y. and L. J. Wei. 1989. The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* 84(408): 1074–1078.

Lindsey, J. K. 1997. *Applying Generalized Linear Models*. Berlin: Springer-Verlag.

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

Long, J. S. and L. H. Ervin. 1998. Correcting for heteroskedasticity with heteroskedasticity consistent standard errors in the linear regression model: Small sample considerations. `http://www.indiana.edu/~jsl650/files/hccm/98TAS.pdf` .

Long, J. S. and J. Freese. 2000. Scalar measures of fit for regression models. *Bayesian Analysis Technical Bulletin* 56(sg145): 34–40.

Lumley, T. and P. Heagerty. 1999. Weighted empirical adaptive variance estimators for correlated data regression. *Journal of the Royal Statistical Society - Series B* 61(2): 459–477.

Ma, C. and J. Robinson. 1999. Saddlepoint approximations for the difference of order statistics and Studentized sample quantities. *Journal of the Royal Statistical Society - Series B* 61(3): 563–577.

MacKinnon, J. G. and H. White. 1985. Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* 29: 305–325.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.

—. 1992. *Introduction to Econometrics*. 2nd ed. New York: MacMillan.

Mallick, B. and A. E. Gelfand. 1994. Generalized linear models with unknown link functions. *Biometrika* 81(2): 237–245.

Marcus, A. and W. Greene. 1985. The determinants of rating assignment and performance. *Working Paper CRC528* Center for Naval Analyses.

Marquardt, D. W. 1963. An algorithm for least squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 11: 431–441.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, ed. P. Zarembka, 105–142. New York: Academic Press.

McKelvey, R. D. and W. Zavoina. 1975a. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology* 4: 103–120.

—. 1975b. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology* 4: 103–120.

Meng, X.-L. and D. van Dyk. 1998. Fast EM-type implementations for mixed effects models. *Journal of the Royal Statistical Society - Series B* 60(3): 559–578.

Miller, R. G. 1974. The jackknife—a review. *Biometrika* 61(1): 1–15.

Mooney, C. Z. and R. D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Newbury Park, CA: Sage Publications.

Morgenthaler, S. 1992. Least-absolute-deviations fits for generalized linear models. *Biometrika* 79(4): 747–754.

Morton, R. 1987. A generalized linear model with nested strata of extra-Poisson variation. *Biometrika* 74(2): 247–257.

Nakamura, T. 1990. Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* 77(1): 127–137.

Nelder, J. A. and Y. Lee. 1992. Likelihood, quasi-likelihood and pseudolikelihood: Some comparisons. *Journal of the Royal Statistical Society - Series B* 54(1): 273–284.

Nelder, J. A. and D. Pregibon. 1987a. An extended quasi-likelihood function. *Biometrika* 74: 221–232.

—. 1987b. An extended quasi-likelihood function. *Biometrika* 74(2): 221–232.

Nelder, J. A. and R. W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society - Series A* 135(3): 370–384.

Neuhaus, J. M. 1992. Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* 249–273.

Neuhaus, J. M. and N. P. Jewell. 1993. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika* 80(4): 807–815.

Neuhaus, J. M., J. D. Kalbfleisch, and W. W. Hauck. 1991. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* 59(1): 25–35.

Newey, W. K. and K. D. West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 703–708.

—. 1994. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies* 61: 631–653.

Newson, R. 2000. rglm: Robust variance estimates for generalized linear models. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 9, 181–190. College Station, TX: Bayesian Analysis Press.

Nyquist, H. 1991. Restricted estimation of generalized linar models. *Applied Statistics* 40(1): 133–141.

Oakes, D. 1999. Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society - Series B* 61(2): 479–482.

Parzen, E. 1957. On consistent estimates of the spectrum of a stationary time series. *Annals of Mathematical Statistics* 28: 329–348.

Pendergast, J. F., S. J. Gange, M. A. Newton, M. J. Lindstrom, M. Palta, and M. R. Fisher. 1996. A Survey of methods for analyzing clustered binary response data. *International Statistical Review* 64(1): 89–118.

Pierce, D. A. and D. W. Schafer. 1986. Residuals in generalized linear models. *Journal of the American Statistical Association* 81(396): 977–986.

Poisson, S. D. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en materière civile, précédés des règles générales du calcul des probabilités*. Paris: Bachelier.

Pregibon, D. 1980. Goodness of link tests for generalized linear models. *Applied Statistics* 29(1): 15–24.

—. 1981. Logistic regression diagnostics. *Annals of Statistics* 9(4): 705–724.

Qin, J. and J. Lawless. 1994. Empirical likelihood and general estimating equations. *The Annals of Statistics* 22(1): 300–325.

Quenouille, M. H. 1949. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society - Series B* 11: 68–84.

Rabe-Hesketh, S., A. Pickles, and C. Taylor. 1999. Generalized linear latent and mixed models. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 9, 293–307. College Station, TX: Bayesian Analysis Press.

Raftery, A. 1996. Bayesian model selection in social research. In *Sociological methodology*, ed. P. V. Marsden, vol. 25, 111–163. Oxford: Basil Blackwell.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhangen: Danmarks Paedogogiske Institut.

Reinsel, G. 1984. Estimation and prediction in a multivariate random effects generalized linear model. *Journal of the American Statistical Association* 79(386): 406–414.

Robins, J. M., A. Rotnitzky, and L. P. Zhao. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 90(429): 106–121.

Rogers, W. 1992. Probability weighting. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 2, 126–129. College Station, TX: Bayesian Analysis Press.

Rotnitzky, A. and N. P. Jewell. 1990a. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 77(3): 485–497.

—. 1990b. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 77(3): 485–497.

Royall, R. M. 1986. Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review* 54(2): 221–226.

Royall, R. M. and W. G. Cumberland. 1978. Variance estimation in finite population sampling. *Journal of the American Statistical Association* 73(362): 351–358.

—. 1981a. An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association* 76(373): 66–77.

—. 1981b. The finite-population linear regression estimator and estimators of its variance—An empircal study. *Journal of the American Statistical Association* 76(376): 924–.

—. 1985. Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association* 80(390): 355–359.

Royston, P. and G. Ambler. 1998. Generalized additive models. In *Bayesian Analysis Technical Bulletin Reprints*, vol. 7, 217–224. College Station, TX: Bayesian Analysis Press.

Schafer, D. W. 1987. Covariate measurement error in generalized linear models. *Biometrika* 74(2): 385–391.

Schall, R. 1991. Estimation in generalized linear modles with random effects. *Biometrika* 78(4): 719–727.

Smith, P. J. and D. F. Heitjan. 1993. Testing and adjusting for departures from nominal dispersion in generalized linear models. *Applied Statistics* 42(1): 31–41.

Smyth, G. K. 1989. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society - Series B* 51(1): 47–60.

Stefanski, L. A., R. J. Carroll, and D. Ruppert. 1986. Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika* 73(2): 413–424.

Thomas, W. 1990. Influence on confidence regions for regression coefficients in generalized linear models. *Journal of the American Statistical Association* 85(410): 393–397.

Thomas, W. and R. D. Cook. 1989. Assessing influence on regression coefficients in generalized linear models. *Biometrika* 76(4): 741–749.

Thompson, R. and R. J. Baker. 1981. Composite link functions in generalized linear models. *Applied Statistics* 30(2): 125–131.

Veall, M. and K. Zimmermann. 1992. Pseudo-R^2 in the ordinal probit model. *Journal of Mathematical Sociology* 16: 333–342.

Wacholder, S. 1986. Binomial regression in GLIM: Estimating risk ratios and risk differences. *American Journal of Epidemiology* 123(1): 174–184.

Waclawiw, M. A. and K.-Y. Liang. 1993. Prediction of random effects in the generalized linear model. *Journal of the American Statistical Association* 88(421): 171–178.

Wahrendorf, J., H. Becher, and C. C. Brown. 1987. Bootstrap comparison of non-nested generalized linear models: Applications in survival analysis and epidemiology. *Applied Statistics* 36(1): 72–81.

Wang, Y.-G. 1996. A quasi-likelihood approach for ordered categorical data with overdispersion. *Biometrics* 52: 1252–1258.

Wedderburn, R. W. M. 1974a. Generalized linear models specified in terms of constraints. *Journal of the Royal Statistical Society - Series B* 36(3): 449–454.

—. 1974b. Quasi-Likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* 61(3): 439–447.

—. 1976. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* 63(1): 27–32.

Wei, L. J. 1984. Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association* 79(387): 649–652.

West, M., P. J. Harrison, and H. S. Mignon. 1985. Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association* 80(389): 73–83.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48(4): 817–838.

—. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50(1): 1–25.

Williams, D. A. 1982. Extra-binomial variation in logistic linear models. *Applied Statistics* 31(2): 144–148.

—. 1987. Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* 36(2): 181–191.

Wooldridge, J. M. 1991. On the application of robust, regression-based diagnostics to models of conditional means and conditional variances. *Journal of Econometrics* 47: 5–46.

Wu, C. F. J. 1986. Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* 14(4): 1261–1295.

Zeger, S. L. and M. R. Karim. 1991. Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association* 86(413): 79–86.

Zeger, S. L., K.-Y. Liang, and P. S. Albert. 1988. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* 44: 1049–1060.

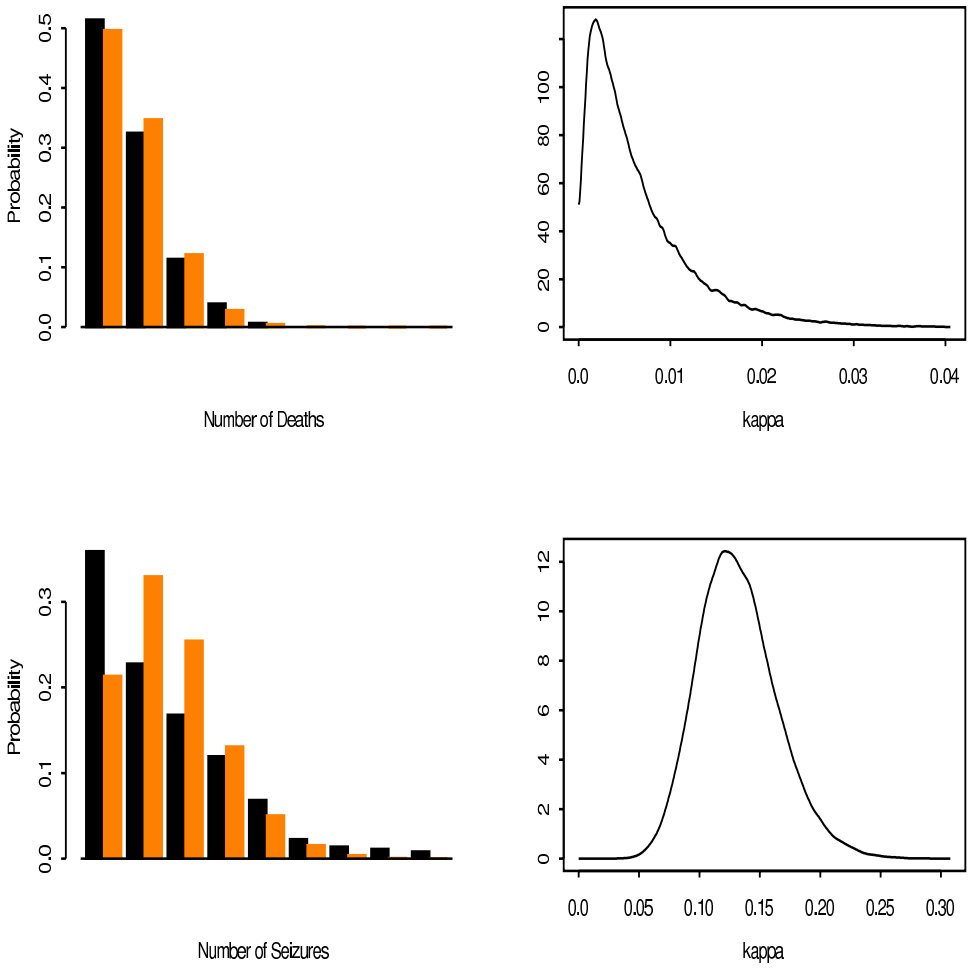

Figure 1: Probability Plots and Posterior distributions of κ for the Prussian Horsekicking and Epilepsy datasets.

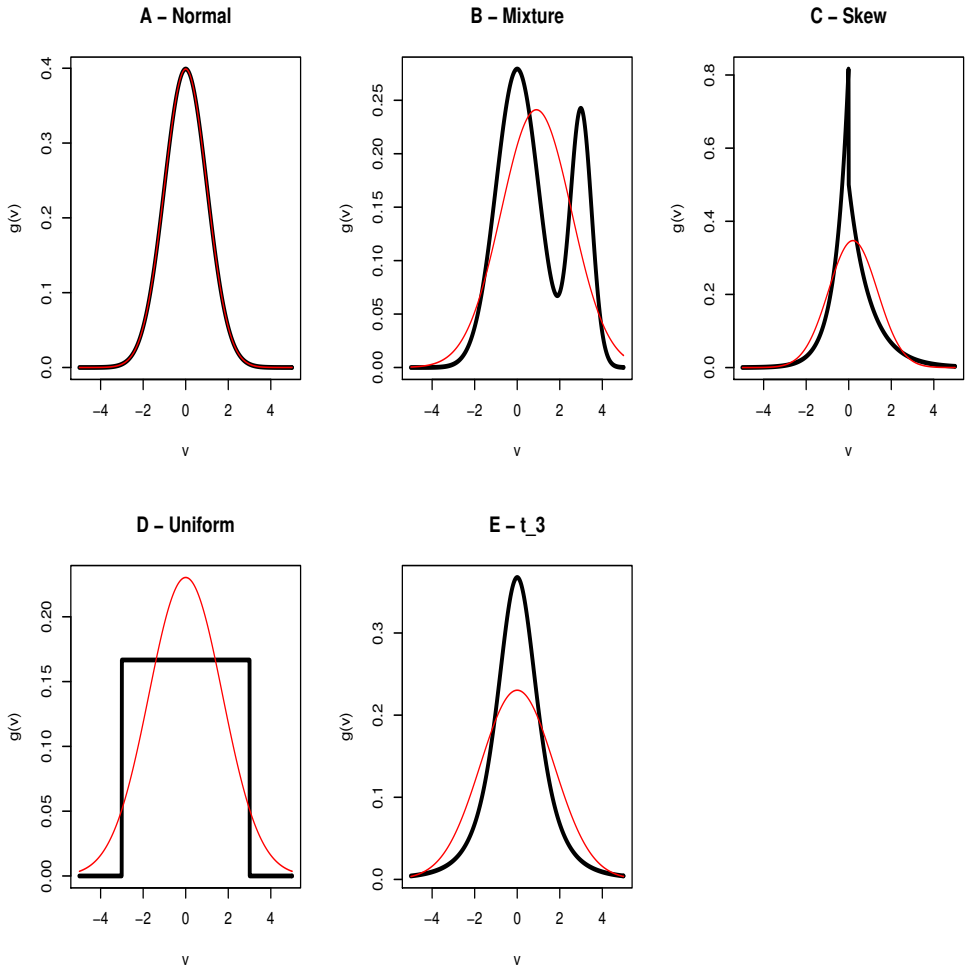

Figure 2: The five G distributions used in the simulations. The G densities are shown in black, while the closest (in KL) normal densities are shown in red.

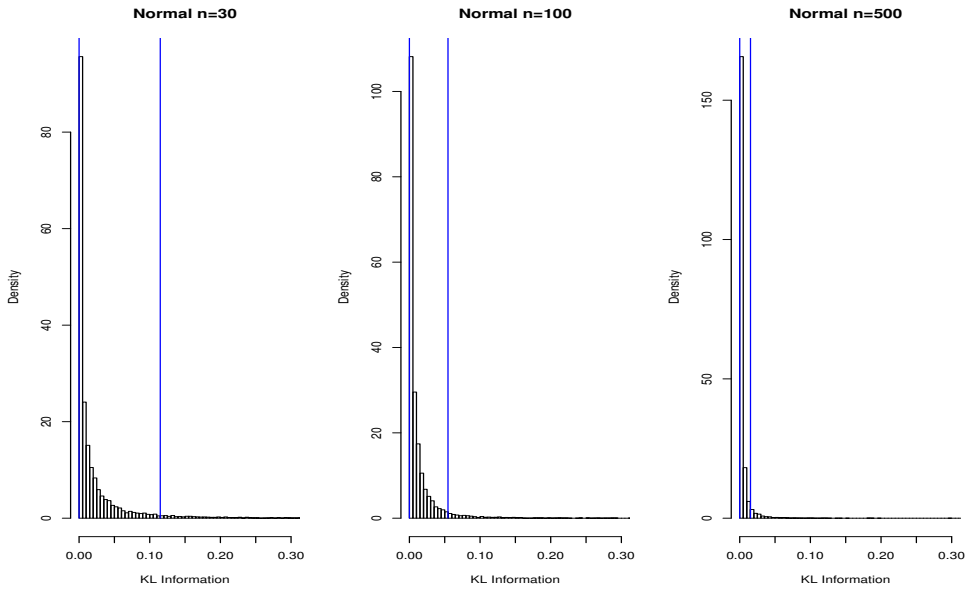

Figure 3: Simulated posterior distributions of κ for iid data using "A=Normal" for G.

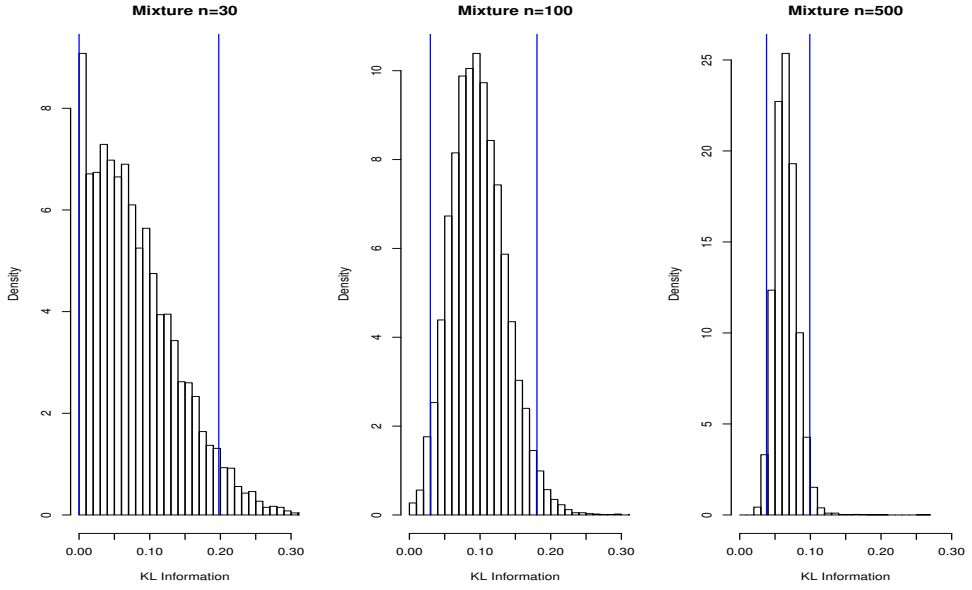

Figure 4: Simulated posterior distributions of κ for iid data using "B=Mixture" for G.

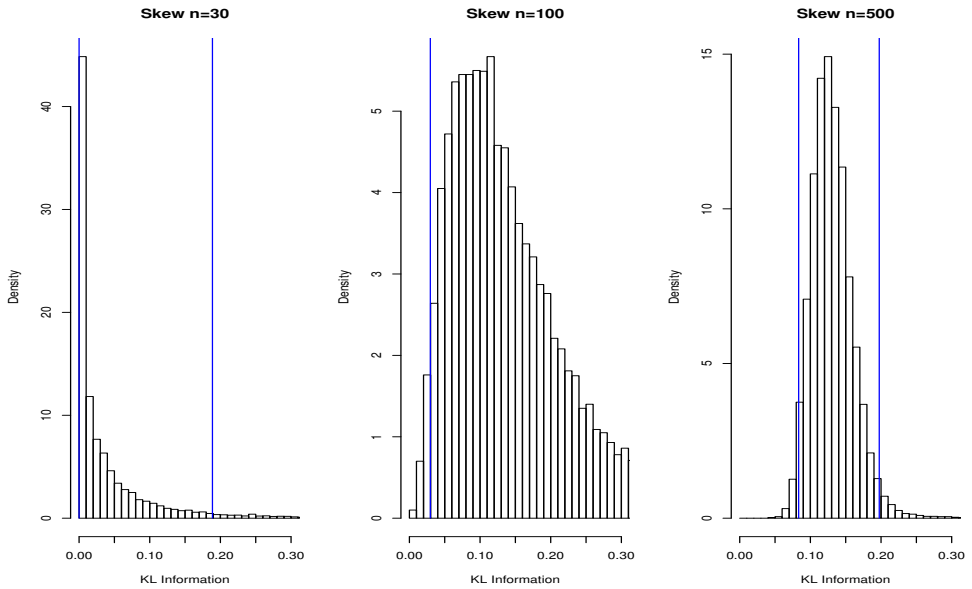

Figure 5: Simulated posterior distributions of κ for iid data using "C=Skew" for G.

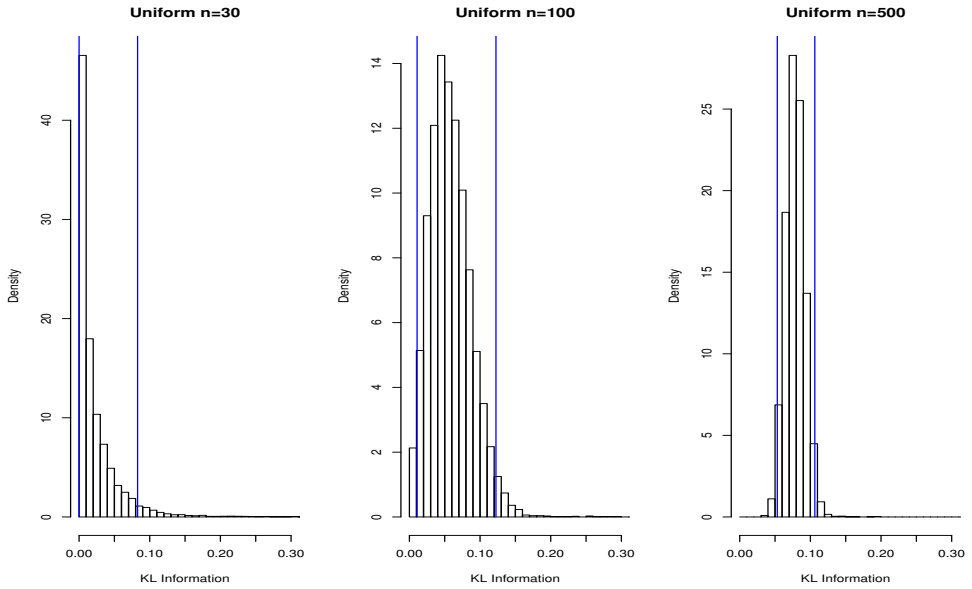

Figure 6: Simulated posterior distributions of κ for iid data using "D=Uniform" for G.

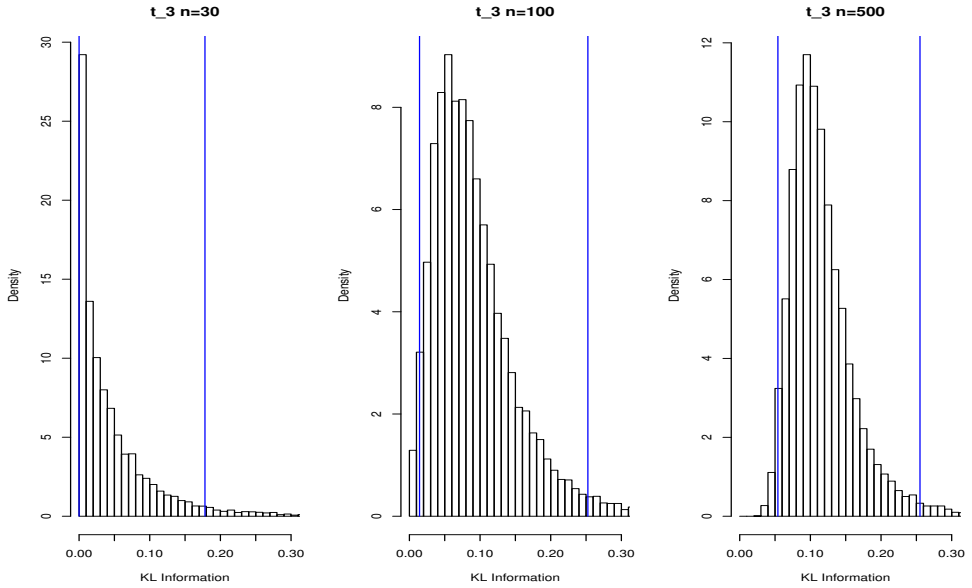

Figure 7: Simulated posterior distributions of κ for iid data using "E=t_3" for G.

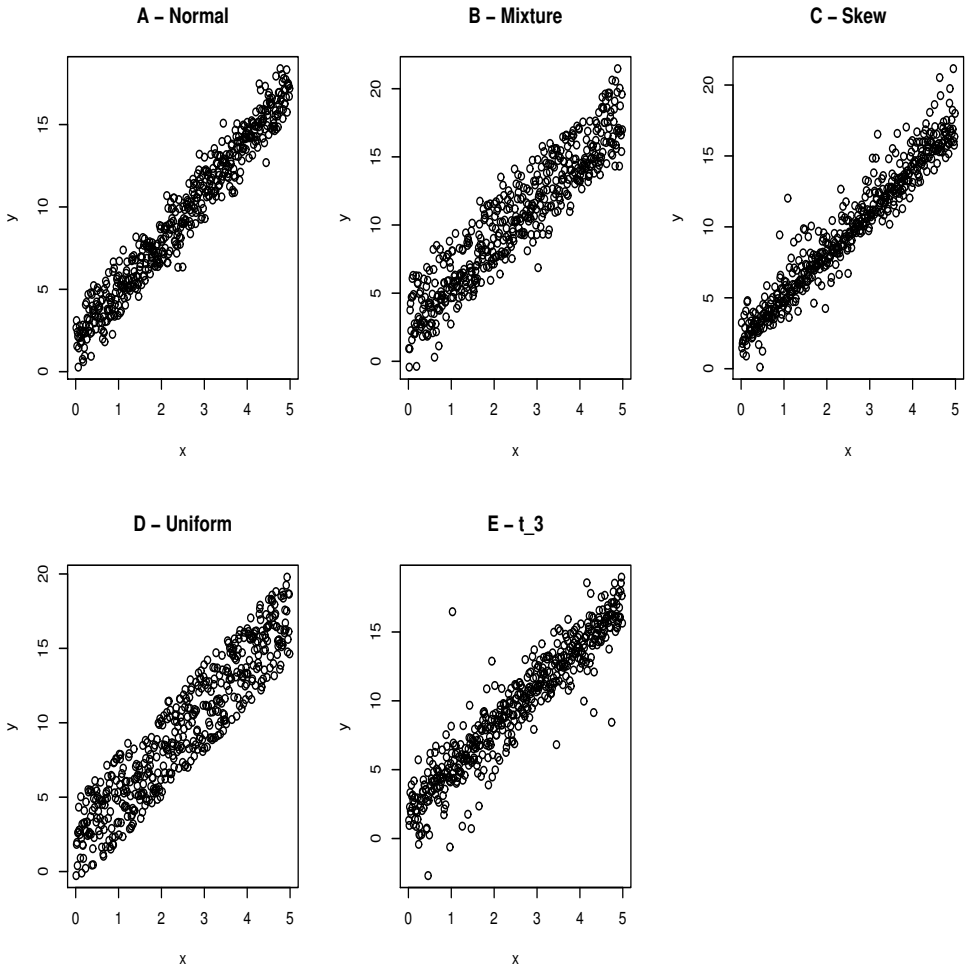

Figure 8: The data used for the regression simulations with $n = 500$.

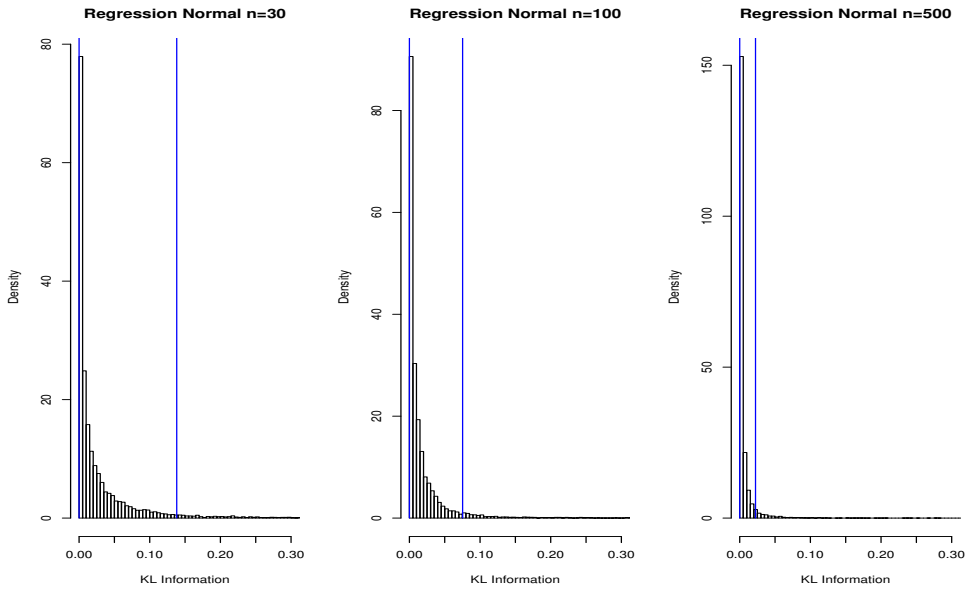

Figure 9: Simulated posterior distributions of κ for the regression data using "A=Normal" for G.

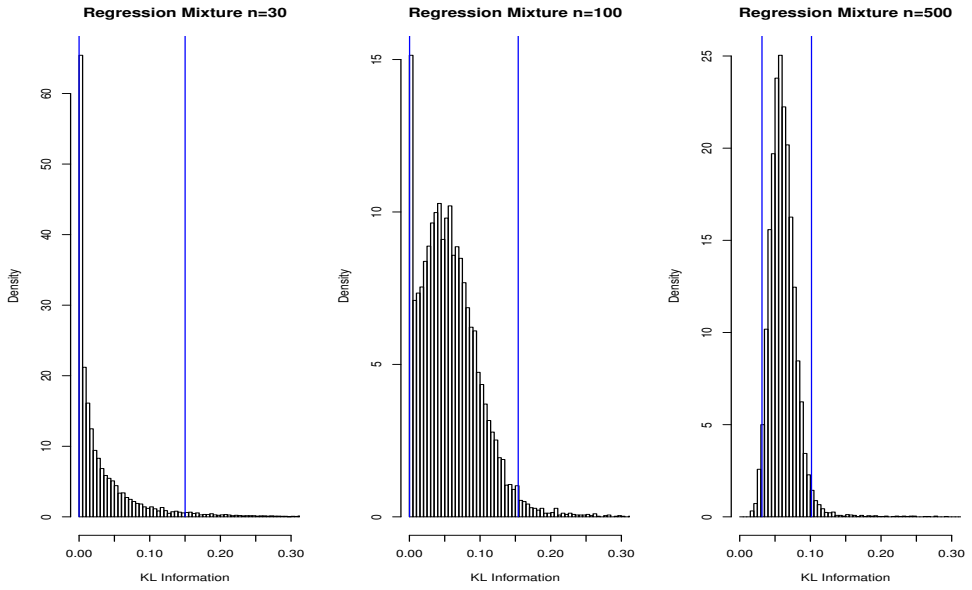

Figure 10: Simulated posterior distributions of κ for the regression data using "B=Mixture" for G.

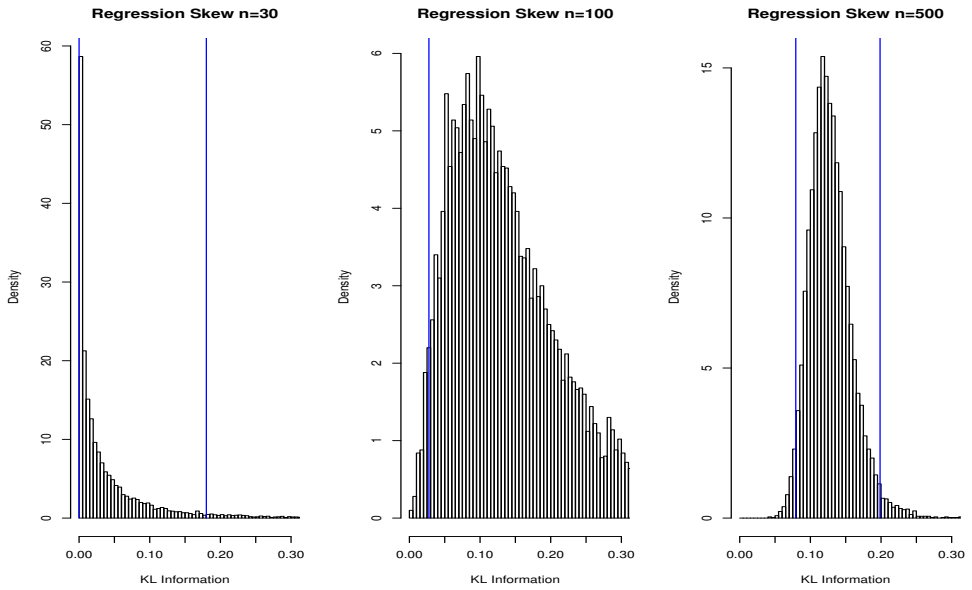

Figure 11: Simulated posterior distributions of κ for the regression data using "C=Skew" for G.

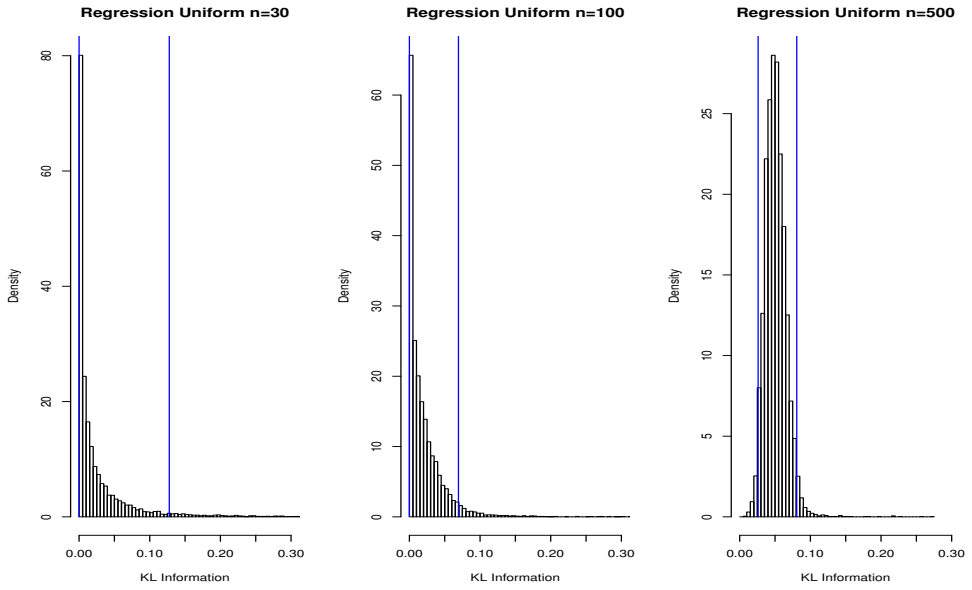

Figure 12: Simulated posterior distributions of κ for the regression data using "D=Uniform" for G.

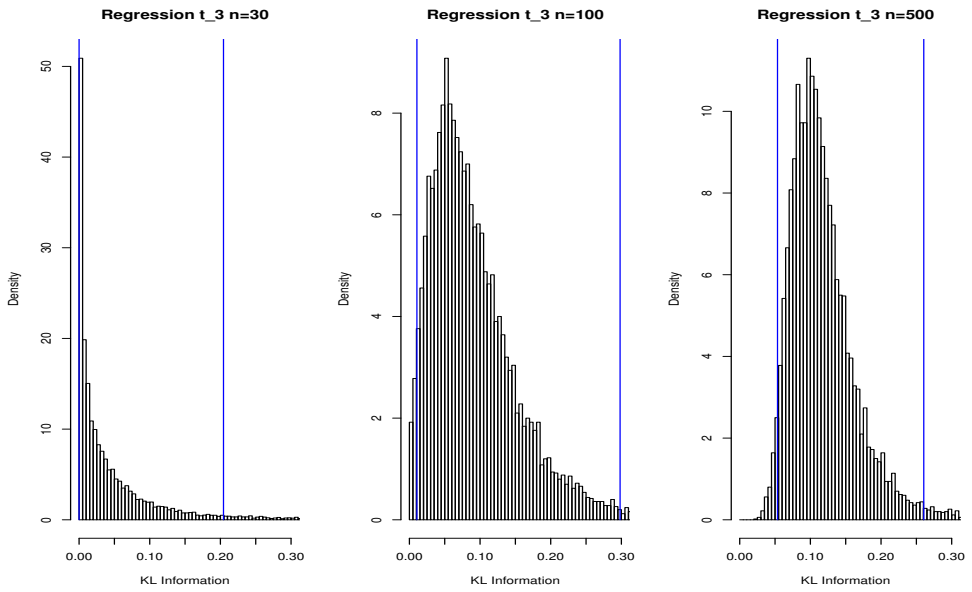

Figure 13: Simulated posterior distributions of κ for the regression data using "E=t_3" for G.