Published in final edited form as:

Stat Modelling. 2013 December ; 13(5-6): 409–429. doi:10.1177/1471082X13494610.

Applications of a Kullback-Leibler Divergence for Comparing Non-nested Models

Chen-Pin Wang and

Department of Epidemiology and Biostatistics, University of Texas Health Science Center, San Antonio TX, 78229, USA

Booil Jo

Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford CA 94305, USA

Chen-Pin Wang: wangc3@uthscsa.edu; Booil Jo: booil@stanford.edu

Abstract

Wang and Ghosh (2011) proposed a Kullback-Leibler divergence (KLD) which is asymptotically equivalent to the KLD by Goutis and Robert (1998) when the reference model (in comparison with a competing fitted model) is correctly specified and when certain regularity conditions hold true. While properties of the KLD by Wang and Ghosh (2011) have been investigated in the Bayesian framework, this paper further explores the property of this KLD in the frequentist framework using four application examples, each fitted by two competing non-nested models.

Keywords

Kullback-Leibler Divergence; Information Criterion; Comparison of Non-nested Models

1 Introduction

The Kullback-Leibler divergence (KLD) is perhaps the most commonly used information criterion for assessing model discrepancy (Shannon, 1948; Kullback and Leibler, 1951; Lindley, 1956; Akaike, 1974; Schwarz, 1978; Bernardo, 1979). A KLD is the expectation of the logarithm of the ratio of the probability density functions (p.d.f.s) of two models, one being a fitted model and the other being the reference model, where the expectation is taken with respect to the reference model. Thus a KLD can be viewed as a measure of the information loss in the fitted model relative to that in the reference model.

KLDs that are suitable for model comparison in the Bayesian framework typically involve the integrated likelihoods of the competing models, where the integrated likelihood under each model is obtained by integrating the likelihood with respect to the prior distribution of model parameters (e.g., Lindley, 1956 and Schwarz, 1978). KLDs based on the ratio of integrated likelihoods however have been challenged by identifying priors that are compatible under the competing models and that produce proper integrated likelihoods. Goutis and Robert (1998) proposed a Kullback-Leibler projection (henceforth G-R KLD)

that overcame the challenges associated with prior elicitation in calculating KLD under the Bayesian framework. More specifically, the G-R KLD is the infimum KLD between the likelihood under the reference model and all possible likelihoods arising from the competing fitted model. When the reference model is correctly specified, the G-R KLD is asymptotically equivalent to the KLD between the reference model and the competing fitted model evaluated at its MLE (ref. Akaike, 1974). The Bayesian estimate of G-R KLD is obtained by integrating the G-R KLD with respect to the posterior distribution of model parameters under the reference model. Thus the Bayesian estimate of G-R KLD is clearly not subject to the drawback due to impropriety of the prior as long as the posterior under the reference model is proper. In addition, G-R KLD is suitable for comparing the predictivity of the competing models since it is calculated with respect to the posterior density of model parameters under the reference model. However, the G-R KLD was originally developed for comparing nested generalized linear models while assuming a known true model, and its extension to general model comparison (e.g., in the frequentist framework) remains limited.

Stemming from the G-R KLD, Wang and Ghosh (2011) proposed a more tractable model discrepancy measure, the Goutis-Robert-Akaike KLD (or G-R-A KLD) – the KLD between the reference model and the competing fitted model evaluated at its maximum likelihood estimate or MLE (Wang and Ghosh, 2011). The G-R-A KLD is equivalent to G-R KLD when the reference model is correctly specified. Unlike G-R KLD, the G-R-A KLD does not require the reference model r to be the true model, nor is the true model to be specified. Also, while G-R KLD has been limited to comparing nested generalized linear models, the G-R-A KLD is more flexible for comparing nested or non-nested models that are broader than generalized linear models. There are two appealing asymptotic properties associated with G-R-A KLD (Wang and Ghosh, 2011). First, under certain regularity conditions, the asymptotic expression of the posterior estimator of G-R-A KLD associated with a statistic T_n (pertinent to the inference or model diagnostics) fitted by two competing models is comprised of a leading term for model discrepancy in the mean of T_n , a term for model discrepancy in the variance of T_n , and a constant to penalize model complexity, plus a smaller order term. The leading terms in the G-R-A KLD estimates are functions of MLE's, which makes it also appropriate for model comparison in the frequentist framework. Second, the G-R-A KLD estimator is intimately related to Bayesian model discrepancy measures based on predictive statistics (Guttman, 1967; Rubin, 1984, Gelman et al., 1996), such as a one-sided weighted posterior predictive p-value (WPPP). The WPPP of T_n evaluates the predictive distribution of T_n under an assumed model f at $T_n^{pred,r}$, where $T_n^{pred,r}$ denotes the prediction of T_n under the reference model r , the posteriors are derived under r , and the weight is used to account for the variation of T_n under r .

While properties of G-R-A KLD have been investigated in the Bayesian framework (Wang and Ghosh, 2011), the goal of this paper is to further explore the utility of G-R-A KLD in the frequentist framework through applying G-R-A KLD to comparing non-nested models in four real application examples. The paper is organized as follows. Section 2 describes the G-R-A KLD and its associated asymptotic properties. Section 3 demonstrates model comparison using G-R-A KLD in four examples.

2 A Proposed Kullback-Leibler Divergence

2.1 Notations

Throughout this paper, we assume that X_i 's are i.i.d. with some common c.d.f. g and parameter(s) θ for $\theta \in \Theta_g$, where Θ_g is a closed set. Denote r for the reference model and f for the fitted model, both governed by θ , where Θ_r and Θ_f are the corresponding parameter spaces. Also, a capital letter is used to denote for random variables (or estimators), and the corresponding lower case denotes its realization (or an estimate). Let $U_n = U(X_1, \dots, X_n)$ be the base for deriving the posterior density of θ under model r . We shall denote the p.d.f.'s of $*$ given θ under model r and f by $r(*|\theta)$ and $f(*|\theta)$, respectively. Also, let $\phi(\mathbf{y}) = (2\pi)^{-l/2} \exp(-\mathbf{y}'\mathbf{y}/2)$ where $\mathbf{y}$ is l -dimensional.

2.2 Define Goutis-Robert-Akaike Kullback-Leibler Divergence

Suppose that model adequacy is evaluated based on its fit for certain statistics $T_n = T(X_1, \dots, X_n)$ that is pertinent to the inference or model diagnostics. Following Wang and Ghosh (2011), we assess the relative fit between models using the KLD of the distribution of T_n under r and f evaluated at $\hat{\theta}_f$, the MLE of θ .

$$\int \log \left(\frac{r(t_n|\theta)}{f(t_n|\hat{\theta}_f)} \right) r(t_n|\theta) dt_n. \quad (2.1)$$

Wang and Ghosh (2011) refer to (2.1) as the Goutis-Robert-Akaike KLD or G-R-A KLD since (2.1) is asymptotically equivalent to the KLD proposed by Goutis and Robert (1998) when the reference model r is the true model (ref. Akaike, 1974). In general, for each $\theta \in \Theta_r$, G-R-A KLD given in (2.1) can be regarded as a measure of the minimal information gain regarding T_n in model r from model f since the minimum information loss under f is achieved at $\hat{\theta}_f$. Note that unlike G-R KLD (Goutis and Robert, 1998), (2.1) does not require the reference model r to be the true model g .

2.3 Estimation of G-R-A KLD

Estimating (2.1) involves approximating the integral with respect to $r(t_n|\theta)$ and estimating unknown model parameters θ . To account for the uncertainty of model parameters, Wang and Ghosh (2011) considered the following estimator:

$$KLD_t(r, f|U_n) = \int \int \log \left(\frac{r(t_n|\theta)}{f(t_n|\hat{\theta}_f)} \right) r(t_n|\theta) \pi_r(\theta|U_n) dt_n d\theta, \quad (2.2)$$

where U_n , as a function of $(X_1, \dots, X_n)$, is used for deriving the posterior of θ , and $\pi_r(\theta|U_n) = r(U_n|\theta) \pi_r(\theta) / \int r(U_n|\theta) \pi_r(\theta) d\theta$ denoting the posterior density of θ under model r . That is, (2.2) is the average discrepancy between r and f , each being weighted by $\pi_r(\theta|U_n)$. Since (2.2) is nonnegative for any given U_n , the closer it is to zero, the less is the information loss by fitting f , instead of r , to the statistic T_n . The utility of (2.2) can be understood from its asymptotic properties described below.

In what follows, let the O statements be interpreted as “almost sure” statements. Also, define $Q(y) \equiv y - \log(y) - 1$ for $y > 0$ which has a unique minimum at $y = 1$. Denote θ for model parameters, and $\hat{\mu}_r(U_n)$, $\hat{\mu}_f(U_n)$, $\hat{\sigma}_r^2(U_n)$ and $\hat{\sigma}_f^2(U_n)$ for the posterior means (or the MLE's) of $\mu_r(\theta)$, $\mu_f(\theta)$, $\sigma_r^2(\theta)$, and $\sigma_f^2(\theta)$, respectively.

Throughout the paper, assume the following regularity conditions (A1)–(A5).

(A1) For each x , both $\log(r(x|\theta))$ and $\log(f(x|\theta))$ are 3 times continuously differentiable in θ . Further, there exist neighborhoods $N_r(\delta) = (\theta - \delta_r, \theta + \delta_r)$ and $N_f(\delta) = (\theta - \delta_f, \theta + \delta_f)$ of θ and integrable functions $H_{\theta, \delta_r}(x)$ and $H_{\theta, \delta_f}(x)$ such that

$$\sup_{\theta' \in N(\delta_r)} \left| \frac{\partial^k}{\partial \theta^k} \log(r(x|\theta)) \right|_{\theta=\theta'} \leq H_{\theta, \delta_r}(x)$$

and

$$\sup_{\theta' \in N(\delta_f)} \left| \frac{\partial^k}{\partial \theta^k} \log(f(x|\theta)) \right|_{\theta=\theta'} \leq H_{\theta, \delta_f}(x)$$

for $k = 1, 2, 3$.

(A2) For all sufficiently large $\lambda > 0$,

$$E_r \left(\sup_{|\theta' - \theta| > \lambda} \log \left(\frac{r(x|\theta')}{r(x|\theta)} \right) \right) < 0$$

and

$$E_f \left(\sup_{|\theta' - \theta| > \lambda} \log \left(\frac{f(x|\theta')}{f(x|\theta)} \right) \right) < 0.$$

(A3)

$$E_r \left(\sup_{\theta' \in (\theta - \delta, \theta + \delta)} \log(r(x|\theta')) \middle| \theta \right) \rightarrow E_r(\log(r(x|\theta)))$$

and

$$E_f \left(\sup_{\theta' \in (\theta-\delta, \theta+\delta)} \log(f(x|\theta')) | \theta \right) \rightarrow E_f (\log(f(x|\theta)))$$

as $\delta \rightarrow 0$.

(A4) The prior density $\pi(\theta)$ is continuously differentiable in a neighborhood of θ and $\pi(\theta) > 0$. Suppose that

(A5) $T_n = T(X_1, \dots, X_n)$ is asymptotically normally distributed under both models r and f with

$$r(T_n|\theta) = \sigma_r^{-1}(\theta) \phi(\sqrt{n}(T_n - \mu_r(\theta))/\sigma_r(\theta)) + O(n^{-1/2}) \quad (2.3)$$

and

$$f(T_n|\theta) = \sigma_f^{-1}(\theta) \phi(\sqrt{n}(T_n - \mu_f(\theta))/\sigma_f(\theta)) + O(n^{-1/2}). \quad (2.4)$$

Property 1.1—

$$2KLD_t(r, f|U_n) = \int \left\{ Q \left(\frac{\sigma_r^2(\theta)}{\hat{\sigma}_f^2(\theta)} \right) + \frac{n(\hat{\mu}_f(\theta) - \mu_r(\theta))^2}{\hat{\sigma}_f^2(\theta)} + O(n^{-1/2}) \right\} \pi_r(\theta|U_n) d\theta, \quad (2.5)$$

which is further reduced to

$$\frac{2KLD_t(r, f|U_n)}{n} - \frac{\{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\}^2}{\hat{\sigma}_f^2(U_n)} = o_p(1) \quad (2.6)$$

when $\mu_f(\theta) = \mu_r(\theta)$, and

$$2KLD_t(r, f|U_n) - Q \left(\frac{\hat{\sigma}_r^2(U_n)}{\hat{\sigma}_f^2(U_n)} \right) = o_p(1) \quad (2.7)$$

when $\mu_r(\theta) = \mu_f(\theta)$ but $\sigma_r^2(\theta) \neq \sigma_f^2(\theta)$.

Property 1.2. Suppose that—(A6) T_n is a p -dimensional statistic ($p > 1$ and p being independent of n) with

$$r(T_n|\theta) = |\sum_r(\theta)|^{-1/2} \phi \left(\sqrt{n} \sum_r^{-1/2}(\theta) (T_n - \mu_r(\theta)) \right) + O(n^{-1/2})$$

and

$$f(T_n|\theta) = |\sum_f(\theta)|^{-1/2} \phi \left(\sqrt{n} \sum_f^{-1/2}(\theta) (T_n - \mu_f(\theta)) \right) + O(n^{-1/2}).$$

Then

$$\frac{2KLD_t(r, f|U_n)}{n} - \{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\}' \sum_f^{-1}(U_n) \{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\} = o_p(1) \quad (2.8)$$

when $\mu_f(\theta) = \mu_r(\theta)$, and

$$2KLD_t(r, f|U_n) - \{\log(|\sum_r(U_n)|/|\sum_f(U_n)|) - \text{tr}(\sum_f^{-1}(U_n) \sum_r(U_n)) + p\} = o_p(1) \quad (2.9)$$

when $\mu_r(\theta) = \mu_f(\theta)$ and $\Sigma_r(\theta) = \Sigma_f(\theta)$.

Proofs of Properties 1.1 and 1.2 can be found in Wang and Ghosh (2011).

From the asymptotic results given in (2.6)–(2.9), it is clear that when n is sufficiently large, G-R-A KLD is appropriate for model comparison in the frequentist framework. This is because the leading terms of G-R-A KLD involve functions of MLE's. Specifically,

$\{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\}^2 / \hat{\sigma}_f^2(U_n)$ in (2.6), $Q(\hat{\sigma}_r^2(U_n) / \hat{\sigma}_f^2(U_n))$ in (2.7),

$\{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\}' \sum_f^{-1}(U_n) \{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\}$ in (2.8), and

$\log(|\sum_r(U_n)|/|\sum_f(U_n)|) - \text{tr}(\sum_f^{-1}(U_n) \sum_r(U_n)) + p$ in (2.9) can all be viewed as a discrepancy between r and f in terms of their predictivity of T_n .

The above results also suggest ways of choosing T_n in $KLD_t(r, f|U_n)$ that would yield effective model diagnostics. For example, suppose that the estimate of $E(X_i)$ is unbiased under both models r and f (i.e., $\mu_f(\theta) = \mu_r(\theta)$), but $\text{Var}(X_i)$ differs between the two models. Then the empirical variance of X_i would be a better choice for T_n compared to the sample mean of X_i since according to (2.6) and (2.7), $KLD_t(r, f|U_n) = O_p(n)$ when T_n is the empirical variance of X_i , while $KLD_t(r, f|U_n) = O_p(1)$ when T_n is the sample mean of X_i .

Regarding the magnitude of G-R-A KLD, since asymptotically, G-R-A KLD differs from the Bayes factor by a term associated with the logarithm of priors under r and f (Lewis and Raftery, 1994), one could use the Jeffreys' rule (Jeffreys, 1961) as guidance to assess model discrepancy: “worth not more than a bare mention” if $(2 * KLD_t(r, f|U_n)) < 2.13$; “strong discrepancy” if $5 < (2 * KLD_t(r, f|U_n)) \leq 10$; and “decisive discrepancy” if $(2 * KLD_t(r, f|U_n)) > 10$.

Property 1.3—Consider a one-sided weighted posterior predictive p-value (or WPPP) with respect to model r :

$$WPPP_r(U_n) \equiv \int \left\{ \int_{-\infty}^{t_n} f(y_n | \hat{\theta}_f) dy_n \right\} r(t_n | \theta) dt_n \pi_r(\theta | U_n) d\theta, \quad (2.10)$$

where r and f are the density functions of T_n under r and f , respectively. Then the asymptotic relationship between $KLD_t(r, f | U_n)$ and $WPPP_r(U_n)$ follows:

$$\frac{2KLD_t(r, f | U_n)}{n} - \frac{\{\Phi^{-1}(WPPP_r(U_n))\}^2}{n} = \left\{ \frac{(\hat{\mu}_r(U_n) - \hat{\mu}_f(U_n))^2}{\hat{\sigma}_f^2(U_n) + \hat{\sigma}_r^2(U_n)} \right\} \frac{\hat{\sigma}_r^2(U_n)}{\hat{\sigma}_f^2(U_n)} + o_p(1) \quad (2.11)$$

when $\mu_f(\theta) = \mu_r(\theta)$; and

$$2KLD_t(r, f | U_n) - Q \left(\frac{\hat{\sigma}_r^2(U_n)}{\hat{\sigma}_f^2(U_n)} \right) = o_p(1) \quad (2.12)$$

and

$$WPPP_r(U_n) - 0.5 = o_p(1) \quad (2.13)$$

when $\mu_r(\theta) = \mu_f(\theta)$ but $\sigma_r^2(\theta) \neq \sigma_f^2(\theta)$.

A Proof of Property 1.3 is given in Wang and Ghosh (2011).

Note that WPPP, developed based on the posterior predictive p-value technique previously proposed for assessing absolute model fit (Guttman, 1967; Rubin, 1984; Gelman et al., 1996), is a model discrepancy measure that evaluates the predictivity of model f under model r . The expression given in (2.10) shows that WPPP calculates the predictive p-value of T_n under an assumed model f should T_n originate from the reference model r . The asymptotic relationship between $KLD_t(r, f | U_n)$ and $WPPP_r(U_n)$ shown above suggests the common nature between WPPP and G-R-A KLD: a measure for comparing model predictivity. More specifically, when $\mu_f(\theta) = \mu_r(\theta)$ (i.e., the estimate of the mean of T_n differs between r and f), both $KLD_t(r, f | U_n)$ and $\Phi^{-1}(WPPP_r(U_n))$ are of order $O_p(n)$. In contrast, suppose that $\mu_r(\theta) = \mu_f(\theta)$ and $\sigma_f^2(\theta) \neq \sigma_r^2(\theta)$ (i.e., the mean of T_n is the same both r and f , but the variance of T_n differs between r and f). Then $\Phi^{-1}(WPPP_r(U_n))$ converges to 0, while $KLD_t(r, f | U_n)$ converges to a positive quantity of $O_p(1)$. This suggests that compared to WPPP, G-R-A KLD could be more effective to reveal the subtle model discrepancy in variance specification.

In terms of practical applications of G-R-A KLD, it is important to note that Properties 1.1 and 1.3 hold under assumptions (A1)–(A5), while Property 1.2 holds under assumptions (A1)–(A4) and (A6). Assumptions (A1)–(A3) are fairly standard, and they can be met by most distributions used in the literature (e.g., the exponential family). Assumption (A4) holds for priors that are commonly used in the literature, such as a (multivariate) normal prior for location parameter(s), a gamma prior for the shape parameter, a Dirichlet prior for the mixture component in a mixture distribution, and an inverse Wishart prior for the covariance, where the extent to which the prior is informative is governed by the hyper-

parameters associated with precision of the prior distribution. Under (A1)–(A4), the asymptotic normality assumption given in (A5) or (A6) is key for the approximation given in (2.5)–(2.12) to hold. When this normality assumption is violated, results in (2.5)–(2.12) may not be credible. In this case, based on (A1)–(A3) and a reviewer's comment, we approximate G-R-A KLD by $\sum_i \log(r(x_i|\hat{\theta}_r)/f(x_i|\hat{\theta}_f)r(x_i|\hat{\theta}_r))$ when n is sufficiently large. The next section demonstrates how G-R-A KLD performs when comparing non-nested models under the violation of the normality assumption.

3 Application of G-R-A KLD in Frequentist Analyses

In this section, we apply G-R-A KLD to compare non-nested models in four real-world studies. In each of these applications, we assess the model fit to the entire dataset; that is, $T_n = (X_1, \dots, X_n)$ is selected as the diagnostic statistic. The choice of T_n was made based on investigators' interest in model fit to all study subjects (instead of one particular statistic). Regarding the choice of model candidates r and f in the G-R-A KLD calculation for these examples, we let model r be the superior model (hence f be the inferior model), based on either our exploratory analyses or scientific plausibility suggested by the literature. Our choice of model r (the superior model) and model f (the inferior model) is consistent with that in G-R KLD, where model r is the true model, and hence the superior model. From the information theory viewpoint, this choice also seems quite sensible since the resulting G-R-A KLD can be interpreted as a measure of the minimal information gain in model r from model f (note that the minimum information loss under model f is achieved at its MLE).

Regarding the calculation of G-R-A KLD in these examples, we first verify the assumptions required for (2.6)–(2.9). According to Cox and Hinkley (1974) and Cramer (1946), assumptions (A1)–(A3) are met in Examples 1 and 2 described below; and they can also be met in Examples 3 and 4 under additional restrictions on the model parameters (McLachlan and Basford, 1998). In Bayesian analyses, assumption (A4) can be met by choosing proper noninformative priors. However, (A4) is not intimately relevant when applying G-R-A KLD to frequentist analyses (similar to that when applying the Bayesian information criterion (Schwartz, 1978) to model comparison in frequentist analyses). The asymptotic normality assumption given in (A5) or (A6) does not hold in any of the examples below. However since n in these examples is sufficiently large, we approximate G-R-A KLD by an empirical estimate: $\sum_i \log(r(x_i|\hat{\theta}_r)/f(x_i|\hat{\theta}_f)r(x_i|\hat{\theta}_r))$.

To assess the performance of G-R-A KLD in these examples, we verify whether the result of G-R-A KLD is consistent with that of alternative model diagnostic tools, such as the quasi-likelihood information criterion (Pan, 2001), the Akaike Information Criterion (Akaike, 1974), and model fit plots.

3.1 Example I: Trend Analysis of Physician vs. Resident Ratio Pre-and Post-TORT Reform In Texas

In the US, it is commonly perceived that medical malpractice litigation is one major source for the country's skyrocketing health care costs and steadily diminishing access to care. In 2003, the state of Texas imposed a cap on the amount of non-economic damages patients could recover from doctors and shielded emergency room doctors from liability except in

cases of “willful and wanton acts of negligence (see <http://bit.ly/nGI1cF> and <http://bit.ly/pGO4qg>). While litigation over malpractice in Texas has plummeted dramatically since the caps were imposed in 2003, it is likely that the temporal trend of the physician vs. resident ratio (i.e., the number of doctors practicing per capita) in Texas might have been affected.

The purpose of our analyses is to assess whether the temporal trend of the number of doctors practicing per capita in Texas between 1995 and 2011 changed after 2003. To this end, we considered statistical models to characterize the temporal trend of the proportion of physicians among Texas residents (per year) between 1995 and 2011. Two non-nested model candidates with the same number of parameters were compared: a generalized linear mixed effect model (GLMM) assuming a binomial distribution and the logit link vs. GLMM assuming a binomial distribution with the probit link. That is, $r(x_i|\theta) : \log\{Pr(x_i = 1|t)/Pr(x_i = 0|t)\} = \theta_0 + \theta_1(t - 1995) + \theta_2 1_{[t \geq 2003]} + \theta_3(t - 2003) 1_{[t \geq 2003]} + \delta_{it}$ vs. $f(x_i|\theta) : \Phi^{-1}(Pr(x_i = 1|t)) = \theta_0 + \theta_1(t - 1995) + \theta_2 1_{[t \geq 2003]} + \theta_3(t - 2003) 1_{[t \geq 2003]} + \delta_{it}$, where x_i is the indicator for being a physician, t for year, and δ_{it} for the random effect following a normal distribution. The MLE's are: $\hat{\theta}_0 = -6.4408$, $\hat{\theta}_1 = 0.0061$, $\hat{\theta}_2 = 0.0373$, and $\hat{\theta}_3 = 0.0087$ (se=0.00005) under f ; and $\hat{\theta}_0 = -2.9493$, $\hat{\theta}_1 = 0.0019$, $\hat{\theta}_2 = 0.0115$, $\hat{\theta}_3 = 0.0027$ (se=0.000025) under r . The empirical estimate of $KLD_t(r, f|U_n) : \sum_i \log(r(x_i|\hat{\theta}_r)/f(x_i|\hat{\theta}_f))$ was calculated as 0.063, suggesting that model r and model f provide similar fit to the data. This subtle difference between the two models is evident in Figure 1, model estimates (in dashed lines) are indistinguishable between r and f . The G-R-A KLD assessment is also consistent with that due to the quasi-likelihood information criterion (QIC; Pan, 2001) – QIC=0.1814 under r and QIC=0.1824 under f . In particular, $\hat{\theta}_3$ was significantly greater than 0 under both models, implying that the mean yearly increase in the number of doctors per capita in Texas has been increased after the 2003 TORT reform. Thus the TORT effect estimated by the two models was similar despite the difference in $\hat{\theta}$ between r (in the logit scale) and f (in the probit scale).

3.2 Example 2: Hospitalization in The National Health and Nutrition Examination Survey (NHANES)

The National Health and Nutrition Examination Survey (NHANES) of 2003–2004 has obtained hospitalization information from 10,117 individuals. Table 1 shows that around 10% of these individuals were hospitalized in the 12-month survey period. The objective of our analyses was to model the distribution of the number of hospitalizations per person in a 12-month period. Based on the skewed distribution of the hospitalization count, and that the mean (0.147) is smaller than the variance (0.302), a zero-inflated Poisson (or ZIP) model could be appropriate: a portion of the population were more likely to be hospitalized, and these count data can be fitted by a Poisson distribution, while the rest of the population had a very low hospitalization incidence. Alternatively, the negative binomial (NB) distribution is also plausible for fitting the skewed distribution of the hospitalization count. Both model candidates are generalization of the Poisson distribution, but they are not nested. To some extent, the NB model is more flexible than the ZIP model since it allows the Poisson rate to follow a continuous mixture distribution while ZIP restricts it to be a two-component mixture. Thus in our G-R-A KLD assessment, we let the ZIP model be model f , which arises as a discrete mixture of a point mass at zero with probability θ_2 and a Poisson distribution

with probability $(1 - \theta_2)$, where the point mass portion is posited to address excess 0's in the distribution. Denote the NB model by r , which arises as a continuous mixture of Poisson distributions (i.e. a compound probability distribution), where the mixing distribution of the Poisson rate is distributed according to $Gamma(\theta_2, \theta_1/\theta_2)$. That is, $r(x_i|\theta) = \Gamma(\theta_2 + x_i)/\Gamma(\theta_2)\Gamma(x_i + 1)\{\theta_1/(\theta_1 + \theta_2)\}^{x_i}\{\theta_2/(\theta_1 + \theta_2)\}^{\theta_2}$ and

$f(X_i=x_i|\theta) = \theta_2 \mathbf{1}_{[x_i=0]} + (1-\theta_2) \exp(-\theta_1) \theta_1^{x_i}/x_i!$. The MLE's under r are $\hat{\theta}_1 = 0.1473$ and $\hat{\theta}_2 = 0.1555$, and $\hat{\theta}_1 = 0.8609$ and $\hat{\theta}_2 = 0.1711$ under f . The empirical estimate of $KLD_f(r, f|U_n)$: $\sum_i \log(r(x_i|\hat{\theta}_r)/f(x_i|\hat{\theta}_f))r(x_i|\hat{\theta}_r)$ between r and f is 40.55, which suggests that the NB model is superior to the ZIP model. The G-R-A KLD result is consistent with that due to the Akaike Information Criterion (AIC; Akaike, 1974): AIC= 8546.0 under the ZIP model (f) and AIC=8383.4 under the NB model (r). The discrepancy between r and f can also be seen in Figure 2. In summary, the superiority of the NB model over the ZIP model suggests the heterogeneity of hospitalization rate in the diverse NHANES population.

3.3 Example 3: The Johns Hopkins PIRC Study

This example used data from a school intervention trial conducted by the Johns Hopkins University Preventive Intervention Research Center (hereafter PIRC) in 1993–1994 (Ialongo et al., 1999). The PIRC study was designed to improve academic achievement and to reduce early behavioral problems. Teachers and first-grade children were randomly assigned to the control and intervention conditions (either the Classroom-Centered intervention or the Family-School Partnership intervention). The TOCA-R or Teacher Observation of Classroom Adaptation-Revised (Werthamer-Larsson et al., 1991) measures, each item with scale ranging from 1 to 6, were assessed at baseline, the 6-month followup, and the 18-month followup.

In this paper, we focused on using growth mixture modeling (GMM; Jo et al., 2009) to identify distinct trajectories of TOCA-R attention deficit (AD) measures in the control group ($n=184$) – this analysis is critical to deriving credible causal inference associated with the intervention effect under GMM framework (Jo et al., 2009). The TOCA-R AD score consists of TOCA-R items that measure hyperactivity, concentration problem, and impulsiveness. Denote X_{i1} , X_{i2} , and X_{i3} for the AD score measured at baseline, the 6-month followup, and the 18-month followup, respectively. Missing data on AD assessments were analyzed under the missing at random (MAR) assumption (Little and Rubin, 2002) using the likelihood of the observed data. Covariates included in the analyses were child's sex ($z_{i1} = 1$ for male, $z_{i1} = 0$ for female), ethnicity ($z_{i2} = 1$ for African-American, $z_{i2} = 0$ other), parent's education ($z_{i3} = 1$ some education beyond high school, $z_{i3} = 0$ high school or less), marital status ($z_{i4} = 1$ for married, $z_{i4} = 0$ for not married), and employment status ($z_{i5} = 1$ for employed, $z_{i5} = 0$ for not employed). GMM with two classes of quadratic AD trajectories were employed (Jo et al., 2009). Model r included covariates as predictors for AD trajectory class membership:

$$\mathbf{X}_i | S_i = s = \theta_{s1} + \theta_{s2} \mathbf{t} + \theta_{s3} \mathbf{t}^2 + \mathbf{e}_{si},$$

$$\log \left(\frac{Pr(S_i=2|z_i)}{Pr(S_i=1|z_i)} \right) = \gamma_0 + \sum_j \gamma_j z_{ij}$$

for $s = 1, 2$, where $\mathbf{e}_{si} \sim N(0, \Omega)$ and $\mathbf{t}$ denotes the design matrix of time points; model f did not include predictors for AD trajectory class membership:

$$\mathbf{X}_i | S_i = s = \theta_{s1} + \theta_{s2}\mathbf{t} + \theta_{s3}\mathbf{t}^2 + \mathbf{e}_{si},$$

$$\log \left(\frac{Pr(S_i=2)}{Pr(S_i=1)} \right) = \gamma_0$$

for $s = 1, 2$, where $\mathbf{e}_{si} \sim N(0, \Omega)$. Note that rigorously speaking, model f is not nested within model r . This is because S is not observed, and that S derived from the two models can be different. MLE's under f are: $(\hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{13}) = (4.037, 0.041, -0.013)$ and $(\hat{\theta}_{21}, \hat{\theta}_{22}, \hat{\theta}_{23}) = (1.824, -0.029, 0.054)$; MLE's under r are: $(\hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{13}) = (4.031, 0.228, -0.064)$ and $(\hat{\theta}_{21}, \hat{\theta}_{22}, \hat{\theta}_{23}) = (1.856, -0.054, 0.06)$. Since S is unobserved, the pseudoclass technique (Wang et al., 2005) was employed to first estimate S , and then obtain the pseudoclass-specific likelihood. More specifically, we first obtained estimates of S for each individual under both r and f : S estimates under r (or f) were obtained by drawing a random sample from a binomial distribution with probabilities being the individual's posterior class probabilities $\{Pr(S_i = s | X_i, z_i, \hat{\theta}_r)\}$ under r (or $\{Pr(S_i = s | X_i, \hat{\theta}_f)\}$ under f). Then the pseudoclass-specific likelihood for each individual $f(x_i | \hat{s}_{fi})$ and $r(x_i | \hat{s}_{ri})$ was used to calculate $KLD_i(r, f | U_n)$, where $\hat{s}$ denotes the pseudoclass estimate. The empirical estimate of $KLD_i(r, f | U_n)$ was obtained as 29.208, suggesting that the GMM adjusting for covariates provided a better fit for the observed AD scores (as well as the underlying S) compared to the GMM without covariate adjustment. That is, the variation in AD trajectory could be attributed to difference in covariates. The calculated G-R-A KLD arrived at the same conclusion as AIC assessment: $AIC=592.704$ under f (i.e., GMM without covariates) and $AIC=575.798$ under r (i.e., GMM with covariates). According to Figure 3, the estimated trajectories between r and f are similar, which suggests the potential orthogonality (Fisher, 1915) between $\hat{\gamma}$'s and $\hat{\theta}$'s. There are two advantages associated with the GMM adjusting for covariates compared to the GMM without covariate adjustments: (i) the former allows a more precise estimation of S ; and (ii) γ_s allows us to assess characteristics associated with each trajectory class. Therefore, in this example, the GMM with covariate adjustments is both statistically and scientifically preferred over the GMM without.

3.4 Example 4: A Longitudinal Study of Substance Use in Adolescents

This example demonstrates analyses of longitudinal tobacco use in a diverse sample of adolescents from age 11 to age 21 derived from a longitudinal study of substance use in adolescents (Dishion and Kavanagh, 2000). The study began in 1996, recruiting all students from 6th grade classrooms in three schools. Half of the classrooms within these schools received a class-wide intervention program. Participants in the study completed questionnaires once a year in their classroom from 6th grade through 9th grade, and then in 11th grade, 12th grade, and at ages 19, 22, and 23 via mail responses (a total of 8 repeated measures).

In this paper, we present analyses of repeated measures tobacco use in 498 subjects in the control classrooms. We consider two model candidates, both are suitable for characterizing count data with excessive zero's: a mixture of six distinct ZIP distributions (henceforth

referred to as ZIP mixture) and a mixture of six distinct 2-part distributions by Olsen and Schafer (2001), which is referred to as 2-part mixture below. Compared to the 2-part mixture model, the ZIP mixture model is scientifically more plausible since (i) ZIP is more natural for characterizing the discrete count data; and (ii) according to the literature on smoking (Dishion and Kavanagh, 2000), “0” tobacco users at a time point can consist of two types of subjects: abstainers and occasional (or infrequent) users, which is consistent with the ZIP model assumption, while the 2-part model limits “0” tobacco users solely to abstainers. Thus in our G-R-A KLD calculation, we let model r be the ZIP mixture, and model f be the 2-part mixture. That is, model r assumes:

$$\begin{aligned} Pr(X_{ij}=0|S_i=s, \theta) &= \frac{\exp(\theta_{1s}+\theta_{2s}t_j)}{1+\exp(\theta_{1s}+\theta_{2s}t_j)} + \frac{\exp\{-e^{(\theta_{3s}+\theta_{4s}t_j+\theta_{5s}t_j^2)}\}}{1+\exp(\theta_{1s}+\theta_{2s}t_j)}, \\ Pr(X_{ij}=x_{ij}|S_i=s, \theta, x_{ij}>0) &= \frac{\exp\{-e^{(\theta_{3s}+\theta_{4s}t_j+\theta_{5s}t_j^2)}\} (e^{(\theta_{3s}+\theta_{4s}t_j+\theta_{5s}t_j^2)})^{x_{ij}}}{x_{ij}!(1+\exp(\theta_{1s}+\theta_{2s}t_j))} \\ Pr(S_i=s) &= \theta_{6s}, \end{aligned}$$

for $s = 1, \dots, 6$; and model f assumes:

$$\begin{aligned} Pr(X_{ij}=0|S_i=s, \theta) &= 1 - Pr(X_{ij}>0|S_i=s, \theta) = \frac{\exp(\theta_{1s}+\theta_{2s}t_j)}{1+\exp(\theta_{1s}+\theta_{2s}t_j)}, \\ X_{ij}|S_i=s, X_{ij}>0 &= \theta_{3s}+\theta_{4s}t_j+\theta_{5s}t_j^2 + e_{s,ij}, \\ Pr(S=s) &= \theta_{6s}, \end{aligned}$$

for $s = 1, \dots, 6$, where $e_{s,ij} \sim N(0, \Omega_{jj})$, with Ω_{jj} being the j th diagonal element of Ω , the covariance matrix of $(e_{s,i1}, \dots, e_{s,i8})$. Following the argument in Example 3, we employed the pseudoclass technique to calculate $KLD_t(r, f|U_n)$. We first calculated estimates of S for each individual under both r and f by drawing a random sample from a multinomial distribution with probabilities being the individual's posterior class probabilities $\{Pr(S_i = s|X_i, \hat{\theta}_r): s = 1, \dots, 6\}$ under r (or $\{Pr(S_i = s|X_i, \hat{\theta}_f): s = 1, \dots, 6\}$ under f). Then the pseudoclass-specific likelihood for each individual was used to calculate $KLD_t(r, f|U_n)$ using $\sum_i \log\{r(x_i|\hat{S}_i = s_i, \hat{\theta}_r)/f(x_i|\hat{S}_i = s_i, \hat{\theta}_f)\}$. This calculation yielded $KLD_t(r, f|U_n) = 43.212$, suggesting that the ZIP mixture model is superior to the 2-part mixture model. The G-R-A KLD assessment is consistent with the AIC result, where $AIC=7417.864$ under the 2-part mixture model (f) and $AIC=7330.652$ under the ZIP mixture model (r). Note that means of each trajectory class are similar between models r and f (see Figures 4 and 5). Thus the discrepancy between model r and model f revealed by G-R-A KLD and AIC could be due to their difference in estimation of mixing component probabilities, which then results in different pseudoclass estimates and predicted trajectory means for certain subjects.

4 Summary

Researchers often face the challenge of choosing between competing models that are scientifically plausible but non-nested. This paper shows that the G-R-A KLD is appropriate for comparing non-nested models in the frequentist framework. Our work draws on asymptotic properties of G-R-A KLD shown in Section 2: the G-R-A KLD quantifies information discrepancy between competing models in terms of their MLE's. Four novel examples given in Section 3 have confirmed the utility of G-R-A KLD as a measure of

model discrepancy in the frequentist framework: the result due to G-R-A KLD was consistent with that due to AIC and model diagnostic plots. We found emerging conclusions between G-R-A KLD and AIC on model comparison in these four examples with moderate to large sample sizes, which could be due to that the similar asymptotics revealed in these examples. As a preliminary investigation of the finite sample property of the proposed G-R-A KLD for scenarios considered in Section 3, we have conducted simulations to assess the extent to which model discrepancy is revealed by G-R-A KLD under various situations. Based on the recommendation by Jeffreys (1961), one may regard that there is a “decisive evidence” of model discrepancy if the expected G-R-A KLD exceeds 5. In our simulations, the result of each scenario was drawn from 100 independently simulated datasets of size ranging from 50 to 2000. For practical reasons, we only consider situations where the reference model is the true model. Findings of our simulations are summarized below.

- To differentiate between models for a binary outcome with a logistic link vs. a probit link using G-R-A KLD, for a given $k_0 > 0$ the minimum sample size required to achieve $KLD_t(r, f|U_n) > k_0$ depends on the true prevalence of the binary outcome (say θ): a larger sample size is required if θ is closer to 0.5 (see Figure 6(a)). For example, $n = 100, 200, 300, 700$ is required to achieve $KLD_t(r, f|U_n) > 5$ for $|\theta - 0.5| = 0.3, |\theta - 0.5| = 0.2, |\theta - 0.5| = 0.15, |\theta - 0.5| = 0.1$, respectively.
- For a continuous outcome with a bell-shaped distribution, to differentiate the fit of a t -distribution from the fit of a 2-normal mixture (with the same mean but different variances between the two mixture components) using G-R-A KLD, it was shown that a larger sample size is required to achieve $KLD_t(r, f|U_n) > k_0$ (where $k_0 > 0$) when there is less discrepancy in the variance between 2 components of the normal mixture (see Figure 6(b)). For example, $n = 50$ is adequate to achieve $KLD_t(r, f|U_n) > 5$ if the more variable class in the normal mixture distribution has a variance (say σ_1^2) that is 5 times as large as the variance of the less variable class (say σ_2^2), or the ratio of variances $r = \sigma_1^2 / \sigma_2^2$ equals to 5, while $n = 100$ is needed if the variance of the more variable class has a variance that is 3.2 times as large as the variance of the less variable class (i.e., $r = \sigma_1^2 / \sigma_2^2 = 3.2$).
- For count data with a right-skewed distribution, to compare the fit between a ZIP model and a negative binomial model using G-R-A KLD, the minimum sample size required to achieve $KLD_t(r, f|U_n) > k_0$ (where $k_0 > 0$) depends on the proportion of “excess zeros (i.e., θ_2 under model f in Example 3.2) and the distribution of non-excess zeros – a larger sample size is required when the proportion of “excess zeros (i.e., θ_2 under model f in Example 3.2) is small, and the mean of non-excess zeros (i.e., θ_1 under model f in Example 3.2) assumes a small positive number. For example, as shown in Figure 6(c), $n = 100$ is sufficient to achieve $KLD_t(r, f|U_n) > 5$ when the proportion of excess zeros is greater than 0.1, and the mean of non-excess zeros is 5; $n = 200$ is adequate when the proportion of excess zeros is greater than 0.5, and the mean of non-excess zeros is equal to 2, while $n = 700$ is required when the mean of non-excess zeros is 2, and the proportion of excess zeros is 0.1; $n = 800$ is required when the proportion of excess zeros is greater than 0.5, and the mean of non-excess zeros is 1, while (not shown in Figure 6(c)) $n > 2000$ is

required when the proportion of excess zeros is 0.1 and the mean of non-excess zeros is 1; and $n > 3000$ is required if the mean of non-excess zeros is less than 0.6 and the proportion of excess zeros is less than 0.7.

To sum up, our preliminary simulations suggest that the finite-sample performance of G-R-A KLD depends on the discrepancy and the underlying distributions of the models under comparison. Further investigation of the consistency between G-R-A KLD and other model fit indices (such as AIC) is needed for the finite sample situation. For example, while G-R-A KLD was originally developed from the Bayesian framework, it is not yet clear the extent to which (2.6)–(2.9), (2.11), and (2.12) would hold under different choices of priors when the sample size is limited.

In general, the G-R-A KLD provides the relative fit between competing models without requiring the true model to be specified (see (2.5)–(2.7) which show that G-R-A KLD estimate only involves MLE's under models r and f). This implies that G-R-A KLD is a practical measure for model discrepancy since it does not need to assume the true model to yield a meaningful interpretation regarding model discrepancy. For example, when the superior model is designated as model r and the inferior model is designated as model f in the G-R-A KLD calculation (like G-R KLD and G-R-A KLD considered herein), then the resulting G-R-A KLD can be interpreted as a measure of the minimum information gain in model r from model f (since the inferior model f is evaluated at the MLE's). In contrast, if the superior model is designated as model f (hence the inferior model is designated as model r) in the G-R-A KLD calculation, then the resulting G-R-A KLD would have been a measure of the maximum discrepancy between model r and model f (since the superior model f is evaluated at the MLE's). Further research is needed to help us better understand this type of KLD.

Finally, for the purpose of assessing model adequacy (rather than relative model fit), G-R-A KLD should be used in conjunction with absolute model departure indices such as residuals. Nevertheless, the G-R-A KLD can be regarded as a measure of the absolute fit (or adequacy) of model f when the superior model r is the true model.

Acknowledgments

Wang's work on this paper was partially supported by K25-DK075092 and R21-CA161180. Jo's work on this paper is partially supported by R01-DA031698 and R01-MH086043. We thank Dr. Ron Stewart, Dr. Nick Ialongo, and Dr. Tom Dishion for generously providing the datasets for Example 1, Example 3, and Example 4, respectively. We thank the Reviewers for their helpful comments.

References

1. Akaike H. A new look at the statistical identification model. *IEEE Transactions on Automatic Control*. 1974; 19(6):716–23.
2. Bernardo JM. Expected Information as Expected Utility. *The Annals of Statistics*. 1979; 7(3):686–90.
3. Cox, DR.; Hinkley, DV. *Theoretical Statistics*. London: Chapman & Hall Ltd; 1974.
4. Cramer, H. *Mathematical Methods of Statistics*. Princeton: Princeton University Press; 1946.
5. Dishion TJ, Kavanagh K. A multilevel approach to family-centered prevention in schools: process and outcome. *Addictive Behaviors*. 2000; 25(6):899–911. [PubMed: 11125778]

6. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*. 1915; 10:507–21.
7. Gelman, A.; Carlin, J.; Stern, HS.; Rubin, D. *Bayesian Data Analysis*. Chapman and Hall; London: 1996.
8. Goutis C, Robert CP. Model Choice in Generalised Linear Models: A Bayesian Approach Via Kullback-Leibler Projections. *Biometrika*. 1998; 85(1):29–37.
9. Guttman I. The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B*. 1967; 29(1):83–100.
10. Ialongo NS, Werthamer L, Kellam SG, Brown CH, et al. Proximal impact of two first-grade preventive interventions on the early risk behaviors for later substance abuse, depression and antisocial behavior. *American Journal of Community Psychology*. 1999; 27:599–642. [PubMed: 10676542]
11. Jeffreys, H. *Theory of Probability*. 3. Oxford: University Press; 1961.
12. Jo B, Wang C-P, Ialongo N. Using Latent Outcome Trajectory Classes in Causal Inference. *Statistics and Its Inference*. 2009; 2(4):403–12.
13. Kullback S, Leibler RA. On Information and Sufficiency. *The Annals of Mathematical Statistics*. 1951; 22(1):79–86.
14. Lewis SM, Raftery AE. Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator. *Journal of the American Statistical Association*. 1994; 92:648–55.
15. Lindley DV. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*. 1956; 27(4):986–1005.
16. Little, RJA.; Rubin, DB. *Statistical analysis with missing data*. New York: Wiley; 2002.
17. McLachlan, GJ.; Basford, KE. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker; 1998.
18. Olsen MK, Schafer JL. A two-part random effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*. 2001; 96:730–45.
19. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics*. 2001; 57(1):120–25. [PubMed: 11252586]
20. Rubin DB. Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician. *Annals of Statistics*. 1984; 12(4):1151–72.
21. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*. 1978; 6(2):461–64.
22. Shannon CE. A mathematical theory of communication. *Bell System Technical Journal*. 1948; 27:379–423. 623–56.
23. Wang CP, Brown CH, Bandeen-Roche K. Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*. 2005; 100:1054–76.
24. Wang CP, Ghosh M. A Kullback-Leibler Divergence for Bayesian Model Diagnostics. *Open Journal of Statistics*. 2011; 1:172–84.
25. Werthamer-Larsson L, Kellam SG, Wheeler L. Effect of first-grade classroom environment on shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*. 1991; 19:585–602. [PubMed: 1755437]

Figure 1.

Plot for Estimates under Probit and Logistic Models vs. Empirical Mean Number of Physicians Per Capita in The TORT Dataset

Figure 2.

Plot for Estimates under ZIP and Negative Binomial Models vs. Empirical Means in the NHANES 2003–2004 Hospitalization Dataset (solid line for the Negative Binomial Model; Broken Line for ZIP Model)

Figure 3.

Plot for 2-Class GMM Estimates Fit to the Control Group in the Johns Hopkins PIRC Study: GMM With Covariate Adjustment (solid lines) vs. GMM Without Covariate Adjustment (dashed lines)

Figure 4.

Plot for Estimates of 6-Class 2-part Mixture Model Fit to the Control Group's Tobacco Use in the Longitudinal Study of Substance Use in Adolescents vs. Empirical Means of Tobacco Use by Class

Figure 5.

Plot for Estimates of 6-Class ZIP Mixture Model Fit to Control Group's Tobacco Use in the Longitudinal Study of Substance Use in Adolescents vs. Empirical Means of Tobacco Use by Class

Figure 6.
Simulation Results for Finite Sample Performance of G-R-A KLD

Table 1

The NHANES 2003–2004 Hospitalization Dataset

Days of Hospitalization	0	1	2	3	4	5	6
Counts	9118	725	165	57	17	14	21