Open Journal of Statistics, 2011, 1, 172-184
doi:10.4236/ojs.2011.13021 Published Online October 2011 (http://www.SciRP.org/journal/ojs)

A Kullback-Leibler Divergence for Bayesian Model Diagnostics

Chen-Pin Wang[1], Malay Ghosh[2]
[1]*Department of Epidemiology and Biostatistics*, *University of Texas Health Science Center*, *San Antonio*, *USA*
[2]*Department of Statistics*, *University of Florida*, *Gainesville*, *USA*
E-mail: *wangc3@uthscsa.edu*, *ghoshm@stat.ufl.edu*
Received August 18, 2011; *revised September* 19, 2011; *accepted September* 29, 2011

Abstract

This paper considers a Kullback-Leibler distance (KLD) which is asymptotically equivalent to the KLD by Goutis and Robert [1] when the reference model (in comparison to a competing fitted model) is correctly specified and that certain regularity conditions hold true (ref. Akaike [2]). We derive the asymptotic property of this Goutis-Robert-Akaike KLD under certain regularity conditions. We also examine the impact of this asymptotic property when the regularity conditions are partially satisfied. Furthermore, the connection between the Goutis-Robert-Akaike KLD and a weighted posterior predictive p-value (WPPP) is established. Finally, both the Goutis-Robert-Akaike KLD and WPPP are applied to compare models using various simulated examples as well as two cohort studies of diabetes.

Keywords: Kullback-Leibler Distance, Model Diagnostic, Weighted Posterior Predictive p-Value

1. Introduction

Information theory provides a general framework of developing statistical techniques for model comparison (Akaike [2]; Shannon [3]; Kullback and Leibler [4]; Lindley [5]; Bernardo [6]; Schwarz [7]). The Kullback-Leibler distance (KLD) is perhaps the most commonly used information criterion for assessing model discrepancy (Akaike [2]; Kullback and Leibler [4]; Lindley [5]; Schwarz [7]). In essence, a KLD is the expected 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 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 [5] and Schwarz [7]). KLDs based on the ratio of integrated likelihoods however have been challenged by identifying priors that are compatible under the competing models and that the resulting integrated likelihoods are proper. As a way to overcome the challenges associated with prior elicitation in calculating KLD under the Bayesian framework, one may consider the Bayesian estimate of the Kullback-Leibler projection by Goutis and Robert [1], henceforth G-R KLD. More specifically, for a given reference model indexed by parameter(s) θ, 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. Thus if the reference model is correctly specified, then 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 [2]). 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. First, 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. Second, 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 remains limited. For example, if the refer-

ence model is not correctly specified, then the G-R KLD is not necessarily reduced to to the KLD between the reference model and the competing fitted model evaluated at its maximum likelihood estimate or MLE (ref. Akaike [2]), referred to as the Goutis-Robert-Akaike KLD or G-R-A KLD a more tractable model discrepancy measure.

This paper proposes to use G-R-A KLD for assessing model discrepancy in terms of the fit of certain statistics T_n that is central to our inference or model diagnostic purpose. That is, we evaluate the G-R-A KLD between the probability density function (p.d.f.) of T_n under the reference model r and that under the assumed model f evaluated at its MLE (see Section 2). We investigate the (asymptotic) property of G-R-A KLD under certain regularity conditions as well as under the violation of some regularities, including non-nested models. Note that unlike G-R KLD (Goutis and Robert [1]), the G-R-A KLD considered herein 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 seems to be more flexible for comparing nested or non-nested models that are broader than generalized linear models (see Sections 3 and 4). Theorem 1 shows that under certain regularity conditions, the asymptotic expression of the posterior estimator of G-R-A KLD 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. Since the first two leading terms in the G-R-A KLD estimator resemble a measure that differentiates the predictability between models r and f, it is natural to study its connection with Bayesian model discrepancy measures based on predictive statistics (Guttman [8], Rubin [9], Gelman *et al.* [10]). In particular, we consider the posterior predictive check technique using a one-sided weighted posterior predictive p-value (WPPP). The WPPP of T_n evaluates the predictive distribution of T_n under f at $T_n^{pred,r}$, where $T_n^{pred,r}$ denotes the prediction of T_n under r, the posteriors are derived under r, and the weight is used to account for the variation of T_n under r. Theorem 2 explicitly shows that for any T_n satisfying certain asymptotic normality and regularity conditions, how the model discrepancy is reflected by the G-R-A KLD in connection to that by WPPP. To verify the results in Theorem 1 and Theorem 2 as well as to evaluate G-R-A KLD under partial violation of the regularity conditions, we examine the (asymptotic) property of G-R-A KLD and WPPP via both simulations and real data applications motivated by two cohort studies of diabetes. These examples include the comparison between nested models as well as non-nested models.

The paper is organized as follows. Section 2 studies the G-R-A KLD for (predictive) p.d.f.s of T_n between two competing models. It also derives the relationship between G-R-A KLD and WPPP. Sections 3 and 4 evaluate model fit using both G-R-A and WPPP for examples that meet all the regularity conditions required in Theorems 1 and 2 as well as for examples that meet only part of these regularity conditions.

2. A Proposed Kullback-Leibler Divergence

2.1. Notations

Throughout this paper, we assume that X_i's originate from model g, and are i.i.d. with some common p.d.f. with 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 the parameter spaces. Also, when a capital letter is used to denote for a random variable (or an estimator), the corresponding lower case is for its realization (or an estimate). Let $U_n = U(X_1, \cdots, X_n)$ be the base for deriving the posterior density of θ under model r. We shall denote $r(*|\theta)$ and $f(*|\theta)$ for the p.d.f.'s of * given θ under model r and f, respectively. Also, let $\varphi(\mathbf{y}) = (2\pi)^{-l/2}\exp(-\mathbf{y}'\mathbf{y}/2)$ where $\mathbf{y}$ is l-dimensional.

2.2. Define Kullback-Leibler Divergence

Consider that model adequacy is evaluated based on its fit for certain statistics $T_n = T(X_1, \cdots, X_n)$ that is pertinent to the inference or model diagnostics. Stemmed from Goutis and Robert [1], we assess the relative fit between models using the KLD of the distribution of T_n under r and that under f evaluated at $\hat{\theta}_f$, the MLE of θ:

$$\int \log\left(\frac{r(t_n|\theta)}{f(t_n|\hat{\theta}_f)}\right) r(t_n|\theta)\,\mathrm{d}t_n, \qquad (1)$$

We shall refer (1) to as the Goutis-Robert-Akaike KLD or G-R-A KLD since (1) is asymptotically equivalent to the KLD proposed by Goutis and Robert [1] when the reference model r is the true model (ref. Akaike [2]).

In general, for each $\theta \in \Theta_r$, G-R-A KLD given in (1) can be regarded as a measure of the minimal information gain 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 [1]), (1) by definition does not require the reference model r to be the true model. To understand the utility of (1) in the Bayes-

ian framework, below we derive its Bayesian estimate and the associated (asymptotic) property under certain regularity conditions (see Sections 2.3 and 2.4). We also study the property of (1) when partial regularity conditions hold true using simulated data (Section 3) as well as two applications of diabetes studies (Section 4).

2.3. Bayesian Estimation of the Proposed KLD

Estimating (1) involves approximating the integral with respect to $r(t_n \mid \theta)$ and estimating unknown model parameters θ. To account for the uncertainty of model parameters, we consider the following estimator:

$$\iint \log\left(\frac{r(t_n \mid \theta)}{f\left(t_n \mid \hat{\theta}_f\right)}\right) r(t_n \mid \theta)\pi_r(\theta \mid U_n)\mathrm{d}t_n\mathrm{d}\theta, \quad (2)$$

where U_n, as a function of $(X_1,\cdots,X_n)$, is used for deriving the posterior of θ, and $\pi_r(\theta \mid U_n) = r(U_n \mid \theta)\pi_r(\theta)/\int r(U_n \mid \theta)\pi_r(\theta)\mathrm{d}\theta$ denoting the posterior density of θ under model r. That is, (2) is the average discrepancy between r and f, each being weighted by $\pi_r(\theta \mid U_n)$. We shall denote (2) by $KLD_t(r, f \mid U_n)$. Since (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 statistic T_n.

To gain further insight about the utility of (2), we derive its asymptotic properties below. The approximation of (2) is tied to the assumptions of T_n under r and f, which will be described in Theorems 1 and 2. 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 attained 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.

Assume the following regularity conditions under Theorems 1 and 2.

(A1) For each x, both $\log(r(x \mid \theta))$ and $\log(f(x \mid \theta))$ 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

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

and

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

for $k = 1, 2, 3$.

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

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

and

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

(A3)

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

as $\delta \to 0$, and

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

as $\delta \to 0$.

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

(A5) Let T_n be asymptotically normally distributed under both models such that

$$r(T_n \mid \theta) = \sigma_r^{-1}(\theta)\varphi\left(\sqrt{n}(T_n - \mu_r(\theta))/\sigma_r(\theta)\right) + O(n^{-1/2}) \quad (3)$$

and

$$f(T_n \mid \theta) = \sigma_f^{-1}(\theta)\varphi\left(\sqrt{n}(T_n - \mu_f(\theta))/\sigma_f(\theta)\right) + O(n^{-1/2}). \quad (4)$$

Theorem 1.

$$\frac{2KLD_t(r, f \mid U_n)}{n} - \frac{\{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\}^2}{\hat{\sigma}_f^2(U_n)} = o_p(1) \quad (5)$$

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

$$2KLD_t(r, f \mid U_n) - Q\left(\frac{\hat{\sigma}_r^2(U_n)}{\hat{\sigma}_f^2(U_n)}\right) = o_p(1) \quad (6)$$

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

The proof of Theorem 1 is given in Appendix 1. Since model comparison in real applications can rely on the relative fit to a multi-dimensional statistic, it is useful to study whether the results in Theorem 1 are applicable to the multivariate case with a fixed dimension. Suppose that T_n is a p-dimensional statistic ($p > 1$ and independent of n) with

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

and

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

Then following the derivation given in Appendix 1, it can be shown that

$$\frac{2KLD_t\left(r, f \mid U_n\right)}{n} - \left\{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\right\}' \Sigma_f^{-1}(U_n) \left\{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\right\} = o_p(1)$$

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

$$2KLD_t\left(r, f \mid U_n\right) - \log\left(\left|\Sigma_r(U_n)\right| \Big/ \left|\Sigma_f(U_n)\right|\right) + tr\left(\Sigma_f^{-1}(U_n)\Sigma_r(U_n)\right) - p = o_p(1)$$

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

The result above implies the importance of choosing T_n to yield an effective model diagnostic based on $KLD_t(r, f \mid U_n)$. For example, T_n is typically chosen to be the sufficient statistic for θ since it contains all the information of θ. Yet, T_n clearly is not the best diagnostic statistic if the estimates of the mean of T_n are unbiased under both models r and f as $KLD_t\left(r, f \mid U_n\right) = O_p(1)$. Also, note that both $\left\{\hat{\mu}_f(U_n) - \hat{\mu}_r(U_n)\right\}^2 \Big/ \hat{\sigma}_f^2(U_n)$ in (5) and $Q\left(\hat{\sigma}_r^2(U_n) \Big/ \hat{\sigma}_f^2(U_n)\right)$ in (6) can be viewed as a discrepancy between r and f in terms of their posterior predictivity of T_n. We show next how $KLD_t\left(r, f \mid U_n\right)$ is related to a weighted posterior predictive p-value, a typical approach for assessing model discrepancy regarding the predictivity of T_n in the Bayesian framework (see Rubin [9]; Gelman *et al.* [10]).

2.4. KLD vs Posterior Predictive p-Value

Consider

$$WPPP_r(U_n) \equiv \int \left\{ \int \left(\int_{-\infty}^{t_n} f\left(y_n \mid \hat{\theta}_f\right) \mathrm{d}y_n \right) r\left(t_n \mid \theta\right) \mathrm{d}t_n \right\} \pi_r\left(\theta \mid U_n\right) \mathrm{d}\theta, \quad (7)$$

where r and f are the density functions of T_n under r and f, respectively. We shall call (7) a one-sided weighted posterior predictive p-value (or WPPP) with respect to model r since as shown above, it is equivalent to the mean predictive p-value of T_n under f over all possible posterior predicted values of T_n arising from r. Thus WPPP can be viewed as a model discrepancy measure since it evaluates the predictivity of f under r. Note that WPPP proposed here differs from the posterior predictive p-value originally proposed in Rubin [9]: WPPP calculates the predictive p-value of T_n under an assumed model f should T_n originate from the reference model r, while Rubin's original proposal [9] assesses the predictive p-value of T_n under an assumed model. WPPP is considered here since it is more coherent with the nature of G-R-A KLD as a discrepancy measure between models r and f.

The connection between $WPPP_r(U_n)$ and $KLD_t(r, f \mid U_n)$ is derived in Theorem 2 below. Here we assume regularity conditions (A1)-(A5) as assumed in Theorem 1.

Theorem 2.

$$\frac{2KLD_t\left(r, f \mid U_n\right)}{n} - \frac{\left\{\Phi^{-1}\left(WPPP_r(U_n)\right)\right\}^2}{n} = \left\{ \frac{\left(\hat{\mu}_r(U_n) - \hat{\mu}_f(U_n)\right)^2}{\hat{\sigma}_f^2(U_n) + \hat{\sigma}_r^2(U_n)} \right\} \frac{\hat{\sigma}_r^2(U_n)}{\hat{\sigma}_f^2(U_n)} + o_p(1) \quad (8)$$

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

$$2KLD_t\left(r, f \mid U_n\right) - Q\left(\frac{\hat{\sigma}_r^2(U_n)}{\hat{\sigma}_f^2(U_n)}\right) = o_p(1) \quad (9)$$

and

$$WPPP_r(U_n) - 0.5 = o_p(1) \quad (10)$$

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

The proof of the theorem is given in Appendix 2.

Remark. Theorem 2 explicitly shows the asymptotic relationship between $KLD_t\left(r, f \mid U_n\right)$ and $WPPP_r(U_n)$. Suppose that $\mu_f(\theta) \neq \mu_r(\theta)$ (*i.e.*, the estimate of the mean of T_n differs between r and f). Then both $KLD_t\left(r, f \mid U_n\right)$ and $\Phi^{-1}\left(WPPP_r(U_n)\right)$ are of order $O_p(n)$. Suppose $\mu_r(\theta) = \mu_f(\theta)$ but $\sigma_f^2(\theta) \neq \sigma_r^2(\theta)$ (*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 $\Phi^{-1}\left(WPPP_r(U_n)\right)$ converges to 0, while $KLD_t\left(r, f \mid U_n\right)$ converges to a positive quantity of $O_p(1)$.

3. Illustrative Examples

This section demonstrates the utility of the $KLD_t\left(r, f \mid U_n\right)$ discussed previously through two sets of examples. Examples 3.1-3.6 demonstrate simulated examples that confirm the results proved in Theorems 1 and 2. All these six examples meet the regularity conditions required, while Examples 3.3, 3.5, and 3.6 involve comparing non-nested models. Example 3.7 studies the departure from Theorems 1 and 2 due to violation of the regularity condition (A5). We let $U_n = \left(X_1, \cdots, X_n\right)$ in all calculations below.

Example 3.1. (Nested) Assume

$$X_i \overset{i.i.d.}{\sim} g\left(x_i \mid \theta\right) = \sqrt{\theta_2}^{-1} \varphi\left(\left(x_i - \theta_1\right) \Big/ \sqrt{\theta_2}\right),$$

where $\theta_2 > 0$. Let $T_n = (\sum_{i=1}^n X_i)/n$. Suppose that $r = g$ and

$$f(x_i \mid \theta) = \sqrt{\kappa}^{-1}\varphi\left((x_i - \theta_1)/\sqrt{\kappa}\right)$$

for some known $\kappa > 0$. Then

$$\mu_r(\theta) = \mu_f \quad (\theta) = \theta_1,\ \sigma_r^2(\theta) = \theta_2,\ \sigma_f^2(\theta) = \kappa,$$
$$\hat{\mu}_r(U_n) = \hat{\mu}_f(U_n) \quad = T_n,$$

and

$$2KLD_t(r, f \mid U_n) - Q\left(\hat{\theta}_2(U_n)/\kappa\right) = o_p(1).$$

That is, $KLD_t(r, f \mid U_n)$ converges in probability to a positive quantity, which suggests the discrepancy between r and f. However the magnitude of the model discrepancy assessed by $KLD_t(r, f \mid U_n)$ depends on $Q\left(\hat{\theta}_2(U_n)/\kappa\right)$. In contrast, $WPPP_r(U_n)$ converges to 0.5, indicating no difference between r and f. Thus the results are consistent with Theorems 1 and 2.

Example 3.2. (Nested) Assume

$$X_i \overset{i.i.d.}{\sim} g(x_i \mid \theta) = \sqrt{\left|\sum_g(\theta)\right|}^{-1}\varphi\left(\sum\nolimits_g^{-1/2}(\theta)\left(x_i - \nu_g(\theta)\right)\right),$$

where $X_i = (X_{i1}, X_{i2})$, $\nu_g(\theta) = (\theta_1, \theta_1 - \theta_2)$, and $\Sigma_g(\theta)$ is a 2×2 matrix with $\Sigma_{g,11}(\theta) = \Sigma_{g,22} = \theta_3$ and $\Sigma_{g,12}(\theta) = \Sigma_{g,21} = \theta_3\theta_4$. Suppose that

$$r(x_i \mid \theta) = \sqrt{\left|\sum_r(\theta)\right|}^{-1}\varphi\left(\sum\nolimits_r^{-1/2}(\theta)\left(x_i - \nu_r(\theta)\right)\right)$$

and

$$f(x_i \mid \theta) = \sqrt{\left|\sum_f(\theta)\right|}^{-1}\varphi\left(\sum\nolimits_f^{-1/2}(\theta)\left(x_i - \nu_f(\theta)\right)\right),$$

where $\nu_r(\theta) = (\theta_1, \theta_1 - \theta_2)$, $\nu_f(\theta) = (\theta_1, \theta_1)$, and both $\Sigma_r(\theta)$ and $\Sigma_f(\theta)$ are 2×2 matrices with

$$\Sigma_{r,11}(\theta) = \Sigma_{r,22} = \theta_3,\ \Sigma_{r,12}(\theta) = \Sigma_{r,21} = 0,$$

$$\Sigma_{f,11}(\theta) = \Sigma_{f,22} = \theta_3, \text{ and } \Sigma_{f,12}(\theta) = \Sigma_{f,21} = 0.$$

Let $T_n = \left\{\sum_{i=1}^n (X_{i1} - X_{i2})\right\}/n$. Then $\mu_r(\theta) = \theta_2$, $\mu_f(\theta) = 0$ and $\sigma_r^2(\theta) = \sigma_f^2(\theta) = 2\theta_3^2$. It can be shown that $\hat{\mu}_r(U_n) = T_n$, $\hat{\mu}_f(U_n) = 0$,

$$\hat{\sigma}_r^2(U_n) = \left\{\sum_{i=1}^n \left(X_{i1} - \hat{\theta}_{1r}(U_n)\right)^2 + \sum_{i=1}^n \left(X_{i2} - \hat{\theta}_{2r}(U_n)\right)^2\right\}/(2n),$$

and

$$\hat{\sigma}_f^2(U_n) = \left\{\sum_{i=1}^n \left(X_{i1} - \hat{\theta}_{1f}(U_n)\right)^2 + \sum_{i=1}^n \left(X_{i2} - \hat{\theta}_{1f}(U_n)\right)^2\right\}/(2n),$$

where

$$\hat{\theta}_{1r}(U_n) = \left(\sum_{i=1}^n X_{i1}\right)/n,\ \hat{\theta}_{2r}(U_n) = \left(\sum_{i=1}^n X_{i2}\right)/n,$$
$$\hat{\theta}_{1f}(U_n) = \left(\sum_{i=1}^n X_{i1} + \sum_{i=1}^n X_{i2}\right)/(2n).$$

Thus

$$2KLD_t(r, f \mid U_n)/n - \hat{\mu}_r^2(U_n)/\hat{\sigma}_f^2(U_n) = o_p(1).$$

Since $\mu_f(\theta) \neq \mu_r(\theta)$ and $\hat{\mu}_r(U_n) - \hat{\mu}_f(U_n) = T_n$, following the arguments given in Appendix 2, we have

$$WPPP_r(U_n) - \Phi\left(\sqrt{n}\hat{\mu}_r(U_n)/\sqrt{\hat{\sigma}_r^2(U_n) + \hat{\sigma}_f^2(U_n)}\right) = o_p(1).$$

Thus both $KLD_t(r, f \mid U_n)$ and the square of the probit transformed $WPPP_r(U_n)$ suggest the discrepancy between r and f by an $O_p(n)$ term.

Example 3.3 (Non-nested). Assume

$$X_i \overset{i.i.d.}{\sim} g(x_i \mid \theta) = (x_i!)^{-1}\exp\left(-\theta/(1-\theta)\right)\left\{\theta/(1-\theta)\right\}^{x_i},$$

where $0 < \theta < 1$. Let $T_n = \bar{X}_n/(1 + \bar{X}_n)$, $r = g$, and $f(x_i \mid \theta) = \theta^{x_i}(1-\theta)$. Then

$$\mu_r(\theta) = \mu_f(\theta) = \theta,\ \sigma_r^2(\theta) = \theta(1-\theta)^3, \text{ and}$$
$$\sigma_f^2(\theta) = \theta(1-\theta)^2.$$

In this case, θ has the same statistical meaning under both f and r since $\theta = E(X_i)/(1 + E(X_i))$. Yet the substantive meaning of θ under r differs from that under f. It can be shown that $\hat{\mu}_r(\theta) = \hat{\mu}_f(\theta) = T_n$, and

$$2KLD_t(r, f \mid U_n) - Q\left(1 - \hat{\theta}(U_n)\right)$$
$$= 2KLD_t(r, f \mid U_n) - Q(1 - T_n) = o_p(1)$$

for $0 < \theta < 1$. Consistent with Theorem 2,

$$KLD_t(r, f \mid U_n) = O_p(n^{-1/2})$$

converges in probability to a positive quantity (*i.e.*, $Q\left(1 - \hat{\theta}(U_n)\right)$), and $WPPP_r(U_n)$ converges in probability to 0.5 (since $\mu_r(\theta) = \mu_f(\theta)$).

Examples 3.4-3.6 below consider situations where T_n is an unbiased estimate for the parameter of interest under both f and r. However T_n is the MLE of certain parameter under f but not under r. In these examples, since MLEs of the mean of T_n under both f and r converge to the same in probability, $KLD_t(r, f \mid U_n)$ is reduced to $Q\left(\hat{\sigma}_r^2(U_n)/\hat{\sigma}_f^2(U_n)\right) + o_p(1)$ in these examples.

Example 3.4 (Nested). As an extension of Example 3.1, consider

$$X_{ji} \overset{i.i.d.}{\sim} g(x_{ji} \mid \theta) = \sqrt{\theta_2\theta_{3j}}^{-1}\varphi\left((x_{ji} - \theta_1)/\sqrt{\theta_2\theta_{3j}}\right),$$

where $\theta_2 > 0$, $\theta_{31} = 1$, $\theta_{3j} > 0$ for $i = 1, \cdots, n_j$,

$j = 2, \cdots, J$ with $J > 1$. Let

$$T_n = \left(\sum_{j=1}^{J}\sum_{i=1}^{n_j} X_{ji}\right) \Big/ \left(\sum_{j=1}^{J} n_j\right), \quad r = g,$$

and

$$f\left(x_{ji} \mid \theta\right) = \sqrt{\theta_2}^{-1} \varphi\left(\left(x_{ji} - \theta\right)_1 / \sqrt{\theta_2}\right)$$

for some $\theta_2 > 0$. Thus $\mu_r(\theta) = \mu_f(\theta) = \theta_1$,

$$\sigma_r^2(\theta) = \theta_2 \left(\sum_{j=1}^{J} w_j \theta_{3j}\right) \Big/ \left(\sum_{j=1}^{J} n_j\right),$$

and $\sigma_f^2(\theta) = \theta_2 \Big/ \left(\sum_{j=1}^{J} n_j\right)$ for $j = 1, \cdots, J$,

where $w_j = n_j \Big/ \left(\sum_{j=1}^{J} n_j\right)$. Here T_n is the MLE of θ_1 under f, while the MLE of θ_1 under r is

$$\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j} X_{ji} \left\{\hat{\theta}_2(U_n)\hat{\theta}_{3j}(U_n)\right\}^{-1}\right] \Big/ \left[\sum_{j=1}^{J} n_j \left\{\hat{\theta}_2(U_n)\hat{\theta}_{3j}(U_n)\right\}^{-1}\right],$$

where $\hat{\theta}_{3j}(U_n)$ is the MLE of θ_{3j} under r. Also,

$$2KLD_t\left(r, f \mid U_n\right) - Q\left(\hat{\sigma}_r^2(U_n) \Big/ \hat{\sigma}_f^2(U_n)\right) = o_p(1).$$

We conduct numerical calculation for the following combination $\theta_2 = 2.25$, $\theta_{32} = 4, 25$ with $(n_1, n_2) = (5000, 5000)$ and $(n_1, n_2) = (2000, 8000)$. Since the 95% percentiles of $\hat{\sigma}_r^2(U_n) \Big/ \hat{\sigma}_f^2(U_n)$ do not exceed 1 under all situations (hence $Pr\left(KLD_t\left(r, f \mid U_n\right) > 0\right) > 0.95$), it implies that in the ory KLD should be able to distinguish f from r. However, the numerical values of KLD only deviate from 0 by $K \times 10^{-6}$ for $1 < K < 10$. Thus in practice, KLD can hardly be differentiated from $WPPP_r(U_n)$ which converges to 0.5 in probability $\mu_r(\theta) = \mu_f(\theta)$.

Example 3.5 (Non-nested). Assume

$$X_i \overset{i.i.d.}{\sim} g\left(x_i \mid \theta\right) = \left(\Gamma\left(\theta_2/2\right)\sqrt{\pi\theta_2}\right)^{-1} \Gamma\left(\left(\theta_2+1\right)/2\right) \left(1 + \left(x - \theta_1\right)^2 / \theta_2\right)^{-(1+\theta_2)/2},$$

where $\theta_2 > 2$. Let $T_n = \bar{X}$. Suppose that $\left(X_1, \cdots, X_n\right)$ are fitted by $r = g$ and $f\left(x_i \mid \theta\right) = \varphi\left(X_i - \theta_1\right)$. Then

$$\mu_f(\theta) = \mu_r(\theta) = \theta_1, \quad \sigma_r^2(\theta) = \theta_2 \Big/ \left(\theta_2 - 2\right),$$

and $\sigma_f^2(\theta) = 1$. Since

$$KLD_t\left(r, f \mid U_n\right) - Q\left(\hat{\theta}_2(U_n) \Big/ \left(\hat{\theta}_2(U_n) - 2\right)\right) = o_p(1)$$

for all θ_2 with $KLD_t\left(r, f \mid U_n\right)$ converging to 0 with probability 1 if and only if $\kappa = \infty$, $KLD_t\left(r, f \mid U_n\right)$ can detect model discrepancy in the dispersion parameter. In contrast, $WPPP_r(U_n)$ converges to 0.5 in probability, which is not sensitive to the discrepancy in the dispersion parameter.

Example 3.6 (Non-nested). Assume

$$X_i \overset{i.i.d.}{\sim} g\left(x_i \mid \theta\right) = \left(\Gamma\left(\theta_2/2\right)\sqrt{\pi\theta_2}\right)^{-1} \Gamma\left(\left(\theta_2+1\right)/2\right) \left(1 + \left(x - \theta_1\right)^2 / \theta_2\right)^{-(1+\theta_2)/2},$$

where $\theta_2 > 2$, $\theta_3 > 0$, and $0 < \theta_4 < 1$. Suppose that the empirical variance of $\left\{X_i : i = 1, \cdots, n\right\}$ is greater than 1. Consider fitting the data by

$$r\left(x_i \mid \theta\right) = \left(\Gamma\left(\theta_2/2\right)\sqrt{\pi\theta_2}\right)^{-1} \Gamma\left(\left(\theta_2+1\right)/2\right) \left(1 + \left(x - \theta_1\right)^2 / \theta_2\right)^{-(1+\theta_2)/2},$$

or

$$f\left(x_i \mid \theta\right) = \left(\sqrt{\theta_2/\left(\theta_2-2\right)}\right)^{-1} \varphi\left(\left(x_i - \theta_1\right) \Big/ \sqrt{\theta_2/\left(\theta_2-2\right)}\right),$$

where $\theta_2 > 2$. Let $T_n = \bar{X}$. Then $\hat{\mu}_f(U_n) = T_n$ and $\hat{\mu}_r(U_n) = \sum_{i=1}^{m} w_i(U_n) X_i$, where

$$w_i(U_n) = \left\{1 + \left(X_i - \hat{\theta}_1(U_n)\right)^2 \Big/ \hat{\theta}_2(U_n)\right\}^{-1} \Bigg/ \left\{\sum_{i'=1}^{n}\left(1 + \left(X_{i'} - \hat{\theta}_1(U_n)\right)^2 \Big/ \hat{\theta}_2(U_n)\right)^{-1}\right\}$$

Since both $\hat{\mu}_r(U_n)$ and $\hat{\mu}_f(U_n)$ are unbiased estimators for $E(T_n)$ and $var\left(\hat{\mu}_r(U_n) - \hat{\mu}_f(U_n)\right)$ converges to 0 in probability,

$$KLD_t\left(r, f \mid U_n\right) = Q\left(\hat{\sigma}_r^2(U_n) \Big/ \hat{\sigma}_f^2(U_n)\right) + o_p(1),$$

where $\hat{\sigma}_f^2(U_n) = \sum_{i=1}^{n}(X_i - T_n)^2 / n$, while $\hat{\sigma}_r^2(U_n)$ can not be obtained in a closed form. Thus $KLD_t\left(r, f \mid U_n\right)$ is evaluated numerically using eight combinations given in the table below. The results suggest that the degree to which $KLD_t\left(r, f \mid U_n\right)$ can distinguish f from r varies by situation: closer $|\theta_3 - 1|$ and $\min\{\theta_4, 1-\theta_4\}$ are to 0, closer is $KLD_t\left(r, f \mid U_n\right)$ to 0. In contrast, $WPPP_f(U_n)$ converges to 0.5 in probability, (**see Table 1**).

Example 3.7 below shows that the asymptotic relationship between $KLD_t\left(r, f \mid U_n\right)$ and $WPPP_r(U_n)$ does not hold in the sense of Theorem 2 due to violation of the asymptotic normality assumption specified in (3) a and (4).

Table 1. Simulation result of example 3.6.

(θ_2, θ_3)	$\theta_4 = 0.5$	$\theta_4 = 0.2$
(6,2)	1.58×10^{-4}	3.08×10^{-6}
(6,10/3)	1.10×10^{-1}	6.27×10^{-5}
(2.5,2)	1.58×10^{-4}	1.47×10^{-6}
(2.5,10/3)	1.12×10^{-1}	6.01×10^{-5}

Example 3.7 (Nested). Consider

$$X_i \overset{i.i.d.}{\sim} g(x_i \mid \theta) = \theta^{-1}\exp(-x_i/\theta).$$

Suppose that $\{X_i : i = 1,\cdots,n\}$ are fitted by $r = g$ and $f(x_i \mid \theta) = \exp(-x_i)$. Let $T_n = \min\{X_1,\cdots,X_n\}$. Then

$$g(t_n \mid \theta) = n\theta^{-1}\exp(-nt_n/\theta)$$

and

$$f(t_n \mid \theta) = n\exp(-nt_n).$$

Then

$$WPPP_r(U_n) = WPPP_r(\bar{X}_n)$$
$$= E^r(Pr^f(T_n^{pred} < T_n \mid \hat{\theta}_f = 1) \mid \bar{X}_n)$$
$$= \int\left[\int\left\{\int_{-\infty}^{t_n} n\exp(-nx)\mathrm{d}x\right\}\frac{n}{\theta}\exp(-nt_n/\theta)\mathrm{d}t_n\right]\pi_r(\theta \mid \bar{X}_n)\mathrm{d}\theta$$
$$= \int\left[\int\{1-\exp(-nt_n)\}\frac{n}{\theta}\exp(-nt_n/\theta)\mathrm{d}t_n\right]\pi_r(\theta \mid \bar{X}_n)\mathrm{d}\theta$$
$$= \int\frac{\theta}{\theta+1}\pi_r(\theta \mid \bar{X}_n)\mathrm{d}\theta$$

and

$$KLD_t(r, f \mid U_n) = KLD_t(r, f \mid \bar{X}_n)$$
$$= \left[\log\left\{\int\frac{n/\theta\exp(-nt_n/\theta)}{n\exp(-nt_n)}\right\}\frac{n}{\theta}\exp(-nt_n/\theta)\mathrm{d}t_n\right]$$
$$\pi_r(\theta \mid \bar{X}_n)\mathrm{d}\theta$$
$$= \int\left[\{-\log(\theta) + n(1-1/\theta)t_n\}\frac{n}{\theta}\exp(-nt_n/\theta)\mathrm{d}t_n\right]$$
$$\pi_r(\theta \mid \bar{X}_n)\mathrm{d}\theta$$
$$= \int Q(\theta)\pi_r(\theta \mid \bar{X}_n)\mathrm{d}\theta$$

That is,

$$WPPP_r(U_n) - \frac{\bar{X}_n}{\bar{X}_n+1} = o_p(1)$$

and

$$KLD_t(r, f \mid U_n) - Q(\bar{X}_n) = o_p(1).$$

In this example, both $KLD_t(r, f \mid U_n)$ and $WPPP_r(U_n)$ can differentiate r from f by an $O_p(n^{-1/2})$ term despite that the asymptotic relationship between $KLD_t(r, f \mid U_n)$ and $WPPP_r(U_n)$ does not hold in the sense of Theorem 2 due to the violation of the asymptotic normality assumption.

4. Application of $KLD_t(r, f \mid U_n)$ to Diabetes Studies

In this section, we apply both $KLD_t(r, f \mid U_n)$ and $WPPP_r(U_n)$ to compare non-nested models in two studies of diabetes. In these applications, we assess the model fit to the entire dataset due to its clinical implication. Here the selected diagnostic statistic is a multivariate statistic of $O_p(1)$ variables, and it does not meet all the regularity conditions specified in Section 3.

4.1. Study I: Analysis of Change in Glucose in Veterans with Type 2 Diabetes

Study I originated from a clinical cohort of 507 veterans with type 2 diabetes who had poor glucose control (indicated by glycosylated hemoglobin A1c or HbA1c greater than 6.5) at the baseline (fiscal year 1999), and were all treated by metformin as the mono oral glucose-lowering agent. As the literature suggested that the glucose-lowering response due to metformin may vary by an individuals obesity status, the goal of our study was to compare models that assessed whether obesity was associated with the net change in glucose level between baseline and the end of 5-year follow-up. In this study, the empirical mean of the net change in HbA1c over the 5 year period was similar between the obese vs. non-obese groups (–0.498 vs. –0.379), yet the empirical variance was greater in the obese group (1.207 vs. 0.865). Also, the distribution of HbA1c was reasonably symmetric. Thus we considered two candidate models for fitting the HbA1c change: a mixture of normals vs. a t-distribution (note that the overall empirical variance of HbA1c change is 1.03); that is,

$$r(x_i \mid \theta) = \theta_4\varphi\left((x_i-\theta_1)\Big/\sqrt{\theta_2/(\theta_2-2)}\right)\Big/$$
$$\sqrt{\theta_2/(\theta_2-2)} + (1-\theta_4)\varphi\left((x_i-\theta_1)\Big/\sqrt{\theta_3\theta_2/(\theta_2-2)}\right)\Big/$$
$$\sqrt{\theta_3\theta_2/(\theta_2-2)}$$

where $\theta_2 > 2$, $\theta_3 > 0$, and $\theta_4 = 0.487$ (% in the obesity group) vs.

$$f(x_i \mid \theta) = \Gamma\left((\theta_2+1)/2\right)\Big/\left\{\Gamma(\theta_2/2)\sqrt{\pi\theta_2}\right\}$$
$$\left(1+(x-\theta_1)^2/\theta_2\right)^{-(1+\theta_2)/2},$$

where $\theta_2 > 1$. The calculated $KLD_t(r, f \mid U_n)$ was 5.96, suggesting that model r provided a modest better fit to the data compared to model f. This result was also consistent with **Figures 1** and **2** which contrasted the empirical quantiles with predicted quantiles under r and f. Note that both $\hat{\mu}_r(U_n)$ and $\hat{\mu}_f(U_n)$ are unbiased estimators for $E(T_n)$, where $\hat{\mu}_f(U_n) = T_n$ and $\hat{\mu}_r(U_n) = \sum_{i=1}^n w_i(U_n)X_i$ with

$$w_i(U_n) = \left\{1+\left(X_i-\hat{\theta}_1(U_n)\right)^2/\hat{\theta}_2(U_n)\right\}^{-1}\Big/$$
$$\left\{\sum_{i'=1}^n\left(1+\left(X_{i'}-\hat{\theta}_1(U_n)\right)^2/\hat{\theta}_2(U_n)\right)^{-1}\right\}.$$

Thus the model discrepancy assessed by $KLD_t\left(r, f \middle| U_n\right)$ is primarily attributed to the difference in the variance assumption between r and f (as evident in **Figures 1** and **2**). In contrast, WPPP = 0.522 suggested that the overall fit were similar between the two models since the estimated net change in HbA1c was similar between these two models.

4.2. Study II: Analysis of the Functioning Score in Older Adults with Diabetes

Study II arose from the subset of 119 participants with diabetes in the San Antonio Longitudinal Study of Aging (SALSA), a community-based study of the disablement process in Mexican American and European American older adults. Details of the SALSA study design, sampling approach, recruitment and field procedures have been described previously (Hazuda *et al.* [11]). The goal of our analyses was to compare models that assessed whether glucose control trajectory class (poorer vs. better) was associated with the lower-extremity physical functional limitation score during the first follow-up period. The lower-extremity physical functional limitation score was measured by the Short Physical Performance Battery or SPPB, which is a well-established, validated measure of physical functioning. SPPB score is a sum of three items: 8-foot walking times, repeated chair stands, and balance scores, each being a 5-point liker scale 0 - 4. Hence the SPPB score is discrete in nature with a range of 0 - 12. Higher SPPB scores indicate better performance and less functional limitation. Exploratory data analyses suggested that the empirical variance of SPPB (15.60 vs. 14.33) was greater than the mean (7.23 vs. 8.02) in both glucose control classes. Also, due to the

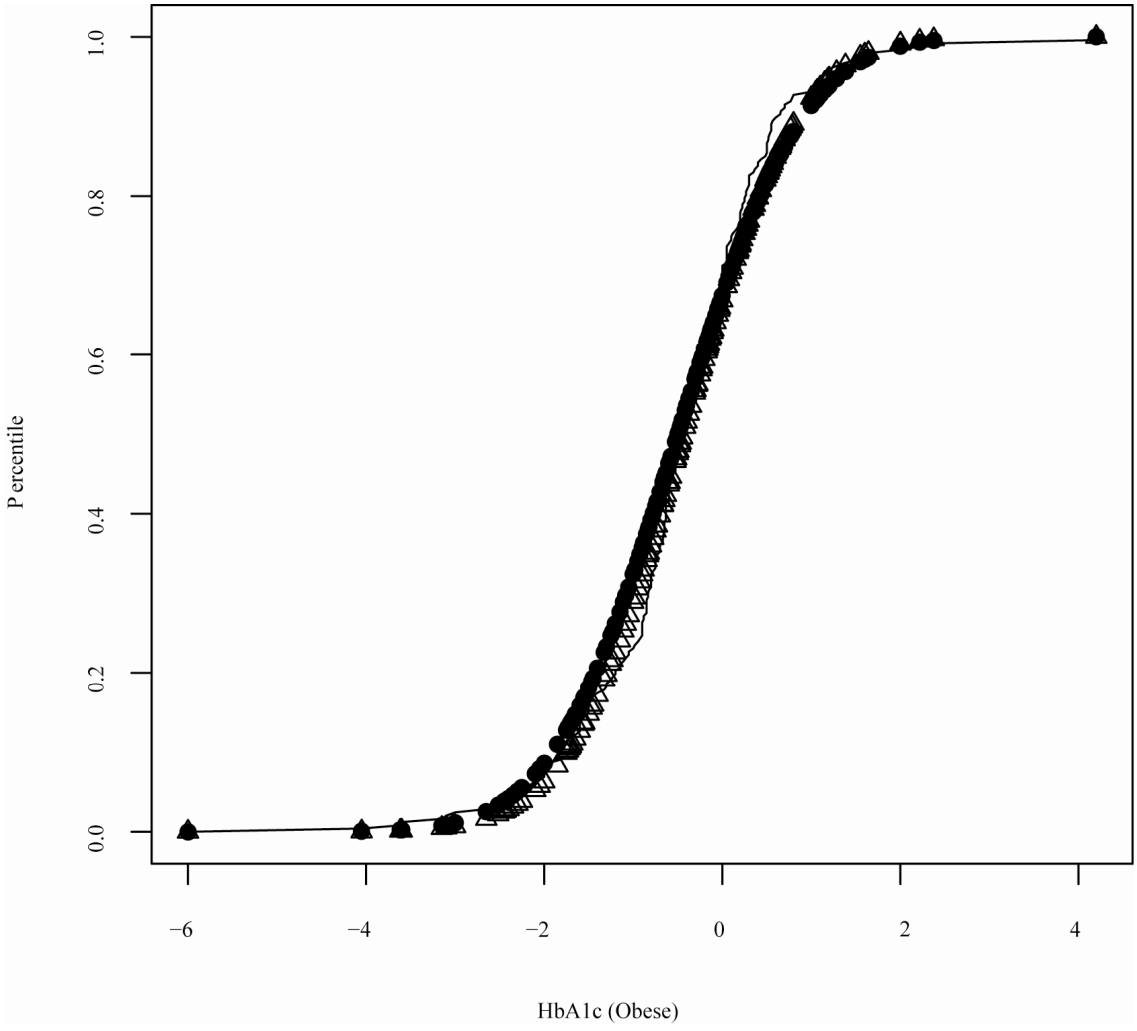

Figure 1. Quantile plot for HbA1c for T2DM participants with obesity in VA (open triangle: quantile estimates based on the model assuming a mixture of normal distributions solid circle: quantile estimates based on the model assuming a t-distribution solid line: empirical quantiles).

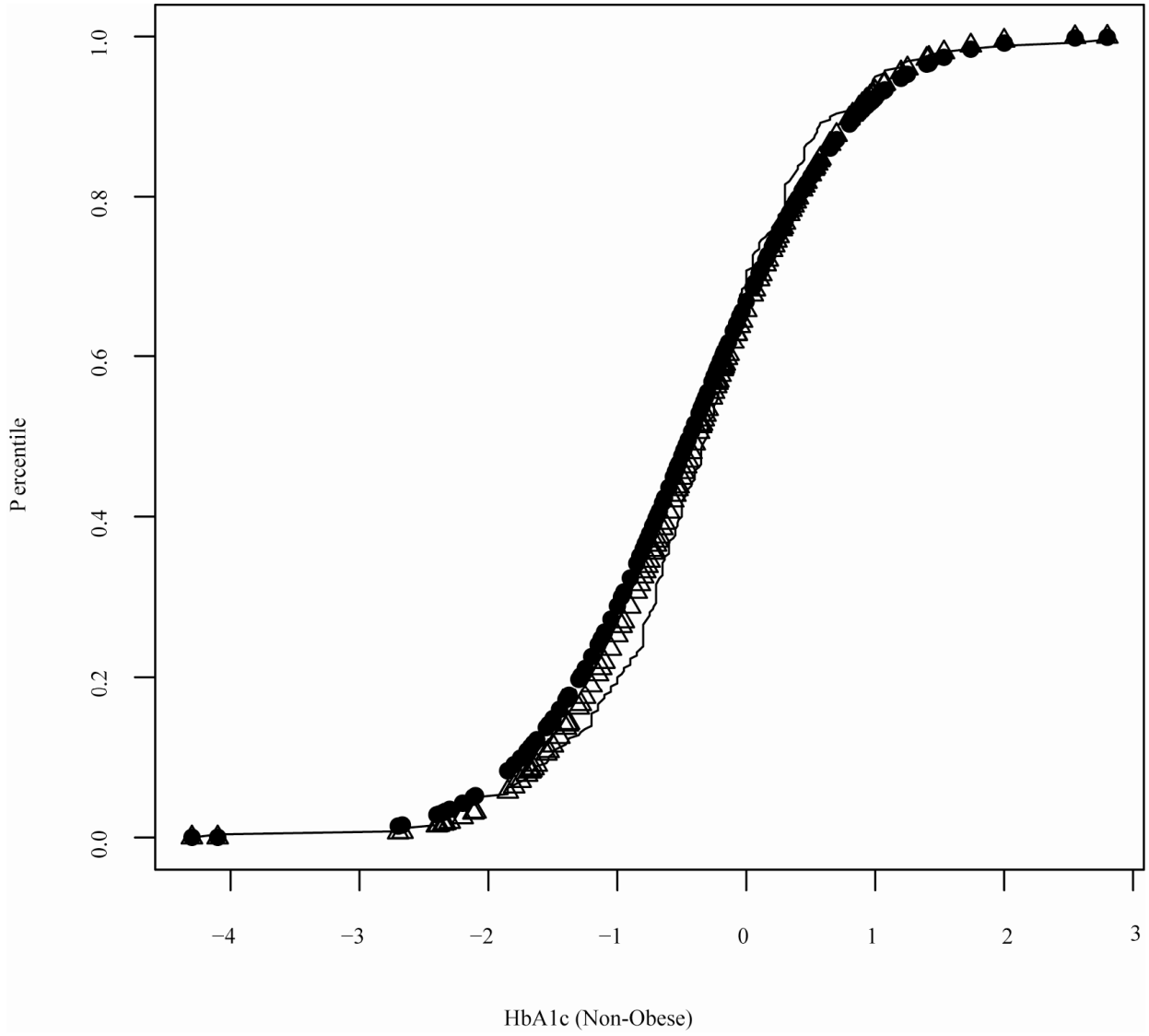

Figure 2. Quantile plot for HbA1c for T2DM participants without obesity in VA (open triangle: quantile estimates based on the model assuming a mixture of normal distributions solid circle: quantile estimates based on the model assuming a t-distribution solid line: empirical quantiles).

left-skewedness of the distribution of SPPB, we considered models fit to the reversed SPPB, *i.e.*, $X_i = (12 - SPPB_i)$. Given the nature of SPPB distribution, we compare two candidate models:

$$r(x_i \mid \theta) = \Gamma\left(\vartheta_\lambda\right)\Gamma(x_i))^{-1}\Gamma(\theta_\lambda + x_i)\theta^{x_i}(1-\theta)^{\theta_\lambda}$$

with $\theta_\lambda > 0$ and

$$f\left(x_i \mid \theta\right) = \exp\left\{-\theta/(1-\theta)\right\}\left\{\theta/(1-\theta)\right\}^{x_i}/(x_i!).$$

The calculated $KLD_t\left(r, f \mid U_n\right)$ is 32.63, suggesting that r has a much better fit than f, which is coherent to the quantile plot shown in **Figures 3** and **4**. Since both r and f yielded similar estimates of $E(X_i)$, the model discrepancy assessed by $KLD_t\left(r, f \mid U_n\right)$ could primarily be attributed to the difference in variance estimation between r and f (as evident in **Figure 2**). In contrast, WPPP = 0.539 suggested similar fit between r and f as expected due to similar estimates of $E(X_i)$ under both r and f.

5. Discussion

This paper considers the G-R-A KLD as given in (2). This KLD is appropriate for quantifying information discrepancy regarding T_n contained in the competing models r and f. We derive the asymptotic property of the G-R-A KLD in Theorem 1, and its relationship to a weighted posterior predictive p-value (WPPP) in Theo-

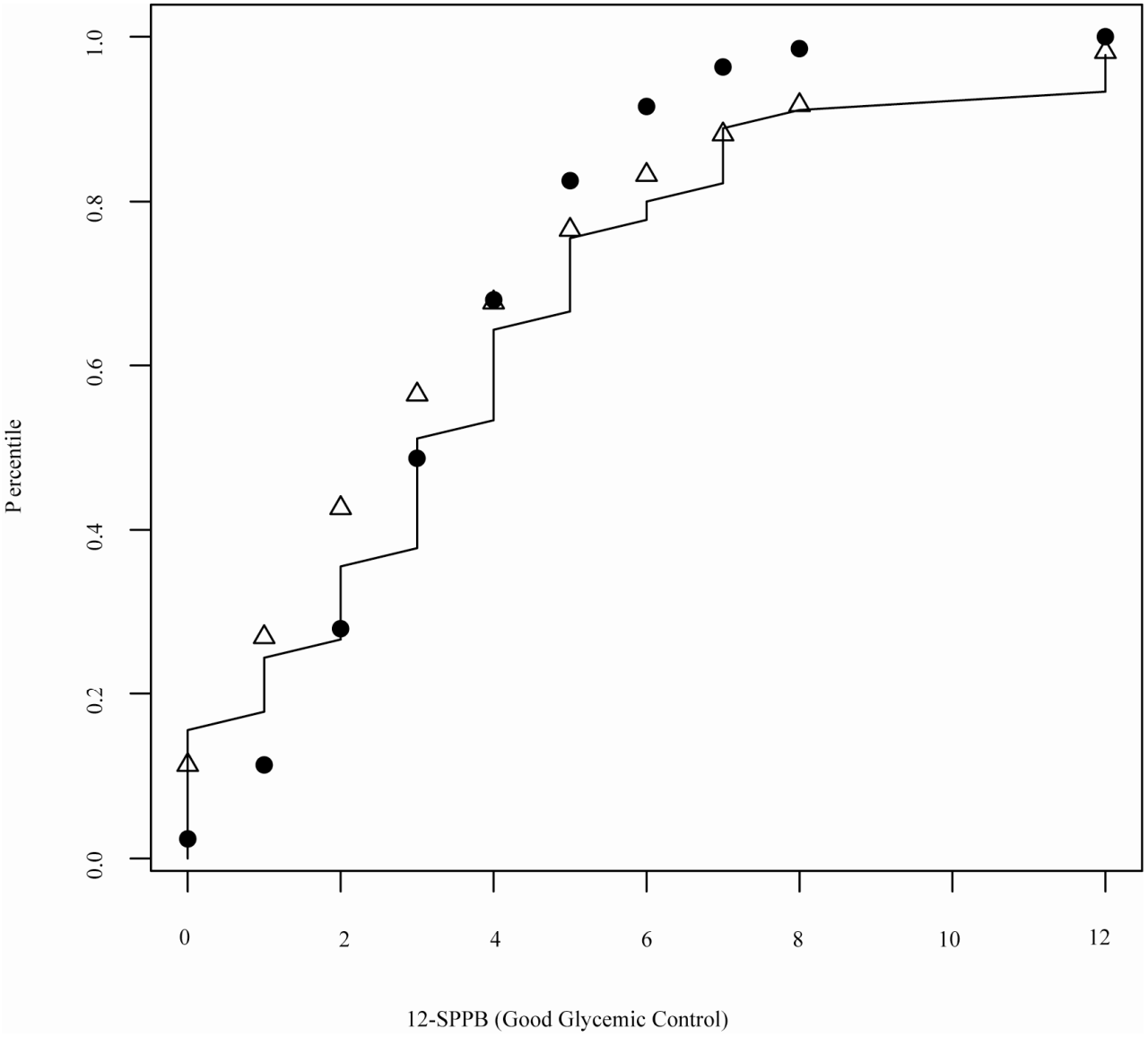

Figure 3. Quantile plot for SPPB for T2DM participants with good glycemic control in SALSA (open triangle: quantile estimates based on the negative binomial model; solid circle: quantile estimates based on the Poisson model; solid line: empirical quantiles).

rem 2. However, our results would need further refinement when the normality assumptions given in (3) and (4) are not suitable (see Example 3.7). As shown in Section 4, model comparison in medical research may rely on the fit to a multidimensional statistic. Although the results in Theorem 1 holds for a multivariate T_n with a fixed dimension, further investigation is needed to assess the property of our proposed KLD for situations when the dimension of T_n increases with n. The KLD proposed herein usually provides the relative fit between competing models. Thus, for the purpose of assessing model adequacy (rather than relative model fit), a KLD should be used in conjunction with absolute model departure indices such as posterior predictive p-values or residuals. Nevertheless, a KLD can be a measure of the absolute fit of model f when the superior model r is the true model, and therefore can be used for checking model adequacy.

6. Acknowledgements

Wang's research is partially supported by NIDDK K25-DK075092; Ghosh's research is partially supported by NSF. We thank Dr. Hazuda for providing data from the SALSA study supported by NIA R01-AG10444 and NIA R01-AG16518.

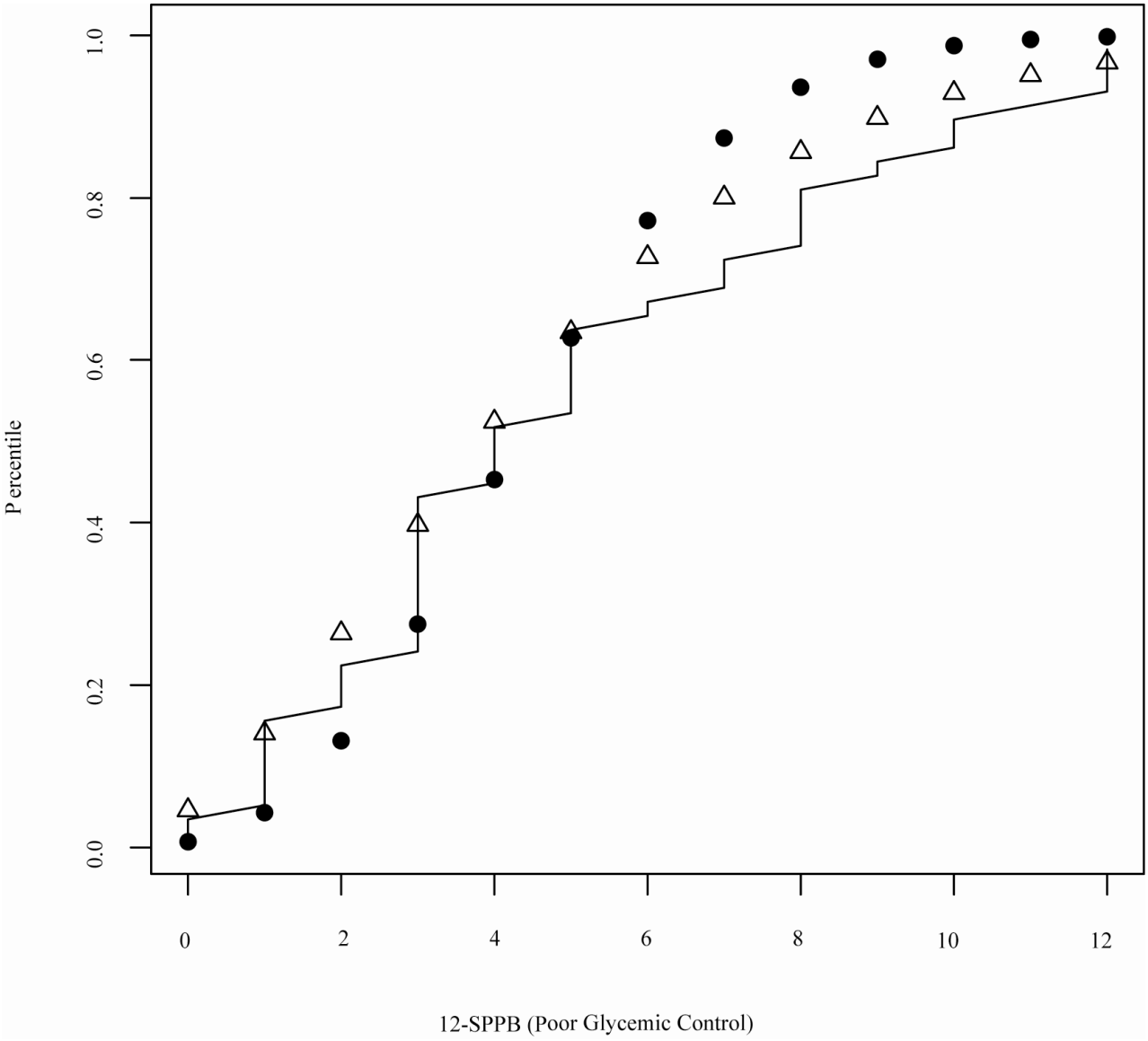

Figure 4. Quantile plot for SPPB for T2DM participants with poor glycemic control in SALSA (open triangle: quantile estimates based on the negative binomial model; solid circle: quantile estimates based on the Poisson model; solid line: empirical quantiles).

7. References

[1] C. Goutis and C. P. Robert, "Model Choice in Generalised Linear Models: A Bayesian Approach via Kullback-Leibler Projections," *Biometrika*, Vol. 85, No. 1, 1998, pp. 29-37. doi:10.1093/biomet/85.1.29

[2] H. Akiaike, "A New Look at the Statistical Identification Model," *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, 1974, pp. 716-723. doi:10.1109/TAC.1974.1100705

[3] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, Vol. 27, 1948, pp. 379-423 and pp. 623-656.

[4] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, Vol. 22, No. 1, 1951, pp. 79-86. doi:10.1214/aoms/1177729694

[5] D. V. Lindley, "On a Measure of the Information Provided by an Experiment," *The Annals of Mathematical Statistics*, Vol. 27, No. 4, 1956, pp. 986-1005. doi:10.1214/aoms/1177728069

[6] J. M. Bernardo, "Expected Information as Expected Utility," *The Annals of Statistics*, Vol. 7, No. 3, 1979, pp. 686-690. doi:10.1214/aos/1176344689

[7] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, Vol. 6, No. 2, 1978, pp. 461-464. doi:10.1214/aos/1176344136

[8] I. Guttman, "The Use of the Concept of a Future Observation in Goodness-of-Fit Problems," *Journal of the Ro-*

yal Statistical Society B, Vol. 29, No. 1, 1967, pp. 83-100.

[9] D. B. Rubin, "Bayesianly Justifiable and Relevant Frequency Calculations for the Applies Statistician," *Annals of Statistics*, Vol. 12, No. 4, 1984, pp. 1151-1172. doi:10.1214/aos/1176346785

[10] A. Gelman, J. Carlin, H. S. Stern and D. Rubin, "Bayesian Data Analysis," Chapman and Hall, London, 1996.

[11] H. P. Hazuda, S. M. Haffner, M. P. Stern and C. W. Eifler, "Effects of Acculturation and Socioeconomic Status on Obesity and Diabetes in Mexican Americans: The San Antonio Heart Study," *American Journal of Epidemiology*, Vol. 128, No. 6, 1988, pp. 1289-1301.

[12] S. Ghosal and T. Samanta, "Expansion of Bayes Risk for Entropy Loss and Reference Prior in Nonregular Cases," *Statistics and Decisions*, Vol. 15, 1997, pp. 129-140.

[13] I. Ibragimov and R. Hasminskii, "Statistical Estimation: Asymptotic Theory," Springler-Verlag, New York, 1980.

Appendix 1. Proof of Theorem 1

Under (3) and (4), it can be shown that

$$\begin{aligned}
&2KLD_t\left(r, f \mid U_n\right)\\
&=\iint \log\left(\frac{r\left(t_n \mid \theta\right)}{f\left(t_n \mid \hat{\theta}_f\right)}\right) r\left(t_n \mid \theta\right)\pi_r\left(\theta \mid U_n\right)\mathrm{d}t_n\mathrm{d}\theta\\
&=\iint\left[\left\{-\log\left(\frac{\sigma_r^2(\theta)}{\hat{\sigma}_f^2(\theta)}\right)-\frac{n\left\{t_n-\mu_r(\theta)\right\}^2}{\sigma_r^2(\theta)}+\frac{n\left\{t_n-\mu_r(\theta)\right\}^2}{\hat{\sigma}_f^2(\theta)}\right.\right.\\
&\left.+\frac{n\left\{\mu_r(\theta)-\hat{\mu}_f(\theta)\right\}^2}{\hat{\sigma}_f^2(\theta)}\right\}\times\sigma_r^{-1}(\theta)\\
&\left.\varphi\left(\frac{\sqrt{n}\left\{t_n-\mu_r(\theta)\right\}}{\sigma_r(\theta)}\right)\mathrm{d}t_n+O(n^{-1/2})\right]\pi_r\left(\theta \mid U_n\right)\mathrm{d}\theta
\end{aligned}$$

Applying an argument similar to that in Ghosal and Samanta [12] (see also Ibragimov and Hasminskii [13]), (A1)-(A5) gives

$$\begin{aligned}
2KLD_t\left(r, f \mid U_n\right)&=\int\left[Q\left(\frac{\sigma_r^2(\theta)}{\hat{\sigma}_f^2(\theta)}\right)\right.\\
&\left.+\frac{n\left(\hat{\mu}_f(\theta)-\mu_r(\theta)\right)^2}{\hat{\sigma}_f^2(\theta)}+O(n^{-1/2})\right]\pi_r\left(\theta \mid U_n\right)\mathrm{d}\theta.
\end{aligned}\tag{11}$$

When $\mu_r(\theta)\neq\mu_f(\theta)$, $\int n(\hat{\mu}_f(\theta)-\mu_r(\theta))^2/\hat{\sigma}_f^2(\theta)\pi_r(\theta \mid U_n)\mathrm{d}\theta$ is the dominating term in (11). Thus by (A1)-(A5) and the arguments similar to those in Ghosal and Samanta [12], one gets

$$\begin{aligned}
&\int\frac{n(\hat{\mu}_f(\theta)-\mu_r(\theta))^2}{\hat{\sigma}_f^2(\theta)}\pi_r\left(\theta \mid U_n\right)\mathrm{d}\theta+O_p(n^{-1/2})\\
&=\int\left\{\frac{n\left(\hat{\mu}_f(U_n)-\hat{\mu}_r(U_n)\right)^2}{\hat{\sigma}_f^2(U_n)}+O_p(n^{1/2})\right\}\\
&\pi_r(\theta \mid U_n)\mathrm{d}\theta+O_p(n^{-1/2})\\
&=\frac{n(\hat{\mu}_f(U_n)-\hat{\mu}_r(U_n))^2}{\hat{\sigma}_f^2(U_n)}+O_p(n^{1/2}).
\end{aligned}\tag{12}$$

If $\mu_f(\theta)=\mu_r(\theta)$, $\sigma_r^2(\theta)\neq\sigma_f^2(\theta)$, then (11) becomes

$$\begin{aligned}
&\int\left\{Q\left(\frac{\sigma_r^2(\theta)}{\hat{\sigma}_f^2(\theta)}\right)+O(n^{-1/2})\right\}\pi_r\left(\theta \mid U_n\right)\mathrm{d}\theta\\
&=Q\left(\frac{\hat{\sigma}_r^2(U_n)}{\hat{\sigma}_f^2(U_n)}\right)+O_p(n^{-1/2}).
\end{aligned}\tag{13}$$

Therefore,

$$2KLD_t\left(r, f \mid U_n\right)/n-\frac{\left\{\hat{\mu}_f(U_n)-\hat{\mu}_r(U_n)\right\}^2}{\hat{\sigma}_f^2(U_n)}=o_p(1)$$

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

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

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

Appendix 2. Proof of Theorem 2

Write z for the realization of Z, a $N(0,1)$ random variable. Then

$$\begin{aligned}
&\iint_{-\infty}^{t_n} f^*\left(y \mid \theta\right) r^*\left(t_n \mid \theta\right)\mathrm{d}y\mathrm{d}t_n\\
&=\int\left\{\Phi\left[\frac{\sqrt{n}\left(t_n-\mu_r(\theta)\right)}{\sigma_r(\theta)}\frac{\sigma_r(\theta)}{\hat{\sigma}_f(\theta)}\right.\right.
\end{aligned}$$

$$+\frac{\sqrt{n}\left(\mu_r(\theta)-\hat{\mu}_f(\theta)\right)}{\hat{\sigma}_f(\theta)}\Bigg]+O(n^{-1/2})\Bigg\}r^*\left(t_n\mid\theta\right)\mathrm{d}t_n$$

$$=\int\Phi\left(\frac{\sigma_r(\theta)}{\sigma_f(\theta)}z+\frac{\sqrt{n}\left(\mu_r(\theta)-\hat{\mu}_f(\theta)\right)}{\hat{\sigma}_f(\theta)}\right)\varphi(z)\mathrm{d}z+O(n^{-1/2}).$$

The second equality follows (A1)-(A5) and the argument similar to that used in Ghosal and Samanta [12]. Let Y be a $N(0,1)$ random variable distributed independently of Z. Then one can rewrite the above integral as

$$\int Pr\left(Y-z\frac{\sigma_r(\theta)}{\hat{\sigma}_f(\theta)}\le\frac{\sqrt{n}\left(\mu_r(\theta)-\hat{\mu}_f(\theta)\right)}{\hat{\sigma}_f(\theta)}\right)\varphi(z)\mathrm{d}z+O(n^{-1/2})$$

$$=Pr\left(Y-Z\frac{\sigma_r(\theta)}{\sigma_f(\theta)}\le\frac{\sqrt{n}\left(\mu_r(\theta)-\hat{\mu}_f(\theta)\right)}{\hat{\sigma}_f(\theta)}\right)+O(n^{-1/2})$$

$$=\Phi\left(\frac{\sqrt{n}\left(\mu_r(\theta)-\hat{\mu}_f(\theta)\right)\big/\hat{\sigma}_f(\theta)}{\sqrt{1+\sigma_r^2(\theta)/\hat{\sigma}_f^2(\theta)}}\right)+O(n^{-1/2})$$

$$=\Phi\left(\frac{\sqrt{n}\left(\mu_r(\theta)-\hat{\mu}_f(\theta)\right)}{\sqrt{\hat{\sigma}_f^2(\theta)+\sigma_r^2(\theta)}}\right)+O(n^{-1/2}).$$

Therefore, applying the arguments similar to those in Ghosal and Samanta [12] under (A1)-(A5) yields

$$WPPP_r(U_n)$$

$$=\iiint_{-\infty}^{t_n}f^*\left(y\mid\theta\right)r^*\left(t_n\mid\theta\right)\pi_r\left(\theta\mid U_n\right)\mathrm{d}y\mathrm{d}t_n\mathrm{d}\theta$$

$$=\int\left\{\Phi\left(\frac{\sqrt{n}\left(\mu_r(\theta)-\hat{\mu}_f(\theta)\right)}{\sqrt{\hat{\sigma}_f^2(\theta)+\sigma_r^2(\theta)}}\right)+O(n^{-1/2})\right\}\pi_r\left(\theta\mid U_n\right)\mathrm{d}\theta$$

$$=\Phi\left(\frac{\sqrt{n}\left(\hat{\mu}_r(U_n)-\hat{\mu}_f(U_n)\right)}{\sqrt{\hat{\sigma}_f^2(U_n)+\hat{\sigma}_r^2(U_n)}}+O_p(1)\right)+O_p(n^{-1/2}) \tag{14}$$

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

$$WPPP_r(U_n)=\int\Phi(z)\varphi(z)\mathrm{d}z+O_p(n^{-1/2})=0.5+O_p(n^{-1/2}) \tag{15}$$

when $\mu_r(\theta)=\mu_f(\theta)$. This completes the proof.

