A Kullback-Leibler Divergence for Bayesian Model Comparison with Applications to Diabetes Studies

Chen-Pin Wang, UTHSCSA
Malay Ghosh, U. Florida

Lehmann Symposium, May 9, 2011

Background

•KLD: the expected (with respect to the reference model) logarithm of the ratio of the probability density functions (p.d.f.'s) of two models.

$$\int \log \left(\frac{r(t_n|\theta)}{f(t_n|\theta)} \right) r(t_n|\theta) \, dt_n$$

•KLD: a measure of the discrepancy of information about θ contained in the data revealed by two competing models (K-L; Lindley; Bernardo; Akaike; Schwarz; Goutis and Robert).

• Challenge in the Bayesian framework:
– identify priors that are compatible under the competing models
– the resulting integrated likelihoods are proper.

G-R KLD

- Remedy: The Kullback-Leibler projection by Goutis and Robert (1998), or G-R KLD: the inf. KLD between the likelihood under the reference model and all possible likelihoods arising from the competing model.

- G-R KLD is the KLD between the reference model and the competing model evaluated at its MLE if the reference model is correctly specified (ref. Akaike 1974).

- G-R KLD overcomes the challenges associated with prior elicitation in calculating KLD under the Bayesian framework.

G-R KLD

- The Bayesian estimate of G-R KLD: integrating the G-R KLD with respect to the posterior distribution of model parameters under the reference model.

– Bayesian estimate of G-R KLD is not subject to impropriety of the prior as long as the posterior under the reference model is proper.

– G-R KLD is suitable for comparing the predictivity of the competing models.

– G-R KLD was originally developed for comparing nested GLM with a known true model, and its extension to general model comparison remains limited.

Proposed KLD

$$\int \log\left(\frac{r(t_n|\theta)}{f(t_n|\widehat{\theta}_f)}\right) r(t_n|\theta)\ dt_n. \tag{1}$$

Bayes estimate of (1):

$$\int \left\{ \int \log\left(\frac{r(t_n|\theta)}{f(t_n|\widehat{\theta}_f)}\right) r(t_n|\theta)\ dt_n \right\} \pi(\theta|U_n). \tag{2}$$

Objective: To study the property of KLD estimate given in (2).

Notations

- X_i's are i.i.d. originating from model g governed by $\theta \in \Theta$.

- $T_n = T(X_1, \cdots, X_n)$: the statistic for model diagnostics.

- Two competing models: r for the reference model and f for the fitted model.

- Assume that prior $\pi_r(\theta)$ leads to proper posterior under r.

Our proposed KLD

- $KLD_t(r, f|\theta)$ quantifies the relative model fit for statistic T_n between models r and f.

- $KLD_t(r, f|\theta)$ is identical to G-R KLD when the reference model r is the correct model.

- $KLD_t(r, f|\theta)$ is not necessarily the same as the G-R KLD.

- $KLD_t(r, f|\theta)$ needs no additional adjustment for non-nested situations.

- $KLD_t(r, f|\theta)$ is more practical than G-R KLD.

Regularity Conditions I

$(A1)$ 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

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

and

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

for k=1, 2, 3.

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

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

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

Regularity Conditions II

$$(A3) \qquad E_r\left[\sup_{\theta'\in(\theta-\delta,\theta+\delta)} \log r(x|\theta')\middle|\,\theta\right] \to E_r[\log r(x|\theta)] \;\; as \;\; \delta \to 0;$$

$$E_f\left[\sup_{\theta'\in(\theta-\delta,\theta+\delta)} \log f(x|\theta')\middle|\,\theta\right] \to E_f[\log f(x|\theta)] \;\; as \;\; \delta \to 0.$$

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

$(A5)$ Suppose that T_n is asymptotically normally distributed under both models such that

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

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

Theorem 1. Assume the regularity conditions (A1)-(A5). Then

$$\frac{2KLD_t(r,f|U_n)}{n} - \frac{\{\widehat{\mu}_f(U_n) - \widehat{\mu}_r(U_n)\}^2}{\widehat{\sigma}_f^2(U_n)} = o_p(1) \quad (3)$$

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

$$2KLD_t(r,f|U_n) - Q\left(\frac{\widehat{\sigma}_r^2(U_n)}{\widehat{\sigma}_f^2(U_n)}\right) = o_p(1) \qquad (4)$$

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

Remarks for Theorem 1

- $KLD_t(r, f|\theta)$ is also a divergence of model parameter estimates

- Model comparison in real applications may rely on the fit to a multi-dimensional statistic. The results in Theorem 1 are applicable to the multivariate case with a fixed dimension.

- $KLD_t(r, f|\theta)$ can be viewed as the discrepancy between r and f in terms of their posterior predictivity of T_n.

- We study how $KLD_t(r, f|\theta)$ is connected to a weighted posterior predictive p-value, a typical Bayesian technique to assess model discrepancy (see Rubin 1984; Gelman et al. 1996).

Weighted Posterior Predictive P-value

$$WPPP_r(U_n) \equiv \int \left\{ \int \int_{-\infty}^{t_n} f^*(y_n|\widehat{\theta}_f) dy_n \, r^*(t_n|\theta) dt_n \right\} \pi_r(\theta|U_n) \, d\theta, \quad (5)$$

where r^* and f^* are the predictive density functions of T_n under r and f, respectively.

• WPPP is equivalent to the weighted posterior predictive p-value of T_n under f with respect to the posterior predictive distribution of T_n under r.

Theorem 2.

$$\begin{aligned}
&\frac{2KLD_t(r,f|U_n)}{n} \\
= \quad &\frac{\{\Phi^{-1}(WPPP_r(U_n))\}^2}{n} \\
&+\left\{\frac{(\widehat{\mu}_r(U_n)-\widehat{\mu}_f(U_n))^2}{\widehat{\sigma}_f^2(U_n)+\widehat{\sigma}_r^2(U_n)}\right\}\frac{\widehat{\sigma}_r^2(U_n)}{\widehat{\sigma}_f^2(U_n)}+o_p(1) \qquad (6)
\end{aligned}$$

when $\mu_f(\theta) \neq \mu_r(\theta)$.

Let $Q(y) = y - log(y) - 1$. Then

$$2KLD_t(r,f|U_n) - Q\left(\frac{\widehat{\sigma}_r^2(U_n)}{\widehat{\sigma}_f^2(U_n)}\right) = o_p(1) \qquad (7)$$

and

$$WPPP_r(U_n) - 0.5 = o_p(1) \qquad (8)$$

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

Remarks of Theorem 2.

- It shows the asymptotic relationship between $KLD_t(r, f|u_n)$ and WPPP.

- Suppose that $\mu_f(\theta) \neq \mu_r(\theta)$.
– Both $KLD_t(r, f|U_n)$ and $\Phi^{-1}(WPPP_r(U_n))$ are of order $O_p(n)$.
– $KLD_t(r, f|U_n)$ is greater than $\Phi^{-1}(WPPP_r(U_n))$ by an $O_p(n)$ term that assumes positive values with probability 1.

- When $\mu_r(\theta) = \mu_f(\theta)$ (i.e., both r and f assume the same mean of T_n) but $\sigma_f^2(\theta) \neq \sigma_r^2(\theta)$,
– $\Phi^{-1}(WPPP_r(U_n))$ converges to 0; $WPPP_r(U_n)$ converges to 0.5
– $KLD_t(r, f|U_n)$ converges to a positive quantity order $O_p(1)$

Example 1. $X_i \stackrel{i.i.d.}{\sim} g_\theta(x_i) = \phi((x_i - \theta_1)/\sqrt{\theta_2})/\sqrt{\theta_2}$, where $\theta_2 > 0$. Let $T_n = \sqrt{n}[(\sum_i X_i)/n - \theta_1]/\sqrt{\kappa}$. Let $r = g$ and $f_\theta(x_i) = \phi((x_i - \theta_1)/\sqrt{\kappa})/\sqrt{\kappa}$. Then

• $\mu_r(\theta) = E_h(T_n) = \mu_f(\theta) = E_f(T_n) = \theta_1$, $\sigma_r^2(\theta) = \theta_2$, $\sigma_f^2(\theta) = \kappa$,

$$\begin{aligned} & 2 \lim_{n\to\infty} \widehat{KLD}_t(r, f|u_n) \\ = \quad & -\log\left(\frac{\widehat{\theta}_2(u_n)}{\kappa}\right) + \frac{\widehat{\theta}_2(u_n)}{\kappa} - 1 \begin{cases} > 0 & if \quad \kappa \neq \theta_2 \\ = 0 & if \quad \kappa = \theta_2 \end{cases} . \end{aligned}$$

•T_n is the MLE for θ_1 under both h and f.

•$\lim_{n\to\infty} WPPP(U_n) = 0.5$

•$WPPP(U_n)$ is asymptotically equivalent to the KLD approaches.

Example 2 Assume $X_i \overset{i.i.d.}{\sim} g_\theta(x_i) = \exp\{-\theta/(1-\theta)\}\{\theta/(1-\theta)\}^{x_i}/x_i!$, where $0 < \theta < 1$. Let $T_n = \bar{X}_n/(1+\bar{X}_n)$, $r = g$, and $f_\theta(x_i) = \theta^{x_i}(1-\theta)$. Then

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

- $\theta = E(X_i)/(1+E(X_i))$.

- T_n is the MLE for θ under both r and f

- $2\lim_{n\to\infty} KLD_t(r, f|u_n) = -\log(1-\widehat{\theta}(u_n)) + (1-\widehat{\theta}(u_n)) - 1 > 0$ for $0 < \theta < 1$.

- $\lim_{n\to\infty} WPPP(U_n) = 0.5$

Example 3 Assume $X_i \overset{i.i.d.}{\sim} g_\theta(x_i) = \frac{\Gamma((\theta_2+1)/2)}{\Gamma(\theta_2/2)\sqrt{\pi\theta_2}}(1 + (x - \theta_1)^2/\theta_2)^{-(1+\theta_2)/2}$, where $\theta_2 > 2$. Let $T_n = \bar{X}$. Let $r = g$ and $f_\theta(x_i) = \phi(X_i - \theta_1)$. Then

- $\mu_f(\theta) = \mu_r(\theta) = \theta_1$, $\sigma_r^2(\theta) = \theta_2/(\theta_2 - 2)$, and $\sigma_f^2(\theta) = 1$

- $2\lim_{n\to\infty} KLD_t(r, f|u_n) = -\log(\theta_2(u_n)/(\theta_2(u_n)-2)) + \theta_2/(\theta_2(u_n)-2) - 1 \geq 0$ for all θ_2 with equality if and only if $\theta_2 = \infty$.

Example 4 Assume $X_i \stackrel{i.i.d.}{\sim} g_\theta(x_i) = \exp(-x_i/\theta)/\theta$. Let $r = g$ and $f_\theta(x_i) = \exp(-x_i)$, $T_n = \min\{X_1, \cdots, X_n\}$. Then

- $r_\theta(t_n) = n \exp(-nt_n/\theta)/\theta$ and $f_\theta(t_n) = n \exp(-nt_n)$

- $WPPP_f(\bar{x}_n) = E^f(Pr(T_n^* < T_n)|\bar{x}_n) \to \frac{\bar{x}_n}{\bar{x}_n+1}$

- $\widehat{KLD}_t(r, f|\bar{x}_n) \to -\log(\bar{x}_n) + n(\bar{x}_n - 1)$

- The asymptotic equivalence between $KLD_t(r, f|u_n)$ and $WPPP_f(u_n)$ does not hold in the sense of Thm. 2 due to the violation of the asym. normality assumption.

A Study of Glucose Change in Veterans with Type 2 Diabetes

- A clinical cohort of 507 veterans with type 2 diabetes who had poor glucose control at the baseline and were then treated by metformin as the mono oral glucose-lowering agent.

- Goal: 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.

- The empirical mean of the net change in HbA1c over time was similar between the obese vs. non-obese groups (-0.498 vs. -0.379). The empirical variance was greater in the obese group (1.207 vs. 0.865).

- Distribution of HbA1c was reasonably symmetric. Considered two candidate models for fitting the HbA1c change: a mixture of normals vs. a t-distribution.

- $KLD_t(r, f|u_n) = 10.75$ suggesting that r was superior to f.

- $KLD_t(r, f|u_n)$ result was consistent with Figures 1 & 2 which contrasted the empirical quantiles with predicted quantiles under r and f. Note that both r and f yielded unbiased estimators of $E(X_i)$. Thus the model discrepancy between r and f assessed by $KLD_t(r, f|u_n)$ is primarily attributed to the difference in the variance assumption between r and f (as evident in Figure 1 which contrasted the empirical quantiles with predicted quantiles under r and f).

- WPPP=0.522 suggested that the overall fit were similar between the two models (the estimated net change in HbA1c was similar between these two models).

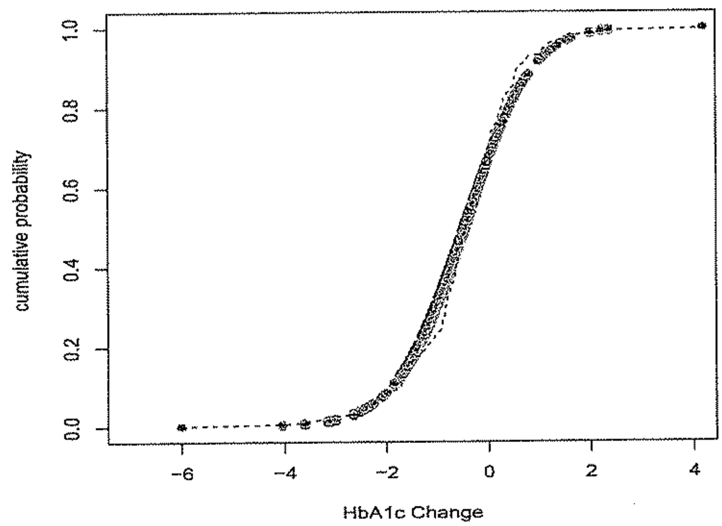

Fig. 1. Quantile Plots: Empirical Quantiles of HbA1c Change vs. Quantiles Uunder Models r and f among Obese Patients in Study I

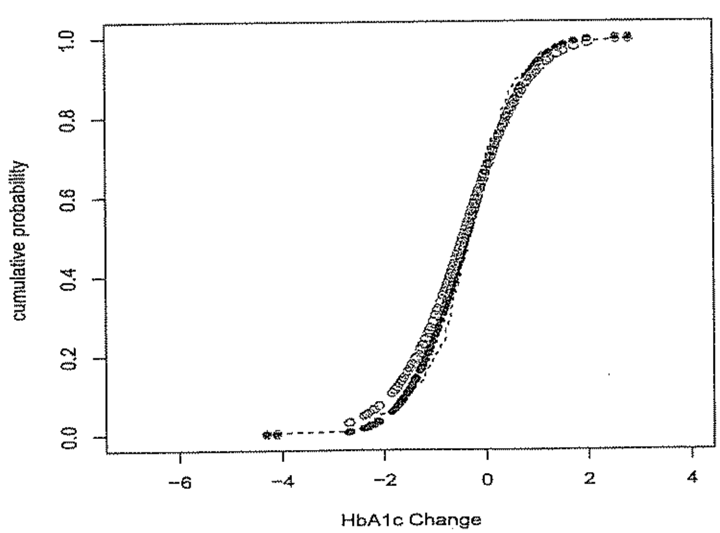

Fig. 2. Quantile Plots: Empirical Quantiles of HbA1c Change vs. Quantiles Uunder Models r and f among Non-Obese Patients in Study I

A Study of Functioning in the Elderly with Diabetes

• The study cohort arisen from the subset of 119 participants with diabetes in the San Antonio Longitudinal Study of Aging, a community-based study of the disablement process in Mexican American and European American older adults.

• Goal: to compare models that assessed whether glucose control trajectory class (poorer vs. better) was associated with the lower-extremity physical functional limitation score (measured by SPPB) during the first follow up period.

• SPPB score is discrete in nature with a range of 0-12. Considered two candidate models for fitting SPPB: a negative binomial vs. a poisson.

• 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.

- $KLD_t(r, f|u_n) = 32.63$ suggested that r was a better fit than f.

- Both r and f yielded similar estimates of $E(X_i)$. The model discrepancy assessed by $KLD_t(r, f|u_n)$ could primarily be attributed to the difference in variance estimation between r and f (as evident in Figures 3 & 4).

- $WPPP(U_n) = 0.539$ suggested similar fit between r and f.

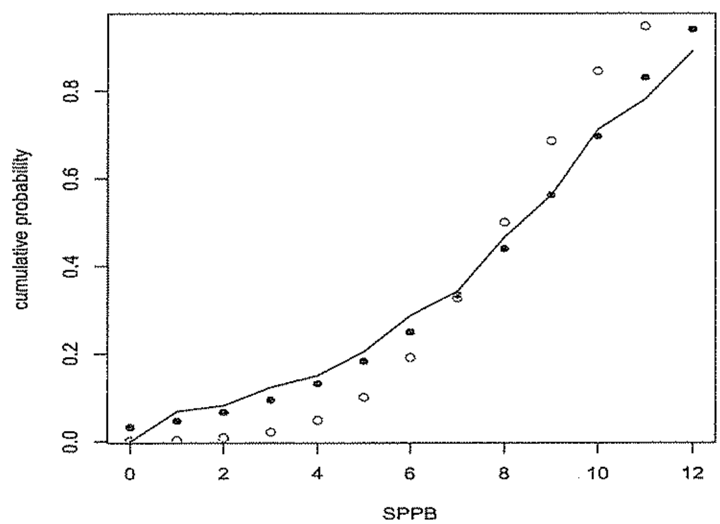

Fig. 3. Quantile Plots: Empirical Quantiles of SPPB vs. Quantiles Uunder Models r and f among Subjects with Poor Glycemic Control in Study II

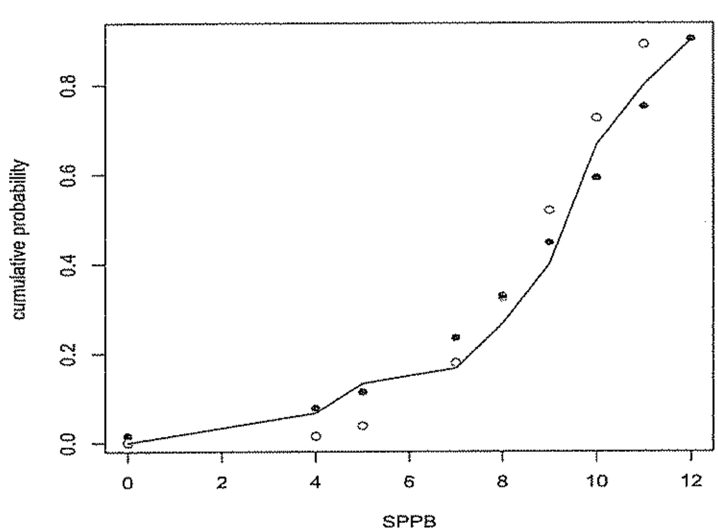

Fig. 4. Quantile Plots: Empirical Quantiles of SPPB vs. Quantiles Uunder Models r and f among Subjects with Better Glycemic Control in Study II

Summary

- This paper considers a Bayesian estimate of the G-R-A KLD as given in (2).

- G-R-A KLD is appropriate for quantifying information discrepancy between 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 Theorem 2.

- Our results need further refinement when the MLE of the mean of T_n differs between r and f , or the normality assumption given in (A5) is not suitable.

- Model comparison in medical research may rely on the fit to a multidimensional statistic. Theorem 1 holds for a multivariate statistic T_n with a fixed dimension. Further investigation is needed to assess the property of our proposed KLD for situation when the dimension of T_n increases with n.

- G-R-A KLD provides the relative fit between competing models. For the purpose of assessing absolute model adequacy, a KLD should be used in conjuction with absolute model departure indices such as posterior predictive p-values or residuals. Nevertheless, a KLD is also a measure of the absolute fit of model f when the reference model r is the true model.