ISSN 1936-5098

CAE Working Paper #09-08

Geometry of the Log-Likelihood Ratio Statistic in Misspecified Models

by

Hwan-sik Choi
and
Nicholas M. Kiefer

May 2009

Geometry of the Log-Likelihood Ratio Statistic in Misspecified Models

Hwan-sik Choi[*]
Texas A&M University

Nicholas M. Kiefer[†]
Cornell University

May, 2009

Abstract

We show that the asymptotic mean of the log-likelihood ratio in a misspecified model is a differential geometric quantity that is related to the exponential curvature of Efron (1978), Amari (1982), and the preferred point geometry of Critchley et al. (1993, 1994). The mean is invariant with respect to reparametrization, which leads to the differential geometrical approach where coordinate-system invariant quantities like statistical curvatures play an important role. When models are misspecified, the likelihood ratios do not have the chi-squared asymptotic limit, and the asymptotic mean of the likelihood ratio depends on two geometric factors, the departure of models from exponential families (i.e. the exponential curvature) and the departure of embedding spaces from being totally flat in the sense of Critchley et al. (1994). As a special case, the mean becomes the mean of the usual chi-squared limit (i.e. the half of the degrees of freedom) when these two curvatures vanish. The effect of curvatures is shown in the non-nested hypothesis testing approach of Vuong (1989), and we correct the numerator of the test statistic with an estimated asymptotic mean of the log-likelihood ratio to improve the asymptotic approximation to the sampling distribution of the test statistic.

AMS 2000 classification: 62F03; 62F05.

Keywords: Differential geometry, log-likelihood ratio, asymptotic mean, exponential curvature, preferred point geometry, non-nested hypothesis.

[*] hwansik.choi@tamu.edu. 3035 Allen, 4228 TAMU, Department of Economics, Texas A&M University, College Station, TX 77843-4228, USA.

[†] nmk1@cornell.edu. 490 Uris Hall, Department of Economics and Department of Statistical Science, Cornell University, Ithaca, NY, 14850, USA, and CREATES, University of Aarhus, DK (Funded by the Danish National Research Foundation.)

1 Introduction

The differential geometrical approach in statistics gives geometrical intuition to the higher order asymptotics in estimation and inference. We show the differential geometrical method is useful in the first order asymptotics for misspecified models.

When a model is misspecified, the first order chi-squared approximation is no longer valid and the departure from the chi-squared distribution appears in the first order term. The first order asymptotic mean of the log-likelihood ratio under the misspecification has the form of the trace of a matrix. Noting the invariance property of the log-likelihood ratio still holds for the misspecified models, we show that the first order asymptotic mean of the log-likelihood ratio of a misspecified model is in fact a geometrical quantity like the Bartlett correction (Bartlett (1937), McCullagh and Cox (1986), DiCiccio et al. (1991)). It is shown that the first order asymptotic mean has two geometrical components. One part is related to the degree of misspecification, and the other is generated by the exponential curvature of the misspecified model. The former is related to the total flatness in the preferred point geometry of Critchley et al. (1993, 1994), and the latter is related to the embedding exponential curvature of Efron (1975, 1978) and Amari (1982). When both curvatures vanish, the mean coincides with the mean of the usual chi-squared limit of correctly specified log-likelihood ratios.

We apply our results to the non-nested hypothesis testing framework of Vuong (1989). The test uses the null hypothesis that competing misspecified models are equidistant from an unknown true distribution with respect to the *Kullback-Leibler Information criterion* (KLIC, relative entropy, Kullback and Leibler (1951)). The numerator of Vuong's test statistic is written in terms of the log-likelihood ratios, and we propose a geometrically motivated mean correction to improve the asymptotic approximation to the sampling distribution of the test statistic. A simple mean correction based on parameter dimensions is sometimes used, but the simple correction is valid under a correct model specification which violates Vuong's conditions. The proposed mean correction is valid under misspecification. We provide a Monte Carlo experiment to show the effect of the curvatures on the mean and the improvement of the mean correction. Our results also give the geometrical (flatness) conditions under which the proposed mean correction is the same as the simple correction even in a misspecified case.

Throughout the paper we will consider i.i.d. samples and assume the regularity conditions in Section 8 of Kent (1982).

2 Differential Geometry for Log-Likelihood Ratios

2.1 Likelihood Ratios in Misspecified Models

Let $y = (y_1, \dots, y_n)$ be i.i.d. data drawn from a distribution $p_0 \equiv p_0(y)$. Consider a parametric family of distributions $p(y|\theta) = \prod_{i=1}^n p(y_i|\theta)$ with a parameter vector $\theta \in \mathbb{R}^k$. Denote $p(y|\theta)$ as $p(\theta)$ and $p(y_i|\theta)$ as $p_i(\theta)$ for simplicity. The model $p(\theta) = \prod_{i=1}^n p_i(\theta)$ is misspecified in the sense that $KLIC(p_0, p(\theta)) > 0$ for all θ, and the maximum likelihood estimator $\hat{\theta} = \operatorname{argmax}_\theta p(\theta)$ is assumed to converge in probability to a pseudo-true value θ^*.

Let $l_i(\theta) = \log p_i(\theta)$ be the log-likelihood function on one observation and $l(\theta) = \sum_{i=1}^n l_i(\theta)$ the log-likelihood on n observations. The dependence on n will not be indicated explicitly. Denote the score and Hessian functions as $s(\theta) = \sum_{i=1}^n s_i(\theta)$ and $h(\theta) = \sum_{i=1}^n h_i(\theta)$ respectively. Let E_0 be the expectation with respect to p_0, and define the expected Hessian $\bar{H}(\theta) = E_0 h_i(\theta)$ and $H(\theta) = E_0 h(\theta) = n\bar{H}(\theta)$. Noting $E_0 s_i(\theta^*) = 0$, let $\bar{J}(\theta^*) = E_0\{s_i(\theta^*)s_i(\theta^*)^T\}$ and $J(\theta^*) = E_0\{s(\theta^*)s(\theta^*)^T\} = n\bar{J}(\theta^*)$ be the variance of the score at θ^*. We analyze the asymptotics of the log-likelihood ratio

$$l(\hat{\theta}) - l(\theta^*), \tag{2.1}$$

using the differential geometrical approach when the true distribution p_0 is not equal to $p(\theta)$ for any $\theta \in \mathbb{R}^k$.

From the Taylor expansion

$$l(\hat{\theta}) - l(\theta^*) = -\frac{1}{2} tr\{\bar{H}(\theta^*)^{-1} s(\theta^*) s(\theta^*)^T\} + o_p(1), \tag{2.2}$$

of the log-likelihood function $l(\hat{\theta})$ around θ^*, the mean of the likelihood ratio is

$$\begin{aligned} E_0(l(\hat{\theta}) - l(\theta^*)) &= -\frac{1}{2} tr\{\bar{H}(\theta^*)^{-1} \bar{J}(\theta^*)\} + o(1) && (2.3) \\ &= -\frac{\lambda(\theta^*)}{2} + o(1), && (2.4) \end{aligned}$$

where $\lambda(\theta^*)$ is defined by

$$\lambda(\theta^*) = tr\{\bar{H}(\theta^*)^{-1} \bar{J}(\theta^*)\}. \tag{2.5}$$

When $p(\theta)$ is correctly specified, we have $p(\theta^*) = p_0$, and the Fisher's identity $\bar{J}(\theta^*) = -\bar{H}(\theta^*)$ holds. Then $-\lambda(\theta^*)/2$ simply becomes $k/2$, where k is the dimension of θ and is irrelevant to the curvature of the model. In this case, the statistical curvatures appear in the higher order terms. In this paper, we show that when $p(\theta)$ is misspecified, the quantity $\lambda(\theta^*)$, which is generally a function of θ^* (or of p_0), is also related to the statistical curvature. We also give flatness conditions under which $\lambda(\theta^*) = -k$ is a constant. This provides geometrical intuitions to the first order asymptotic mean of the likelihood ratio for misspecified models.

It is well known that $\lambda(\theta^*)$ is a tensor. Moreover it is reparameterization invariant therefore a geometric object. Specifically, let θ be the original parameterization and $\xi(\theta)$ be a locally one-to-one reparameterization of θ with $\xi^* = \xi(\theta^*)$. Then

$$tr\{\bar{H}(\theta^*)^{-1}\bar{J}(\theta^*)\} = tr\{\tilde{H}(\xi^*)^{-1}\tilde{J}(\xi^*)\}, \tag{2.6}$$

where $\tilde{H}(\xi)$ and $\tilde{J}(\xi)$ are defined for the new parameterization ξ as $\bar{H}(\theta)$ and $\bar{J}(\theta)$ for θ respectively. This implies that we can use any convenient parameterization for the calculation of $\lambda(\theta^*)$. We use a locally affine parameterization in which the Fisher information becomes an identity matrix at the pseudo-true distribution $p(\theta^*)$, i.e.

$$E_{p(\theta^*)}\{s_i(\theta^*)s_i(\theta^*)^T\} = I_k, \tag{2.7}$$

where I_k is a $(k \times k)$ identity matrix. A globally affine parameterization in which the information matrix is an identity matrix for all θ does not generally exist except in one-dimensional parameter models. When such a global reparameterization exists, the model makes a Euclidean (or 0-flat) manifold. See Amari (1985) for details.

In the next section, we give an interpretation of $\lambda(\theta^*)$ using differential geometrical quantities for the exponential families and show how to extend the approach to general families of distributions.

2.2 Geometry of Log-Likelihood Ratios

A curved exponential family (CEF) is an embedded sub-manifold of an exponential family. It is obtained from an exponential family by reducing the parameter dimension through restrictions.

Let $p(y|\eta) = \exp\left[n\left\{\bar{y}^T\eta - \psi(\eta)\right\}\right] f(y)$ be a density function of an exponential family of i.i.d. observations $y = (y_1, y_2, \ldots, y_n)$ with an m-dimensional parameter vector η, a vector $\bar{y}$ of sufficient statistics, and a cumulant generating function $\psi(\eta)$. The Fisher information matrix of one observation with respect to the natural parameterization is $\psi''(\eta)$. A CEF is obtained by a lower dimensional reparameterization θ of η,

$$p(y|\theta) \equiv p(y|\eta(\theta)) = \exp\left[n\left\{\bar{y}^T\eta(\theta) - \psi(\eta(\theta))\right\}\right] f(y), \tag{2.8}$$

where θ is a $k < m$ dimensional parameter vector. If $\eta(\theta)$ is affine, $p(y|\theta)$ becomes a lower dimensional (full) exponential family.

Let $\eta_{ab}(\theta) = \partial^2\eta(\theta)/\partial\theta_a\partial\theta_b$, where θ_a $(a = 1, 2, ..., k)$ is the a^{th} parameter, and $i_{ab}(\theta) = \left(\partial\eta(\theta)/\partial\theta_a\right)^T g(\eta(\theta))\left(\partial\eta(\theta)/\partial\theta_b\right)$, where $g(\eta(\theta))$ is the Fisher information at $\eta(\theta)$. Suppose $\eta = \phi$ gives the true distribution $p(\phi)$ and denote $\mu = E_{p(\phi)}\bar{y}$. Then we can calculate $\lambda(\theta^*) = tr\{\bar{H}(\theta^*)^{-1}\bar{J}(\theta^*)\}$

from

$$\bar{J}(\theta^*) = \eta'(\theta^*)^T g(\phi)\eta'(\theta^*)), \tag{2.9}$$

and $\bar{H}(\theta^*)$ with the (a, b) element given by $(\mu - \psi'(\eta(\theta^*)))^T \eta_{ab}(\theta^*) - i_{ab}(\theta^*)$.

To represent $\lambda(\theta^*)$ in geometrical quantities, we define the differentials

$$\partial_a = \frac{\partial l(\theta)}{\partial \theta_a},\ \partial_{ab} = \frac{\partial^2 l(\theta)}{\partial \theta_a \partial \theta_b}. \tag{2.10}$$

Using Einstein's summation convention, where the repeating upper and lower indices imply summation over that index, the score function ∂_a can be represented as

$$\partial_a = B_a^i \partial_i, \tag{2.11}$$

where $B_a^i = \partial \eta^i / \partial \theta_a$ and ∂_i is the i^{th} element of the score vector $\partial l(\eta)/\partial \eta = n(\bar{y} - \psi'(\eta))$ of the natural parameterization η.

The (embedding) m-dimensional full exponential family can be reparameterized with the $m - k$ dimensional parameter ν in addition to the m-dimensional parameter vector θ. Thus (θ, ν) is a new diffeomorphic reparameterization of η. Moreover we can choose the parameterization (θ, ν) such that the score functions ∂_γ, where γ are indices of ν, are locally orthonormal to ∂_a. Then the coefficients of the *Euler-Schouten curvature tensor* or the *embedding curvature* with respect to 1-connection (exponential curvature, 1-curvature. See Amari (1982)) of the CEF in the full exponential family is given by

$$H_{ab\gamma}(\theta) = E_{p(\theta)} \left\{ (\partial_{ab} - E_{p(\theta)} \partial_{ab}) \partial_\gamma \right\}. \tag{2.12}$$

We can decompose $(\partial_{ab} - E_{p(\theta)}\partial_{ab})$ with the tangential component and the normal component to the space spanned by the scores ∂_a of θ. Then the tangential and the normal components can be represented with the orthonormal bases $(\partial_\kappa, \partial_\gamma)$ for (θ, ν) respectively. Using the relationship

$$\partial_{ab} - E_{p(\theta)}\partial_{ab} = n(\bar{y} - \psi'(\eta(\theta)))^T \eta_{ab}(\theta), \tag{2.13}$$

we have the decomposition

$$\begin{aligned} n(\bar{y} - \psi'(\eta(\theta)))^T \eta_{ab}(\theta) &= \Gamma_{ab}^\kappa \partial_\kappa + H_{ab}^\gamma \partial_\gamma & (2.14) \\ &= \Gamma_{ab}^\kappa B_\kappa^i \partial_i + H_{ab}^\gamma B_\gamma^i \partial_i, & (2.15) \end{aligned}$$

where Γ_{ab}^κ and H_{ab}^γ are the coefficients of the projected component onto the space spanned by the basis vectors ∂_κ and ∂_γ respectively. The last equality is from equation (2.11). Note that we have $H_{ab}^\gamma = H_{ab\gamma}$ since the bases $(\partial_\kappa, \partial_\gamma)$ are orthonormal.

Lemma 2.1 *Consider* $E_{p(\phi)}(\partial_{ab} - E_{p(\theta^*)}\partial_{ab}) = n(\mu - \psi'(\eta(\theta^*)))^T\eta_{ab}(\theta^*)$, *where* $E_{p(\phi)}$ *is the expectation with respect to a true distribution* $p(\phi)$.
(a) We have

$$E_{p(\phi)}(\partial_{ab} - E_{p(\theta^*)}\partial_{ab}) = A_i B^i_\gamma H^\gamma_{ab}, \tag{2.16}$$

where A_i *is the* i^{th} *element of* $(\mu - \psi'(\eta(\theta^*)))$, *and the quantities* B^i_γ *and* H^γ_{ab} *are defined in eq. (2.11) and (2.14) respectively.*
(b) If the model has zero embedding curvature with respect to 1*-connection at* θ^*, *then*

$$E_{p(\phi)}(\partial_{ab} - E_{p(\theta^*)}\partial_{ab}) = 0. \tag{2.17}$$

Proof. For (a), let ∂^i be the i^{th} element of the score function with respect to the mean parameterization. The score functions of mean and natural parameterizations have the relationship

$$\partial^i = g^{ij}\partial_j, \tag{2.18}$$

where g^{ij} is the (i,j) element of $g(\eta(\theta))^{-1}$. Then we have

$$\begin{aligned}
& E_{p(\phi)}(\partial_{ab} - E_{p(\theta^*)}\partial_{ab}) && (2.19)\\
&= n(\mu - \psi'(\eta(\theta^*)))^T\eta_{ab}(\theta^*) && (2.20)\\
&= E_{p(\theta^*)}\{(\mu - \psi'(\eta(\theta^*)))^T g(\eta(\theta^*))^{-1} n(\bar{y} - \psi'(\eta(\theta^*)))\}\{n(\bar{y} - \psi'(\eta(\theta^*)))^T\eta_{ab}(\theta^*)\} && (2.21)\\
&= E_{p(\theta^*)}\{A_i\partial^i\}\{\Gamma^\kappa_{ab}B^i_\kappa\partial_i + H^\gamma_{ab}B^i_\gamma\partial_i\} && (2.22)\\
&= E_{p(\theta^*)}\{A_i\partial^i\}\left(H^\gamma_{ab}B^i_\gamma\partial_i\right) && (2.23)\\
&= A_i B^i_\gamma H^\gamma_{ab}. && (2.24)
\end{aligned}$$

The fourth equality is from the zero expected score,

$$(\mu - \psi'(\eta(\theta^*)))^T\eta'(\theta^*) = 0. \tag{2.25}$$

The result for (b) is obvious since $H^\gamma_{ab} = 0$ if the exponential curvature of the embedding model vanishes at θ^*. ■

Definition 2.2 (Critchley et al. (1994)) *For a fixed (true) distribution* ϕ, *define*

$$\mu^\phi(\eta) = E_\phi(s(\eta)), \tag{2.26}$$
$$g^\phi(\eta) = Var_\phi(s(\eta)), \tag{2.27}$$

where $s(\eta)$ *is the score function and the expectations are taken with respect to the fixed model* $\eta = \phi$, *then the preferred point geometry,* $(M, \mu^\phi(\eta), g^\phi(\eta))$ *is* g^ϕ*-flat if there exits a coordinate system* η

for which g^{ϕ} is constant for all η. The η coordinates are called g^{ϕ}-affine. M is totally flat, if there exists a coordinate system η for which g^{ϕ} is a constant for all η and μ^{ϕ} is a linear function of $\eta - \phi$.

When an exponential family is totally flat, the natural parameterization is α-affine for all real α in the sense of Amari (1982). The total flatness assumption is quite restrictive. An example would be a normal model with a known variance matrix. Therefore the total flatness is not a condition to expect to hold in general, but a reference or a benchmark for a real problem. Of course, the condition would look more reasonable as a sample size grows, since we consider the local structure of a model asymptotically.

Theorem 2.3 *For a k-dimensional CEF embedded in an exponential family,*

$$\lambda(\theta^*) = tr\{\bar{H}(\theta^*)^{-1}\bar{J}(\theta^*)\} \tag{2.28}$$

is calculated from

$$\bar{J}(\theta^*) = \eta'(\theta^*)^T g(\phi)\eta(\theta^*), \tag{2.29}$$

$\bar{H}(\theta^*)$ with the (a, b) element $\bar{H}_{ab}(\theta^*) = A_i B^i_{\gamma} H^{\gamma}_{ab} - \delta^b_a$, and $\delta^b_a = 1$ for $(a = b)$ and $\delta^b_a = 0$ for $(a \neq b)$. When the model has a locally vanishing embedding curvature with respect to 1-connection (exponential curvature) at θ^*, we have

$$\lambda(\theta^*) = -tr(\eta'(\theta^*)^T g(\phi)\eta(\theta^*)), \tag{2.30}$$

and if the embedding exponential family is totally flat as well, we have

$$\lambda(\theta^*) = -k. \tag{2.31}$$

Proof. Since $\lambda(\theta^*)$ is invariant with respect to a reparameterization, we use the locally 0-affine parameterization such that the Fisher information

$$i(\theta^*) = \eta'(\theta^*)^T g(\eta(\theta^*))\eta(\theta^*) \tag{2.32}$$

becomes a $(k \times k)$-dimensional identity matrix without loss of generality. The existence of such local parameterization at the pseudo-true distribution is sufficient for our results. Therefore we have

$$\bar{H}_{ab}(\theta^*) = A_i B^i_{\gamma} H^{\gamma}_{ab} - \delta^b_a \tag{2.33}$$

using Lemma 2.1 (a). If the CEF has a vanishing exponential curvature, from Lemma 2.1 (b) we get the second result,

$$\lambda(\theta^*) = -tr(\eta'(\theta^*)^T g(\phi)\eta(\theta^*)). \tag{2.34}$$

Also, when the embedding exponential family is totally flat as well, $g(\eta)$ is constant, i.e. $g(\phi) = g(\eta(\theta))$ for all θ (Theorem 4 in Critchley et al. (1994)). Therefore we have

$$\lambda(\theta^*) = -tr(\eta'(\theta^*)^T g(\eta(\theta^*))\eta(\theta^*)) \tag{2.35}$$
$$= -tr(i(\theta^*)) = -k \tag{2.36}$$

from $g(\phi) = g(\eta(\theta^*))$. ■

As discussed earlier, for a general k-dimensional model ($k > 1$), there does not exist a reparameterization that makes the Fisher information matrix an identity matrix for all θ, but there always exists a local parameterization (locally 0-affine) that makes the information matrix an identity matrix at a particular point.

2.3 Summary and Extension

For a curved exponential family embedded in the full exponential family, $\lambda(\theta^*) = tr\{H(\theta^*)^{-1}J(\theta^*)\}$ is related to two factors, total flatness and the exponential curvature at θ^*. Using an 0-affine parameterization, we showed, if the embedding exponential family is totally flat in the sense of Critchley et al. (1994), we have $J(\theta^*) = I_k$, where I_k is a $(k \times k)$ identity matrix where k is the dimension of the parameter vector. If the embedded CEF has a vanishing exponential curvature at θ^*, we also have $H(\theta^*) = -I_k$.

We consider the extension of the results to general parametric families by approximating the model with a curved exponential model around the pseudo-true distribution. The approximating CEF is expanded to include a true distribution by the exponential link. The exponential link between two distributions q_1 and q_2 defines a one dimensional exponential family $\log q(\alpha) = c(\alpha)\{\log q_1 + (1-\alpha)\log q_2\}$ with a parameter α and a normalizing constant $c(\alpha)$. Note that the exponential link is only an example of connecting two distributions. It is a convenient way of representing our geometrical idea, since it creates an embedding exponential family.

Let $l_0 = \log p_0$ where p_0 is the true distribution. We first consider the curved exponential approximation $\tilde{l}(\theta)$ of Efron (1975) for a general log-likelihood function $l(\theta)$, then we include the true distribution. The approximating log-likelihood function $\tilde{l}(\theta)$ around θ^* is a one-dimensional curved exponential family embedded in the $(m+1)$-dimensional exponential family $\tilde{l}(\eta)$ with $\eta = (\eta_0, \eta_1, \eta_2, \ldots, \eta_m)$, and is given by

$$\tilde{l}(\theta) = \tilde{l}(\eta(\theta)) = \eta_0(l_0 - l(\theta^*)) + l(\theta^*) + \sum_{r=1}^{m} \eta_r l^{(r)}(\theta^*) - \psi(\eta), \tag{2.37}$$

where

$$\eta(\theta) = (\eta_0, \eta_1(\theta), \eta_2(\theta), \ldots, \eta_m(\theta)) = \left(0, (\theta - \theta^*), \frac{1}{2}(\theta - \theta^*)^2, \cdots, \frac{1}{m!}(\theta - \theta^*)^m\right), \tag{2.38}$$

$$l^{(r)}(\theta^*) = \frac{\partial^r}{\partial \theta^r} l(\theta)\bigg|_{\theta=\theta^*}, \tag{2.39}$$

and $\psi(\eta)$ is a normalizing constant. The true distribution in the embedding family $\tilde{l}(\eta)$ is given by

$$(\eta_0, \eta_1, \eta_2, \ldots, \eta_m) = (1, 0, 0, \ldots, 0)\,. \tag{2.40}$$

With the approximating CEF and its embedding exponential family, we can apply our geometrical interpretation directly.

3 Non-nested Hypothesis Testing

3.1 Application to Vuong's Test

Non-nested hypothesis testing considers two separate parametric families of distributions. Unlike nested hypothesis testing, where a smaller (restricted) model is typically a natural candidate for a null model, defining a null hypothesis or a true model is a subtle issue in non-nested testing.

Vuong (1989) proposed to test the null hypothesis that competing models are equidistant in KLIC from an unknown true distribution. The test is based on the difference in KLIC for candidate models 1 and 2, given by

$$KLIC_1 - KLIC_2 = E_0(l_0 - l_1) - E_0(l_0 - l_2) \tag{3.1}$$

$$= E_0(l_2 - l_1), \tag{3.2}$$

where l_0, l_1, and l_2 are the log likelihood functions of the true distribution and the pseudo-true distributions of the competing models 1 and 2 respectively. Under the null that $E_0(l_2 - l_1) = 0$, Vuong (1989) proposed a normalized sample mean version of equation (3.2) for the test statistic. The test statistic t_n is given by

$$t_n = \frac{n^{-1/2}(l_2(\hat{\theta}_2) - l_1(\hat{\theta}_1))}{\sqrt{\widehat{V}_n}}, \tag{3.3}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the maximum likelihood estimators (MLEs), and denoting

$$l_j(\theta_j) = \sum_{i=1}^{n} l_{ji}(\theta_j), \tag{3.4}$$

$$\bar{l}_j(\theta_j) = \frac{1}{n}\sum_{i=1}^{n} l_{ji}(\theta_j) \tag{3.5}$$

for $j = 1, 2$, the variance of the numerator V_n is estimated by

$$\widehat{V}_n = \frac{1}{n}\sum_{i=1}^{n} \left\{ (l_{2i}(\hat{\theta}_2) - \bar{l}_2(\hat{\theta}_2)) - (l_{1i}(\hat{\theta}_1) - \bar{l}_1(\hat{\theta}_1)) \right\}^2 . \tag{3.6}$$

This test statistic requires that no model contains the true distribution, i.e. they are misspecified. If it does, the test statistic degenerates. See Vuong (1989) for a test for the degenerating variance. Under Vuong's null hypothesis (with i.i.d. data), the test statistic t_n is asymptotically standard normal. Our results indicate that the numerator of the test statistic has a mean that depends on the curvatures, thus the standard normal approximation is expected to deteriorate when the curvature is large.

The finite sample properties of this test statistic are not studied comprehensively. Vuong's test is extended for stationary time series data by Rivers and Vuong (2002) and Choi and Kiefer (2008). Choi and Kiefer (2008) also studied the finite sample properties of the test statistic for dynamic models and proposed to use the new asymptotic approximation, called the fixed-b asymptotics, developed by Kiefer and Vogelsang (2002) and Kiefer and Vogelsang (2005). The fixed-b asymptotics improves the approximation of the denominator when the heteroskedasticity autocorrelation consistent (HAC) estimator was used. They compared the performance of the fixed-b asymptotic approximation with bootstrap approaches. That approach uses a different asymptotic approximation and allows quite general autocorrelation.

In this paper, we propose to correct the mean in the test statistic rather than changing the approximating distribution to get better finite sample performance. We use the first order mean discussed in the previous sections to correct the numerator. Under the null, the mean correction becomes of order $O(1/\sqrt{n})$ because of the normalization in Vuong's test statistic. The proposed mean correction term can be estimated consistently. We study the magnitude of the asymptotic mean with nonlinear regression models and demonstrate the improvements in the asymptotic approximation of the sampling distribution of the mean corrected test statistic.

3.2 Misspecified Nonlinear Regressions

Consider a linear model

$$y_i = \alpha + \beta x_i + u_i \ (i = 1, \ldots, n), \tag{3.7}$$

where $x_i \sim$ i.i.d. $\mathbf{N}(0,1)$ and $u_i \sim$ i.i.d. $\mathbf{N}(0,\sigma^2)$. The true DGP is $(\alpha,\beta,\sigma^2) = (0,\ 0,\ 0.04)$, and two competing misspecified models M_1 and M_2 are given by two nonlinear restrictions (half circles),

$$M_1 : (\alpha+2)^2 + \beta^2 = 1 \text{ with } \alpha \geq -2, \tag{3.8}$$

$$M_2 : (\alpha+1)^2 + \beta^2 = 4 \text{ with } \alpha \geq -1. \tag{3.9}$$

The pseudo-true distributions are $\theta_1^* = (\alpha^*,\beta^*) = (-1,0)$ for M_1, and $\theta_2^* = (\alpha^*,\beta^*) = (1,0)$ for M_2. The estimated asymptotic mean $\hat{b}$ is calculated from

$$\hat{b} = -\frac{1}{2}\left(tr\{\hat{H}_2(\hat{\theta}_2)^{-1}\hat{J}_2(\hat{\theta}_2)\} - tr\{\hat{H}_1(\hat{\theta}_1)^{-1}\hat{J}_1(\hat{\theta}_1)\}\right). \tag{3.10}$$

The mean adjusted test statistic t_2 is given by

$$t_2 = \frac{n^{-1/2}(l_1(\hat{\theta}_1) - l_2(\hat{\theta}_2) - \hat{b})}{\sqrt{\hat{V}_n}}, \tag{3.11}$$

where

$$\hat{V}_n = \frac{1}{4n}\sum_{i=1}^{n} v_i^2, \tag{3.12}$$

$$v_i = \hat{u}_{1i}^2/\hat{\sigma}_1^2 - \hat{u}_{2i}^2/\hat{\sigma}_2^2, \tag{3.13}$$

and $\{\hat{u}_{ji}\}_{i=1}^n$ are residuals using MLEs $(\hat{\beta}_j, \hat{\sigma}_j^2)$ from model $j = 1,2$. See Lien and Vuong (1987) p.10 for the details of approximating the variance of the numerator using v_i.

We set $n = 20$, and the number of simulation repetition is $3,000$. The nonparametric kernel density estimator of simulated t_2 is compared with the density estimate of Vuong's original test statistic in Figure 1. The sample mean of simulated Vuong's test statistics is 0.340, whereas the simulated values of t_2 have the sample mean of 0.029.

4 Conclusion

When a model is misspecified, the first order chi-squared asymptotic approximation to the log-likelihood ratio is no longer valid and the mean of the limit of the likelihood ratio depends on the pseudo-true values of parameters. We showed that the mean has a differential geometrical interpretation and its value is determined by the exponential curvature of the model and the total flatness of the embedding family. When both the curvatures vanish, the mean becomes a trivial constant and the chi-squared approximation becomes valid.

As an example, the effect of the curvatures on the asymptotic approximation of the non-nested

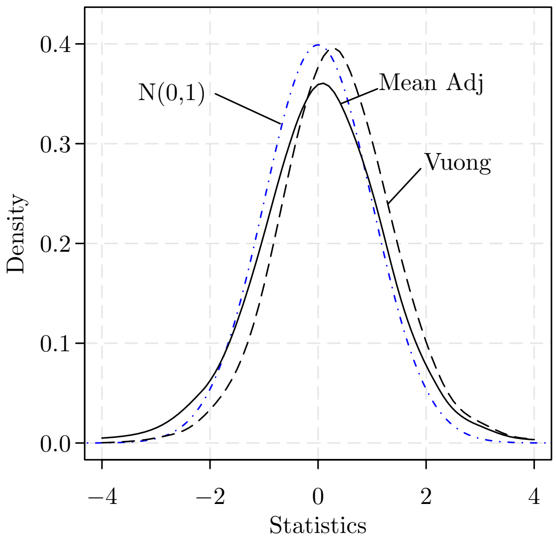

Figure 1: Comparison of kernel density estimators (Gassian kernel, bandwidth 0.3) for simulated Vuong's test statistics ("Vuong") and the mean adjusted statistics ("Mean Adj"). The mean of the test statistic is reduced to 0.029 from 0.340.

hypothesis test of Vuong (1989) was presented. The numerator of the test statistic was modified with a higher order mean correction term calculated by plugging in the MLEs. The results showed that the improvement of the asymptotic approximation from the mean correction could be significant when the curvatures are large.

References

Amari, S. I. (1982). Differential geometry of curved exponential families - curvatures and information loss. *The Annals of Statistics*, **10** 357–385.

Amari, S. I. (1985). *Differential-geometrical methods in statistics*. Lecture Notes in Statistics, Springer-Verlag, Berlin.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **160** 268–282.

Choi, H.-s. and Kiefer, N. M. (2008). Improving robust model selection test for dynamic models with an application to comparing predictive accuracy. *Working paper, Texas A&M University and Cornell University.*

Critchley, F., Marriott, P. and Salmon, M. (1993). Preferred point geometry and statistical manifolds. *The Annals of Statistics*, **21** 1197–1224.

Critchley, F., Marriott, P. and Salmon, M. (1994). Preferred point geometry and the local differential geometry of the kullbank-leibler divergence. *The Annals of Statistics*, **22** 1587–1602.

DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist*, **19** 1053–1061.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, **3** 1189–1217.

Efron, B. (1978). The geometry of exponential families. *The Annals of Statistics*, **6** 362–376.

Kent, J. T. (1982). Robust properties of likelihood ratio test. *Biometrika*, **69** 19–27.

Kiefer, N. M. and Vogelsang, T. J. (2002). Heteroskedasticity-autocorrelation robust standard errors using the bartlett kernel without truncation. *Econometrica*, **70** 2093–2095.

Kiefer, N. M. and Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, **21** 1130–1164.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22** 79–86.

Lien, D. and Vuong, H. Q. (1987). Selecting the best linear regression model: A classical approach. *Journal of Econometrics*, **35** 3–23.

McCullagh, P. and Cox, D. R. (1986). Invariants and likelihood ratio statistics. *The Annals of Statistics*, **14** 1419–1430.

Rivers, D. and Vuong, H. Q. (2002). Model selection tests for nonlinear dynamic models. *Econometrics Journal*, **5** 1–39.

Vuong, H. Q. (1989). Likelihood ratio tests for model selection and non-nested hypothesis. *Econometrica*, **57** 307–333.