Differential Geometry and Bias Correction in Nonnested Hypothesis Testing

Hwan-sik Choi* Cornell University

Nicholas M. Kiefer† Cornell University

April, 2006

Abstract

Many model selection or non-nested hypothesis testing procedures are based implicitly or explicitly on Kullback-Leibler Information Criterion (KLIC). We consider the non-nested hypothesis testing of Vuong (1989) for which the null hypothesis is that the candidate models are equidistant in KLIC from an unknown true model. The test statistic is asymptotically standard normal. We propose a higher order asymptotic bias correction of the test statistic and show that it is invariant with respect to reparametrization. Thus, the simplest possible parameterization can be used when calculating the test statistic. The reparametrization invariance leads to the differential geometrical approach where coordinate system invariant quantities like curvature are useful for understanding the corrections. The relationship of the correction factor with the preferred point geometry of Critchley et al. (1993, 1994) and the expected geometry of Amari (1982) is illustrated.

AMS 2000 classification: 62F03; 62F05.

Keywords: non-nested models, Kullback-Leibler information criterion, higher order asymptotics, curvature.

*hc269@cornell.edu. 404 Uris Hall, Department of Economics, Cornell University, Ithaca, NY, 14850, USA.

†nmk1@cornell.edu. 490 Uris Hall, Department of Economics and Department of Statistical Science, Cornell University, Ithaca, NY, 14850, USA.

1 Introduction

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. The true model can lie in one of the competing models, but it is not clear which model should be given the role of the null and which the alternative. However many non-nested tests are based on this approach. This includes the pioneering work of Cox (1961, 1962) based on log likelihood ratios, and the popular J-test of Davidson and MacKinnon (1981) based on the artificial nesting approach. On the other hand, Vuong (1989) proposed to test the null hypothesis that competing models are equidistant from an unknown true model.

Vuong's test treats the two competing models symmetrically and the divergence from the true model to the candidate models is measured by *Kullback-Leibler Information criterion (KLIC)* (relative entropy, Kullback and Leibler (1951)) between the unknown true model ϕ and a pseudo-true model. A pseudo-true model is defined as the closest member of the candidate parametric family in KLIC. The function KLIC, or more generally a divergence function, is always non-negative and equal to zero if and only if the two models have identical distributions (see Csiszár (1967a,b, 1975)). Noting that KLIC is not metric, we clarify that the divergence in Vuong's test is based on KLIC from the true model to the pseudo-true models not vice versa. Vuong's approach does not require specifying a true model ϕ, since the difference in KLIC for candidate models 1 and 2 is given by

$$KLIC_1 - KLIC_2 = E_\phi(l_\phi - l_1) - E_\phi(l_\phi - l_2) \tag{1.1}$$

$$= E_\phi(l_2 - l_1), \tag{1.2}$$

where l_ϕ, l_1, and l_2 are log likelihood functions of the true model ϕ, and the pseudo-true models of the competing models 1 and 2 respectively. Under the null that $E_\phi(l_2 - l_1) = 0$, Vuong (1989) proposed a normalized sample mean version of equation (1.2) for the test statistic.

Finite sample properties of this test statistic are not studied comprehensively. Recently, Rivers and Vuong (2002) and Choi and Kiefer (2005) extended the idea to dynamic models. Choi and

Kiefer (2005) also studied the finite sample properties of their test statistics for dynamic models and proposed to use the fixed-b asymptotics developed by Kiefer and Vogelsang (2005). 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 test statistic to get better finite sample performance in the case of independent observations. Our approach is related to the idea of the Bartlett correction, extended to cover misspecified models. See Kent (1982) also for the properties of likelihood ratio statistics in misspecified models. We correct the bias of order $O(1/\sqrt{n})$ from the numerator of Vuong's test statistic. The proposed bias correction term can be estimated consistently. A similar approach to bias correction was used in Takeuchi's Information Criterion (TIC, Takeuchi (1976)) which is a variant of Akaike's Information Criterion (AIC, Akaike (1973)) for possible misspecification of the models.

The bias correction term is shown to be invariant with respect to reparameterization, hence differential geometrical approaches are used to understand the effect of the correction factor. Differential geometrical quantities like curvatures can describe parameterization invariant statistical quantities such as the Bartlett correction. See Barndorff-Nielsen and Cox (1984) and McCullagh and Cox (1986) for the Bartlett correction for correctly specified models. For exponential family models, we show that our bias correction factor can be decomposed into two parts. One part is related to the degree of misspecification and the other is generated by the curvatures of the candidate models. The former is related to the preferred point geometry of Critchley et al. (1993, 1994) and is a model-independent constant when the statistical manifold is totally flat as defined in Critchley et al. (1994). The latter is related to the embedding curvature of Efron (1975, 1978) and Amari (1982). The embedding curvature vanishes if the model is a linear exponential family. Throughout the paper we will consider i.i.d. samples and assume the regularity conditions in Amari (1985) p.16.

2 Higher order bias correction of the test statistic

2.1 Main Results

Consider two candidate models $p_1(y|\theta_1)$ and $p_2(y|\theta_2)$ with log likelihood functions $l_1(\theta_1)$ and $l_2(\theta_2)$ (we will denote $p_j(y|\theta_j)$ as $p(\theta_j)$, and $l_j(\theta_j)$ as $l(\theta_j)$ for models $j = 1, 2$, when it does not cause confusion). When the models are misspecified, the probability limits θ_1^* and θ_2^* of the MLEs $\hat{\theta}_1$ and $\hat{\theta}_2$ are called the pseudo-true values and the distributions $p(\theta_1^*)$ and $p(\theta_2^*)$ are pseudo-true models. The pseudo-true values also minimize KLIC from the true model. The non-nested test of Vuong (1989) is based on the difference in KLIC from the true model p_0 to the pseudo-true models $p(\theta_1^*)$ and $p(\theta_2^*)$. The null hypothesis is that they are equidistant, i.e.

$$KLIC(p_0, p(\theta_1^*)) = KLIC(p_0, p(\theta_2^*)), \tag{2.1}$$

or equivalently,

$$KLIC(p_0, p(\theta_1^*)) - KLIC(p_0, p(\theta_2^*)) = E_0\left\{(l(\theta_2^*) - l_0) - (l(\theta_1^*) - l_0)\right\} \tag{2.2}$$

$$= E_0(l(\theta_2^*) - l(\theta_1^*)) = 0, \tag{2.3}$$

where E_0 is the expectation with respect to p_0. We consider whichever closest to the true model in this criterion as a better model.

Under Vuong's null hypothesis, the test statistic t_n (with i.i.d. data) is asymptotically normal and given by

$$t_n = \frac{(l(\hat{\theta}_2) - l(\hat{\theta}_1))/\sqrt{n}}{\sqrt{\widehat{V}_n}}, \tag{2.4}$$

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

$$l(\theta_j) = \sum_{i=1}^{n} l_i(\theta_j), \tag{2.5}$$

$$\bar{l}(\theta_j) = \frac{1}{n} \sum_{i=1}^{n} l_i(\theta_j), \tag{2.6}$$

for $j = 1, 2$, the variance V is estimated by

$$\widehat{V}_n = \frac{1}{n} \sum_{i=1}^{n} \left\{ (l_i(\hat{\theta}_2) - \bar{l}(\hat{\theta}_2)) - (l_i(\hat{\theta}_1) - \bar{l}(\hat{\theta}_1)) \right\}^2. \tag{2.7}$$

This test statistic requires that no model contains the true model. If one does, the other model must also contain the true model to be equidistant in KLIC from the true model. Thus they are identical and the test makes no sense. We also assume that the pseudo-true models are not identical, i.e. $p(\theta_1^*) \neq p(\theta_2^*)$, in which case the test statistic also degenerates. See Vuong (1989) for a discussion of testing the degeneracy of the test statistic. In this paper, the two models are non-nested in the sense that their pseudo-true models are not identical to exclude the degenerate case. But they can generally intersect at other parameter values since we are interested in the local behavior around the pseudo-true models.

We develop a higher order bias correction for the numerator of the test statistic in equation (2.4) decomposing the term $l(\hat{\theta}_2) - l(\hat{\theta}_1)$ by

$$l(\hat{\theta}_2) - l(\hat{\theta}_1) = (l(\theta_2^*) - l(\theta_1^*)) + (l(\hat{\theta}_2) - l(\theta_2^*)) - (l(\hat{\theta}_1) - l(\theta_1^*)) \tag{2.8}$$

$$= S_1 + S_2, \tag{2.9}$$

where $S_1 = l(\theta_2^*) - l(\theta_1^*)$ is the log likelihood ratio of the pseudo-true models, and $S_2 = (l(\hat{\theta}_2) - l(\theta_2^*)) - (l(\hat{\theta}_1) - l(\theta_1^*))$ is the remainder coming from the estimation of the pseudo-true models. The null hypothesis implies

$$E_0(S_1) = E_0(l(\theta_2^*) - l(\theta_1^*)) = 0. \tag{2.10}$$

Therefore the numerator (under the null) has a bias equal to $E_0(S_2)$. Using the expansion

$$l(\hat{\theta}_j) - l(\theta_j^*) = -\frac{1}{2} tr\{H(\theta_j^*)^{-1} s(\theta_j^*) s(\theta_j^*)^T\} + O_p(1/\sqrt{n}), \tag{2.11}$$

for $j = 1, 2$, where $H(\theta_j^*) = E_0 h(\theta_j^*) = \sum_{i=1}^n E_0 h_i(\theta_j^*)$ is the sum of the expected Hessians $h_i(\theta_j)$, and $s(\theta_j^*) = \sum_{i=1}^n s_i(\theta_j^*)$ is the score function of the model j, the bias $E_0(S_2)$ can be calculated by

$$\begin{aligned} E_0(S_2) &= E_0 \left\{ (l(\hat{\theta}_2) - l(\theta_2^*)) - (l(\hat{\theta}_1) - l(\theta_1^*)) \right\} && (2.12) \\ &= -\frac{1}{2} tr\{H(\theta_2^*)^{-1} J(\theta_2^*)\} + \frac{1}{2} tr\{H(\theta_1^*)^{-1} J(\theta_1^*)\} + O(1/\sqrt{n}), && (2.13) \\ &= -\frac{1}{2} tr\{\bar{H}(\theta_2^*)^{-1} \bar{J}(\theta_2^*)\} + \frac{1}{2} tr\{\bar{H}(\theta_1^*)^{-1} \bar{J}(\theta_1^*)\} + O(1/\sqrt{n}), && (2.14) \end{aligned}$$

where

$$\begin{aligned} J(\theta_j^*) &= E_0(s(\theta_j^*) s(\theta_j^*)^T) && (2.15) \\ \bar{J}(\theta_j^*) &= \frac{J(\theta_j^*)}{n}, && (2.16) \\ \bar{H}(\theta_j^*) &= \frac{H(\theta_j^*)}{n}, && (2.17) \end{aligned}$$

for $j = 1, 2$. We propose the correction from the first order term in equation (2.14),

$$b = -\frac{1}{2} tr\{\bar{H}(\theta_2^*)^{-1} \bar{J}(\theta_2^*)\} + \frac{1}{2} tr\{\bar{H}(\theta_1^*)^{-1} \bar{J}(\theta_1^*)\}. \tag{2.18}$$

The term $tr\{\bar{H}(\theta_j^*)^{-1} \bar{J}(\theta_j^*)\}$ in b can be quite large when many parameters are used, and can be zero if the model is defined as a point, say $\theta = \theta^*$.

Theorem 2.1. *Let the bias correction* $\hat{b}$ *be*

$$\hat{b} = -\frac{1}{2} tr\{\bar{H}(\hat{\theta}_2)^{-1} \bar{J}(\hat{\theta}_2)\} + \frac{1}{2} tr\{\bar{H}(\hat{\theta}_1)^{-1} \bar{J}(\hat{\theta}_1)\}, \tag{2.19}$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are (quasi) MLEs. The bias-corrected test statistic $\tilde{t}_n$ is given by

$$\tilde{t}_n = \frac{(l(\hat{\theta}_2) - l(\hat{\theta}_1) - \hat{b})/\sqrt{n}}{\sqrt{\widehat{V}_n}}, \tag{2.20}$$

and the bias of the numerator is of order $O(1/n)$.

Proof. The order of the bias of the numerator immediately follows from $\hat{\theta}_j - \theta_j^* = O_p(1/\sqrt{n})$ for $j = 1, 2$ and

$$\hat{b} = b + O_p(1/\sqrt{n}).$$

∎

The proposed bias correction can be shown to be a part of the higher $(1/\sqrt{n})$ order term in the Edgeworth expansion of the test statistic. The other part is related to the skewness of the numerator.

The following theorem shows that the bias b in equation (2.18) is reparameterization invariant and therefore a geometric object.

Theorem 2.2. *Let θ be the original parameterization and $\xi(\theta)$ be a locally one-to-one reparameterization of θ with $\xi^* = \xi(\theta^*)$. Then*

$$tr\{H(\theta^*)^{-1}J(\theta^*)\} \tag{2.21}$$

in equation (2.18) is invariant with respect to reparameterization $\xi(\theta)$, i.e.

$$tr\{H(\theta^*)^{-1}J(\theta^*)\} = tr\{H(\xi^*)^{-1}J(\xi^*)\}. \tag{2.22}$$

Proof. Let the matrix $D(\xi) = \partial\theta(\xi)^T/\partial\xi$. Since the transformation is locally isomorphic, $D(\xi^*)$ is invertible. The score function is

$$s(\xi) = D(\xi)s(\theta(\xi)), \tag{2.23}$$

and its variance $J(\xi)$ is given by

$$J(\xi) = D(\xi)J(\theta(\xi))D(\xi)^T, \tag{2.24}$$

showing that $J(\theta^*)$ is a tensor. The (a, b) element $h_{ab}(\xi)$ of the Hessian $h(\xi) = [h_{ab}(\xi)]$ is

$$h_{ab}(\xi) = \sum_{k,l} D_{ak}(\xi)h_{kl}(\theta(\xi))D_{bl}(\xi) + \sum_k \partial D_{ak}(\xi)/\partial\xi_b \; s_k(\theta(\xi)), \tag{2.25}$$

and the second summation in the equation (2.25) above has zero expectation at ξ^* since $E_0\{s_k(\theta(\xi^*))\} = 0$ by definition of the pseudo-true value. Therefore we have

$$E_0(h_{ab}(\xi^*)) = H_{ab}(\xi^*) = \sum_{k,l} D_{ak}(\xi^*)H_{kl}(\theta(\xi^*))D_{bl}(\xi^*), \tag{2.26}$$

which also can be written as

$$H(\xi^*) = D(\xi^*)H(\theta(\xi^*))D(\xi^*)^T, \tag{2.27}$$

showing that $H(\theta^*)$ is also a tensor. From the invertibility of $D(\xi^*)$, we have

$$tr\{H(\xi^*)^{-1}J(\xi^*)\} = tr\left[\left\{D(\xi^*)H(\theta(\xi^*))D(\xi^*)^T\right\}^{-1} D(\xi^*)J(\theta(\xi^*))D(\xi^*)^T\right] \tag{2.28}$$

$$= tr\{H(\theta^*)^{-1}J(\theta^*)\}. \tag{2.29}$$

∎

The theorem above makes it possible to use any convenient parameterization for calculation of the bias. We use locally affine parameterizations in which the Fisher information becomes an identity matrix at a particular point of interest (in our case, the pseudo-true models). A globally affine parameterization in which the information matrix is identity everywhere does not generally exist except in one-dimensional parameter models. See Amari (1985) for details.

The invariance leads to the interpretation of the bias correction term using differential geometrical quantities. We next study the bias-corrected test statistic in exponential families and highlight

the differential geometrical interpretation. Extensions of the interpretation to general families of distributions are discussed.

2.2 Curved exponential families

Curved exponential family (CEF) distributions are obtained from (linear) exponential family distributions by reducing the parameter dimension through restriction (Efron (1975)). The dimension of the sufficient statistic is unchanged, unless the restricted model is also linear. Efron (1975) notes that MLE entails an information loss by summarizing the sufficient statistic with a lower dimensional statistic. Efron defined the statistical curvature as a measure of how far the model is from the full exponential family where no information loss occurs. His curvature is invariant to reparameterization and has crucial implications for the information loss in using the MLE rather than the sufficient statistic to summarize the data. The applications of curvature to the higher-order efficiency for one dimensional parameter were studied by Efron (1975, 1978), and Eguchi (1984). Multi-dimensional parameter CEFs were studied in Amari (1982) and Amari and Kumon (1988b). The differential geometrical theory of higher-order asymptotics of statistical test and interval estimators was developed in Amari and Kumon (1983) and Amari and Kumon (1988a). Kass and Vos (1997) summarize the developments in this area. See Barndorff-Nielsen (1978), Barndorff-Nielsen et al. (1986), and Brown (1986). Many econometric models, including simultaneous equations models, finite order AR models, and linear regression models with nonlinear restrictions on parameters are known to be CEFs (see Van Garderen (1996, 1997)).

The density $p_0(y|\eta)$ of a full exponential family distribution in its canonical (or natural), linear, parameterization η can be written as

$$p_0(y|\eta) = \exp\left[n\left\{\bar{y}^T\eta - \psi(\eta)\right\}\right] f(y), \tag{2.30}$$

where n is the number of i.i.d. observations, $\bar{y}$ is the k-dimensional vector of sufficient statistics, η is the k-dimensional parameter vector, and y is the n-dimensional vector of observations. The function $\psi(\eta)$, the log of the normalizing constant, is the cumulant generating function. The cumulants of

one observation y_1 are obtained by differentiating $\psi(\eta)$. The Fisher information matrix of one observation with respect to the natural parameterization is $\psi''(\eta)$.

A curved exponential family (CEF) is a lower dimensional reparameterization θ of η, and the density is given by

$$p(y|\theta) = \exp\left[n\left\{\bar{y}^T\eta(\theta) - \psi(\eta(\theta))\right\}\right] f(y), \tag{2.31}$$

where θ is an $m < k$ dimensional parameter vector. If $\eta(\theta)$ is affine, $p(y|\theta)$ becomes a lower dimensional full exponential family. Efron (1975) defined the statistical curvature $\kappa(\theta)$ at θ for an one-dimensional CEF ($m = 1$) by

$$\kappa(\theta) = \|\eta'(\theta)\|_{\eta(\theta)}^{-3}\left[\|\eta'(\theta)\|_{\eta(\theta)}^2\|\eta''(\theta)\|_{\eta(\theta)}^2 - \langle\eta'(\theta),\eta''(\theta)\rangle_{\eta(\theta)}^2\right]^{1/2}, \tag{2.32}$$

where $g(\eta(\theta)) = \partial^2\psi\left(\eta(\theta)\right)/\partial\eta\partial\eta^T$ is the (Fisher) information matrix of the full exponential family, $\langle x_1, x_2\rangle_{\eta(\theta)} = x_1' g(\eta(\theta))x_2$ is the inner product of x_1 and x_2 with respect to the metric $g(\eta(\theta))$, and $\|x_1\|_{\eta(\theta)}^2 = \langle x_1, x_1\rangle_{\eta(\theta)}$ is the norm of x_1. Intuitively, it is the standardized (rescaled to be parameterization invariant) norm of $\eta''(\theta)$ projected onto the space orthogonal to the space spanned by $\eta'(\theta)$ with respect to the metric defined by the Fisher information matrix. The curvature is invariant with respect to a reparameterization of θ and is equal to zero for a full exponential family. Efron (1975) showed this curvature has an important implication in the higher order efficiency of estimators, especially MLEs. The curvature for a multi-dimensional CEF is more complicated. Amari (1982) generalized the notion of the Efron's curvature. He called the Efron's curvature the 1-curvature (among more general α-curvatures). It is also called the exponential curvature since it vanishes in linear exponential families.

We consider two CEFs $p(\theta_1)$ and $p(\theta_2)$, where θ_1 and θ_2 are $m_1, m_2 < k$ dimensional parameter vectors respectively, as in equation (2.31) in a k-dimensional full exponential family of equation (2.30). These two families are the candidates for the non-nested test. Let $p_0(y|\eta = \phi)$ be the true model in the full exponential family which does not lie in either of the candidate models, and θ_1^* and θ_2^* be the pseudo-true values of model 1 and 2. Thus $\eta(\theta_1) \neq \phi$ and $\eta(\theta_1) \neq \phi$ for any value of θ_1 and θ_2. The sufficient statistic is $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$,and $\mu = E_0\left(\bar{y}\right)$ is the mean parameter vector of

the true model. (Note that $\phi = \eta(\mu)$ is the natural parameter vector of the true model and they have the relationship $\mu = \psi'(\phi)$). The (uncorrected) test statistic t_n in equation (2.4) is given by

$$t_n = \frac{l(\hat{\theta}_2) - l(\hat{\theta}_1)/\sqrt{n}}{\sqrt{\widehat{V}_n}} \tag{2.33}$$

$$= \frac{\sqrt{n}\left[\bar{y}^T(\eta(\hat{\theta}_2) - \eta(\hat{\theta}_1)) - \left\{\psi(\eta(\hat{\theta}_2) - \psi(\eta(\hat{\theta}_1)))\right\}\right]}{\sqrt{\widehat{V}_n}}, \tag{2.34}$$

where $\hat{\theta}_1$, $\hat{\theta}_2$ are MLEs, and

$$\widehat{V}_n = (\eta(\hat{\theta}_2) - \eta(\hat{\theta}_1))^T g(\eta(\bar{y}))(\eta(\hat{\theta}_2) - \eta(\hat{\theta}_1)), \tag{2.35}$$

is the variance estimator. The estimator $g(\eta(\bar{y}))$ of the information matrix $g(\eta(\mu))$ for one observation at the true model $\eta = \phi$ is calculated by

$$g(\eta(\bar{y})) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T, \tag{2.36}$$

or using the Hessian function $g(\eta(\bar{y})) = \psi''(\eta(\bar{y}))$.

2.3 Bias correction for one-dimensional curved exponential models

For one-dimensional parameter CEFs, the score and Hessian functions become

$$s(\theta) = n\left\{\bar{y} - \psi'(\eta(\theta))\right\}^T \eta'(\theta), \tag{2.37}$$

$$h(\theta) = n\left[\left\{\bar{y} - \psi'(\eta(\theta))\right\}^T \eta''(\theta) - \eta'(\theta)^T\psi''(\eta(\theta))\eta'(\theta)\right] \tag{2.38}$$

$$= n\left[\left\{\bar{y} - \psi'(\eta(\theta))\right\}^T \eta''(\theta) - i(\theta)\right], \tag{2.39}$$

where $i(\theta) = \eta'(\theta)^T\psi''(\eta(\theta))\eta'(\theta)$ is the Fisher information of one observation at $\eta(\theta)$. We will write $\psi'(\eta(\theta)) = \psi'(\theta)$ and $\psi''(\eta(\theta)) = \psi''(\theta)$ for simplicity. The expected score $E\{s(\theta^*)\}$ and the

average of the expected Hessian $\bar{H}(\theta^*) = H(\theta^*)/n$ at the pseudo-true value θ^* are

$$E\left\{s(\theta^*)\right\} = n(\mu - \psi'(\theta^*))^T \eta'(\theta^*) = 0, \tag{2.40}$$

$$\bar{H}(\theta^*) = \left[(\mu - \psi'(\theta^*))^T \eta''(\theta^*) - i(\theta^*)\right]. \tag{2.41}$$

Note that when the CEF contains the true model, we have $\mu = \psi'(\theta^*)$, and equation (2.41) becomes

$$\bar{H}(\theta^*) = -\eta'(\theta^*)^T \psi''(\theta^*)\eta'(\theta^*) = -i(\theta^*). \tag{2.42}$$

When the CEF is misspecified, we have $\mu - \psi'(\theta^*) \neq 0$, but by the orthogonality of $\mu - \psi'(\theta^*)$ and $\eta'(\theta^*)$, equation (2.40) still holds. However, we do not have the Fisher information equality in this case. The variance of the score $\bar{J}(\theta^*)$ of one observation is

$$\bar{J}(\theta) = \eta'(\theta^*)^T g(\phi)\eta'(\theta^*), \tag{2.43}$$

where $\eta = \phi$ is the true model.

When the parameter θ satisfies

$$i(\theta) = \|\eta'(\theta)\|^2_{\eta(\theta)} = 1, \text{ for all } \theta, \tag{2.44}$$

the parameterization is called an arclength parameterization or 0-affine. Since the bias correction is invariant, we are free to use the arclength parameterization.

If we decompose $\eta''(\theta)$ into a tangential component $(\eta''(\theta))_T$ and a normal component $(\eta''(\theta))_N$ to $\eta'(\theta)$ with respect to the metric $g(\eta(\theta))$, i.e.

$$\eta''(\theta) = (\eta''(\theta))_T + (\eta''(\theta))_N, \tag{2.45}$$

and

$$\langle \eta'(\theta), (\eta''(\theta))_N \rangle_{\eta(\theta)} = 0, \tag{2.46}$$

then, with the arclength parameterization, there exists a useful relationship

$$\kappa(\theta) = \| \left(\eta''(\theta)\right)_N \|_{\eta(\theta)}, \tag{2.47}$$

between the curvature $\kappa(\theta)$ and the norm of $\left(\eta''(\theta)\right)_N$.

Lemma 2.3. *Using the arclength parameterization, the bias in equation (2.18) can be calculated from*

$$tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\} = \frac{\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*)}{\left\langle (\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}, \left(\eta''(\theta_j^*)\right)_N \right\rangle_{\eta(\theta)} - 1}, \tag{2.48}$$

for model $j = 1, 2$. If model j is exponential flat, we have

$$tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\} = -\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*). \tag{2.49}$$

Proof. Using equation (2.41), (2.43) and $i(\theta_j^*) = 1$, we have

$$tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\} = \frac{\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*)}{(\mu - \psi'(\theta_j^*))^T\eta''(\theta_j^*) - 1}, \tag{2.50}$$

for each model $j = 1, 2$. The term $(\mu - \psi'(\theta_j^*))^T\eta''(\theta_j^*)$ in the denominator can be rewritten as

$$(\mu - \psi'(\theta_j^*))^T\eta''(\theta_j^*) = \left\langle (\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}, \eta''(\theta_j^*) \right\rangle_{\eta(\theta)}. \tag{2.51}$$

Since the orthogonality condition in equation (2.40) implies $(\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}$ is orthogonal to $\eta'(\theta_j^*)$, i.e.

$$(\mu - \psi'(\theta^*))^T\eta'(\theta^*) = \left\langle (\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}, \eta'(\theta_j^*) \right\rangle_{\eta(\theta)} = 0, \tag{2.52}$$

we have

$$(\mu - \psi'(\theta_j^*))^T\eta''(\theta_j^*) = \left\langle (\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}, \left(\eta''(\theta_j^*)\right)_N \right\rangle_{\eta(\theta)}, \tag{2.53}$$

from equation (2.45) and (2.51).

When the model is exponential flat, $\kappa(\theta_j^*) = \left\| \left(\eta''(\theta_j^*)\right)_N \right\|_{\eta(\theta)} = 0$ gives the second result. ∎

We showed that the denominator of equation (2.48) is related to the curvature $\kappa(\theta_j^*)$ at the pseudo-true model, and the numerator is related to the information matrix $g(\phi)$ at the true model ϕ. In general, $g(\phi)$ is different from $g(\eta(\theta_j^*))$ because of misspecification ($\eta(\theta_j^*) \neq \phi$). But if the information matrix of the full exponential family is constant, we have $g(\eta(\theta_j^*)) = g(\phi)$, which implies that the numerator

$$\eta'(\theta_j^*)^T g(\phi) \eta'(\theta_j^*) = 1, \tag{2.54}$$

by the arclength parameterization. The condition $g(\eta(\theta)) = g(\phi)$ is satisfied by a totally flat manifold in exponential families.

Definition 2.4 (Critchley et al. (1994))**.** *For a fixed (true) model ϕ, define*

$$\mu^\phi(\eta) = E_\phi(s(\eta)), \tag{2.55}$$

$$g^\phi(\eta) = Var_\phi(s(\eta)), \tag{2.56}$$

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, $g(\eta)$ is constant (see Theorem 4 in Critchley et al. (1994)) and 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. We have the following theorem about the relationship between the geometry of the models and the bias.

Theorem 2.5. *For one dimensional curved exponential family, the log of $-tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\}$ can be decomposed by*

$$\ln\left(-tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\}\right) = P + K, \tag{2.57}$$

where $P = \ln\left\{\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*)\right\}$ *and* $K = -\ln\left\{1 - (\mu - \psi'(\theta_j^*))^T \eta''(\theta_j^*)\right\}$ *for the candidate models* $j = 1, 2$.. *If the model is correctly specified, then* $P = K = 0$. *When the model is misspecified,* $P = 0$ *if the full exponential family is totally flat as defined in Critchley et al. (1994), and* $K = 0$ *if the exponential curvature of Efron (1975) is zero at the pseudo-true model.*

Proof. The decomposition directly follows from equation 2.48 using the arclength parameterization. If the model is correctly specified ($\phi = \eta(\theta_j^*)$), we have $\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*) = \eta'(\theta_j^*)^T g(\eta(\theta_j^*))\eta'(\theta_j^*) = \left\|\eta'(\theta_j^*)\right\|^2_{\eta(\theta)} = 1$ which implies $P = 0$, and $K = 0$ from $\mu = \psi'(\theta_j^*)$. If the model is misspecified, $\phi \neq \eta(\theta_j^*)$, and if the exponential family is totally flat, the information matrix $g(\eta)$ is constant from the Theorem 4 in Critchley et al. (1994), therefore $g(\eta(\theta_j^*)) = g(\phi)$ gives $P = 0$. Also if the model has zero exponential curvature, $K = 0$ from Lemma 2.3. ∎

2.4 Multi-parameter CEFs

When the parameter θ is m-dimensional and η is k-dimensional ($k > m$), the score vector at θ is given by

$$s(\theta) = n\eta'(\theta)^T(\bar{y} - \psi'(\theta)), \tag{2.58}$$

where $\eta'(\theta)$ is now the $k \times m$ matrix $\partial\eta(\theta)/\partial\theta' = \left[\partial\eta(\theta)/\partial\theta_1 \ \ ... \ \ \partial\eta(\theta)/\partial\theta_m\right]$, and the variance $\bar{J}(\theta_j^*) = J(\theta_j^*)/n$ of the score vector $s(\theta_j^*)$ at the pseudo-true model for models $j = 1, 2$, is given by

$$\bar{J}(\theta_j^*) = \eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*). \tag{2.59}$$

The Hessian matrix $h(\theta)$ has (a, b) elements

$$h_{ab}(\theta) = n\left[(\bar{y} - \psi'(\eta(\theta)))^T \eta_{ab}(\theta) - i_{ab}(\theta)\right], \tag{2.60}$$

where $\eta_{ab}(\theta) = \partial^2\eta(\theta)/\partial\theta_a\partial\theta_b$, and $i_{ab}(\theta) = \left(\partial\eta(\theta)/\partial\theta_a\right)^T g(\eta(\theta))\left(\partial\eta(\theta)/\partial\theta_b\right)$, and the average expected Hessian matrix, $\bar{H}(\theta_j^*) = [\bar{H}_{ab}(\theta_j^*)] = [H_{ab}(\theta_j^*)]/n$, has elements

$$\bar{H}_{ab}(\theta_j^*) = (\mu - \psi'(\eta(\theta_j^*)))^T \eta_{ab}(\theta_j^*) - i_{ab}(\theta_j^*). \tag{2.61}$$

Using equation (2.59), (2.61), the bias term,

$$b = -\frac{1}{2}tr\{\bar{H}(\theta_2^*)^{-1}\bar{J}(\theta_2^*)\} + \frac{1}{2}tr\{\bar{H}(\theta_1^*)^{-1}\bar{J}(\theta_1^*)\}, \tag{2.62}$$

can be calculated from

$$tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\} = tr(\left[\bar{H}_{ab}(\theta_j^*)\right]^{-1}\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*)) \tag{2.63}$$

$$= tr(\left[(\mu - \psi'(\eta(\theta_j^*)))^T\eta_{ab}(\theta_j^*) - i_{ab}(\theta_j^*)\right]^{-1}\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*)), \tag{2.64}$$

for $j = 1, 2$.

To represent the term $tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\}$ in geometrical quantities, we consider a differentiable (smooth) manifold of probability densities of the full exponential family as considered in Amari (1982). The parameter η serves as a coordinate system on the manifold. The curved exponential family is the imbedded sub-manifold. We briefly summarize the differential geometrical approach of Amari (1982). We define the differential operator

$$\partial_a = \frac{\partial}{\partial\theta_a}, \tag{2.65}$$

$$\partial_a\partial_b = \frac{\partial^2}{\partial\theta_a\partial\theta_b}, \tag{2.66}$$

where θ_a is the a^{th} parameter for $a = 1, 2, ..., m$. The inner product of ∂_a and ∂_b is defined by

$$\langle\partial_a, \partial_b\rangle = Cov_\theta\left(\partial_a l(\theta), \partial_b l(\theta)\right) \tag{2.67}$$

$$= i_{ab}. \tag{2.68}$$

Note that $\bar{J}_{ab} = Cov_\phi\left(\partial_a l(\theta), \partial_b l(\theta)\right) \neq i_{ab}$ for misspecified models. The differential operators $\{\partial_1, \partial_2, ..., \partial_m\}$ span the tangent space at θ with the metric defined in the equation (2.67). Using the Einstein summation convention where the repeating upper and lower indices imply summation

over that index, the score function ∂_a can be represented as

$$\partial_a = B_a^i \partial_i, \tag{2.69}$$

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

The (imbedding) k-dimensional full exponential family can be reparameterized with the $k-m$ dimensional parameter ν in addition to the m-dimensional parameter vector θ. Thus (θ, ν) is a new (diffeomorphic) parameterization of η. Moreover we can choose the parameterization ν such that the score functions are locally orthonormal to ∂_a, i.e.

$$\langle \partial_a, \partial_\gamma \rangle = 0 \quad \text{for } a = 1, ..., m \text{ and } \gamma = 1, ..., k-m, \tag{2.70}$$

$$\langle \partial_\gamma, \partial_\zeta \rangle = \delta_\gamma^\zeta \quad \text{for } \gamma = 1, ..., k-m, \text{ and } \zeta = 1, ..., k-m, \tag{2.71}$$

where $\partial_\gamma = \partial/\partial\nu_\gamma$, and $\delta_\gamma^\zeta = 1$ for $\zeta = \gamma$, zero otherwise. The *Euler-Schouten curvature tensor* or the *imbedding curvature* of the CEF in the full exponential family is given by

$$H_{ab\gamma}(\theta) = \langle \partial_a \partial_b, \partial_\gamma \rangle \tag{2.72}$$

$$= E\left\{ (\partial_a \partial_b - E\partial_a \partial_b)\partial_\gamma \right\}. \tag{2.73}$$

The *Euler-Schouten curvature* $H_{ab\gamma}(\theta)$ is an important geometrical quantity for the higher order asymptotic analysis It depends on the imbedding space which means it is extrinsic, whereas the *Riemann-Christoffel curvature* is intrinsic. For example, the surface of a cylinder in three dimensional Euclidean space has zero *Riemann-Christoffel curvature* since one can unroll it to two dimensional Euclidean space without destroying its geometrical structure. But the *Euler-Schouten curvature tensor* is not zero since its tangent space changes around the cylinder.

The mean zero random variable $(\partial_a \partial_b - E\partial_a \partial_b)$ in equation (2.72) is called a covariant derivative

with respect to 1-connection, and from equation (2.60), we have

$$\partial_a\partial_b - E\partial_a\partial_b = n(\bar{y} - \psi'(\eta(\theta)))^T\eta_{ab}(\theta). \tag{2.74}$$

We can decompose $(\partial_a\partial_b - E\partial_a\partial_b)$ with the tangential component and the normal component to the space spanned by $\{\partial_1, \partial_2, ..., \partial_m\}$. The tangential and the normal components can be represented with the orthonormal bases ∂_c and ∂_γ respectively. We have

$$\partial_a\partial_b - E\partial_a\partial_b = n(\bar{y} - \psi'(\eta(\theta)))^T\eta_{ab}(\theta) \tag{2.75}$$

$$= \Gamma^c_{ab}\partial_c + H^\gamma_{ab}\partial_\gamma \tag{2.76}$$

$$= \Gamma^c_{ab}B^i_c\partial_i + H^\gamma_{ab}B^i_\gamma\partial_i, \tag{2.77}$$

where Γ^c_{ab} and H^γ_{ab} are the coefficients of the projected component onto the space spanned by the basis vectors ∂_c and ∂_γ respectively. The last equality is from equation (2.69). When the bases $\{\partial_\gamma\}$ are orthonormal to $\{\partial_c\}$, we have $H^\gamma_{ab} = H_{ab\gamma}$, and the coefficients H^γ_{ab} represents the coefficients of the imbedding curvatures.

Theorem 2.6. *The term $(\mu - \psi'(\eta(\theta^*_j)))^T\eta_{ab}(\theta^*_j)$ in equation (2.61) is given by*

$$n(\mu - \psi'(\eta(\theta^*_j)))^T\eta_{ab}(\theta^*_j) = A_iB^i_\gamma H^\gamma_{ab}, \tag{2.78}$$

where A_i be i^{th} element of $(\mu - \psi'(\eta(\theta^*_j)))$, and B^i_γ and H^γ_{ab} are defined in equation (2.69) and (2.76) respectively. If the model is 1-flat, or equivalently, has zero Euler-Schouten curvature (with respect to 1-connection) at θ^*_j, $n(\mu - \psi'(\eta(\theta^*_j)))^T\eta_{ab}(\theta^*_j) = 0$.

Proof. 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

$$\partial^i = g^{ij}\partial_j, \tag{2.79}$$

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

$$n(\mu - \psi'(\eta(\theta_j^*)))^T \eta_{ab}(\theta_j^*) = E\left\{(\mu - \psi'(\eta(\theta_j^*)))^T g(\eta(\theta_j^*))^{-1} n(\bar{y} - \psi'(\eta(\theta_j^*)))\right\}\left\{n(\bar{y} - \psi'(\eta(\theta_j^*)))^T \eta_{ab}(\theta_j^*)\right\} \tag{2.80}$$

$$= E\left\{A_i \partial^i\right\}\left\{\Gamma^c_{ab} B^i_c \partial_i + H^\gamma_{ab} B^i_\gamma \partial_i\right\} \tag{2.81}$$

$$= E\left\{A_i \partial^i\right\}\left(H^\gamma_{ab} B^i_\gamma \partial_i\right) \tag{2.82}$$

$$= A_i B^i_\gamma H^\gamma_{ab}, \tag{2.83}$$

where E is the expectation with respect to the distribution at $\eta(\theta_j^*)$. The third equality is from the zero expected score,

$$E_0 \partial_c = (\mu - \psi'(\eta(\theta_j^*)))^T \eta'(\theta_j^*) \tag{2.84}$$

$$= \left\langle A_i \partial^i, B^i_c \partial_i \right\rangle \tag{2.85}$$

$$= E\left\{A_i \partial^i\right\}\left\{B^i_c \partial_i\right\} = 0. \tag{2.86}$$

Note that the expectation E_0 is with respect to the true model $\eta = \phi$. Therefore we have the duality of the mean and natural parameterization showing that the coefficients A_i of the score functions of the mean parameterization ∂^i and the coefficients B^i_a of the score functions of the natural parameterization ∂_i which is called a dual parameterization of the mean parameterization, are orthogonal. When the curvature of the embedding model vanishes at θ_j^*, i.e. $H^\gamma_{ab} = 0$, we have $n(\mu - \psi'(\eta(\theta_j^*)))^T \eta_{ab}(\theta_j^*) = 0$. ∎

In the general m-dimensional parameter case ($m > 1$), there does not exist a reparameterization that makes the 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. The existence of such parameterization at the pseudo-true model is sufficient for our results. If we use a locally 0-affine parameterization such that $i_{ab}(\theta^*) = \delta^b_a$, then the bias can

be calculated from

$$tr\{H(\theta_j^*)^{-1}J(\theta_j^*)\} = tr(\left[A_iB_\gamma^iH_{ab}^\gamma - \delta_a^b\right]^{-1}\eta'(\theta_j^*)^Tg(\phi)\eta'(\theta_j^*)), \tag{2.87}$$

for $j = 1, 2$ using Theorem 2.6. When the model j is exponential flat ($H_{ab}^\gamma = 0$), we have

$$tr\{H(\theta_j^*)^{-1}J(\theta_j^*)\} = -tr(\eta'(\theta_j^*)^Tg(\phi)\eta'(\theta_j^*)). \tag{2.88}$$

Moreover if the full exponential family is totally flat ($g(\phi) = g(\eta(\theta_j))$), then $\eta'(\theta_j^*)^Tg(\phi)\eta'(\theta_j^*)$ is also a $(m_j \times m_j)$ identity matrix since θ_j is 0-affine and we have

$$tr\{H(\theta_j^*)^{-1}J(\theta_j^*)\} = -m_j, \tag{2.89}$$

where m_j is the dimension of the parameter vector in model j.

2.5 Summary and Extension

The term $tr\{H(\theta_j^*)^{-1}J(\theta_j^*)\}$ in equation (2.18) can be used for the general form of the higher order bias of the numerator of the test statistic. For one dimensional curved exponential families embedded in a full exponential family, the bias can be decomposed into two parts $(P + K)$ as shown in Theorem 2.5. The first part (P) vanishes when the imbedding model is totally flat and the other part (K) vanishes when the curved exponential model has zero Efron's curvature. For multiparameter curved exponential families, if the embedding exponential model is totally flat, we have $J(\theta_j^*) = I_{m_j}$, where I_{m_j} is an $(m_j \times m_j)$ identity matrix and m_j is the number of parameters in model j. If the model j has zero imbedding curvature with respect to 1-connection we have $H(\theta_j^*) = -I_{m_j}$.

We consider the extension of the results to general parametric families by approximating the models with exponential models around the pseudo-true models. We illustrate the idea for general (non-exponential) one-parameter models. Let $l_j = l_j(\theta_j)$ be a log likelihood function of model j. As proposed in Efron (1975), the log likelihood function $\tilde{l}(\eta)$ of the m-dimensional approximate

exponential model around θ_j^* is

$$\tilde{l}(\eta) = l_j^* + \sum_{k=1}^{m} \eta_k l^*_{j/\theta_j^k} - \psi(\eta), \tag{2.90}$$

where

$$l_j^* = l_j(\theta_j^*), \tag{2.91}$$

$$l^*_{j/\theta_j^k} = \left. \frac{\partial^k}{\partial \theta_j^k} l_j(\theta_j) \right|_{\theta_j = \theta_j^*}, \tag{2.92}$$

and $\psi(\eta)$ is a normalizing constant. The model $\tilde{l}(\theta_j)$ is a one-dimensional curved exponential model imbedded in $\tilde{l}(\eta)$ with

$$\eta(\theta_j) = \left((\theta_j - \theta_j^*), \frac{1}{2}(\theta_j - \theta_j^*)^2, \cdots, \frac{1}{m!}(\theta_j - \theta_j^*)^m \right)^T. \tag{2.93}$$

To approximate two separate families of models, we propose to consider an $(m_1 + m_2)$-dimensional exponential model

$$\tilde{l}(\eta) = l_1^* + \sum_{k=1}^{m_1} \eta_k l^*_{1/\theta_1^k} \tag{2.94}$$

$$+ l_2^* + \sum_{k=1}^{m_2} \eta_{m_1+k} l^*_{2/\theta_2^k} - \psi(\eta). \tag{2.95}$$

The model $j = 1, 2$ are given by two curved exponential families with

$$\eta(\theta_1) = \left((\theta_1 - \theta_1^*), \frac{1}{2}(\theta_1 - \theta_1^*)^2, \cdots, \frac{1}{m_1!}(\theta_1 - \theta_1^*)^{m_1}, 0, 0, \cdots, 0 \right)^T, \tag{2.96}$$

and

$$\eta(\theta_2) = \left(0, 0, \cdots, 0, (\theta_2 - \theta_2^*), \frac{1}{2}(\theta_2 - \theta_2^*)^2, \cdots, \frac{1}{m_2!}(\theta_2 - \theta_2^*)^{m_2} \right)^T, \tag{2.97}$$

respectively. The true model $\eta = \phi$ is given with respect to the mean parameterization $\mu(\eta)_{\eta=\phi}$,

$$\mu(\phi) = E_0 \left(l^*_{1/\theta^1_1}, l^*_{1/\theta^2_1}, \cdots, l^*_{1/\theta^{m_1}_1}, l^*_{2/\theta^1_2}, l^*_{2/\theta^2_2}, \cdots, l^*_{2/\theta^{m_2}_2} \right)^T, \tag{2.98}$$

where E_0 is the expectation with respect to the true model. Using the approximate embedding exponential model $\tilde{l}(\eta)$ and the approximate true model $\mu(\phi)$ on it, we can generalize the differential geometrical intuition to general families of models.

3 Fisher's circles

We consider an example with Fisher's circle models. The embedding space is a two-dimensional exponential family with identity Fisher information matrix in the natural parameterization.

Let y_1 and y_2 be independent normal random variables with variance one and mean η_1 and η_2 respectively. We define two models M_1 and M_2 by two nonlinear restrictions on the mean (η_1, η_2) of the random vector (y_1, y_2). The models are given by,

$$M_1 : (\eta_1 + 2)^2 + \eta_2^2 = 1, \tag{3.1}$$

$$M_2 : (\eta_1 + 0.5)^2 + \eta_2^2 = 1.5^2. \tag{3.2}$$

Figure 1 shows the models in $(\eta_1,\ \eta_2)$ plane. The true model $\eta = \mu = (\eta_1,\ \eta_2)$ is assumed to be $\mu = (0,\ 0)$ and the observed data are $y = (y_1,\ y_2)$. These two models have constant curvatures $\kappa_1 = 1$ $(radius = 1)$ and $\kappa_2 = 2/3$ $(radius = 1.5)$. The pseudo-true models are $\eta(\theta_1 = 0) = (-1, 0)^T$, $\eta(\theta_2 = 0) = (1, 0)^T$ and MLEs are given by the closest models $\eta(\hat{\theta}_1),\ \eta(\hat{\theta}_2)$ from y. For simplicity, we parameterize the models by the counter-clockwise arclength $\theta_1 \in [0, 2\pi)$, $\theta_2 \in [0, 6\pi)$ from the pseudo-true models. We can easily see the pseudo-true models of the two CEF circles have the same divergence in KLIC from the true model since KLIC can be directly calculated from the Euclidean distance in Fisher's setting.

We compare the original Vuong test statistic (from equations (2.34) and (2.35))

$$t_1 = \frac{\{\eta(\hat{\theta}_2) - \eta(\hat{\theta}_1)\}^T y - \{\eta(\hat{\theta}_2)^T \eta(\hat{\theta}_2) - \eta(\hat{\theta}_2)^T \eta(\hat{\theta}_2)\}/2}{\parallel \eta(\hat{\theta}_2) - \eta(\hat{\theta}_1) \parallel} \tag{3.3}$$

and the bias corrected test statistic

$$t_2 = t_1 - \frac{\hat{b}}{\parallel \eta(\hat{\theta}_2) - \eta(\hat{\theta}_1) \parallel}, \tag{3.4}$$

where

$$\hat{b} = -\frac{1}{2}\left(\frac{1}{\eta''(\hat{\theta}_2)^T(y - \eta(\hat{\theta}_2)) - 1} - \frac{1}{\eta''(\hat{\theta}_1)^T(y - \eta(\hat{\theta}_1)) - 1}\right), \tag{3.5}$$

and

$$\kappa_1 = \parallel \eta''(\hat{\theta}_1) \parallel, \quad \kappa_2 = \parallel \eta''(\hat{\theta}_2) \parallel . \tag{3.6}$$

Since the embedding space is totally flat, the bias correction term is driven by the curvatures only. Figure 2 is the density and the cumulative density function (CDF) of the two test statistics t_1 and t_2 from $3,000$ iterations. We can see that the original test statistic is biased toward model 2 (positive t_1) and the bias corrected test statistic is closer to the standard normal distribution. The first graph in Figure 3 shows the empirical CDF of the squared test statistics with compared to the CDF of $\chi^2(1)$. The 45 degree line implies exact match of the two CDFs. The bias corrected test statistic is closer to the chi-square distribution. This means it performs better in two tail tests. The second graph shows the empirical CDFs of t_1 and t_2 with respect to the standard normal CDF. The size approximation of the bias corrected test statistic especially improves in the left tail area and it is better than the original test statistic at all levels of tests.

To see the effect of curvatures of the models, we consider different radii (curvatures) $R =$ 1.1 (0.909) or 1.4 (0.714) or 2 (0.5) or 3 (0.333) for the model 2. Figure 4 shows the CDF comparisons from the different radii of Model 2. As the curvature of model 2 increases the improvement from the bias correction increases, as expected from our geometric analysis.

4 Conclusion

We showed that the numerator of the test statistic of the non-nested hypothesis test of Vuong (1989) can be modified with a higher order bias correction term that can be calculated by plugging in the MLEs. The bias correction term is shown to be reparameterization invariant. For a curved exponential family, we have shown that it is influenced by two geometrical factors, the total flatness of the embedding full exponential family and the Efron's curvatures of the candidate models. When the full exponential model is totally flat and the Efron's curvature is zero (no exponential curvature), the correction term is a simple function of the number of parameters used. In a simulation, bias correction clearly improved the performance of the test statistic.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (B. N. Petrov and F. Csáki, eds.). Akadémia Kiadó, Budapest, 267–281.

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

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

Amari, S. I. and Kumon, M. (1983). Differential geometry of edgeworth expansions in curved exponential family. *Annals Of The Institute Of Statistical Mathematics*, **35** 1–24.

Amari, S. I. and Kumon, M. (1988a). Differential geometry of testing hypothesis - a higher order asymptotic theory in multi-parameter curved exponential family. *Journal of The Faculty of Engineering, The University of Tokyo (B)*, **39** 241–273.

Amari, S. I. and Kumon, M. (1988b). Estimation in the presence of infinitely many nuisance parameters–geometry of estimating functions. *Annals of Statistics*, **16** 1044–1068.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, New York, NY.

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

Barndorff-Nielsen, O. E., Cox, D. R. and Reid, N. (1986). The role of differential geometry in statistical theory. *International Statistical Review*, **54** 83–96.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families*. IMS Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward, CA.

Choi, H.-S. and Kiefer, N. M. (2005). Robust model selection in dynamic models. *Working paper, Cornell University.*

Cox, D. (1961). Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. 105–123.

Cox, D. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B*, **24** 406–424.

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.

Csiszár, I. (1967a). Information type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, **2** 299–318.

Csiszár, I. (1967b). On topological properties of f-divergence. *Studia Scientiarum Mathematicarum Hungarica*, **2** 329–339.

Csiszár, I. (1975). i-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, **3** 146–158.

Davidson, R. and MacKinnon, J. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica*, **49** 781–793.

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

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

Eguchi, S. (1984). A characterization of second order efficiency in a curved exponential family. *Ann. Inst. Statist. Math.*, **36** 199–206.

Kass, R. E. and Vos, P. W. (1997). *Geometrical foundations of asymptotic inference*. John Wiley & Sons, New York, NY.

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

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.

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.

Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, **153** 12–18. In Japanese.

Van Garderen, K. (1996). Exact geometry of autoregressive models. *Journal of time series analysis*, **20** 1–21.

Van Garderen, K. (1997). Curved exponential models in econometrics. *Econometric Theory*, **13** 771–790.

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

A Figures

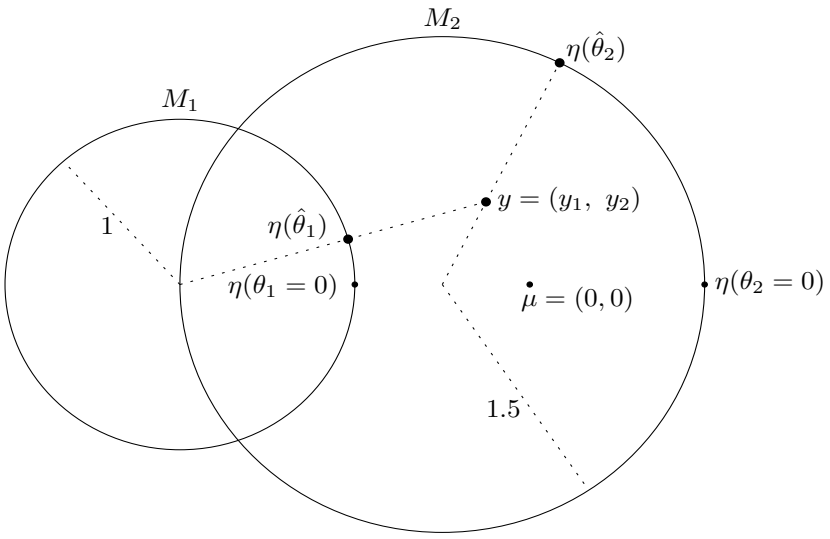

Figure 1: Two competing Fisher's circles

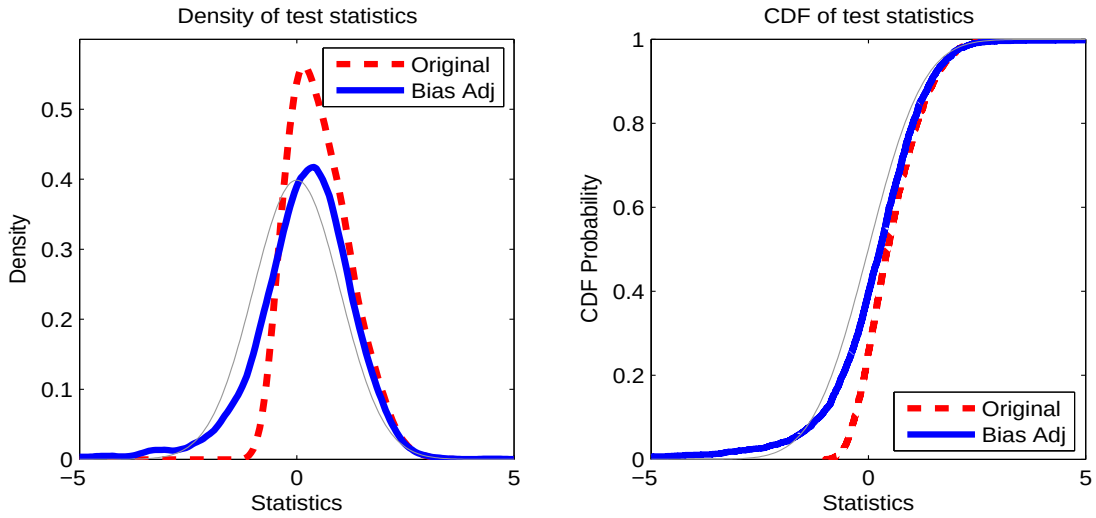

Figure 2: Comparison of the distributions of the original Vuong's and the bias corrected test statistics with the standard normal distribution. The thin lines are from $N(0,1)$

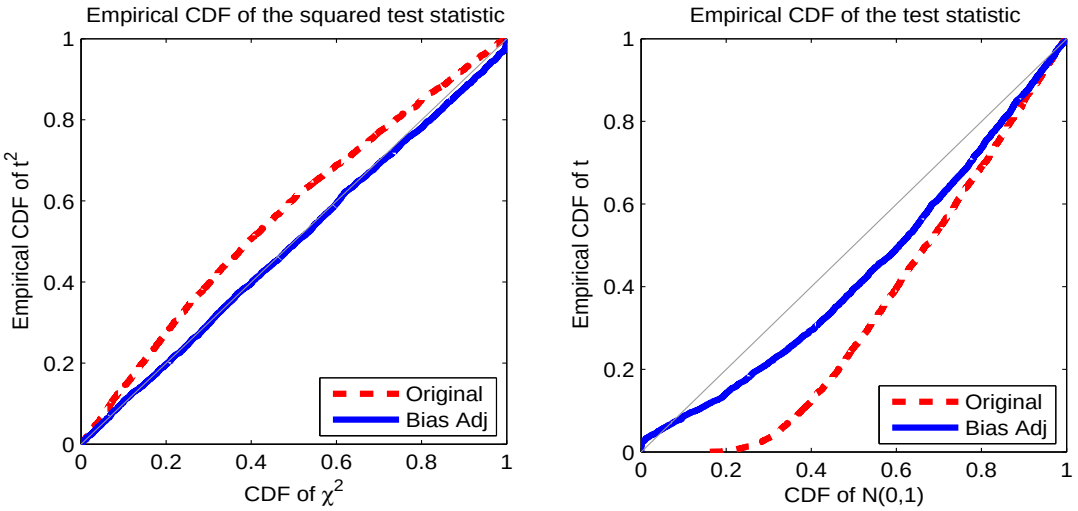

Figure 3: Empirical CDF of the squared test statistics with respect to the Chi-square CDF, and the empirical CDF of the test statistics with respect to the standard normal CDF. 45 degree lines imply exact match to the comparing CDF.

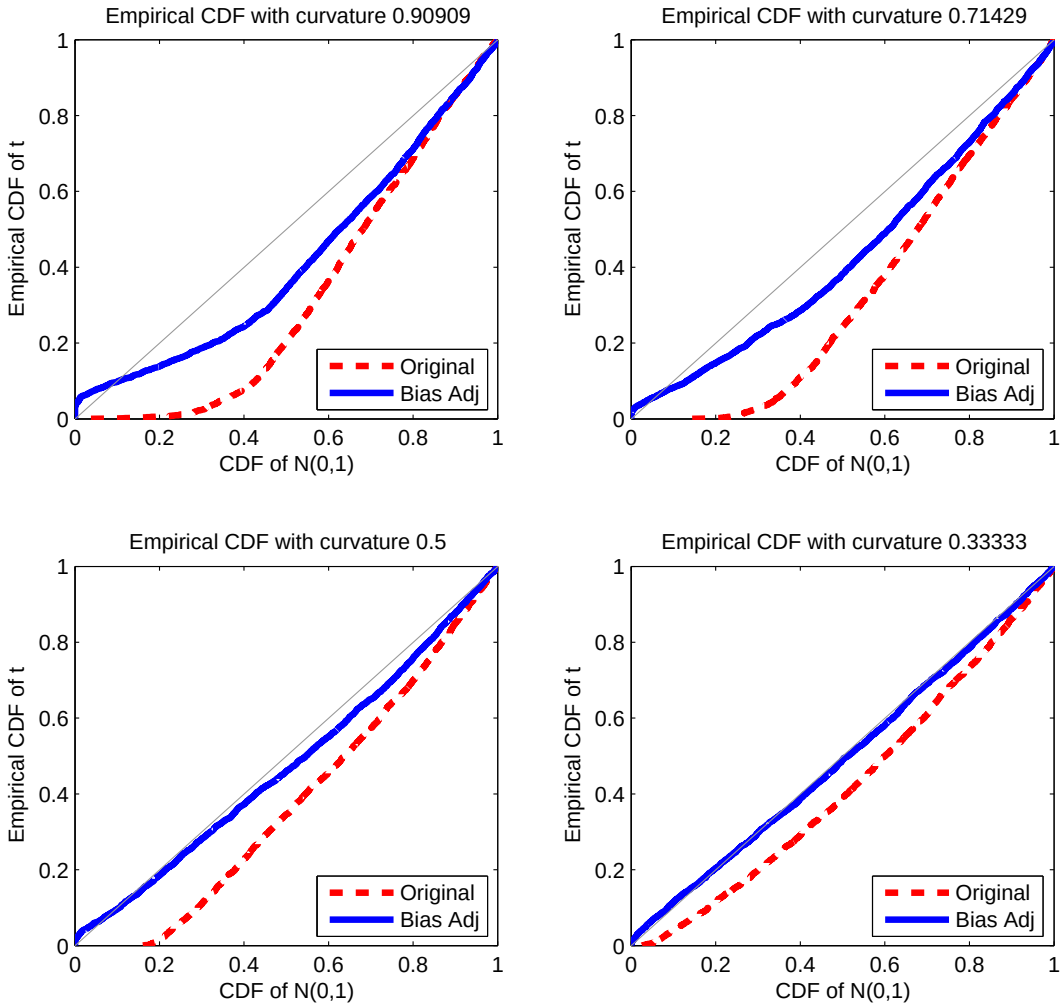

Figure 4: Comparison of CDFs of the test statistics from different curvatures for Model 2. The 45 degree line is the exact match of CDFs.