ESSAYS ON MODEL SELECTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Hwan-sik Choi

August 2007

ESSAYS ON MODEL SELECTION

Hwan-sik Choi, Ph.D.

Cornell University 2007

Model selection and nonnested hypothesis testing procedures are considered in three papers. The papers generalize the existing testing procedures and propose methods to improve approximations to the sampling distribution of the test statistics. The first paper proposes robust tests which generalizes the J test (Davidson and MacKinnon (1981)) and the F test (Deaton (1982) and Dastoor (1983)) for non-nested dynamic models with unknown serial correlation and conditional heteroskedasticity in errors. In the second paper, a model selection procedure based on a general criterion function, with an example of the Kullback-Leibler Information Criterion (KLIC) using quasi-likelihood functions, is considered for dynamic non-nested models. I propose a robust test which generalizes Lien and Vuong's (1987) test with a Heteroskedasticity/ Autocorrelation Consistent (HAC) variance estimator. In both papers, I use the fixed-b asymptotics developed in Kiefer and Vogelsang (2005) to improve the asymptotic approximation to the sampling distributions of the test statistics. The fixed-b approach is compared with a bootstrap method and the standard normal approximation in Monte Carlo simulations. The third chapter considers the nonnested hypothesis testing of Vuong (1989) for which the null hypothesis is that the candidate models are equidistant in KLIC from an unknown true model. I propose a higher order asymptotic bias correction of the test statistic and show that it is invariant with respect to reparameterization. The

reparameterization 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.

BIOGRAPHICAL SKETCH

Hwan-sik Choi holds B.S. in engineering and M.A. in economics from Seoul National University, Seoul, Republic of Korea. He also obtained M.A. in economics and M.S. in statistics from Cornell University, Ithaca, NY. His work has been published or is forthcoming in Journal of Business and Economic Statistics and International Journal of Forecasting. His research interest includes model selection, differential geometrical methods in statistics, Bayesian approaches, semiparametric methods, empirical finance, Basel II, and banking system. He also has worked for the Office of the Comptroller of the Currency, US Department of the Treasury, as a research assistant.

To My Family

ACKNOWLEDGEMENTS

It has been a privilege to talk with and be under the guidance of Nicholas M. Kiefer. It is hard to express my adoration and excitement from his insightful comments such as "Statistics reject and do not accept." His words including rather sophisticated jokes last long in my mind, and it is and will be always great to think with him about the science, life, and people.

Looking back in my life, I am lucky to have met two professors, Nick Kiefer and Joon Park, who influenced my thoughts deeply and pleasantly. I would like to send my utmost gratitude to them.

My parents and my wife, Kyongson, have been always with me with their eternal support. I am indebted greatly to their care, and would like to devote my work to them.

April 25, at home in Syracuse.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

A statistical or econometric model is often defined as a family of distributions. Model selection issues arise when we want to choose a member in the family of distributions or when we would like to selection a family from many families of distributions. If the candidate models are not nested, conventional testing procedures are not applicable since they are typically based on a contrast between restricted and unrestricted versions of a nesting model. Since Cox (1961), many developments were made in nonnested hypothesis testing and model selection. I consider generalizations of two approaches in nonnested hypothesis testing.

I propose new test statistics based on the J test of Davidson and MacKinnon (1981) and the Vuong (1989)'s test in Chapter 2 and 3. The new tests are robust to unknown serial correlation and conditional heteroskedasticity in errors and regressors. Approximations to the sampling distributions of the test statistics are provided by the fixed-b asymptotics developed by Kiefer and Vogelsang (2005) and the bootstrap methods. The fixed-b asymptotics and the bootstrap methods are shown to be superior to the standard normal approximations through Monte Carlo experiments.

In Chapter 4, I propose an asymptotic bias correction for the Vuong (1989)'s test statistic, motivated by the differential geometrical methods in statistics. When the primary objective of a statistical inference is identifying the true data generating model, and if we assume that the data are generated from a member of a parametric family described by a finite dimensional parameter vector, the search for the true data generating model becomes the estimation of the true parameter vector that fully describes the probability distribution. If the information from the

data accumulates by, for example, observing more samples that are marginally informative, the true model can be estimated more precisely. This suggests that the local analysis around the true model or the true parameter vector would be a good approximation when we have large amount of data. The idea of defining statistical manifold of probability distributions and the study of local geometry around a model provide an elegant framework to be used for the analysis of asymptotic behavior of statistical estimation and inference. The geometrical quantity appears in higher order terms in the Edgeworth expansions, higher order asymptotic bias and variance, and higher order local power. The following sections give an introduction to the differential geometrical methods in statistics summarizing Amari (1985) and Amari and Nagaoka (2000).

1.1 Motivation to differential geometrical methods

A manifold is a locally euclidean topological space and can serve as a useful tool to study many abstract objects like probability distributions in the statistics or spacetime in the general relativity theory. Defining a geometrical shape of a manifold essentially starts with defining what is curved, or equivalently, what is straight. The following example shows that this task is not simple as it first looks.

A geometrical approach on the space of probability distributions was first recognized by Bhattacharyya (1943, 1946), and Rao (1945). They considered a Riemannian manifold using Fisher information as the Riemannian metric. The length of a curve $c(\theta)$ connecting two probability distributions A and B (Figure 1.1) is defined by

$$\int_c I(\theta)^{1/2} d\theta,$$

where $I(\theta)$ is the Fisher information at θ. Among all the curves connecting A and

B, the shortest curve is the geodesic on this manifold and its distance is called the information distance (Bhattacharyya distance or Rao's distance). Even though

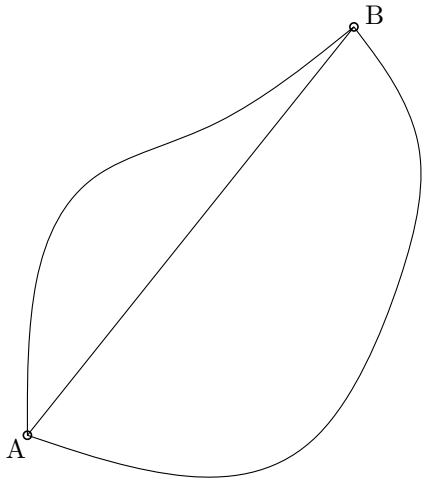

Figure 1.1: Information distance between probability distributions. The shortest curve is the geodesic.

the geodesics look straight to an observer living on the manifold, the manifold can generally be curved. For example, the shortest path between two locations on the earth may look straight to us, it is actually curved in 3-dimensional space. Then there arises the question of how we can measure whether a manifold is straight or flat intrinsically or not. On a curved space, parellel translation of a tangent vector at a point A to another point B depends on the curve that it has been taken along. In Figure 1.2, translation of the tangent vector v_0 at point A to B along two different paths $A \to B$ and $A \to A' \to B$ results in two different vectors v_1 and v_2 respectively. The Riemannian-Christoffel curvature tensor measures the degree of the infinitesimal difference of this kind around a point. But the crucial question still remains in calculating the Riemannian-Christoffel curvature tensor. What does it mean by parellel translation? It requires the concept of

'parellel transport'. To check if two vectors v_0 and v are parellel, v_0 and v should

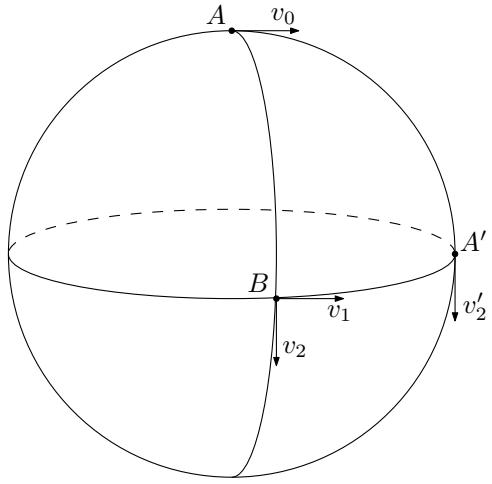

Figure 1.2: Parellel translations of a vector v_0 along two different path $A \to B$ and $A \to A' \to B$

lie in a same vector space. When a space is curved, the displacement of the vector v_0 in T_0 will be in another vector space T_1. In Figure 1.3, the vector v_0 in the tangent space T_0 at A_0 is moving along the curve from the south to the north. The displaced vector v_1 at A_1 lies in another tangent vectore space T_1. Therefore we can not compare v_0 and v_1 for the parellel translation. A mapping C: $T_1 \to T_0'$ makes it possible to compare the two vectors. This mapping is called a *conenction*. There are infintie number of connections that we can define. Among them, exponential connection or 1-connecion is defined as the connection with respect to which the transport of a tangent vector along a curve defined by a one-dimensional exponential family is considered as a parellel transport. Thus the exponential connection recognizes exponential family as a straight or flat space. The curvature of Efron (1975) implicitly assumes exponential connection since the curvature vanishes for exponential families. It's known that the Hilbert space

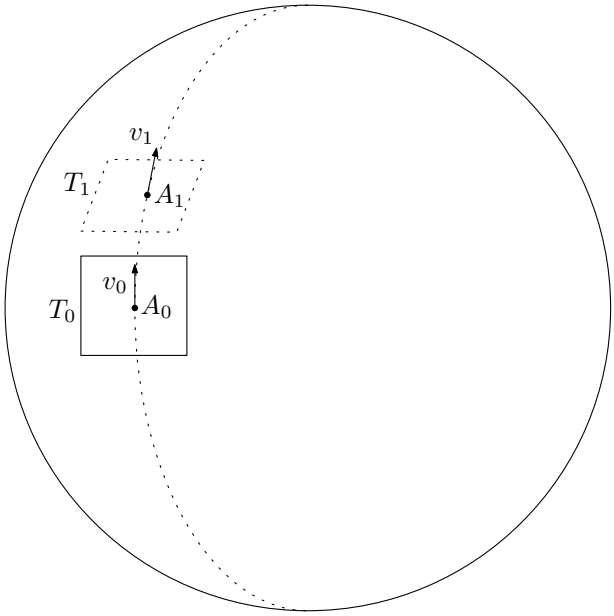

Figure 1.3: Change of a tangent vector

formed by the functions $2c\sqrt{p}$, where p is a probability density, makes a flat space under metric connection or 0-connection. In the Hilbert space, the probability densities form the sphere of radius 2. The transport from p_0 to p_1 along the curve $L = 2\sqrt{p_0}(1-t) + 2\sqrt{p_1}t$ is considered as a parellel transport and the curve is a straight line. Noting that the curve is not probability densities unless $t = 0$ or 1, we can consider a projection of the straight line L onto the sphere of probability densities. The projected line is given by

$$G = c(t)\{2\sqrt{p_0}(1-t) + 2\sqrt{p_1}t\},$$

where $c(t)$ is normalizing constants that makes the function probability densities. Figure 1.4 shows the sphere S of probability densities in the hilbert space. The line L is the cord and the arc G is the geodesic connecting $A = 2\sqrt{p_0}$ and $B = 2\sqrt{p_1}$.

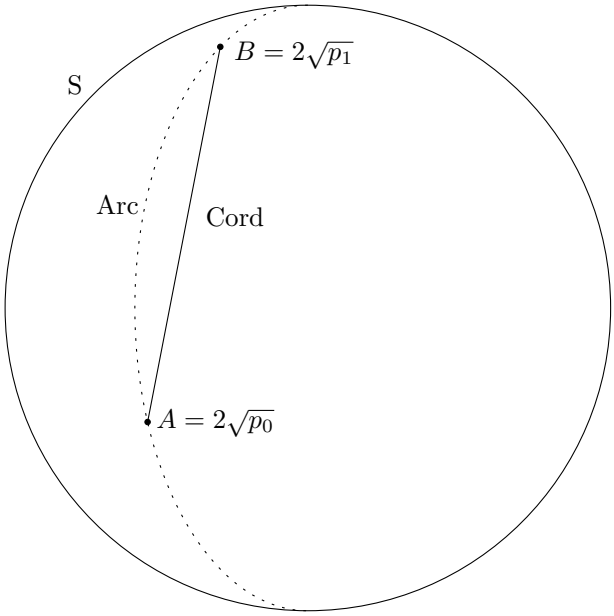

Figure 1.4: Geometry of metric connection

The *half* of the squared length of the cord L,

$$D_0(p_0, p_1) = 2\int(\sqrt{p_0} - \sqrt{p_0})^2 dx = 4\left(1 - \int\sqrt{p_0 p_1}dx\right),$$

is called 0-divergence and $H(p_0, p_1) = \sqrt{2D_0}$ is called Hellinger distance. The information distance $d(p_0, p_1)$ of Rao and Bhattacharyya is the length of arc G given by a simple function of the Hellinger distance or 0-divergence,

$$\begin{aligned} d(p_0, p_1) &= 4\arcsin H/2 \\ &= 2\arccos(1 - D_0/4) \\ &= 2\arccos\left(\int\sqrt{p_0 p_1}dx\right), \end{aligned}$$

from the relationship between the cord and arc. We can see that the Rao's geodesic is curved in this geometry.

Now we consider more detailed treatment of connections.

1.2 Manifolds and Tangent Spaces

A finite dimensional parametric family of probability distributions can be considered as a finite dimensional submanifold embedded in the space of all probability distributions. When this manifold is "smooth" or "differentiable", the differential geometry can play an important role on this manifold. Under suitable regularity conditions, a family of distributions makes a differentiable manifold and the local approximation of the manifold gives important insights to the large sample theory of estimation and statistical inference. Especially we consider an approximation around a point $p \in M$ on a q-dimensional differentiable manifold M by a linear space T_p spanned by differential operators

$$\partial_k = \partial/\partial_k, \tag{1.2.1}$$

where $k = 1, ..., q$ is the coordinate (or chart) index for q dimensional manifold. (We consider simple manifolds that require only a single chart. When we need multiple charts to cover the manifold, we call the collection of charts an atlas.) This linear (vector) space is called a *tangent space* and an element of the tangent space is a *tangent vector*. The collection of the tangent spaces $\{T_p : p \in M\}$ of all points of M is called a *tangent bundle*. A vector field is the mapping from $p \in M$ to a tangent vector $v \in T_p$.

The tangent space is the space of operators. To consider a manifold of probability distributions we use the isomorphism between the operators and random variables

$$\partial_k \sim \partial_k l(x, \theta), \tag{1.2.2}$$

where $l(x, \theta) = \ln p(x, \theta)$ is the log likelihood function of the probability distribution $p(x, \theta)$ and θ is the q dimensional parameter vector. Then the tangent space

$T_p^{(1)}$ generated by $\partial_k l(x,\theta)$ is the linear space of random variables spanned by the score functions and the elements have mean zero with respect to the distribution $p(x,\theta)$. Since we mainly deal with manifolds of probability distributions, we will denote $T_p^{(1)}$ as T_p for convenience. We will consider more general class of transformations of $p(x,\theta)$ other than $l(x,\theta)$. Amari (1985) called the log likelihood function $l(x,\theta)$ the 1-representation among the class of α-representation.

Definition 1.2.1 (α-representation, Amari (1985))**.** *Random variable* $l^{(\alpha)}$ *defined as*

$$l^{(\alpha)} = \begin{cases} \frac{2}{1-\alpha} p^{\frac{1-\alpha}{2}} & \text{for } \alpha \neq 1 \\ \ln p & \text{for } \alpha = 1 \end{cases}, \tag{1.2.3}$$

is α*-representation of* $p = p(x,\theta)$

A tangent space spanned by $\partial_k l^{(\alpha)}$ is useful in its particular setting to be described later.

When we endow a metric on the tangent space, the manifold is a *Riemannian manifold* and the metric is called the *Riemannian metric*. It is also called the *information metric* since it is defined from the Fisher information matrix. First order asymptotic properties of statistical inference can be characterized by the tangent spaces endowed with the Riemannian metric defined by Fisher information matrix $\{g_{ij}\}$ and

$$g_{ij} = E_p \partial_i \partial_j, \tag{1.2.4}$$

where ∂_i is a score function

$$\partial_i = \partial_i l(x,\theta), \tag{1.2.5}$$

and the E_p is the expectation with respect to $p(x,\theta)$. This implies the first order asymptotics is related to the first order approximation (or the tangent space) of the manifold around the point $p \in M$. Therefore if we want to study larger

scale properties such as higher order asymptotics, we need to look at higher order structure of the manifold such as the change of the tangent space. These higher order properties of a manifold are the main subject of this paper.

Change of a tangent space is closely related to the definition of geometrical structure of a manifold such as how curved a manifold is. (For example consider the tangent plane at a point on a sphere. The more curved the surface is, the more the tangent space changes.) So the question is how we can define geometrical structure in a meaningful way on a manifold of probability distributions. This can be done by defining the local change of tangent vector $v \in T_p$ at $p \in M$ to $v' \in T_{p'}$ at its neighbor $p' \in M$. Since v' is not in T_p but in $T_{p'}$, we can not compare two vectors v and v' by vector operations such as $v' - v$. Specifically, a score function of $p' = p(x, \theta')$ is not a mean zero random variable with respect to E_p thus it can not be in the tangent space T_p. A *connection* is a mapping L of a vector $v' \in T_{p'}$ into T_p. Therefore we have $L(v') \in T_p$ and the mean of random variable $L(v')$ is zero under E_p. See Amari (1982) for details about an example of how to achieve this. With a connection, we can define the change of tangent vector using $L(v') - v$. But there are infinite number of possible connections. Among which, a class of affine connections is useful. Affine connections were studied by Čenčov (1972). Efron (1975) introduced the idea "statistical curvature" which was shown to be an application of a special member of affine connections of Čenčov (1972). Efron (1975, 1978) recognized the importance of geometrical approach in the higher order efficiency of estimators, in particular the maximum likelihood estimators (MLEs). Amari (1982) defined a class of affine connections indexed by a real number, namely α-connections where $\alpha \in \mathbb{R}$. It turns out to be same as the affine connections of Chentsov.

1.3 Connections

Change of a tangent space is defined by *covariant derivative* $\nabla_X Y$. Covariant derivative represents the infinitesimal change of a tangent vector $Y \in T_p$ along the direction of a vector $X \in T_p$. Naturally $\nabla_X Y$ is defined through the choice of a connection. The covariant derivative $\nabla_X Y$ can be defined in a coordinate form using the basis vectors $\partial_k \in T_p$. The coordinate form of covariant derivative of ∂_j in direction of ∂_i is a tensor with q^2 components and denoted by $\nabla_{\partial_i}\partial_j$ or shortly $\nabla_i \partial_j$ (Schouten (1954)).

Definition 1.3.1 (Coordinate form of covariate derivative)**.** *The covariant derivative vector $\nabla_{\partial_i}\partial_j$ is defined as*

$$\nabla_{\partial_i}\partial_j = \Gamma_{ij}^k \partial_k, \tag{1.3.1}$$

in its coordinates Γ_{ij}^k with respect to a basis $\{\partial_k\}$ of T_p.

The repeating upper and lower index implies the summation over that index following Einstein's summation convention, i.e.

$$\Gamma_{ij}^k \partial_k = \sum_{k=1}^{q} \Gamma_{ij}^k \partial_k. \tag{1.3.2}$$

Equivalently we can define $\nabla_{\partial_i}\partial_j$ with inner products $\langle \nabla_{\partial_i}\partial_j, \partial_m \rangle$ with basis vectors $\{\partial_m\}$.

Definition 1.3.2 (Coefficients of covariate derivative)**.** *The coefficients of covariant derivative Γ_{ijm} is defined by*

$$\Gamma_{ijm} = \langle \nabla_{\partial_i}\partial_j, \partial_m \rangle \tag{1.3.3}$$

$$= \left\langle \Gamma_{ij}^k \partial_k, \partial_m \right\rangle \tag{1.3.4}$$

$$= \Gamma_{ij}^k g_{km}, \tag{1.3.5}$$

where $\{g_{km}\}$ *is the metric on the tangent space* T_p.

These two forms Γ^k_{ij} and Γ_{ijm} are different representation of the same object $\nabla_{\partial_i}\partial_j$, and $\nabla_X Y$ can be calculated with

$$\nabla_X Y = X^i \partial_i Y^j \partial_j + Y^j \nabla_{X^i \partial_i} \partial_j \tag{1.3.6}$$

$$= (X^i \partial_i Y^j + X^i Y^j \nabla_{\partial_i})\partial_j, \tag{1.3.7}$$

where X^j and Y^j are coordinates of X and Y,

$$X = X^i \partial_i, \tag{1.3.8}$$

$$Y = Y^j \partial_j. \tag{1.3.9}$$

The first term in eq. (1.3.7) is the change of coordinates of Y in direction of X, and the second term represents the change of the origin (the frame of reference) in direction of X.

Although the connections can be defined by defining Γ_{ijm} or Γ^k_{ij} for a given coordinate system, Amari's α-connections are defined with Γ_{ijm} in a convenient form.

Definition 1.3.3 (α-connection)**.** *A class of affine connection* $\overset{(\alpha)}{\nabla}_{\partial_i}\partial_j = \overset{(\alpha)}{\Gamma}{}^k_{ij}\partial_k$ *is called* α*-connection and defined by*

$$\overset{(\alpha)}{\nabla}_{\partial_i}\partial_j = \partial_i\partial_j l + \left[\left(\frac{1-\alpha}{2}\right)\partial_i l \partial_j l - \left(\frac{1+\alpha}{2}\right) E_p \partial_i \partial_j l\right], \tag{1.3.10}$$

or equivalently,

$$\overset{(\alpha)}{\Gamma}{}_{ijk} = \langle \overset{(\alpha)}{\nabla}_{\partial_i}\partial_j, \partial_k \rangle \tag{1.3.11}$$

$$= E_p\left\{\left(\partial_i\partial_j l + \frac{1-\alpha}{2}\partial_i l \partial_j l\right)\partial_k l\right\} \tag{1.3.12}$$

$$= E_p(\partial_i\partial_j l \partial_k l) + \frac{1-\alpha}{2} T_{ijk}, \tag{1.3.13}$$

where T_{ijk} is a skewness tensor,

$$T_{ijk} = E_p \partial_i l \partial_j l \partial_k l. \tag{1.3.14}$$

The term,

$$\left(\frac{1-\alpha}{2}\right) \partial_i l \partial_j l - \left(\frac{1+\alpha}{2}\right) E_p \partial_i \partial_j l \tag{1.3.15}$$

in the equation (1.3.10), defines a class of the connections (affine mappings $\partial_i \partial_j l \mapsto v \in T_p$) needed to make $\overset{(\alpha)}{\nabla}_{\partial_i} \partial_j$ be on the tangent space where every vector should have expection zero. It makes sure that

$$E_p \overset{(\alpha)}{\nabla}_{\partial_i} \partial_j = 0. \tag{1.3.16}$$

The first term in eq. (1.3.13) is the covariance between Hessian and score functions, and the skewness tensor T_{ijk} measures the skewness of score functions. Also note that the connection coefficient is symmetric with respect to the first two indices,

$$\overset{(\alpha)}{\Gamma}_{ijk} = \overset{(\alpha)}{\Gamma}_{jik}. \tag{1.3.17}$$

We called this kind of connection a symmetric connection or torsion-free connection. All members of α-connection are torsion-free.

Definition 1.3.4 (α-flat manifold)**.** *When the coefficients of α-connection vanish, $\overset{(\alpha)}{\Gamma}_{ijk} = 0$, the manifold is a α-flat manifold or α-affine manifold. If a statistical model whose denormalization is an α-affine manifold is called α-family. See Amari and Nagaoka (2000) p.47 for the definition of denormalization.*

Example 1.3.5 (1-flat manifold)**.** *Let p_1 and p_2 be two probability densities. Then a family of distributions $p(\theta) = C p_1^{\theta} p_2^{1-\theta}$, where C is a normalizing constant, is one parameter exponential family such that $p(1) = p_1$ and $p(0) = p_2$. We have for 1-connection ($\alpha = 1$),*

$$\overset{(1)}{\Gamma}_{ijk} = E_p(\partial_i \partial_j l \partial_k l) = 0. \tag{1.3.18}$$

Therefore $p(\theta)$ is a 1*-flat family of distributions. In general, any exponential family is 1-flat, and* 1*-connection is the connection that exponential families are understood as flat. For this reason* 1*-connection is also called* exponential connection. *In this example, we connected two distributions p_1 and p_2 with a* straight line *with respect to the exponential connection.*

Example 1.3.6 (-1-flat manifold)**.** *Let p_1 and p_2 be two probability densities. Then $p(\theta) = \theta p_1 + (1-\theta)p_2$ makes one parameter mixture family. With* -1*-connection, we have*

$$\overset{(-1)}{\Gamma}_{ijk} = E_p(\partial_i \partial_j l \partial_k l) + T_{ijk} = 0. \tag{1.3.19}$$

Therefore $p(\theta)$ is -1*-flat.* -1*-connection is the connection that mixture families are understood as flat, and* -1*-connection is also called* mixture connection *for this reason. We connected two distributions p_1 and p_2 with another straight line with respect to the mixture connection.*

Lemma 1.3.7. *There is a useful relationship between the derivative of the metric $\{g_{ij}\}$ and the connection coefficient $\overset{(\alpha)}{\Gamma}_{ijk}$,*

$$\partial_k g_{ij} = \overset{(\alpha)}{\Gamma}_{ikj} + \overset{(-\alpha)}{\Gamma}_{jki}. \tag{1.3.20}$$

Moreover when the manifold of α-flat, we have

$$\partial_k g_{ij} = \overset{(-\alpha)}{\Gamma}_{jki}. \tag{1.3.21}$$

Proof. Using the chain rule and interchangeability of differentiation and integra-

tion, we have

$$\partial_k g_{ij} = \partial_k E_p \partial_i l \partial_j l \tag{1.3.22}$$

$$= E_p \partial_k \partial_i l \partial_j l + E_p \partial_i l \partial_k \partial_j l + \int \partial_i l \partial_j l \partial_k p(x,\theta)\ dx \tag{1.3.23}$$

$$= E_p \partial_k \partial_i l \partial_j l + E_p \partial_i l \partial_k \partial_j l + \int \partial_i l \partial_j l \frac{\partial_k p(x,\theta)}{p(x,\theta)}\ p(x,\theta)\ dx \tag{1.3.24}$$

$$= E_p \partial_k \partial_i l \partial_j l + E_p \partial_i l \partial_k \partial_j l + \int \partial_i l \partial_j l \partial_k l\ p(x,\theta)\ dx \tag{1.3.25}$$

$$= E_p \partial_k \partial_i l \partial_j l + E_p \partial_k \partial_j l \partial_i l + T_{ijm} \tag{1.3.26}$$

$$= \left(E_p \partial_i \partial_k l \partial_j l + \frac{1-\alpha}{2} T_{ijk} \right) + \left(E_p \partial_j \partial_k l \partial_i l + \frac{1+\alpha}{2} T_{ijk} \right) \tag{1.3.27}$$

$$= \overset{(\alpha)}{\Gamma}_{ikj} + \overset{(-\alpha)}{\Gamma}_{jki}. \tag{1.3.28}$$

and when the manifold is α-flat, $\overset{(\alpha)}{\Gamma}_{ikj} = 0$ gives the second result. □

Corollary 1.3.8. 0-*connection* $\overset{(0)}{\Gamma}_{ijk}$ *can be represented with the derivatives of the metric* $\{g_{ij}\}$,

$$\overset{(0)}{\Gamma}_{ijk} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}) \tag{1.3.29}$$

Proof. By Lemma 1.3.7 for $\alpha = 0$, we have

$$\partial_i g_{jk} = \overset{(0)}{\Gamma}_{ijk} + \overset{(0)}{\Gamma}_{ikj}, \tag{1.3.30}$$

$$\partial_j g_{ik} = \overset{(0)}{\Gamma}_{ijk} + \overset{(0)}{\Gamma}_{jki}, \tag{1.3.31}$$

$$\partial_k g_{ij} = \overset{(0)}{\Gamma}_{ikj} + \overset{(0)}{\Gamma}_{jki}. \tag{1.3.32}$$

Therefore we have $\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij} = 2\overset{(0)}{\Gamma}_{ijk}$. □

Since 0-connection can be defined with the metric or Riemannian metric, we have the following.

Definition 1.3.9 (Metric connection)**.** *The* 0*-connection is also called the* metric connection *or* Riemannian connection *or* Levi-Civita connection*. The connection*

coefficient $\overset{(0)}{\Gamma}_{ijk}$ *of the metric connection is called the* Christoffel symbol of the first kind*, and* $\overset{(0)}{\Gamma}{}^{k}_{ij}$ *is called the* Christoffel symbol of the second kind.

Definition 1.3.10 (Information distance, Rao (1945))**.** *The information distance* $d_0(p_1, p_2)$ *between two probability distributions* p_1 *and* p_2 *along a curve* $c(\theta)$ *indexed by a scalar parameter* θ *with* $c(0) = p_1$ *and* $c(1) = p_1$*, is defined by*

$$d_0(p_1, p_2) = \int_0^1 I(\theta)^{1/2} d\theta, \tag{1.3.33}$$

where $I(\theta)$ *is the Fisher information at* θ*.*

Information distance represents the "length" of a curve with respect to the metric defined with Fisher information and it is symmetric,

$$d_0(p_1, p_2) = d_0(p_2, p_1). \tag{1.3.34}$$

Among all the curves connecting p_1 and p_2, the curve $c^*(\theta)$ that minimizes the information distance is the *0-geodesic* or *straight line* with respect to the information metric. The metric connection is the connection that the line $c^*(\theta)$ is understood as a flat or a straight line. Therefore if we consider $c^*(\theta)$ as one dimensional probability distribution family, it is 0-flat.

The coefficients of α-connections appear in higher order terms in the Edgeworth expansion, higher order asymptotic bias and variance, and higher order local power. Therefore if the model is α-flat, α-affine coordinate system makes those terms vanish. If the model is not flat we can identify the geometrical sources of the terms in the higher order asymptotic expansions by measuring the coefficients.

CHAPTER 2

ROBUST NONNESTED TESTING AND THE DEMAND FOR MONEY

2.1 Introduction

Distinguishing nonnested or separate families of hypotheses for model selection has been an important and active area of formal research since Cox (1961, 1962). Cox used centered log likelihood ratios between two nonnested models. Goldfeld and Quandt (1972) noted the importance of model selection tests for choosing multiplicative or additive errors, and Quandt (1974) considered a nonnested test of $\lambda = 0$ or $\lambda = 1$ in an artificial compound model $\lambda p_1 + (1 - \lambda)p_2$ from a mixture of two competing models with distributions p_1 and p_2. Pesaran (1974) introduced a test for nonnested linear regression models. Davidson and MacKinnon (1981) proposed the popular J test. Deaton (1982) and Dastoor (1983) proposed an F test. These approaches require nesting the competing models in a more general model.

McAleer (1995) compared 9 different nonnested testing procedures and found that the J test (especially a paired comparison) is most popular in empirical papers in journals he considered. The J test is computationally straightforward and easy to interpret. But he also noted that the J test is based on i.i.d. errors, and a diagnostic test to validate this assumption was rarely performed.

The finite sample properties of the J test are known to be poor in some cases even with considerably large samples (Godfrey and Pesaran (1983), McAleer (1995)). In general, the J test is known to reject a correct null hypothesis more

[0]Coauthored with Nicholas M. Kiefer. We are grateful to James MacKinnon and Tim Vogelsang for discussions and comments.

often than a specified level of the test. Fisher and McAleer (1981) and Godfrey and Pesaran (1983) suggested the J_A test using a bias correction of the numerator of the statistic. But the J_A test often has much lower power than the J test. Fan and Li (1995), Godfrey (1998) and Davidson and MacKinnon (2002) used bootstrap methods to approximate the sampling distribution of the J test statistic.

Dynamic models have not been extensively considered in the J test literature. Davidson and MacKinnon (2002) mentioned the possibility of the use of the bootstrap method for the J test in dynamic models. This paper relaxes the i.i.d. error assumption of the J test and considers a generalized J test with serially dependent observations or dynamic models. We propose a robust version of the J test, namely the J_D test, using a Heteroskedasticity/Autocorrelation Consistent (HAC) estimator. A HAC version of the nonnested F test (F_D test) of Deaton (1982) and Dastoor (1983) is also proposed, and its size and power properties are compared with the proposed J_D test.

Since HAC estimators require choosing a bandwidth parameter M, the finite sample performance crucially depends on M. We use Kiefer-Vogelsang-Bunzel (KVB) fixed-b asymptotics (Kiefer et al. (2000), Kiefer and Vogelsang (2002a,b), Kiefer and Vogelsang (2005)) to approximate the sampling distribution of our test statistics. We also consider the semiparametric i.i.d. and block bootstrap methods and compare the performance with the fixed-b asymptotics approximations, as well as with the standard normal approximation.

We present an empirical application to a money demand function. The question of whether the relevant scale variable in money demand is income or consumption was raised by Mankiw and Summers (1986) and subsequently examined by Elyasiani and Nasseh (1994). We revisit this question using our new technique

and data through 2005. Our results support consumption as the better measure of economic activity as far as money demand is concerned.

2.2 KVB fixed-b asymptotics

Kiefer et al. (2000), Kiefer and Vogelsang (2002a,b, 2005) proposed a new asymptotic approximation to the sampling distribution of a HAC test statistic. We illustrate the KVB fixed-b approach with a linear model with a single regressor.

Let $\{y_t\}$ be generated by

$$y_t = \alpha + x_t\beta + u_t,\ (t = 1, 2, ..., T), \tag{2.2.1}$$

where $\{x_t\}$ is a weakly stationary regressor, $E(u_t|x_t) = 0$ for all t, and $\{u_t\}$ is a weakly stationary process with autocovariance function $\gamma(j)$ $(j = 0, \pm1, \pm2, \ldots)$ with possible conditional heteroskedasticity. Let $\hat{\alpha}$ and $\hat{\beta}$ be OLS estimators of α and β respectively. For testing the null hypothesis $\beta = 0$, we consider the following statistic

$$Y_T = \frac{\sum_{t=1}^{T} \tilde{x}_t y_t/\sqrt{T}}{\sqrt{\widehat{V}_T}}, \tag{2.2.2}$$

where $\tilde{x}_t = x_t - \bar{x}$, $\bar{x} = \sum_{t=1}^{T} x_t/T$, and $\widehat{V}_T$ is given by

$$\widehat{V}_T = \sum_{j=1-T}^{T-1} K\left(\frac{j}{M}\right)\hat{\gamma}(j), \tag{2.2.3}$$

where $K(x)$ is the kernel of the non-parametric estimator $\widehat{V}_T$, $M \leq T$ is the bandwidth used in the kernel estimator, and

$$\hat{\gamma}(j) = \frac{1}{T}\sum_{t=|j|+1}^{T} (\hat{v}_t - \bar{v})(\hat{v}_{t-|j|} - \bar{v}), \tag{2.2.4}$$

where $\hat{v}_t = \tilde{x}_t\hat{u}_t$, $\bar{v} = \sum_{t=1}^{T} \hat{v}_t/T = 0$, and $\hat{u}_t = y_t - \hat{\alpha} - x_t\hat{\beta}$. Under conventional asymptotics, $M/T \to 0$ as $T \to \infty$, the HAC variance estimator $\widehat{V}_T$ is consistent for the long run variance of the numerator of the test statistic, and we have asymptotic normality of the test statistic Y_T. The KVB approach models $M/T \to b$ in the $T \to \infty$ conceptual experiment. We assume that $\{v_t\} = \{\tilde{x}_t u_t\}$ satisfies the following.

Assumption 2.2.1 (Functional Central Limit Theorem (FCLT))**.**

$$T^{-1/2} \sum_{t=1}^{[rT]} v_t \Rightarrow \lambda W(r), \tag{2.2.5}$$

where $W(r)$ is a standard Brownian motion defined on $C[0,1]$ and

$$\lambda^2 = \sum_{j=-\infty}^{\infty} \gamma(j) < \infty.$$

As shown in Phillips and Durlauf (1986), this assumption holds under slightly weaker conditions than the assumptions for a consistent estimation in the HAC variance estimation literature (Andrews (1991); weaker assumptions than Andrews (1991) can be found in Hansen (1992) and de Jong and Davidson (2000)). It allows conditional heteroskedasticity but excludes unconditional heteroskedasticity. See Kiefer and Vogelsang (2005, p.1135) for further discussion.

Under Assumption 2.2.1 and some regularity conditions such as $\hat{\beta} \xrightarrow{p} \beta$ as in the OLS example above, and if we use the Bartlett kernel for example, Kiefer and Vogelsang (2005) showed that we have the limiting distribution of the test statistic Y_T with $b = \lim_{T\to\infty} M/T \in (0,1]$,

$$Y_T \Rightarrow \frac{W(1)}{\sqrt{\frac{2}{b}\left[\int_0^1 \widetilde{W}(r)^2\, dr - \int_0^{1-b} \widetilde{W}(r+b)\widetilde{W}(r)\, dr\right]}}. \tag{2.2.6}$$

where $\widetilde{W}(r) = W(r) - rW(1)$.

See Kiefer and Vogelsang (2005, p.1137) for another example with Monte Carlo experiments and Kiefer and Vogelsang (2005, Theorem 3) for other kernel functions. Critical values of this non-standard limiting distribution must be simulated in practice, and they are tabulated in Kiefer and Vogelsang (2005, p.1146) for some popular kernels. We work with the Bartlett kernel throughout.

2.3 The Dynamic J test

We present the main idea with the problem of choosing between a pair of linear regression models. Our approach essentially applies to non-linear models also as the original J test. Let H_1 and H_2 be two competing nonnested linear models given, for $t = 1, ..., T$, by

$$H_1 : y_t = x_t'\beta_1 + u_{1t}, \tag{2.3.1}$$

$$H_2 : y_t = z_t'\beta_2 + u_{2t}, \tag{2.3.2}$$

where x_t, z_t are k_1 and k_2 dimensional (exogenous) regressors respectively and u_{1t}, u_{2t} are i.i.d. errors with mean zero and variance σ^2. For convenience we sweep out common regressors and intercepts in H_1 and H_2 by projection, leading to eq. (2.3.1) and (2.3.2). Under the hypothesis that H_1 is the true model, the J test uses the artificial model

$$H_0 : y_t = x_t'\beta + \hat{y}_t\theta + u_t, \tag{2.3.3}$$

where $\hat{y} = \{\hat{y}_1, ..., \hat{y}_T\}'$ is the fitted value of the dependent variable $y = \{y_1, ..., y_T\}'$ from the regression H_2. The J test statistic is the t-statistic for θ,

$$J = \frac{\hat{\theta}}{\sqrt{\hat{\sigma}^2 \left(\hat{y}' M_X \hat{y}\right)^{-1}}}, \tag{2.3.4}$$

where $M_X = I - P_X$, P_X is the projection matrix onto the space spanned by the $(T \times k_1)$ regressor matrix X in the model H_1, $\hat{\theta} = (\hat{y}' M_X \hat{y})^{-1} \hat{y}' M_X y$, $\hat{\sigma}^2 = \sum_{t=1}^{T} \hat{u}_t^2 / T$ and $\hat{u}_t$ is the residual from the regression equation (2.3.1) or (2.3.3). The sampling distribution of the test statistic J is approximated by the standard normal distribution. This test is widely used because of its simplicity and intuitive appeal (see McAleer (1995)).

This paper generalizes the J test by relaxing the assumptions on the errors in the true model. First, we formally assume the non-orthogonality of regressors in H_1 and H_2 as in Davidson and MacKinnon (1981), and we assume stationarity.

Assumption 2.3.1. *The regressors x_t in H_1 and z_t in H_2 are weakly stationary processes with unknown serial correlation and conditional heteroskedasticity, and they satisfy*

$$\operatorname*{plim}_{T \to \infty} T^{-1} \sum_{t=1}^{T} x_t z_t' \neq \mathbf{0}. \tag{2.3.5}$$

When the correlation between x_t and z_t is weak, it is known that the J test statistic shows over-rejection (Godfrey and Pesaran (1983), Godfrey (1998)). Michelis (1999) proposed a new asymptotic approximation under near population orthogonality (NPO) in which $\operatorname{plim}_{T \to \infty} T^{-1/2} \sum x_t z_t' = \Delta$, where Δ is a matrix of constants. When the regressors are orthogonal, we have $\hat{\beta}_2 \xrightarrow{p} 0$ and the estimated H_2 converges to a nested model in H_1. We can consider in this case an F test such as testing $\beta_2 = \mathbf{0}$ in

$$y_t = x_t' \beta + z_t' \beta_2 + u_t. \tag{2.3.6}$$

Under Assumption 2.3.1, we have non-degenerating $\hat{y}$.

We generalize the model by allowing serial correlation in the error process $\{u_{1t}\}$ of the true model H_1 with possible conditional heteroskedasticity.

Assumption 2.3.2. *The error process* $\{u_{1t}\}$ *of the true model* H_1 *is weakly stationary with unknown serial correlation and conditional heteroskedasticity.*

Assumption 2.3.2 does not specify a form of serial correlation or conditional heteroskedasticity. The test statistic proposed in this paper is robust to unknown serial correlation or heteroskedasticity, thus an empirical researcher need not test the existence of serial correlation. We exclude unconditional heteroskedasticity in the errors. The following assumption restricts the serial correlation under certain situation.

Assumption 2.3.3. *The error process* $\{u_{it}\}$ *for* $i = 1, 2$ *is assumed to satisfy*

$$E(u_{it}|x_t, z_t) = 0. \tag{2.3.7}$$

Assumption 2.3.3 excludes, for example, having both lagged dependent variables and serial correlation in the errors. It also excludes endogenous variables. But conditional heteroskedasticity in the errors is still allowed.

We propose the dynamic J test (J_D test) using a HAC estimator. Our statistic is given by

$$J_D = \frac{\hat{y}' M_X y/\sqrt{T}}{\sqrt{\widehat{V}_T}}, \tag{2.3.8}$$

where

$$\widehat{V}_T = \sum_{j=1-T}^{T-1} K\left(\frac{j}{M}\right)\hat{\gamma}(j), \tag{2.3.9}$$

$K(\cdot)$ is the kernel of the non-parametric estimator $\widehat{V}_T$, the bandwidth $M \leq T$ is the number of lags used for $\widehat{V}_T$, and

$$\hat{\gamma}(j) = \frac{1}{T}\sum_{t=|j|+1}^{T} (\hat{v}_t - \bar{v})(\hat{v}_{t-|j|} - \bar{v}), \tag{2.3.10}$$

where $\hat{v}_t = \hat{u}_t (M_X \hat{y})_t$, $\bar{v}$ is the sample mean of $\{\hat{v}_t\}$, and $\{\hat{u}_t\}$ is the residual vector from the null model or the artificial model. Although using $\{\hat{u}_t\}$ from either the

null model or the artificial model are asymptotically equivalent, Davidson and MacKinnon (1985) and Ligeralde and Brown (1995) demonstrated that using the null model can reduce the size problem (i.e. over-rejection) in HAC tests. But using the residuals from the null model can result in lower power when the imposed null is not the truth. We use the residuals from the null model in examples in the next section.

To apply the fixed-b asymptotic approximation to the sampling distribution of the test statistic J_D, we assume the FCLT holds for the partial sum of a product process $v_t = u_{1t}w_t$, where $\{w_t\}_{t=1}^T$ is asymptotically equivalent to the vector $M_X\hat{y} = M_X P_Z y$, where P_Z is the projection matrix onto the space spanned by the $(T \times k_2)$ regressor matrix Z in the model H_2. Specifically, as $T \to \infty$, w_t is the limit of the t-th observation $(M_X\hat{y})_t$ of $M_X\hat{y}$ given by

$$w_t = \operatorname*{plim}_{T\to\infty}(M_X\hat{y})_t \tag{2.3.11}$$

$$= (z_t' - x_t'Q_{xx}^{-1}Q_{xz})Q_{zz}^{-1}Q_{zx}\beta_1, \tag{2.3.12}$$

where

$$Q_{xx} = \operatorname*{plim}_{T\to\infty} X'X/T,$$

$$Q_{zz} = \operatorname*{plim}_{T\to\infty} Z'Z/T$$

and

$$Q_{zx} = Q_{xz}' = \operatorname*{plim}_{T\to\infty} Z'X/T.$$

The vector $\beta_2^* = Q_{zz}^{-1}Q_{zx}\beta_1$ in eq. (2.3.12) is called a pseudo-true value of β_2.

Assumption 2.3.4. *The process $\{v_t\} = \{u_{1t}w_t\}$ satisfies the FCLT,*

$$T^{-1/2}\sum_{t=1}^{[rT]} v_t \Rightarrow \lambda W(r), \tag{2.3.13}$$

where λ^2 is the long run variance of $\{v_t\}$, and

$$T^{-1/2}\sum_{t=1}^{[rT]}(\hat{v}_t - v_t) \xrightarrow{p} 0, \text{ uniformly in } r \in [0,1]. \tag{2.3.14}$$

This high level assumption allows us to apply the FCLT to $\{\hat{v}_t\}$. The following theorem gives the asymptotic approximation to the sampling distribution of our test statistic J_D.

Theorem 2.3.5. *Under Assumption 2.3.1-2.3.4, and if $M/T \to b \in (0,1]$, the limiting distribution of the J_D test statistic in eq. (2.3.8) is given by the KVB fixed-b asymptotics of Kiefer and Vogelsang (2005, Theorem 3),*

$$J_D \Rightarrow \frac{W(1)}{\sqrt{Q_1(b)}}, \text{ as } T \to \infty, \tag{2.3.15}$$

where $Q_1(b)$ is defined in Kiefer and Vogelsang (2005, Definition 1) depending on the kernel function.

Proof. The FCLT applies to $\{\hat{v}_t\}$ by Assumption 2.3.4. Under Assumption 2.3.1-2.3.3, we get eq. (2.3.15) from Theorem 3 in Kiefer and Vogelsang (2005). □

We also consider a HAC robust version of the nonnested F test of Deaton (1982) and Dastoor (1983). Consider the artificial model

$$y_t = x_t'\beta + z_t'\beta_2 + u_t. \tag{2.3.16}$$

Let $(k_2 \times 1)$ vector process $\{\hat{v}_t\} = \{\hat{u}_t(Z'M_X)_t\}$, where $(Z'M_X)_t$ is t-th column of $Z'M_X$ and $\hat{u}_t$ is the residual from the null or the artificial model. The HAC robust nonnested F test statistic F_D for $\beta_2 = \mathbf{0}$ is given by

$$F_D = \frac{Ty'(M_XZ)(\widehat{V}_T)^{-1}(Z'M_X)y}{k_2} \tag{2.3.17}$$

where

$$\widehat{V}_T = \sum_{j=1-T}^{T-1} K\left(\frac{j}{M}\right)\widehat{\Gamma}(j), \tag{2.3.18}$$

and

$$\widehat{\Gamma}(j) = \frac{1}{T}\sum_{t=j+1}^{T}(\hat{v}_t - \bar{v})(\hat{v}_{t-j} - \bar{v})' \text{ for } j \geq 0, \tag{2.3.19}$$

$$\widehat{\Gamma}(j) = \widehat{\Gamma}'(-j) \text{ for } j < 0. \tag{2.3.20}$$

Noting that $Z'M_X = Z' - Z'X(X'X)^{-1}X'$, we define the limit $\tilde{w}_t$ of the t-th observation $(Z'M_X)_t$ of $Z'M_X$ as

$$\tilde{w}_t = \operatorname*{plim}_{T\to\infty}(Z'M_X)_t = z_t - Q_{zx}Q_{xx}^{-1}x_t. \tag{2.3.21}$$

We assume the FCLT for the product process $v_t = u_{1t}\tilde{w}_t$, and the process $\{\hat{v}_t\}$ is asymptotically equivalent to $\{v_t\}$.

Assumption 2.3.6. *For $\{v_t\} = \{u_{1t}\tilde{w}_t\}$, we have*

$$T^{-1/2}\sum_{t=1}^{[rT]} v_t \Rightarrow \Lambda W_{k_2}(r), \text{ as } T \to \infty, \tag{2.3.22}$$

where $\Lambda\Lambda' = \sum_{j=-\infty}^{\infty}\Gamma(j)$, $\Gamma(j) = E(v_t v_{t-j}')$, and $W_{k_2}(r)$ is a $(k_2 \times 1)$ vector standard Brownian motion, and

$$T^{-1/2}\sum_{t=1}^{[rT]}(\hat{v}_t - v_t) \xrightarrow{p} \mathbf{0}, \text{ uniformly in } r \in [0,1]. \tag{2.3.23}$$

Theorem 2.3.7. *Under the Assumption 2.3.1-2.3.3 and 2.3.6, if $M/T \to b \in (0,1]$, the limiting distribution of the F_D test statistic in eq. (2.3.17) is given by the KVB fixed-b asymptotics of Kiefer and Vogelsang (2005, Theorem 3),*

$$F_D \Rightarrow W_{k_2}(1)'Q_{k_2}(b)^{-1}W_{k_2}(1)/k_2, \tag{2.3.24}$$

where $Q_{k_2}(b)$ is defined in Kiefer and Vogelsang (2005, Definition 1) depending on the kernel function.

Proof. The FCLT applies to $\{\hat{v}_t\}$ by Assumption 2.3.6. Eq. (2.3.24) directly follows from Kiefer and Vogelsang (2005, Theorem 3) under Assumption 2.3.1-2.3.3. □

In the next section, we consider the finite sample properties of the proposed J_D and F_D test. The performance of the fixed-b approach is compared to the bootstrap methods and the standard normal approximation.

2.4 Monte Carlo Study

2.4.1 Specification of simulations

We present two examples. One is with serially correlated errors with conditional heteroskedasticity, and the other is with a lagged dependent variable and conditional heteroskedasticity only in the errors.

- Case I (serially correlated errors with conditional heteroskedasticity)

$$H_1 : y_t = x_t'\beta_1 + u_{1t} = x_{1t} + 0.5x_{2t} + u_{1t}, \tag{2.4.1}$$

$$H_2 : y_t = z_t'\beta_2 + u_{2t} = z_{1t} + 0.5z_{2t} + 0.5z_{3t} + 0.5z_{4t} + u_{2t}, \tag{2.4.2}$$

- Case II (a lagged dependent variable and conditional heteroskedasticity in errors)

$$H_1 : y_t = y_{t-1}\delta + x_{1t} + 0.5x_{2t} + u_{1t}, \tag{2.4.3}$$

$$H_2 : y_t = y_{t-1}\delta + z_{1t} + 0.5z_{2t} + 0.5z_{3t} + 0.5z_{4t} + u_{2t}. \tag{2.4.4}$$

In both cases we assume H_1 is the true model. Godfrey and Pesaran (1983) noted that J test may have poorer finite sample performance when there are different numbers of regressors in the models. Thus we define x_t and z_t to be (2×1)

and (4×1) regressor vectors respectively. The regressors are generated from a vector autoregressive (VAR) process

$$W_t = \Phi W_{t-1} + \zeta_t, \tag{2.4.5}$$

where $W_t = (x_t', z_t')'$ is (6×1) vector, and Φ is the autoregressive coefficient matrix. The autoregressive coefficient matrix is given by a symmetric Toeplitz matrix

$$\Phi = \begin{pmatrix} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \\ v_2 & v_1 & v_2 & v_3 & v_4 & v_5 \\ v_3 & v_2 & v_1 & v_2 & v_3 & v_4 \\ v_4 & v_3 & v_2 & v_1 & v_2 & v_3 \\ v_5 & v_4 & v_3 & v_2 & v_1 & v_2 \\ v_6 & v_5 & v_4 & v_3 & v_2 & v_1 \end{pmatrix}. \tag{2.4.6}$$

The error $\{\zeta_t\}$ is a mean zero Gaussian i.i.d. process with variance matrix Ω. The variance Ω of the error $\{\zeta_t\} = \{\zeta_{1t} \ ... \ \zeta_{6t}\}'$ has upper triangular (j,k)-elements given by

$$E\zeta_{jt}\zeta_{kt} = \begin{cases} 1, & \text{for } j = k, \\ 0.8, & \text{for } j = 1 \text{ and } k = 2, \\ 0.7, & \text{for } j < k \text{ and } j, k = 3, ..., 6, \\ C_{xz}, & \text{for } j < k \text{ and } j = 1, 2, \ k = 3, ..., 6. \end{cases} \tag{2.4.7}$$

We consider the following two specifications for x_t and z_t for CASE I ensuring the stationarity of $\{W_t\}$.

- Specification 1 (Strong $\{x_t, z_t\}$): $(v_1, v_2, v_3, v_4, v_5, v_6) = (-0.3,\ 0.1,\ 0.3, -0.2,\ 0.1,\ -0.3)$ and $C_{xz} = 0.8$,
- Specification 2 (Weak $\{x_t, z_t\}$): $(v_1, v_2, v_3, v_4, v_5, v_6) = (0.8,\ 0,\ 0,\ 0,\ 0,\ 0)$ and $C_{xz} = 0.2$.

The specification 2 makes weaker correlation between x_t and z_t processes than the specification 1. For CASE II, we consider the specification 1 (Strong $\{x_t, z_t\}$) only.

We generated the error process $\{u_{1t}\}_{t=1}^{T}$ from an AR(1) process with GARCH(1,1),

$$u_{1t} = \alpha u_{1(t-1)} + \varepsilon_t, \tag{2.4.8}$$

$$\varepsilon_t = \sigma_t \xi_t, \tag{2.4.9}$$

$$\xi_t \sim i.i.d.\ N(0,1), \tag{2.4.10}$$

$$\sigma_t^2 = 0.04 + 0.86\sigma_{t-1}^2 + 0.1\varepsilon_{t-1}^2. \tag{2.4.11}$$

We initialize $\{u_{1t}\}$ with variance $\sigma_1^2 = 1$ and $u_{11} = \varepsilon_1$. We use $\alpha = 0,\ 0.5,\ 0.9,\ -0.5$ for CASE I, and $\alpha = 0$ only for CASE II. For CASE II, $\{y_t\}$ is generated by setting $\delta = 0.5$ or 0.9.

The GARCH(1,1) coefficients satisfy the stationarity conditions and moment conditions for $\{\varepsilon_t\}$ required for our test statistics. See Ling (1999) and Ling and McAleer (2002) for necessary and sufficient conditions on the parameters for the existence of higher moments in GARCH models.

We generated 100 observations of y_t then took the last $T = 50$ observations. The total number of iterations was $5,000$ with $B = 399$ bootstrap resamplings for each iteration. In all examples, we used the Bartlett kernel. The bandwidths were $M = 1, 5, 10, 25, 50$. Even though we don't have serial correlation ($\alpha = 0$) in CASE II, we used bandwidths $M > 1$ to capture serial correlation in finite sample.

All tests are run at the 5% level, and we compare the J_D test, and the F_D test. The sampling distribution of the J_D and F_D tests was approximated by four methods.

1. The standard normal approximation for the J_D test, and $\chi^2(k_2)/k_2$ for the

F_D test ($k_2 = 4$ in our example),

2. The fixed-b asymptotic approximation in eq. (2.3.15) for the J_D test, and eq. (2.3.24) for the F_D test,

3. The semi-parametric i.i.d. bootstrap: Using the residuals $\{\hat{u}_t\}$ from the null model H_1, we bootstrap $\{u_t^*\}$ with i.i.d. resampling, and normalize u_t^* with

$$\sqrt{\frac{T}{T-4}}\,(u_t^* - \bar{u}^*)\,, \tag{2.4.12}$$

where $\bar{u}^* = \sum_{t=1}^{T} u_t^*/T$. This step makes sure the bootstrap residuals have mean zero. The factor $\sqrt{\frac{T}{T-4}}$ is used to correct the smaller-variance problem in the bootstrap residuals in small samples. See Davidson and MacKinnon (2002) for more details. Then $\{y_t^*\}$ is generated by using $\{u_t^*\}$ and the estimated parameters from H_1. We calculate the bootstrap J_D^* and F_D^* test statistics $B = 399$ times. Critical values are from the quantiles of the empirical distribution of J_D^* and F_D^*.

4. The semi-parametric (overlapping) block bootstrap: We use the overlapping block bootstrap with block size five to get $\{u_t^*\}$, then follow the same procedure as above.

For power comparisons, we can not compare size corrected powers of different approximations to the sampling distribution of the test statistic since they use the same test statistic for a given kernel and a bandwidth. But we can compare the size corrected powers of J_D and F_D tests with different kernels and bandwidths. Asymptotically speaking, while all the approximations provide correct size under an appropriate conceptual experiment, either $M/T \to 0$ or $M/T \to b$, it is notable that under the conventional standard normal approximation, the asymptotic local

power of a HAC test statistic is exactly same no matter which kernel function or bandwidth were used, but under the fixed-b asymptotics, the local power depends on the kernel and the bandwidth. Kiefer and Vogelsang (2005) found that the Bartlett kernel has good asymptotic local power comparing to other popular kernels, and using large bandwidths decreases local power in all the kernels. Our Monte Carlo study also supported their asymptotic results.

2.4.2 Size properties

Table 2.1−2.2 show the rejection rates of the 5% level (two tail for J_D) tests for four different AR(1) error correlation coefficients α for CASE I. Table 2.1 shows the size performance of different asymptotic approximations of the strong correlation in the regressors (specification 1), and Table 2.2 is from the weak regressor correlation (specification 2).

The bootstrap tests showed the best performance in both J_D and F_D tests. The block bootstrap with large bandwidths shows robust performance in all settings we considered. The block bootstrap gave relatively good performance in the worst scenario (Table 2.2, $\alpha = 0.9$). The i.i.d. and block bootstrap methods showed similar performance in many cases (Table 2.1). See Gonçalves and Vogelsang (2006) for an explanation of the good performance of the i.i.d. bootstraps with serially correlated errors. They show that the bootstrap methods (both block and i.i.d.) have the same limiting distribution as the fixed-b asymptotics.

The fixed-b asymptotic approach provides a clear improvement over the standard or chi-square approximations, although it overrejects in small bandwidths. When the bandwidths are small, the fixed-b limiting distributions are "close" to the standard normal (or chi-square) distribution. Therefore they perform similarly.

Table 2.1: Size Comparison (CASE I, level=0.05). x_t and z_t are strongly correlated, and α is the AR(1) coefficient of the errors

Strong $\{x_t, z_t\}$ and $\alpha = 0$								
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.1250	0.1136	0.0490	0.0478	0.0892	0.0658	0.0512	0.0500
5	0.1538	0.0966	0.0480	0.0514	0.2080	0.0468	0.0476	0.0490
10	0.1926	0.0902	0.0480	0.0484	0.3820	0.0452	0.0484	0.0480
25	0.3022	0.0828	0.0456	0.0508	0.7260	0.0470	0.0504	0.0492
50	0.4482	0.0816	0.0466	0.0496	0.8864	0.0476	0.0482	0.0500
Strong $\{x_t, z_t\}$ and $\alpha = 0.5$								
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.1138	0.1024	0.0540	0.0576	0.0632	0.0454	0.0368	0.0484
5	0.1426	0.0864	0.0506	0.0530	0.1760	0.0348	0.0328	0.0414
10	0.1808	0.0798	0.0566	0.0532	0.3498	0.0324	0.0350	0.0398
25	0.2868	0.0766	0.0528	0.0522	0.6948	0.0370	0.0382	0.0444
50	0.4322	0.0782	0.0566	0.0552	0.8668	0.0362	0.0388	0.0422
Strong $\{x_t, z_t\}$ and $\alpha = 0.9$								
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.1494	0.1364	0.0644	0.0548	0.0734	0.0560	0.0466	0.0572
5	0.1580	0.0976	0.0538	0.0458	0.1376	0.0222	0.0230	0.0346
10	0.2042	0.0882	0.0574	0.0468	0.2928	0.0188	0.0204	0.0306
25	0.3350	0.0852	0.0556	0.0476	0.6658	0.0210	0.0202	0.0316
50	0.4812	0.0844	0.0550	0.0476	0.8608	0.0202	0.0210	0.0306
Strong $\{x_t, z_t\}$ and $\alpha = -0.5$								
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.2404	0.2228	0.0838	0.0574	0.2050	0.1654	0.1428	0.0702
5	0.2300	0.1484	0.0522	0.0494	0.2538	0.0676	0.0666	0.0552
10	0.2736	0.1262	0.0456	0.0436	0.4402	0.0624	0.0608	0.0522
25	0.3912	0.1136	0.0464	0.0434	0.7708	0.0612	0.0604	0.0542
50	0.5392	0.1142	0.0448	0.0428	0.9100	0.0614	0.0598	0.0518

Table 2.2: Size Comparison (CASE I, level=0.05). x_t and z_t are weakly correlated, and α is the AR(1) coefficient of the errors

Weak $\{x_t, z_t\}$ and $\alpha = 0$								
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.0924	0.0826	0.0480	0.0520	0.0886	0.0644	0.0514	0.0562
5	0.1272	0.0770	0.0446	0.0516	0.2106	0.0394	0.0438	0.0494
10	0.1666	0.0678	0.0432	0.0474	0.3940	0.0410	0.0442	0.0494
25	0.2668	0.0624	0.0406	0.0440	0.7410	0.0446	0.0450	0.0490
50	0.4052	0.0642	0.0426	0.0442	0.8994	0.0434	0.0446	0.0496
Weak $\{x_t, z_t\}$ and $\alpha = 0.5$								
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.2736	0.2560	0.1648	0.0772	0.3770	0.3200	0.2816	0.0968
5	0.2078	0.1386	0.0734	0.0522	0.3348	0.0970	0.1004	0.0650
10	0.2450	0.1146	0.0602	0.0482	0.5066	0.0784	0.0864	0.0574
25	0.3610	0.1108	0.0612	0.0512	0.8152	0.0900	0.0936	0.0612
50	0.4992	0.1110	0.0566	0.0532	0.9370	0.0916	0.0912	0.0612
Weak $\{x_t, z_t\}$ and $\alpha = 0.9$								
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.6572	0.6416	0.4810	0.1554	0.7560	0.7170	0.6860	0.2318
5	0.4986	0.3768	0.1482	0.0714	0.5172	0.1874	0.1954	0.0912
10	0.5038	0.2814	0.1012	0.0580	0.6542	0.1410	0.1454	0.0770
25	0.6024	0.2466	0.1056	0.0626	0.8934	0.1574	0.1584	0.0806
50	0.7316	0.2430	0.1004	0.0588	0.9670	0.1484	0.1476	0.0766
Weak $\{x_t, z_t\}$ and $\alpha = -0.5$								
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.0236	0.0190	0.0170	0.0392	0.0130	0.0076	0.0058	0.0368
5	0.0802	0.0446	0.0384	0.0472	0.1230	0.0168	0.0194	0.0452
10	0.1170	0.0484	0.0388	0.0460	0.2898	0.0196	0.0216	0.0432
25	0.2252	0.0462	0.0394	0.0418	0.6486	0.0190	0.0198	0.0404
50	0.3714	0.0472	0.0394	0.0448	0.8514	0.0192	0.0212	0.0408

The fixed-b asymptotics works better in F_D tests than in J_D tests. Since the J test has a finite sample bias from using fitted values $\hat{y}$ from the alternative model with the same data set, J_D will also carry this problem. Although the J_D test with the fixed-b asymptotics mitigates this problem when large bandwidths were used, it does not remove the problem because the fixed-b asymptotics corrects the sampling distribution of the asymptotic variance (denominator) of the test statistic. When the regressors are weakly correlated (the specification 2, Table 2.2), this bias problem becomes more serious (Compare Table 2.1 and 2.2). The reason that the bootstrap methods work better than the fixed-b approach is that they correct both the bias and the variance.

Table 2.3 shows the size of J_D and F_D tests in CASE II with the AR(1) coefficients $\delta = 0.5$ and 0.9 for the lagged dependent variable y_{t-1}. For Case II, even though the serial correlation in the errors is not present, using large bandwidths will capture the serial correlation in finite sample and gives better performance. The bootstrap (especially the block bootstrap) works best, and the fixed-b asymptotics in the J_D tests overrejects in small bandwidths but is a clear improvement over the standard normal or chi-square approximations as the bandwidth increases.

2.4.3 Power comparison

Table 2.4 (CASE I) and Table 2.5 (CASE II) show the size corrected (at 5%) power comparison of the J_D and F_D test statistics for various bandwidths. It is known that the Davidson and MacKinnon's J test has better local power than the F test in a paired comparison (Dastoor and McAleer (1989)). In our simulation, the J_D test also gave better size-adjusted power than the F_D test especially when high bandwidths were used. The power decreases as the bandwidth increases in

Table 2.3: Size Comparison (CASE II, level=0.05). δ is the AR(1) coefficient of y_{t-1}

	$\delta = 0.5$							
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.1304	0.1182	0.0500	0.0476	0.0866	0.0622	0.0480	0.0450
5	0.1584	0.1024	0.0480	0.0466	0.2102	0.0470	0.0442	0.0446
10	0.1980	0.0926	0.0454	0.0460	0.3820	0.0444	0.0438	0.0438
25	0.3070	0.0904	0.0484	0.0446	0.7222	0.0434	0.0408	0.0424
50	0.4438	0.0864	0.0482	0.0456	0.8848	0.0452	0.0434	0.0416
	$\delta = 0.9$							
	J_D				F_D			
	SN	Fixed-b	Boot(1)	Boot(5)	Chi-sq	Fixed-b	Boot(1)	Boot(5)
$M = 1$	0.1392	0.1260	0.0536	0.0528	0.0938	0.0686	0.0498	0.0492
5	0.1646	0.1046	0.0482	0.0492	0.2110	0.0482	0.0484	0.0478
10	0.2028	0.0976	0.0488	0.0480	0.3952	0.0462	0.0450	0.0458
25	0.3126	0.0862	0.0456	0.0458	0.7308	0.0478	0.0436	0.0442
50	0.4514	0.0878	0.0452	0.0488	0.8858	0.0474	0.0444	0.0438

both J_D and F_D tests, but the power decreased less in J_D tests. The decrease in power with large bandwidths is consistent with the asymptotic local power results of Kiefer and Vogelsang (2005).

Therefore we have a trade-off between size and power in choosing bandwidth. It is recommended to use the J_D test rather than the F_D test for better power, and the block bootstrap for good size performance on the basis of our simulation results.

Table 2.4: Power Comparison (CASE I). Size is controlled to be 0.05. α is the AR(1) coefficient of the errors.

	Strong $\{x_t, z_t\}$				Weak $\{x_t, z_t\}$			
	$\alpha = 0$		$\alpha = 0.9$		$\alpha = 0$		$\alpha = 0.9$	
	J_D	F_D	J_D	F_D	J_D	F_D	J_D	F_D
$M = 1$	0.9962	0.9756	0.8746	0.7430	1.0000	0.9994	0.8778	0.7458
5	0.9820	0.8724	0.9128	0.7914	0.9916	0.8792	0.9202	0.7954
10	0.9380	0.7218	0.8686	0.7002	0.9600	0.7246	0.8780	0.6988
25	0.8050	0.6710	0.7448	0.6536	0.8314	0.6700	0.7428	0.6496
50	0.7852	0.6916	0.7210	0.6646	0.8024	0.6948	0.7240	0.6610
	$\alpha = 0.5$		$\alpha = -0.5$		$\alpha = 0.5$		$\alpha = -0.5$	
	J_D	F_D	J_D	F_D	J_D	F_D	J_D	F_D
$M = 1$	0.9918	0.9700	0.9652	0.8534	0.9934	0.9712	0.9620	0.8510
5	0.9874	0.8968	0.9022	0.7400	0.9876	0.8912	0.9040	0.7290
10	0.9516	0.7694	0.8168	0.5858	0.9512	0.7670	0.8190	0.5806
25	0.8252	0.7142	0.6372	0.5602	0.8268	0.7102	0.6382	0.5654
50	0.8062	0.7368	0.6376	0.5624	0.8040	0.7254	0.6418	0.5658

2.5 Money Demand

We test the idea of Mankiw and Summers (1986) that consumption (or personal expenditure) rather than income (Gross National Product, GNP) is the right scale variable for money demand (for M1 or M2). Elyasiani and Nasseh (1994) used vari-

Table 2.5: Power Comparison (CASE II). Size is controlled to be 0.05. δ is the AR(1) coefficient of y_{t-1}

	$\delta = 0.5$		$\delta = 0.9$	
	J_D	F_D	J_D	F_D
$M = 1$	0.9956	0.9746	0.9956	0.9772
5	0.9732	0.8478	0.9786	0.8662
10	0.9142	0.6952	0.9296	0.7274
25	0.7714	0.6740	0.7848	0.6758
50	0.7532	0.6712	0.7656	0.6816

ous nonnested tests indicating that the consumption measure seems to be the right scale variable. They considered different measures for consumption and income. We consider their model,

$$y_t = \beta_1 + \beta_2 r_t + \beta_3 r_{t-1} + \beta_4 r_{t-2} + \beta_5 z_t + \beta_6 z_{t-1} + \beta_7 z_{t-2} + \varepsilon_t, \tag{2.5.1}$$

where y_t is the difference in log of real money stock $M2$, r_t is the difference in log of the 3-month treasury bill rate, z_t is the difference in log of *real personal expenditure* (for a consumption measure) or *real GNP* (for an income measure). Elyasiani and Nasseh (1994) adjusted for serial correlation in the errors though the Cochrane-Orcutt procedure which probably helps but may not remove the serial correlation completely. They used the original J test which is valid only under no serial correlation and no heteroskedasticity in the errors. Our approach does not require this step and the J_D test is directly applicable to the data. We use quarterly data from 1959.I (Jan) $\sim$ 2005.III (July) (187 observations) from the Federal Reserve Bank. We used the GNP implicit price deflator to get real M2, the 3-month treasury rates are from the secondary market rate, and the real personal

consumption expenditures and the real GNP are in year-2000-dollars. Table 2.6 shows the results from OLS regressions when the consumption and the income measures were used. We can see that the consumption measure gives a better fit. Figure $2.1 - 2.2$ shows the results from the J_D and F_D tests. The solid lines are the values of the J_D and F_D test statistics. The other lines are critical values of various asymptotic approximations. "Boot(1)" and "Boot(5)" are critical values from the i.i.d. and the block bootstrap respectively. The bootstrap critical values were calculated from $B = 3999$ resamplings. In the tests of the null hypothesis that the income measure (GNP) is the scale variable against the consumption measure, we could reject the null at 5% level for all bandwidths with the J_D tests. With the F_D test, we reject the null on small bandwidths but could not reject with large bandwidths. When the null hypothesis is the consumption measure, we did not reject the null at 5% level for all bandwidths with both the J_D and the F_D tests. Our J_D tests supported the idea of Mankiw and Summers (1986) that consumption is the more appropriate scale variable in the money demand function. With the F_D tests, it's unclear whether the result for large bandwidths in Figure 2.1 is from the low power of the F_D tests or not. On balance, our results support the conclusion that consumption is the better scale variable in the money demand equation.

2.6 Conclusion

Robust J tests and F tests are proposed for comparing nonnested dynamic models. We generalized the test statistics to HAC robust versions (J_D and F_D test). We have shown by Monte Carlo simulations that the bootstrap approaches and the KVB fixed-b asymptotics correct the size distortion known to be a problem with the normal approximation. The usual standard normal asymptotic approximation

Table 2.6: Money (M2) demand function estimation with a consumption measure and an income measure. (Quarterly data from 1959.I to 2005.III)

	z	
Regressor	GNP	Consumption (Expenditure)
constant	0.0024 (2.2225**)	−0.0016 (−1.3040)
r_t	−0.0164 (−3.4405**)	−0.0135 (−3.0514**)
r_{t-1}	−0.0226 (−4.6843**)	−0.0206 (−4.5532**)
r_{t-2}	−0.0092 (−1.9734*)	−0.0071 (−1.6712*)
z_t	0.3041 (4.0340**)	0.4363 (5.0398**)
z_{t-1}	0.2067 (2.6866**)	0.3836 (4.2471**)
z_{t-2}	0.1268 (1.6493)	0.2239 (2.4932**)
R^2	0.29239	0.40581
$\bar{R}^2$	0.2684	0.38567
DW	1.0896	1.2254

** significant at the 5% level, * significant at the 10% level.

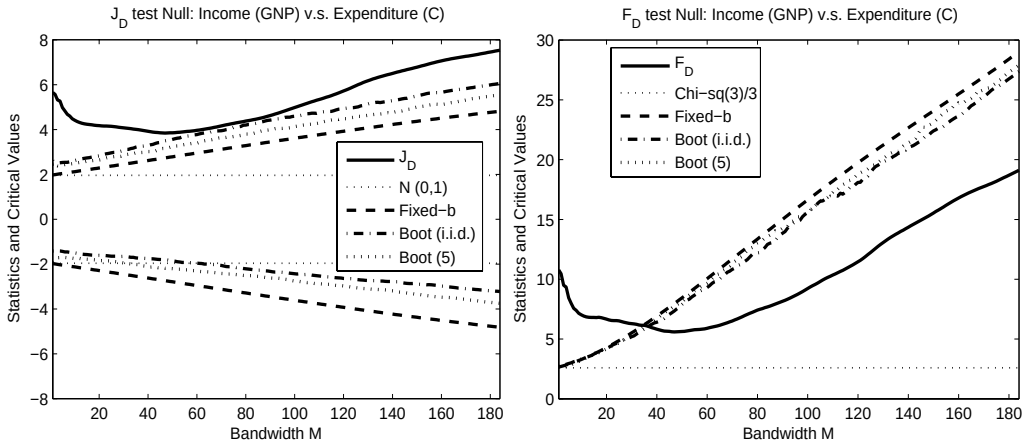

Figure 2.1: Testing the null of income for the scale variable for U.S. money demand. The solid lines are J_D and F_D statistics for different bandwidths and the other lines show the critical values from different asymptotic approximations. "Boot(5)" shows the critical values from the block bootstrap with the block size five.

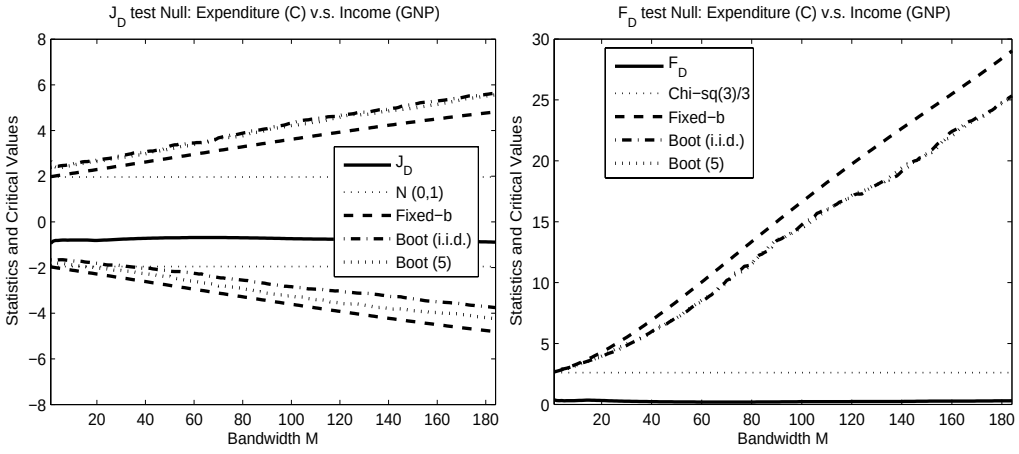

Figure 2.2: Testing the null of consumption for the scale variable for U.S. money demand. The solid lines are J_D and F_D statistics for different bandwidths and the other lines show the critical values from different asymptotic approximations. "Boot(5)" shows the critical values from the block bootstrap with the block size five.

had the worst performance. The fixed-b approach provides a great improvement on the standard normal approximation. The i.i.d. and block bootstrap showed similar size properties, typically better than the fixed-b asymptotic approximation. The block bootstrap method showed robust results.

When the regressors are weakly correlated, the standard normal approximation overrejects seriously and the fixed-b asymptotics also overrejects although it performs better than the standard normal approximation. This overrejection is reduced as bandwidth increases in the fixed-b approach. The block bootstrap works best in this case.

In size controlled power experiments, the J_D test showed better power than the F_D test especially when large bandwidths were used. Strong serial correlation in the errors affected both size and power, and especially decreased power significantly.

In an application to the money demand function in the US, we find that aggregate consumption provides a better scale variable than income.

The overall finding is that the HAC testing of nonnested hypotheses based on the J and F test is feasible and reliable providing a sensible approximation to the sampling distribution is used. The fixed-b and the bootstrap methods are significant improvements on the normal approximation, with the semiparametric block bootstraps providing further improvement especially when the regressors are weakly correlated.

CHAPTER 3

ROBUST MODEL SELECTION IN DYNAMIC MODELS WITH AN APPLICATION TO COMPARING PREDICTIVE ACCURACY

3.1 Introduction

Since Cox (1961, 1962), many methods for distinguishing separate families of hypotheses for model selection have been developed. Model selection is quite different from nested hypothesis testing. The null hypothesis in nested hypothesis testing is well defined but the alternative hypothesis can be arbitrarily close to, though different from, the null, and therefore difficult to detect. Further, these close alternatives may not be importantly different from the null in any practical sense. In contrast, nonnested hypothesis testing has clear separation between candidate models but presents the difficulty of choosing a sensible null hypothesis. Cox used centered log likelihood ratios between two nonnested models under the null hypothesis that one of the models is true. A test for nonnested linear regression models was developed in Pesaran (1974). Along the tradition of the nesting approach of Atkinson (1970) which sets up a general model that contains the candidate models, the J test of Davidson and MacKinnon (1981) is popular (McAleer (1995)). See Gourieroux and Monfort (1999) for a summary.

There is a different approach that does not assume the true model is among the candidates. Vuong (1989) considered a selection criterion based on the difference in the *Kullback-Leibler Information Criterion* (*KLIC*, Kullback and Leibler (1951)) between the (unknown) true model and the competing models, and the null hypothesis is that two models are equivalent in *KLIC*. This approach has the

[0]Coauthored with Nicholas M. Kiefer.

advantage of treating two competing models symmetrically and it does not require the specification of a nesting model. Vuong's approach is sometimes called *model selection* in contrast to nonnested hypothesis testing (Davidson and MacKinnon (2004)). It has recently been extended for dynamic models using different criterion functions (see Rivers and Vuong (2002)).

In nonnested hypothesis testing, the usual asymptotic approximation to the distribution of the J test statistic is known to be poor even with large samples (Godfrey and Pesaran (1983), McAleer (1995)) and the bootstrap is an attractive alternative in these cases (see Fan and Li (1995), Godfrey (1998), Davidson and MacKinnon (2002), and Choi and Kiefer (2005b)). But in model selection, less is known about the performance of the asymptotic approximations of the Vuong (1989) and Rivers and Vuong (2002) test.

This paper proposes a generalized model selection test for dynamic models using a Heteroskedasticity/Autocorrelation Consistent (HAC) estimator of the long run variance as in Rivers and Vuong (2002), and using Kiefer-Vogelsang-Bunzel (KVB) fixed-b asymptotics (Kiefer et al. (2000), Kiefer and Vogelsang (2002a,b, 2005)) to approximate the finite sample distribution of our test statistic. Our approach is applicable to general criterion functions and robust to unknown (nonparametric) serial correlation in the data. Specifically, we represent the idea using a model selection criterion based on quasi-likelihood functions and the resulting test statistic forms a difference-in-KLIC measure. Many general criterion functions can be interpreted as quasi-likelihood functions. The quasi-likelihood functions were used for Monte Carlo study of performance of our test statistic. Our method is compared with a bootstrap method and the conventional standard normal approximation and shown to be remarkably superior to the standard normal approximation.

We also considered a prediction accuracy measure for an empirical application. Our approach was used for two competing exchange rate forecasting models. In the forecasting model comparison literature, a bootstrap method is also used by White (2000). White considers a "benchmark" model and a group of alternative models. The null is that none of the other models dominates the benchmark. The differences between the forecast errors from the benchmark model and all alternatives are arranged in a vector. Then, the test is that the maximum of these differences is negative, so no model dominates the benchmark. The distribution of this maximum is obtained by the stationary bootstrap of Politis and Romano (1994). Thus, this test is like ours, but the null is different, favoring a benchmark model, and of course there is no HAC estimator or fixed-b approximation. Hansen (2005b) also considers comparing a benchmark model with a number of alternatives. He tests the superiority of the benchmark model and uses the stationary bootstrap methods as in White (2000). Hansen (2005b) differs from White (2000) in that he studentizes the statistic before taking the maximum. White is essentially using the null that is closest to the alternative. Hansen estimates the null mean, rather than using zero.

Instead of testing a superiority of prediction accuracy, the idea of testing equivalence in a criterion function is used in Diebold and Mariano (1995) (DM test). The DM test compares forecast accuracy of two competing models, where the accuracy is measured by some criterion function (such as a goodness of fit measure) and the null is that the forecasts are equally accurate. It is similar to Vuong (1989), except the likelihood is not used, rather a fairly general function of the fit. The variance estimator in the DM test is also a HAC estimator. Harvey et al. (1997) attempted to improve finite sample performance of the DM test by using a correction factor

to the DM test statistic (MDM (Modified DM) test).

Our approach is applicable to the DM test. An empirical application for the DM test is presented for testing equality of predictive accuracy of the foreign exchange rate forecasting models considered in Diebold and Mariano (1995) using USD/EURO and YEN/USD exchange rate data. Although we aim to improve the finite sample properties of the DM test statistic, our approach is different from the MDM test in two aspects. First, our test considers a better approximation to the whole distribution of the test statistic whereas the MDM test considers the *scaled* normal approximations only. Second, our approximation depends on the kernel function and bandwidth used in a HAC estimator whereas MDM is derived for a particular kernel function (the uniform kernel) and a bandwidth (a forecasting horizon) used in the DM test.

3.2 Dynamic Model Selection Testing

3.2.1 The test statistic and limiting distributions

Let $p_1(z_1, \theta_1)$ and $p_2(z_2, \theta_2)$ be two models to compare, and (z_i, θ_i) are the variables and the parameter vector used in the model $i = 1, 2$.

Assumption 3.2.1. *The stochastic process* $z_i = \{z_{it}\}_{t=-\infty}^{\infty}$ *is weakly stationary for* $i = 1, 2$.

We consider $\{z_{1t}, z_{2t}\}_{t=1}^{T}$ are the available data used for the model comparison (and estimation of the parameter vectors).

Assumption 3.2.2. *For* $i = 1, 2$*, the estimator* $\hat{\theta}_i$ *of* θ_i *converges to a fixed vector* θ_i^* *in probability, i.e.*

$$\hat{\theta}_i \xrightarrow{p} \theta_i^*. \tag{3.2.1.1}$$

The limits θ_i^* $(i = 1, 2)$ are called pseudo-true values when the models are misspecified. This high-level assumption can itself be based on assumptions about the objective function (for example identification) and the parameter space (for example compactness in the case of an extremum estimator). We assume 3.2.2 directly, noting that there are many routes to the result such as (quasi) maximum likelihood estimation (MLE), generalized method of moments (GMM), minimum divergence estimators (MDE), generalized empirical likelihood (GEL), and other parametric, semiparametric methods.

We consider a model selection procedure that compares Q_i (of a criterion function) from model $i = 1, 2$, then chooses the model that has the smallest Q_i. We assume that Q_i satisfies the following assumption.

Assumption 3.2.3 (Weak law of large numbers)**.** *Let the value of the model selection criterion at pseudo-true values θ_i^* be $Q_i = Q_i(z_i, \theta_i^*)$ for models $i = 1, 2$. We have a function $\widehat{Q}_{iT} = Q_{iT}(\{z_{it}\}_{t=1}^T, \hat{\theta}_i)$ of the data available that satisfies*

$$Q_i = \operatorname*{plim}_{T\to\infty} \widehat{Q}_{iT}. \tag{3.2.1.2}$$

Denoting $\widehat{Q}_{it}^T = Q_i(t, \{z_{is}\}_{s=1}^T, \hat{\theta}_i)$, we have

$$\operatorname*{plim}_{T\to\infty} \sum_{t=1}^{T} \widehat{Q}_{it}^T/T \xrightarrow{p} Q_i,$$

and when $Q_1 = Q_2$, we also have an approximation of $\sqrt{T}(\widehat{Q}_{2T} - \widehat{Q}_{1T})$ given by

$$\left[\sqrt{T}(\widehat{Q}_{2T} - \widehat{Q}_{1T}) - \sum_{t=1}^{T}(\widehat{Q}_{2t}^T - \widehat{Q}_{1t}^T)/\sqrt{T}\right] \xrightarrow{p} 0. \tag{3.2.1.3}$$

Assumption 3.2.3 allows us to calculate the asymptotic variance of $(\widehat{Q}_{2T} - \widehat{Q}_{1T})$ using $\{\widehat{Q}_{2t}^T - \widehat{Q}_{1t}^T\}_{t=1}^T$ under $Q_1 = Q_2$. This assumption is satisfied in many model selection criterion including lack of fit measures such as the mean squared error

($\widehat{Q}_{it}^T = (y_{it} - \hat{y}_{it})^2$) or mean absolute error ($\widehat{Q}_{it}^T = |y_{it} - \hat{y}_{it}|$). When the criterion function is a quasi-likelihood, we use first order Taylor expansion of $\widehat{Q}_{iT} = \ln \hat{\sigma}_i^2$ around the pseudo-true value $(\sigma_i^*)^2$ and get

$$\widehat{Q}_{it}^T = \left[\ln(\sigma_i^*)^2 + \frac{\hat{u}_{it}^2}{(\sigma_i^*)^2} - 1 \right], \tag{3.2.1.4}$$

where $\hat{\sigma}_i^2 = \sum_{t=1}^T \hat{u}_{it}^2/T$ and $\hat{u}_{it}$ are residuals from the quasi-maximum likelihood estimation (QMLE) of the models $i = 1, 2$. This approach was used in Lien and Vuong (1987).

We introduce an additional assumption on $\widehat{Q}_{it}^T$ for asymptotic approximation of the sampling distribution of our test statistic to be described later.

Assumption 3.2.4 (Functional Central Limit Theorem)**.** *Let* $\{\hat{v}_t\}_{t=1}^T = \{\widehat{Q}_{2t}^T - \widehat{Q}_{1t}^T\}_{t=1}^T$. *We have*

$$T^{-1/2} \sum_{t=1}^{[rT]} \hat{v}_t \Rightarrow \lambda W(r), \tag{3.2.1.5}$$

where $W(r)$ *is a standard Brownian motion defined on* $C[0,1]$ *and* λ^2 *is the long run variance of* $\{\hat{v}_t\}$.

Assumption 3.2.4 holds under a variety of regularity conditions and permits conditional heteroskedasticity in $\{\hat{v}_t\}$ but rules out most form of unconditional heteroskedasticity. A set of sufficient conditions can be found in Phillips and Durlauf (1986) which require that the process $\{\hat{v}_t\}$ is weakly stationary, satisfies α-mixing conditions, and each element $\hat{v}_t$ has a finite moment greater than two. The condition holds for stationary and invertible ARMA processes with innovations with finite fourth moments (Hall and Heyde (1980), see Kiefer et al. (2000) for further discussion).

Our null hypothesis is that the competing models are asymptotically "equal",

i.e.

$$Q_1 - Q_2 = 0, \tag{3.2.1.6}$$

and the test statistic is given by

$$\tau_T = \frac{\sum_{t=1}^{T}(\widehat{Q}_{2t}^{T} - \widehat{Q}_{1t}^{T})/\sqrt{T}}{\sqrt{\widehat{V}_T}}, \tag{3.2.1.7}$$

where $\widehat{V}_T$ is the HAC variance estimator of the serially correlated process $\{\hat{v}_t\} = \{\widehat{Q}_{2t}^{T} - \widehat{Q}_{1t}^{T}\}$ given by

$$\widehat{V}_T = \sum_{j=1-T}^{T-1} K\left(\frac{j}{M}\right)\hat{\gamma}(j), \tag{3.2.1.8}$$

where $K(x)$ is the kernel function, M is the bandwidth used in the kernel estimation and the autocovariance function estimator $\hat{\gamma}(j)$ is given by

$$\hat{\gamma}(j) = \frac{1}{T}\sum_{t=|j|+1}^{T}(\hat{v}_t - \bar{v})(\hat{v}_{t-|j|} - \bar{v}), \tag{3.2.1.9}$$

where

$$\bar{v} = \frac{1}{T}\sum_{t=1}^{T}\hat{v}_t. \tag{3.2.1.10}$$

This approach does not specify a correct model and treats two competing models symmetrically. Also it is directional, under an alternative, favoring the model 1 when $\tau_T \overset{a.s.}{\to} +\infty$ and vice versa, if we exclude the cases where Q_i is not defined under the alternative.

Theorem 3.2.5. *Under assumptions 3.2.1–3.2.4, the limiting distribution of the test statistic τ_T under $M/T = b$ is given by the KVB fixed-b asymptotics,*

$$\tau_T \Rightarrow \frac{W(1)}{\sqrt{Q_1(b)}}, \tag{3.2.1.11}$$

where $Q_1(b)$ depends on the kernel function used in $\widehat{V}_T$ and is given by Definition 1 in Kiefer and Vogelsang (2005).

Proof. From assumptions 3.2.1−3.2.4, we get the result by applying Theorem 3 in Kiefer and Vogelsang (2005) to τ_T. □

Different kernels give different denominators in the limiting distribution, thus our approximating distribution depends both on the kernel and bandwidths used. Critical values can be obtained by simulating the fixed-b approximating distributions in practice. For popular kernel functions, they are tabulated in Kiefer and Vogelsang (2005, p.1146).

3.2.2 Quasi-likelihood criterion

In general, the selection criterion Q_i should also be the objective function used in estimation, but this is not necessary. See Rivers and Vuong (2002) for a discussion of using a different model selection criterion than the estimation criterion. See also Pötscher (1991) and Hansen (2005a) for how a model selection step can affect the inference for the models.

We consider the quasi-likelihood function for both the estimation and selection criteria as an example (Many other estimation methods have QMLE interpretation). The quasi-likelihood we specify is the likelihood under normality with independent observations (Heyde (1997)). The quasi-likelihood method leads to consistent parameter estimation under certain conditions (for example, OLS with exogenous regressors and serially correlated errors is consistent). When it is not consistent, its probability limits are pseudo-true values. Using quasi-likelihood also gives our model selection criterion a KLIC interpretation. See Vuong (1989).

We define the model selection criterion $Q_{iT} = -2 \ln p_i(\theta_i)$. The test statistic

τ_T is based on the quasi-log likelihood ratio

$$\ln p_1(\hat{\theta}_1) - \ln p_2(\hat{\theta}_2) = \frac{T}{2} \ln(\hat{\sigma}_2^2/\hat{\sigma}_1^2), \tag{3.2.2.1}$$

and given by

$$\tau_T = \frac{\sqrt{T} \ln(\hat{\sigma}_2^2/\hat{\sigma}_1^2)}{\sqrt{\widehat{V}_T}}, \tag{3.2.2.2}$$

The HAC variance estimator $\widehat{V}_T$ for

$$\hat{v}_t = \widehat{Q}_{2t}^T - \widehat{Q}_{1t}^T \tag{3.2.2.3}$$

$$= \left[\ln(\sigma_2^*)^2 - \ln(\sigma_1^*)^2 + \frac{\hat{u}_{2t}^2}{(\sigma_2^*)^2} - \frac{\hat{u}_{1t}^2}{(\sigma_1^*)^2} \right], \tag{3.2.2.4}$$

is given by plugging $\hat{v}_t$ into eq. (3.2.1.9) and using the estimated $\hat{\sigma}_i^2$ for $(\sigma_i^*)^2$ in eq. (3.2.2.4). The sampling distribution of the test statistic τ_T is approximated by different fixed-b asymptotic approximations depending on the kernel function and the bandwidth.

If the data are i.i.d., this test can be implemented easily (see Lien and Vuong (1987) and Vuong (1989)). Our approach is similar to Lien and Vuong (1987), but we consider serial correlation in $\hat{v}_t$. Our approach includes Lien and Vuong (1987) as a special case $M = 1$. It should be noted that our quasi-likelihood function is applicable to nonlinear models, and our approach in general can be used for any model selection criteria satisfying the assumption 3.2.3 such as the lack of fit criterion, mean squared prediction error (used in Rivers and Vuong (2002)) or mean absolute error (used in Diebold and Mariano (1995)). Our test statistic is also similar to the one considered in Rivers and Vuong (2002). But we use a different approximate distribution given by the KVB approach. We use the quasi-likelihood criterion and show by Monte Carlo simulations that the KVB fixed-b approach gives a superior approximation to the standard normal approximation based on the usual HAC asymptotics.

3.2.3 The bootstrap method

Bootstrap methods are popular alternatives to the conventional asymptotic approximation in econometrics. In the nonnested hypothesis context, the bootstrap is known to improve the approximation of the sampling distributions of test statistics. See Fan and Li (1995), Godfrey (1998), Davidson and MacKinnon (2002), and Choi and Kiefer (2005b).

We used a bootstrap method for our test statistic in a similar way to the method in Hall and Horowitz (1996) and White (2000). In the fixed-b asymptotics, the leading term in the asymptotic expansion is not normal, and the validity of the bootstrap is an open question. Recently, Gonçalves and Vogelsang (2006) showed that the "naive" block bootstrap has the same limiting distribution as the fixed-b asymptotics. The argument proceeds by writing the test statistic and the bootstrap test statistic as the same functions of the data and the bootstrap data respectively. Using appropriate assumptions on the bootstrap data and the continuous mapping theorem gives the result that the limit distributions are identical. Showing that the resulting distribution is an improvement on the normal approximation is more difficult. Gonçalves and Vogelsang (2006) are able to obtain this result for a special case (estimation of a normal mean). See also Jansson (2004) who shows that the fixed-b asymptotics can improve on the normal approximation in terms of rate of error in rejection probability (ERP). Our simulation results indicate that the bootstrap is practically useful in our settings.

Our null hypothesis does not assume a specific form of the true model, therefore we can not use the explanatory variables as given and generate bootstrap samples. This implies that since neither of the candidate models is correct, we should not bootstrap from one particular model. Instead, we should bootstrap

from the joint empirical distribution of the dependent variable and the explanatory variables (sampling together (y_t, x_t) for example). Consequently, when the bootstrap samples are drawn from the original samples which may happen to be a realization in favor of one model over the other, the distribution of the bootstrap test statistic will be biased and give inaccurate critical values. This happens because it is hard to implement the null hypothesis in generating bootstrap samples in our setting. We correct the bootstrap test statistics using the statistics from the original sample as standard in bootstrap literature. We use the quasi-likelihood criterion and the bootstrap test statistics is given by

$$t_b = \frac{\sqrt{T}\left(\ln(\tilde{\sigma}_2^2/\tilde{\sigma}_1^2) - C_0\right)}{\sqrt{\widetilde{V}_T}}, \tag{3.2.3.1}$$

where $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$ are the variance estimators calculated with the bootstrap samples, and

$$C_0 = \ln \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}, \tag{3.2.3.2}$$

where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are variance estimators from the original sample, and $\widetilde{V}_T$ is calculated from eq. (3.2.1.8) and (3.2.1.9) with

$$\tilde{v}_t = \left[\frac{\tilde{u}_{2t}^2}{\hat{\sigma}_2^2} - \frac{\tilde{u}_{1t}^2}{\hat{\sigma}_1^2} - D_1\right], \tag{3.2.3.3}$$

where

$$D_1 = \frac{\tilde{\sigma}_2^2}{\hat{\sigma}_2^2} - \frac{\tilde{\sigma}_1^2}{\hat{\sigma}_1^2}. \tag{3.2.3.4}$$

We have applied the bootstrap method to our examples in the simulation section of this paper. Direct (without modification) bootstrap is not recommended in any case. For all examples, we considered block bootstraps with the block sizes one (the i.i.d. bootstrap) and five.

We also emphasize that a special concern is required for the candidate models with lagged variables. Since the bootstrap cannot be semi-parametric for the

nature of the problem, it is hard to generate the bootstrap $\{y_t\}$ sequentially. We propose to use non-parametric bootstrap with $\{y_t, y_{t-j}, x_t\}$, where y_{t-j} is the vector of all the lagged variable used as explanatory variables in the candidate models and x_t is the vector of all the other explanatory variables, and we drop the first J observation where J is the highest lagged number used. We used this method for our MA(2) example later in this paper.

3.2.4 Linear models: A Curious Result

We consider a special case in which the true model is linear when the quasi-likelihood criterion is used. The true model is

$$y_t = w_t'\delta + x_t'\alpha_1 + z_t'\alpha_2 + u_t \ (t = 1, ..., T), \tag{3.2.4.1}$$

where $\{u_t\}$ is a mean zero weakly stationary process with autocovariance function $\gamma(j)$, and w_t, x_t, z_t are weakly stationary and correlated each other. The competing models are

$$H_1 : y_t = w_t'\delta_1 + x_t'\beta_1 + u_{1t}, \tag{3.2.4.2}$$

$$H_2 : y_t = w_t'\delta_2 + z_t'\beta_2 + u_{2t}, \tag{3.2.4.3}$$

where $t = 1, ..., T$ (T is the number of observations), w_t is the $(l \times 1)$ vector of common regressors, and x_t, z_t are $(k_1 \times 1)$ and $(k_2 \times 1)$ explanatory variables respectively. The parameters $(\delta_i, \beta_i, \sigma_i^2)$ are conditional mean and variance parameter vectors for model H_i. As typical for economic data, w_t, x_t and z_t are serially correlated and the unknown true model's errors are also serially correlated. We rule out data generating processes (DGPs) for which the models H_1 and H_2 are identical ($\delta_1 = \delta_2$ and $\beta_1 = \beta_2 = 0$ for our example), as in this case there is no real

testing problem. Let (δ_1^*, δ_2^*) be the pseudo true values of (δ_1, δ_2) and (β_1^*, β_2^*) be the pseudo true values of (β_1, β_2).

Assumption 3.2.6. *With $\xi_t = (w_t, x_t, z_t)$, two processes, $\{u_t\}$ and $\{\xi_t\}$, are independent.*

Assumption 3.2.7. *The regressors w_t, x_t, and z_t are serially uncorrelated but possibly correlated contemporaneously.*

Assumption 3.2.8. *Two competing models are equal in quasi-likelihood criterion from the true model, i.e. plim $\hat{\sigma}_1^2$ = plim $\hat{\sigma}_2^2$.*

We have the following theorem under the above assumptions.

Theorem 3.2.9. *Under assumptions 3.2.6, 3.2.7 and 3.2.8, the autocovariance function $Cov(U_t, U_{t-j})$ of $U_t = u_{2t}^2 - u_{1t}^2$ is zero for all $j \neq 0$.*

Proof. We have

$$
\begin{aligned}
U_t &= u_{2t}^2 - u_{1t}^2 \\
&= (y_t - w_t'\delta_2^* - z_t'\beta_2^*)^2 - (y_t - w_t'\delta_1^* - x_t'\beta_1^*)^2 \\
&= \{w_t'(\delta_1^* - \delta_2^*) + x_t'\beta_1^* - z_t'\beta_2^*\}\,[2y_t - \{w_t'(\delta_1^* + \delta_2^*) + x_t'\beta_1^* + z_t'\beta_2^*\}] \\
&= \{w_t'(\delta_1^* - \delta_2^*) + x_t'\beta_1^* - z_t'\beta_2^*\} \\
&\quad \times [2(u_t + w_t'\delta + x_t'\alpha_1 + z_t'\alpha_2) - \{w_t'(\delta_1^* + \delta_2^*) + x_t'\beta_1^* + z_t'\beta_2^*\}] \\
&= \{w_t'(\delta_1^* - \delta_2^*) + x_t'\beta_1^* - z_t'\beta_2^*\} \\
&\quad \times [2u_t + w_t'\{2\delta - (\delta_1^* + \delta_2^*)\} + x_t'(2\alpha_1 - \beta_1^*) + z_t'(2\alpha_2 - \beta_2^*)] \\
&= 2u_t\{w_t'(\delta_1^* - \delta_2^*) + x_t'\beta_1^* - z_t'\beta_2^*\} \\
&\quad + \{w_t'(\delta_1^* - \delta_2^*) + x_t'\beta_1^* - z_t'\beta_2^*\} \\
&\quad \times [w_t'\{2\delta - (\delta_1^* + \delta_2^*)\} + x_t'(2\alpha_1 - \beta_1^*) + z_t'(2\alpha_2 - \beta_2^*)].
\end{aligned}
$$

Under assumption 3.2.8 we have

$$E\left(U_t\right) = 0,$$

therefore

$$Cov\left(U_t, U_{t-j}\right) = E\left(U_t, U_{t-j}\right).$$

If we put

$$\begin{aligned} A_t &= \left\{w_t'\left(\delta_1^* - \delta_2^*\right) + x_t\beta_1^* - z_t\beta_2^*\right\}, \\ B_t &= \left\{w_t'\left(\delta_1^* - \delta_2^*\right) + x_t'\beta_1^* - z_t'\beta_2^*\right\} \\ &\quad \times \left[w_t'\left\{2\delta - \left(\delta_1^* + \delta_2^*\right)\right\} + x_t'(2\beta_1 - \beta_1^*) + z_t'(2\beta_2 - \beta_2^*)\right], \end{aligned}$$

we have

$$\begin{aligned} E\left(U_t, U_{t-j}\right) &= E\left(2u_tA_t + B_t\right)\left(2u_{t-j}A_{t-j} + B_{t-j}\right) \\ &= 4E\left(u_tu_{t-j}A_tA_{t-j}\right) + E\left(B_tB_{t-j}\right) \\ &= 4\gamma(j)\gamma_A(j) + \gamma_B(j), \end{aligned} \tag{3.2.4.4}$$

from the independence between u_t and A_t by the assumption 3.2.6. Assumption 3.2.7 implies $\gamma_A(j) = 0$ and $\gamma_B(j) = 0$ for all $j \neq 0$. Therefore we have $Cov(U_t, U_{t-j}) = 0$ for $j \neq 0$. □

Theorem 3.2.9 implies that under the assumptions 3.2.6 and 3.2.8, autocorrelation in u_t does not affect the asymptotic variance of the numerator of our statistic unless the regressors are autocorrelated. The Monte Carlo simulations supported this.

3.2.5 Power of the test

For the comparison of two different test statistics, size corrected power is often used. Since we have proposed different approximations to the distribution of the same

test statistic, size corrected power comparisons are not applicable. To check the finite sample power properties of the fixed-b approximation, we did the following experiments.

- For different sample sizes, compare the powers as a function of the levels implied by the fixed-b approximating distributions given a fixed alternative, a kernel function, and a bandwidth. We considered $T = 50,\ 100,\ 200$.

- For the different sample sizes, compare the local powers given a level, a kernel function, and a bandwidth.

- For different kernel functions, compare the local powers given a level, a bandwidth, and a sample size. We considered five different kernels, Bartlett, Parzen, Quadratic spectral, Daniell, and Bohman. In the fixed-b approach, different kernels give different approximating distributions. We calculated the critical values using the formula given in Kiefer and Vogelsang (2005) for each kernel.

Note that the asymptotic power was not available for the traditional standard normal approximation, since the test statistic's (traditional) limiting distribution under the local alternative is identical regardless of the choice of kernels and bandwidths. The fixed-b asymptotics makes possible comparison of the asymptotic powers for different kernels and bandwidths as shown in Kiefer and Vogelsang (2005). Our finite sample power comparison showed that the fixed-b asymptotic power comparison can be useful in understanding the actual difference in the finite sample powers among kernels and bandwidth choices. The simulation in the next section showed that the fixed-b approximation has reasonable power.

3.3 Monte Carlo Study

We consider two data generating processes. An MA(2) model, and linear regression with autocorrelated regressors and errors.

3.3.1 Size Comparison

MA(2) model

Consider the following MA(2) true data generating process

$$y_t = \varepsilon_t + 0.5\varepsilon_{t-1} + \varepsilon_{t-2} \ (t = 1, ..., T), \tag{3.3.1.1}$$

where $\varepsilon_t \sim$ i.i.d. $N(0, 1)$. The competing models are AR models

$$H_1 : y_t = \alpha_1 + \beta y_{t-1} + \varepsilon_{1t}, \tag{3.3.1.2}$$

$$H_2 : y_t = \alpha_2 + \delta y_{t-2} + \varepsilon_{2t}, \tag{3.3.1.3}$$

where ε_{1t} and ε_{2t} are assumed to be white noises. The true model has $\gamma(1) = \gamma(2)$, and we know $\hat{\beta} \xrightarrow{p} \gamma(1)/\gamma(0)$ and $\hat{\delta} \xrightarrow{p} \gamma(2)/\gamma(0)$. Thus we have the same pseudo true values, $\beta^* = \delta^*$. From this fact we can easily show

$$\text{plim } \hat{\sigma}_1^2 = \text{plim } \hat{\sigma}_2^2, \tag{3.3.1.4}$$

which implies they are equivalent in our quasi-log likelihood criterion function. The variance $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ were calculated based on $T - 1$ observations in H_1 and $T - 2$ observations in H_2 respectively, and the HAC denominator was based on $T - 2$ residuals from H_1 and H_2 (we dropped out the first residual from H_1). The test statistic is given by

$$\tau_T = \frac{\sqrt{T-2}\ln\left(\hat{\sigma}_2^2/\hat{\sigma}_1^2\right)}{\sqrt{\widehat{V}_T}}. \tag{3.3.1.5}$$

The number of iteration of the simulation was $5,000$. We used four sample sizes $T = 12, 27, 52, 102$ for the convenience of the bootstrap. The test are 5% level two tail tests. For the bootstrap tests, we resampled the lagged variables $\{y_t, y_{t-1}, y_{t-2}\}$ together, dropping the first two observations. Therefore the bootstrap sample size is $T-2$, and our choice of sample sizes makes the block bootstrap simple. We used two different block sizes, one (the i.i.d. bootstrap) and five. Of course the i.i.d. bootstrap ignores the serial dependence in the data. The bootstrap critical values were obtained from the 2.5% and 97.5% quantiles of the empirical distribution of the $1,200$ bootstrap iterations. The empirical rejection rates of the standard normal, fixed-b, i.i.d. bootstrap ('boot(1)'),and block bootstrap ('boot(5)') are shown in Figure 3.1.

The fixed-b asymptotics showed great improvement upon the standard normal approximation especially when a large M is used for all sample sizes considered. Also the i.i.d. bootstrap approach was better than the block bootstrap and similar to, but a little bit worse than, the fixed-b approximation. For large sample sizes ($T = 52, 102$) the block bootstrap improves, but in all cases the fixed-b approximation was better than the others. We surmise that a more sophisticated bootstrap approach is required in this setting.

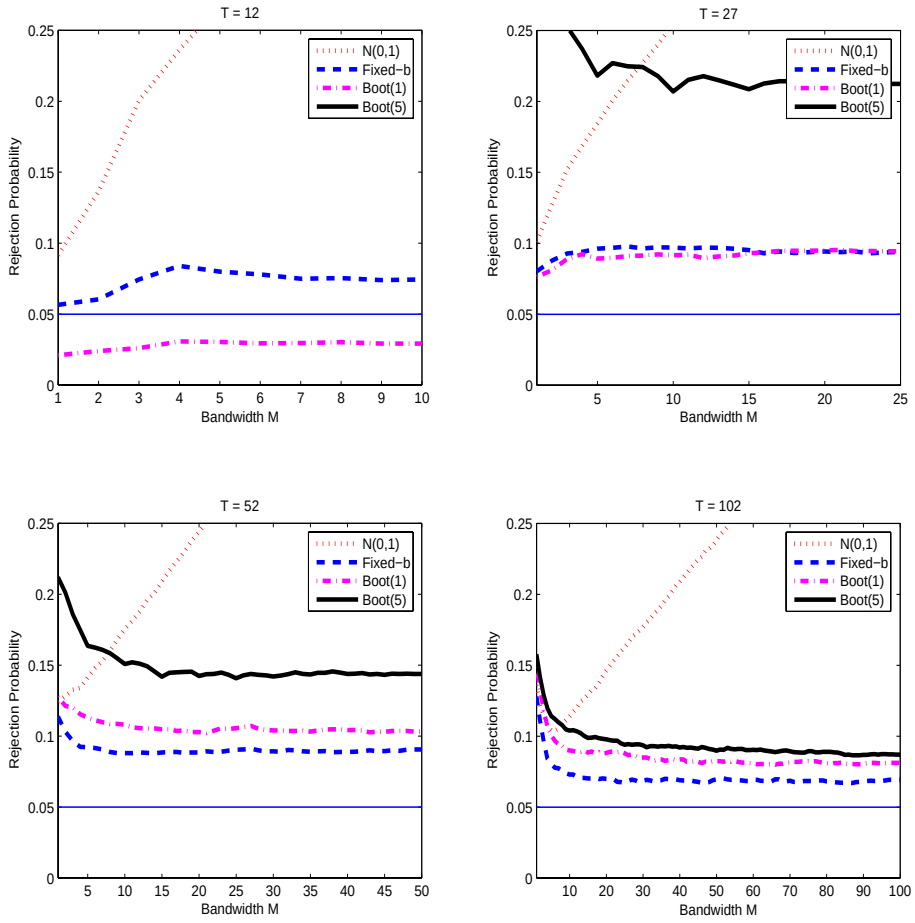

Figure 3.1: MA(2) DGP with two competing AR(1) models

Linear regression model

We generated the following variables for $t = 1, ..., T$

$$u_t = \alpha u_{t-1} + \varepsilon_t,\ \varepsilon_t \sim i.i.d.N\ (0, 1), \tag{3.3.1.6}$$

$$w_t = \rho w_{t-1} + \zeta_{1t}, \tag{3.3.1.7}$$

$$x_t = \rho x_{t-1} + \zeta_{2t}, \tag{3.3.1.8}$$

$$z_t = \rho z_{t-1} + \zeta_{3t}, \tag{3.3.1.9}$$

and

$$\begin{pmatrix} \zeta_{1t} \\ \zeta_{2t} \\ \zeta_{3t} \end{pmatrix} \sim i.i.d.\ N \left(0, \begin{bmatrix} 1 & \kappa_1 & \kappa_1 \\ \kappa_1 & 1 & \kappa_2 \\ \kappa_1 & \kappa_2 & 1 \end{bmatrix} \right), \tag{3.3.1.10}$$

where $\alpha, \rho, \kappa_1, \kappa_2$ are parameters we choose for the simulation. We consider two cases as true models

$$Case\ I : y_t = w_t + 0.5x_t + 0.5z_t + u_t, \tag{3.3.1.11}$$

$$Case\ II : y_t = w_t + 0.5x_t + 0.5z_t + 0.5y_{t-1} + u_t. \tag{3.3.1.12}$$

Note that we have lagged dependent variable in the second case. The competing models are

$$H_1 : y_t = \alpha_1 + w_t\delta_1 + x_t\beta_1 + u_{1t}, \tag{3.3.1.13}$$

$$H_2 : y_t = \alpha_2 + w_t\delta_2 + z_t\beta_2 + u_{2t}. \tag{3.3.1.14}$$

In *Case I*, our competing models are missing one variable, but in *Case II*, both models are missing one variable and one lagged dependent variable. We generated $T = 50$ observations. We have chosen $\kappa_1 = \kappa_2 = 0.5$ and $\rho = 0, \pm 0.5, 0.9,\ \alpha = 0, \pm 0.5, 0.9$. The number of iterations was $5,000$ for each case. For the bootstrap

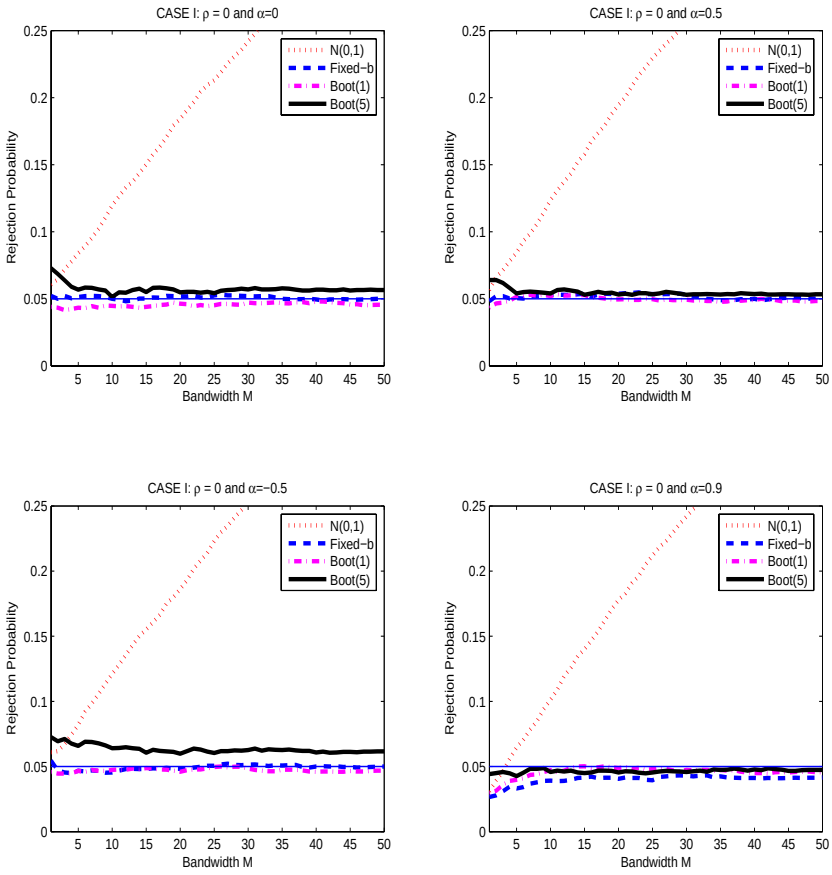

Figure 3.2: (CASE I) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0$, of the errors is α.

tests we have used the modified bootstrap we proposed with block size one and five as in the previous example. We performed 5% level two tail tests. But note that the test can be directional. For example, in the right tail test, the rejection favors the model H_1 over H_2. The results under *Case I* are shown in Figures $3.2 - 3.5$. The results under *Case II* are in Figures $3.6 - 3.9$.

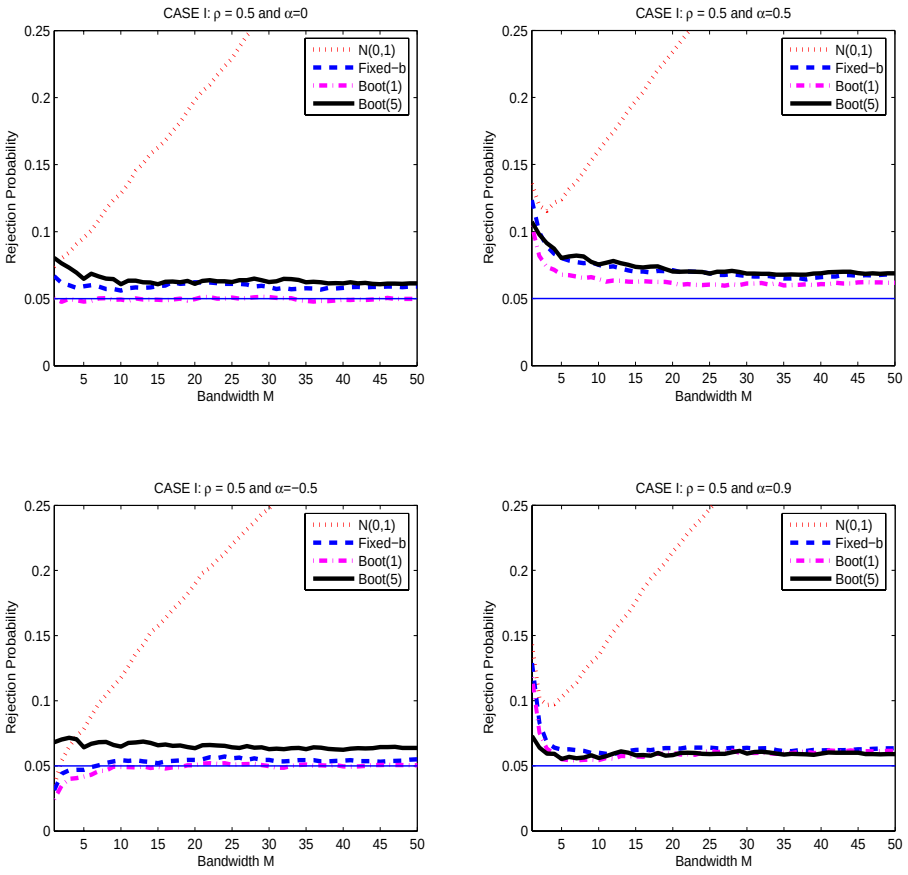

Figure 3.3: (CASE I) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0.5$, of the errors is α.

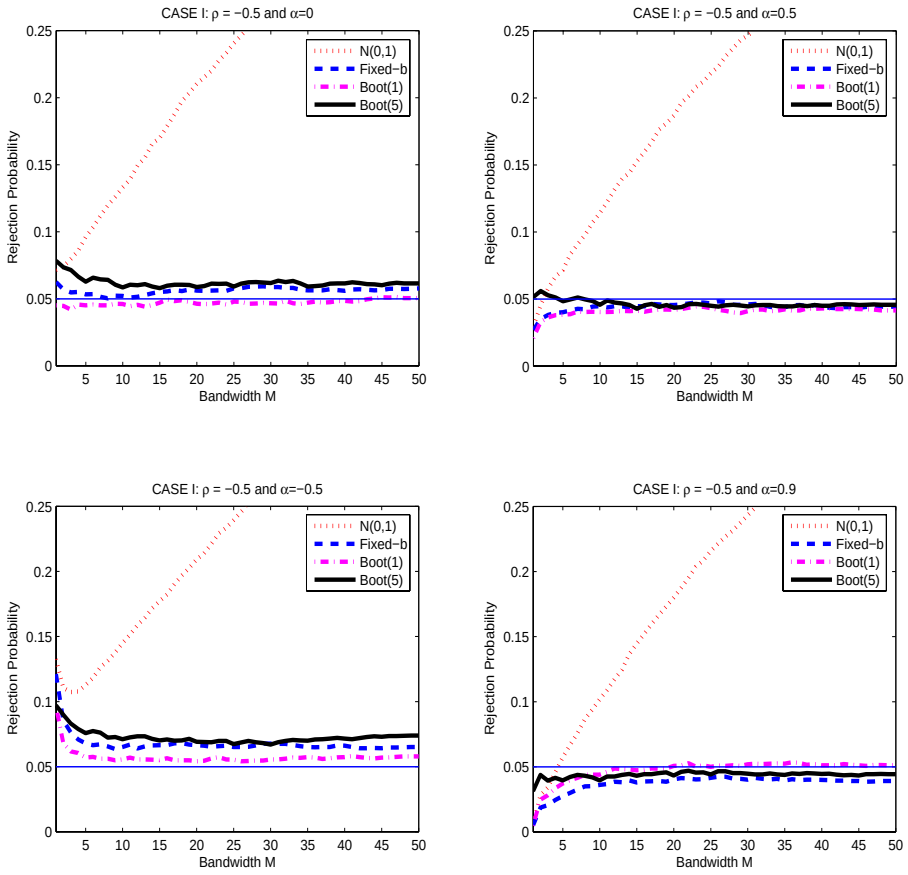

Figure 3.4: (CASE I) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = -0.5$, of the errors is α.

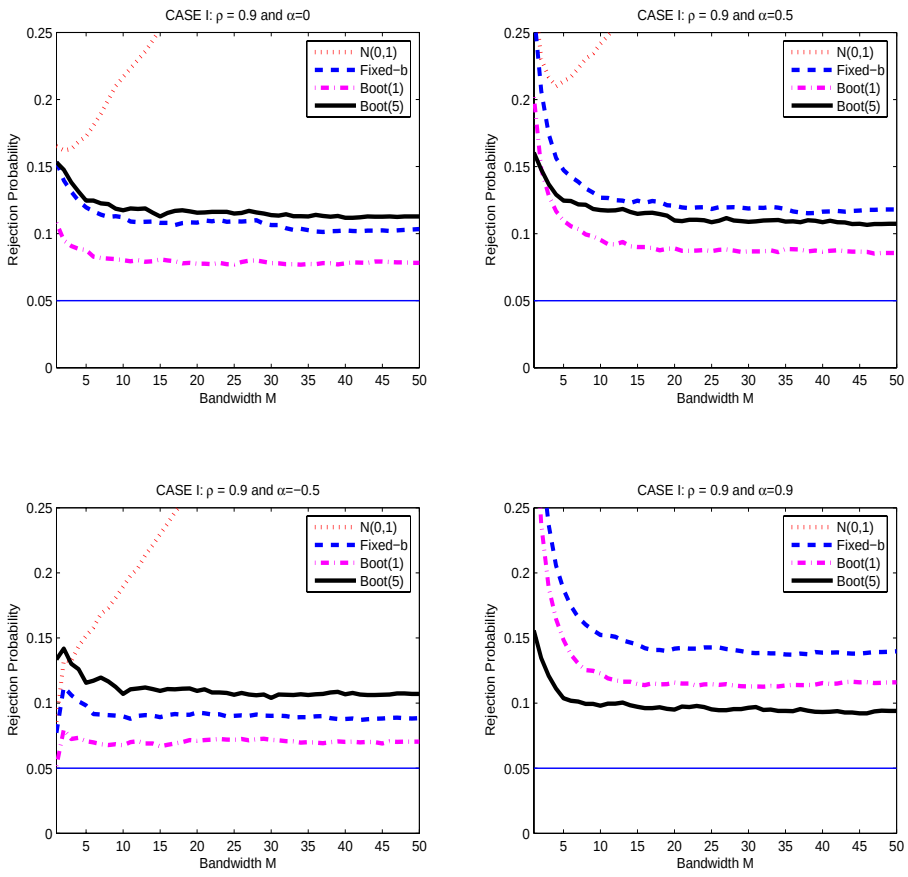

Figure 3.5: (CASE I) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0.9$, of the errors is α.

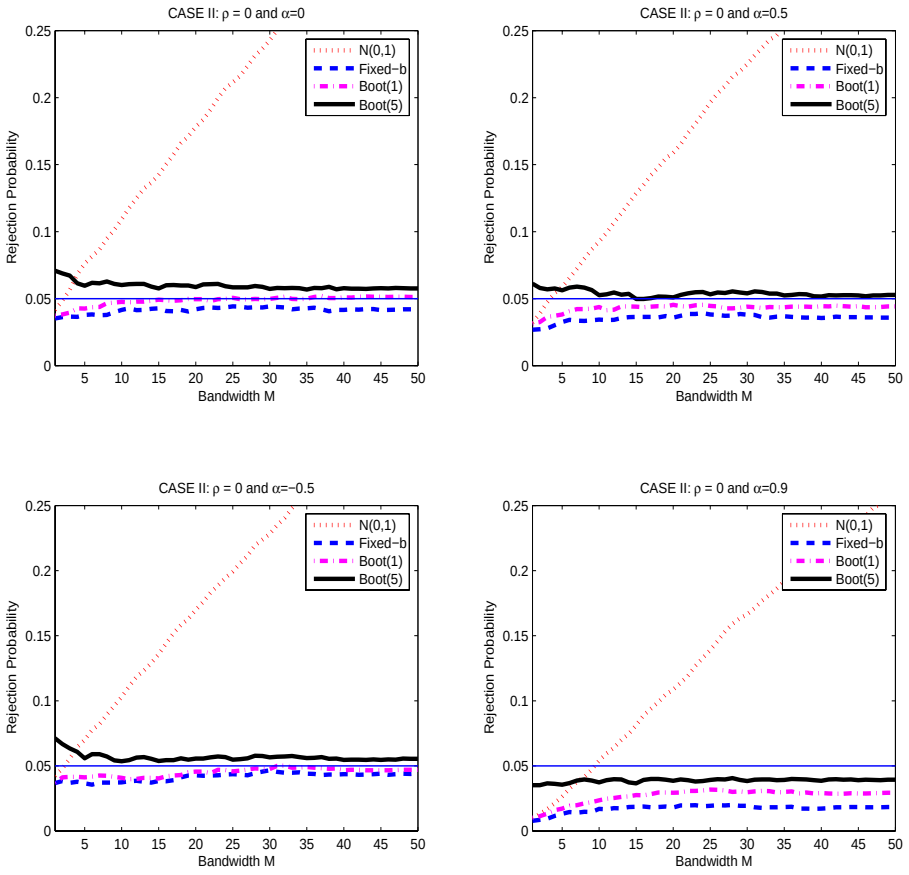

Figure 3.6: (CASE II) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0$, of the errors is α.

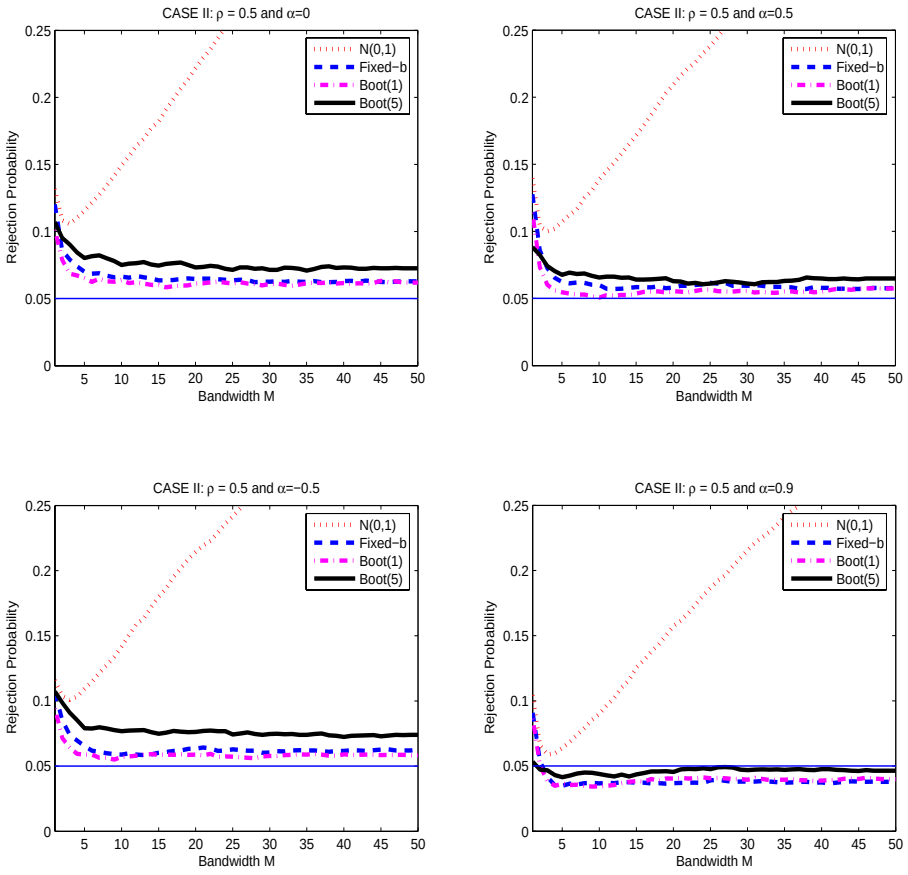

Figure 3.7: (CASE II) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0.5$, of the errors is α.

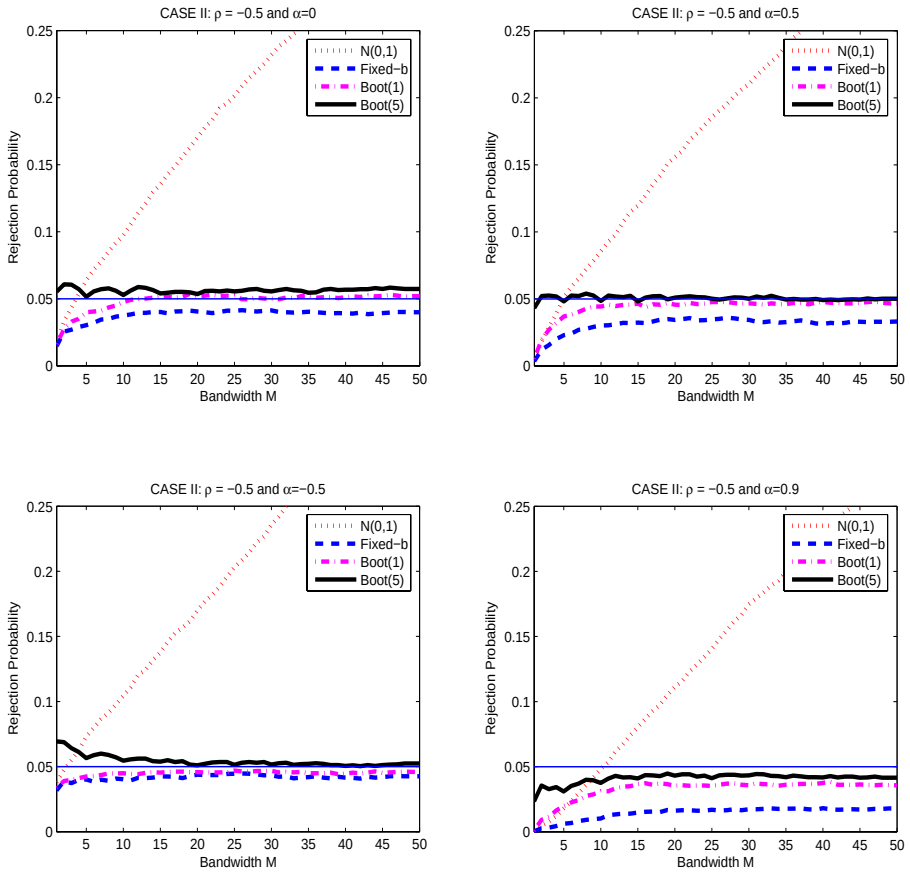

Figure 3.8: (CASE II) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = -0.5$, of the errors is α.

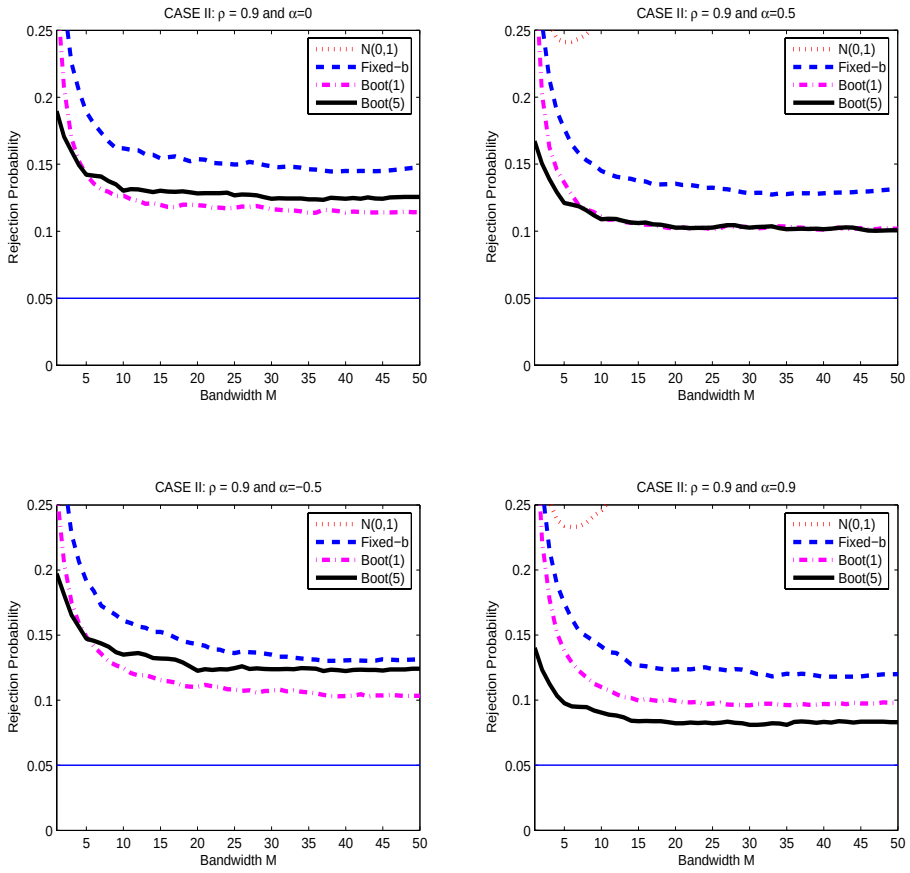

Figure 3.9: (CASE II) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0.9$, of the errors is α.

As shown in the theorem 3.2.9, if regressors are serially uncorrelated ($\rho = 0$), the value of α does not make much difference in the distribution of the test statistic although there were cases with under rejection due to the fact that we have to estimate the pseudo-true values. For $\rho = 0$ of course, it's perhaps best to ignore possible autocorrelation. In all cases, if a robust test is used when unnecessary ($\rho = 0$) then the normal approximation is a disaster and both fixed-b approximation and bootstrap method are better, with very similar performance, although i.i.d. bootstrap seems a little better than block bootstrap. With positive autocorrelation in regressors and errors (the expected case), the robust test is required and the normal approximation is bad. The fixed-b and bootstrap methods beat the normal approximation and are about the same, except when both correlations are quite strong ($\rho = \alpha = 0.9$), in which case the bootstrap methods outperform the fixed-b approach. The ranking of the i.i.d. and block bootstrap when the regressors are highly autocorrelated depends on the actual value of the error autocorrelation, with the block bootstrap performing better with high autocorrelation. Perhaps this is understandable, since the block bootstrap was designed for this case. However it is interesting that the i.i.d. bootstrap is better with moderate error autocorrelation ($\alpha = 0.5$). This is true with and without lagged dependent variables (*Case II and I* respectively). With negative regressor autocorrelation, as might arise from differencing the regressors, the bootstrap and fixed-b methods perform similarly and dominate the normal approximation. In the case of lagged dependent variables and strong positive error autocorrelation as well, the fixed-b tends to under-reject relative to both i.i.d. and block bootstraps. In all cases of negative error autocorrelation, the fixed-b and bootstrap methods perform similarly and dominate the normal approximation.

Although not shown in the figures, we found that when the common regressor w_t is strongly correlated with the other regressors the power of the test is reduced since the wrong model still contains much information through w_t about the true model.

3.3.2 Power Comparison

MA(2) model

For the power comparison, we used the same candidate models as in the size comparison and the true DGP,

$$y_t = \varepsilon_t + 0.5(1+c)\varepsilon_{t-1} + \varepsilon_{t-2}, \tag{3.3.2.1}$$

where $c \in [0,1]$ is the deviation parameter ($c = 0$ gives the null hypothesis) and the errors are from the i.i.d. standard normal distribution. We generated 300 observations and truncated the first 100 observations. Figures $3.10 - 3.12$ are the power comparisons from 5,000 iterations for each of following experiments.

Experiment 1 (Figure 3.10): For the sample sizes $T = 50,\ 100,\ 200$, compare the powers as a function of the levels implied by the fixed-b approximating distributions given a fixed alternative $c = 0.6$, Bartlett kernel, and bandwidths $b = 0.02,\ 0.25,\ 0.5,\ 1$.

Experiment 2 (Figure 3.11): For the sample sizes $T = 50,\ 100,\ 200$, compare the local powers given 5% level, Bartlett kernel, and bandwidths $b = 0.02,\ 0.25,\ 0.5,\ 1$.

Experiment 3 (Figure 3.12): For the five different kernel functions, Bartlett, Parzen, Quadratic spectral, Daniell, and Bohman, compare the local powers

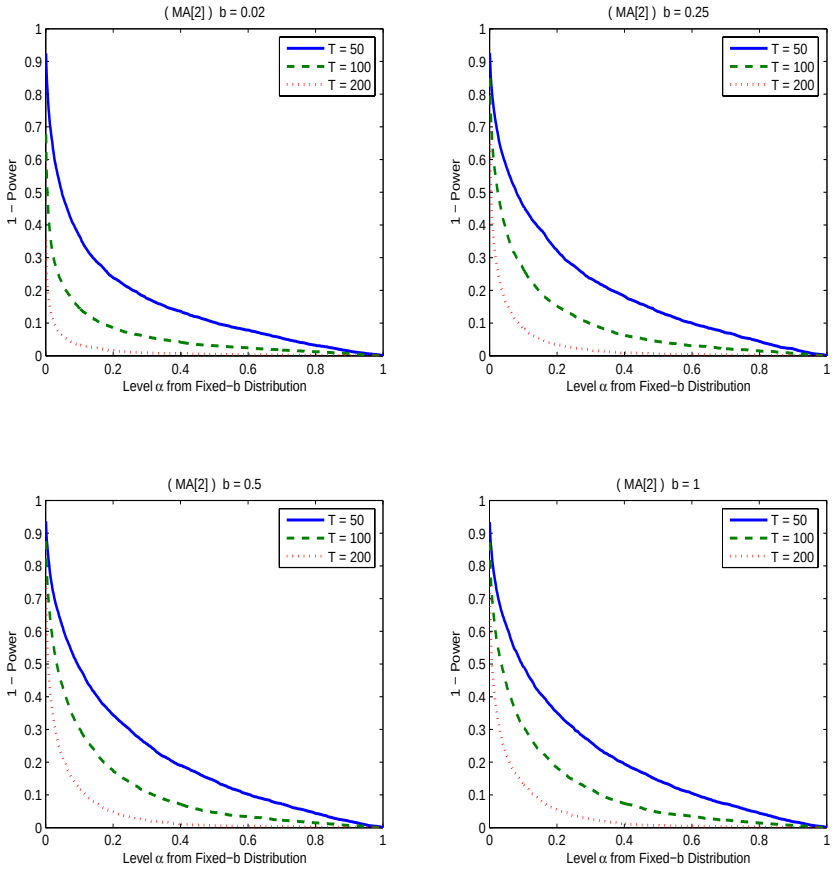

Figure 3.10: (MA(2)) Type II error $(1 - Power)$ as a function of the level α implied by the fixed-b approximating distribution. $b = M/T$ is the bandwidth and T is the sample size.

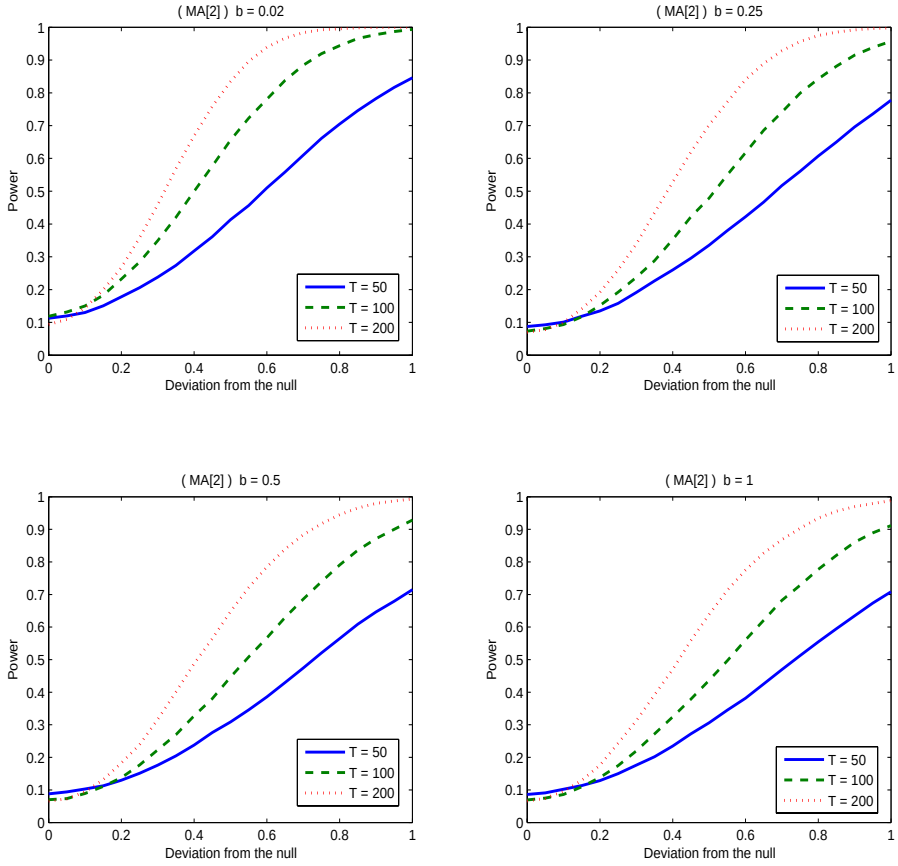

Figure 3.11: (MA(2)) Local power curves for different sample sizes T. $b = M/T$ is the bandwidth.

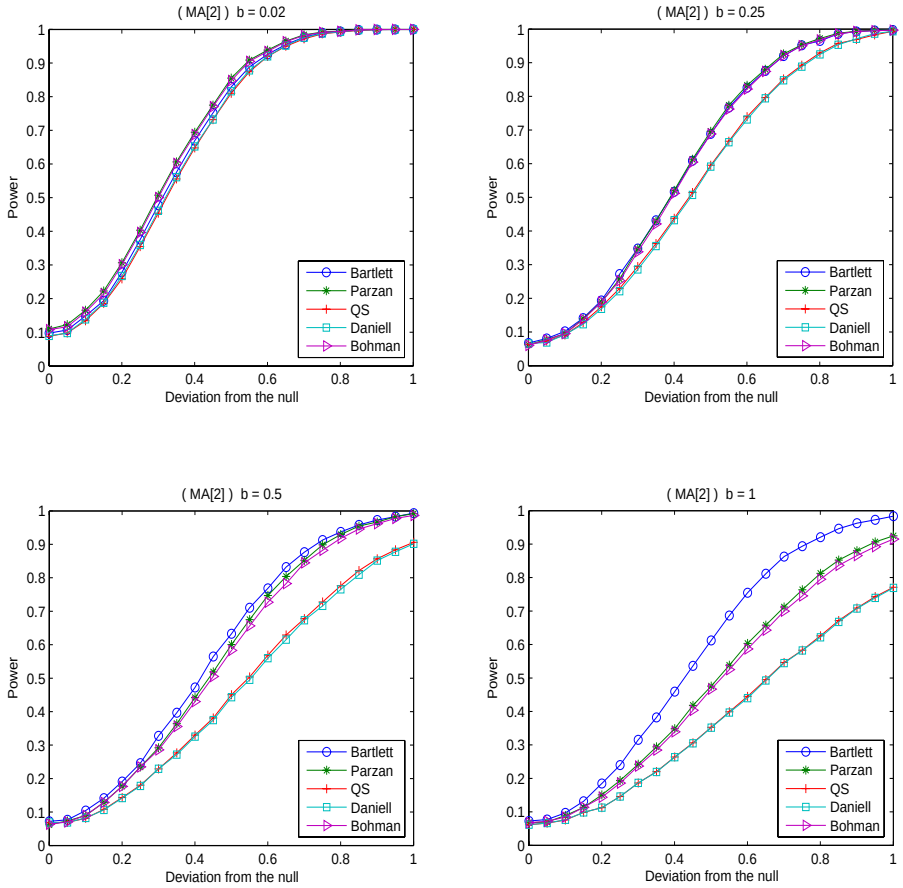

Figure 3.12: (MA(2)) Local power curves for different kernel functions. $b = M/T$ is the bandwidth.

given 5% level, bandwidths $b = 0.02,\ 0.25,\ 0.5,\ 1$, and the sample size $T = 200$.

The first experiment showed the type II errors $(1 - Power)$ for various levels of the test given by fixed-b asymptotic distributions. The power improves as the sample size increases. Larger bandwidths decreased the power but they gave better size behavior. Note that the critical values from the standard normal approximations are smaller than the fixed-b asymptotics critical values thus they will imply larger power at the cost of larger actual size.

The second experiment showed the local power curves with respect to the deviation parameter c ranging from zero to one. Clearly, the power curves are steeper with larger sample sizes. We could also see that smaller bandwidths gave better powers.

In the third experiment, we can see the clear difference between two groups of kernels. The quadratic spectral (QS) and Daniell kernels behaved very similarly and the Bartlett, Parzen, and Bohman kernels gave similar results. The local power curves from the QS and Daniell kernels are sensitive to the bandwidth and large bandwidth decreases the power more than the other kernels. But they showed good size. The power curve of the Bartlett kernel was robust to the bandwidth, and the Parzen and Bohman were also robust but less than the Bartlett kernel. This supports the asymptotic power comparison given in Kiefer and Vogelsang (2005). Small bandwidths increased power as also shown in Kiefer and Vogelsang (2005).

Linear regression model

We use the same candidate models as in the size comparison and the power of the tests was compared with the true DGP

$$Case\ I : y_t = w_t + 0.5(1+c)x_t + 0.5(1-c)z_t + u_t, \tag{3.3.2.2}$$

$$Case\ II : y_t = w_t + 0.5(1+c)x_t + 0.5(1-c)z_t + 0.5y_{t-1} + u_t, \tag{3.3.2.3}$$

where $c \in [0,1]$ is the deviation parameter. We set $\rho = 0.5, \alpha = 0.5$ for the regressors and the error DGP specification in the size comparison section and the other settings are the same. We generated 300 observations and dropped the first 100 observations. Figures $3.13 - 3.15$ (CASE I) and Figures $3.16 - 3.18$ (CASE II) are the power comparisons from 5,000 iterations for the three experiments as in the MA(2) model power comparison.

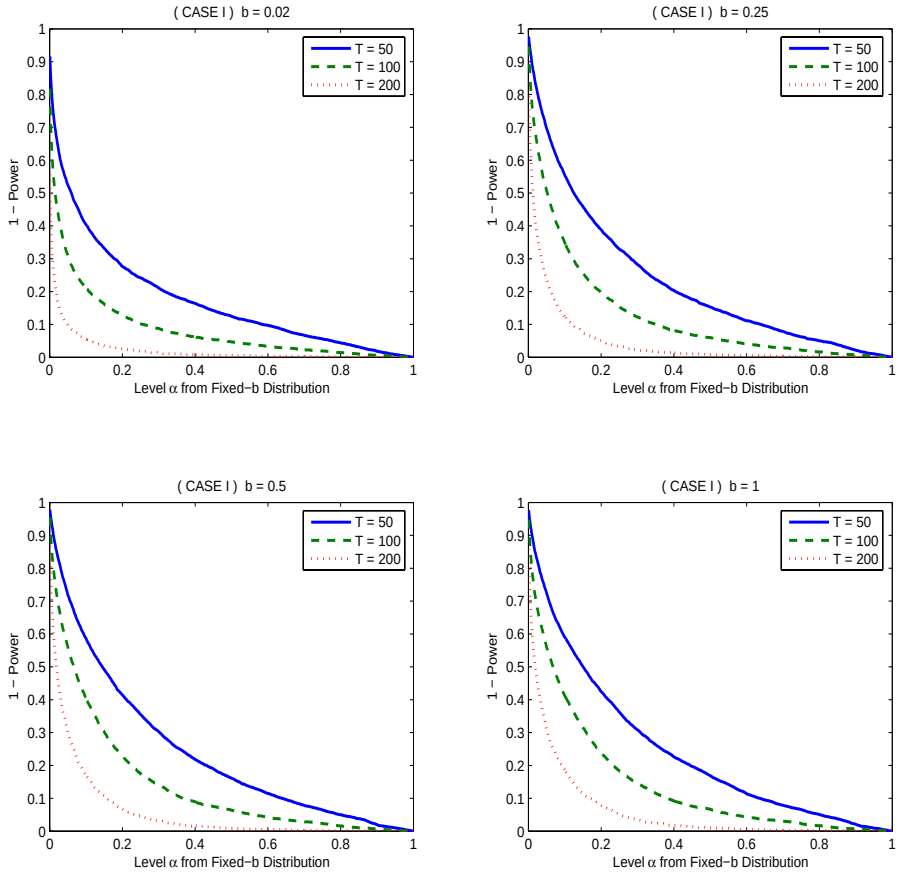

Figure 3.13: (CASE I) Type II error $(1 - Power)$ as a function of the level α implied by the fixed-b approximating distribution. $b = M/T$ is the bandwidth and T is the sample size.

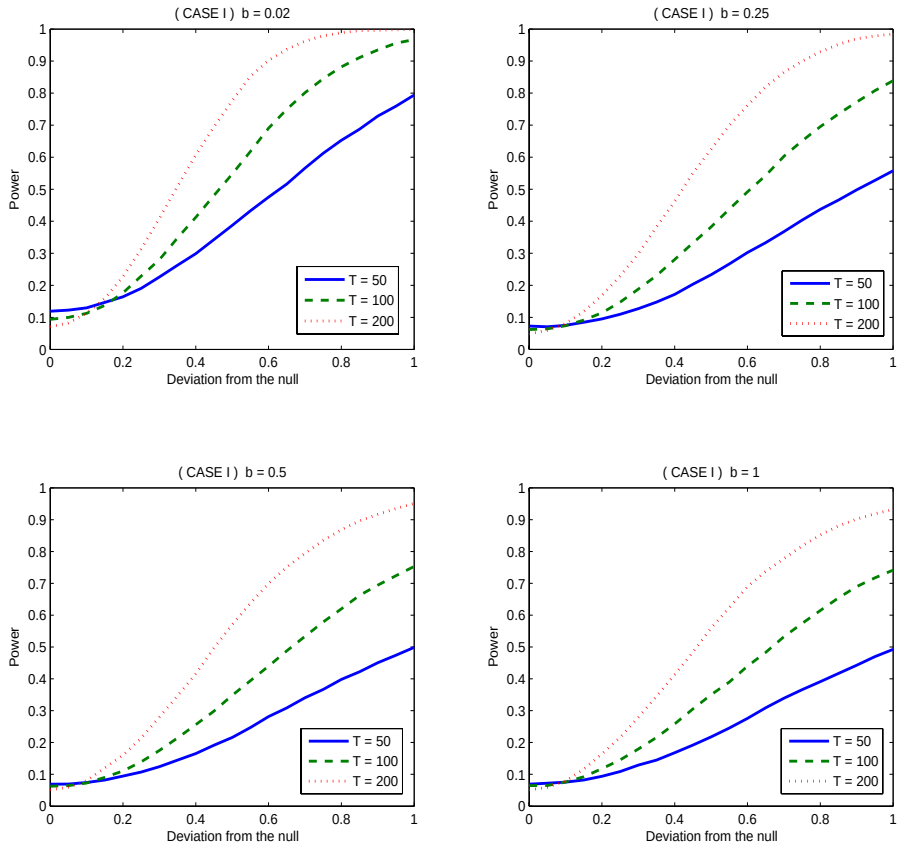

Figure 3.14: (CASE I) Local power curves for different sample sizes T. $b = M/T$ is the bandwidth.

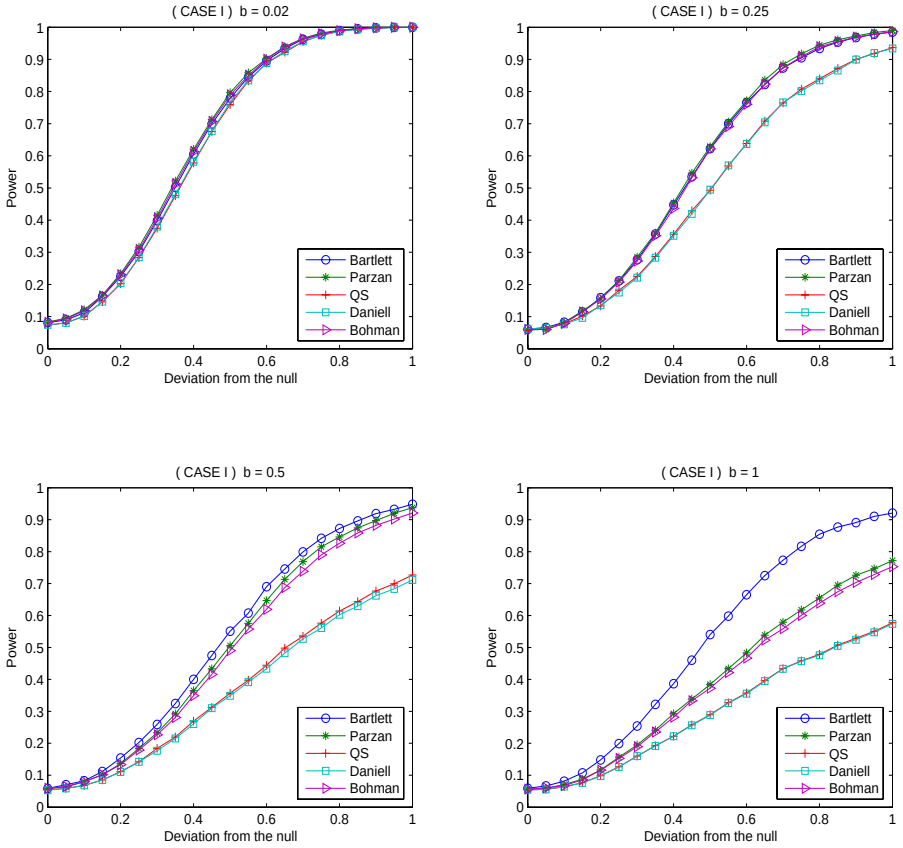

Figure 3.15: (CASE I) Local power curves for different kernel functions. $b = M/T$ is the bandwidth.

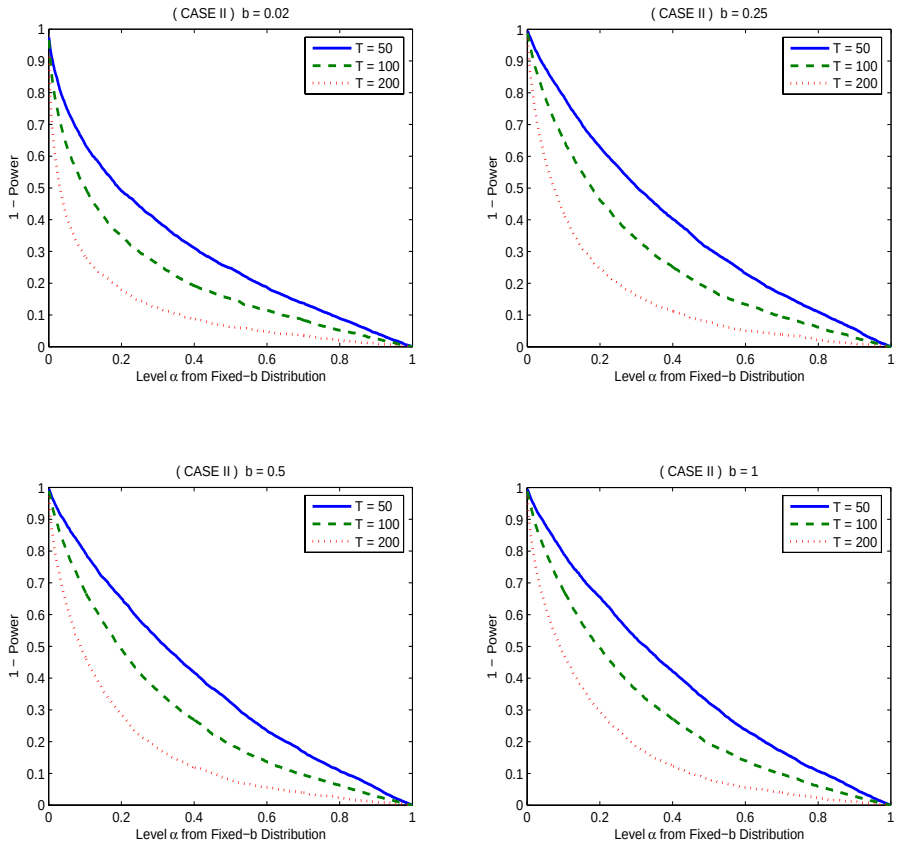

Figure 3.16: (CASE II) Type II error $(1 - Power)$ as a function of the level α implied by the fixed-b approximating distribution. $b = M/T$ is the bandwidth and T is the sample size.

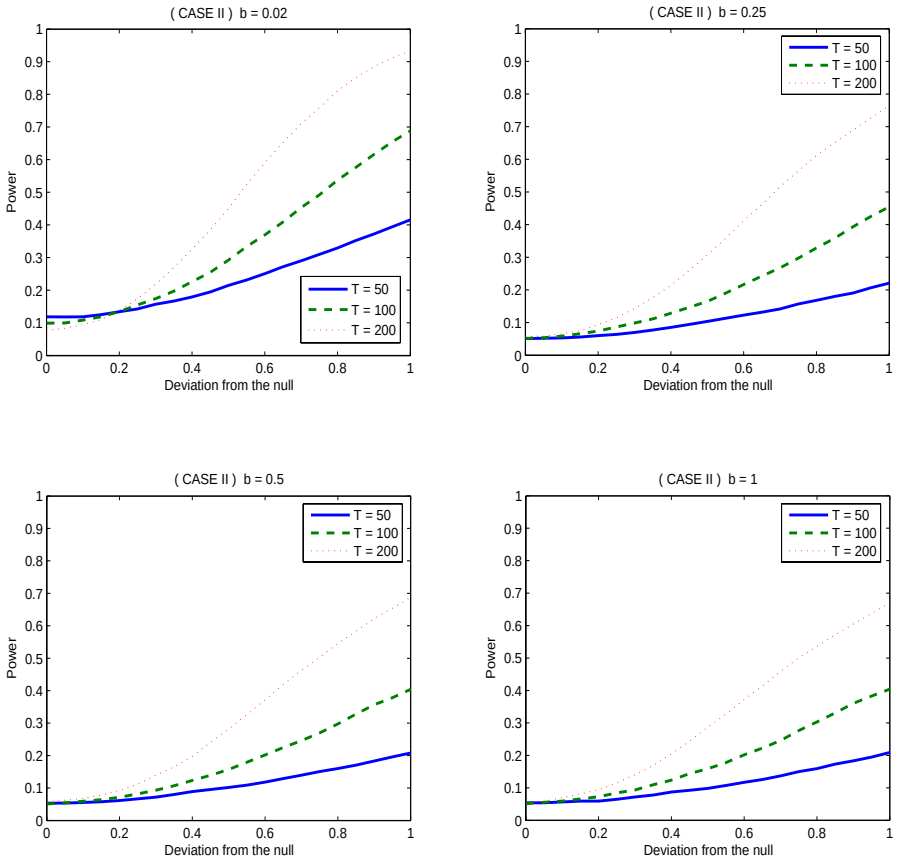

Figure 3.17: (CASE II) Local power curves for different sample sizes T. $b = M/T$ is the bandwidth.

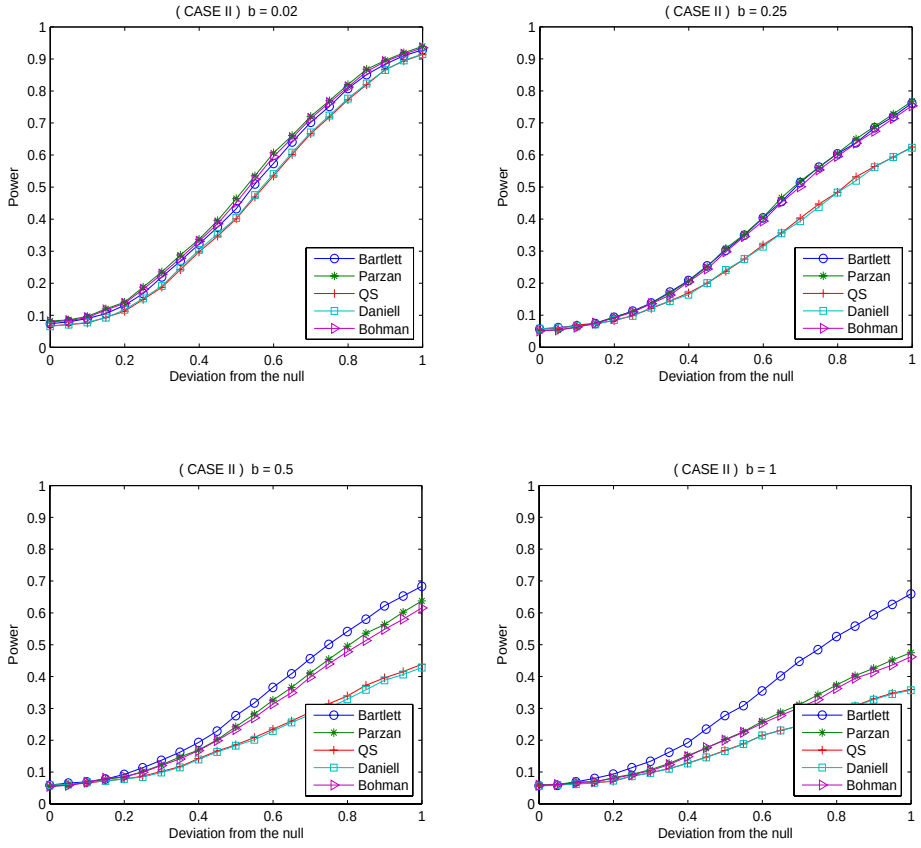

Figure 3.18: (CASE II) Local power curves for different kernel functions. $b = M/T$ is the bandwidth.

We got similar results to the MA(2) power results. The first experiment showed the type II error decreases (the power increases) as sample size becomes larger and small bandwidths give better powers. In the second experiment, larger sample size and smaller bandwidths give better local power. The third experiment shows the Bartlett kernel is robust to the bandwidth for detecting the local alternatives. The QS and Daniell kernels had low local powers when the bandwidth is close to one, but they showed good size in small bandwidths. It is notable that the QS and Daniell kernels behave very similarly and the Parzen and Bohman kernels show close power curves. If we compare CASE I and II, in CASE II where the candidate models are missing the lagged dependent variable y_{t-1}, the powers decreased in all experiments. The power decrease is more severe when we increase the AR(1) coefficient for y_t. We found that the Bartlett kernel has a reasonably good size property with very robust power behavior. Choosing small bandwidth leads to good power but larger size distortion, and a large bandwidth reduces size distortion but lowers the power. The power decrease can be mitigated by using the Bartlett kernel.

Though not shown in figures in the paper, we found that the regressor and the error serial correlation ρ and α affect the power. The power gets worse as serial correlation gets stronger and the effect of α is greater than that of ρ.

3.4 Exchange Rates

Diebold and Mariano (1995) considered a test for equality of predictive accuracy of two exchange rate models in forecasting 3-months ahead spot rates. They considered a random walk model (no difference in 3 months) and forward exchange rate model (current 3-months forward rate). The accuracy is compared with mean

absolute error criterion. We revisit their analysis using New York Federal reserve bank's USD/EURO and YEN/USD, end of month, noon-buying rates (spot rates) and 3-months forward rates. The data range from 1999.1 to 2006.7 and all changes are measured with difference in logs of exchange rates.

The selection criterion is the mean absolute error,

$$E|e_{it}| = E|y_{t+3} - \hat{y}_{it}|, \text{ for } i = 1, 2, \tag{3.4.1}$$

where $y_{t+3} = \log(s_{t+3}/s_t)$ is the change in (actual) spot rates in 3 months, $\{s_t\}$ is the spot exchange rate process, and $\hat{y}_{it}$ is the prediction from model $i = 1, 2$. The prediction from the model 1 is $\hat{y}_{1t} = \log(f_t/s_t)$, where f_t is 3-months forward rate at t, and the model 2 gives a random walk prediction $\hat{y}_{2t} = 0$. The null hypothesis is $E[d_t] = E[|e_{1t}| - |e_{2t}|] = 0$ and our HAC robust test statistic is the same as the DM test given by

$$\tau_T = \frac{\sqrt{T}\bar{d}}{\sqrt{\widehat{V}_T}}, \tag{3.4.2}$$

where $\bar{d}$ is the sample mean of $\{d_t\}$ and $\widehat{V}_T$ is the HAC variance estimator for $\{d_t\}$ with $\hat{v}_t = |e_{1t}| - |e_{2t}|$ in eq. (3.2.1.9), but we use the fixed-b approximation.

Figure 3.19 shows the actual changes of USD/EURO and YEN/USD rates, predictions from the forward rate and the random walk models. The average absolute error in the forward rate model for USD/EURO (YEN/USD) is 0.0194 (0.0173) and for the random walk model, 0.0187 (0.0163). In both currencies, the random walk model wins. We test the statistical significance of the superiority of the random walk model.

Figure 3.20 is the autocovariance function for the $\{d_t\}$ showing a strong serial correlation in low lags and varying degree of correlation in higher order lags. The DM test uses $(h-1)$ as a choice of bandwidths for the h-step ahead forecasting

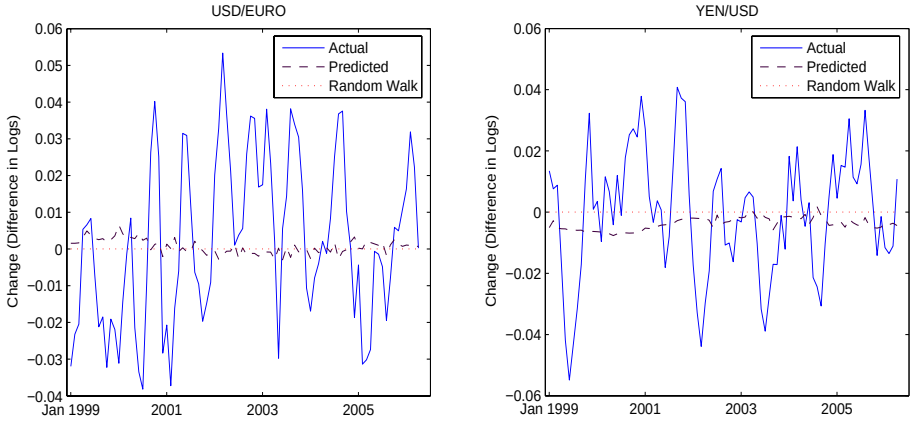

Figure 3.19: Three months change of exchange rates (monthly data). The solid line is actual changes, the dashed is from the 3-months forward rate model, and the dotted is from the random walk prediction (no change).

problem (in our case, $h = 3$) and the uniform kernel. We use the Bartlett kernel and explore all bandwidths.

Figure 3.21 shows the values of our test statistic for a range of bandwidths and the critical values from the fixed-b approximations with 5% level two sided tests. For USD/EURO, we could reject the null for small bandwidths but could not reject for large bandwidths at 5% level (two sided). The tests for YEN/USD could not reject the null for most of the bandwidths. We can see that if we used the standard normal approximation, using large bandwidths will reject the null in the both currencies, and this rejection may have come from the size distortion of the conventional approximation. Also for YEN/USD, the standard normal approximation rejects the null for very low bandwidth but could not reject the null for a wide range of bandwidths up to about $M/T = 1/2$, then rejects the null again for large bandwidths. This confirms the fact that the DM test shows over rejection as the forecasting horizon h gets larger since it uses a bandwidth equal

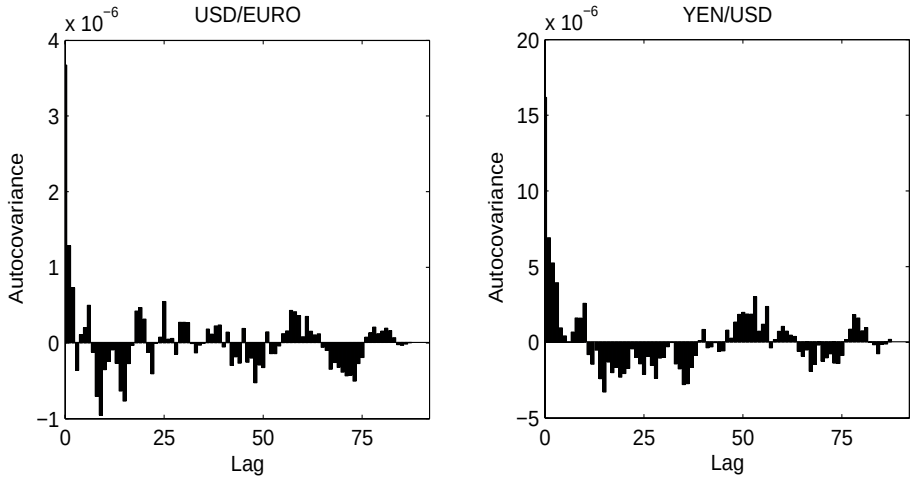

Figure 3.20: Autocovariance function of the difference in absolute prediction error, $\{|e_{1t}| - |e_{2t}|\}$, where e_{1t} =actual change−forward rate model and e_{2t} =actual change−random walk model.

to $(h-1)$ (Harvey et al. (1997)). The fixed-b approximation properly addresses the size distortion problem by giving larger critical values for larger bandwidths.

3.5 Conclusion

For comparing nonnested dynamic models, a robust test statistic was proposed based on a general criterion function or a quasi-log likelihood ratio using a HAC variance estimator. The test treats two competing models symmetrically and does not assume a true model. The test procedure is directional, favoring one over the other. In the special cases of linear models where regressors are serially uncorrelated, serial correlation in the errors has little impact on the distribution of the test statistic. An important improvement in the finite sample properties was made by using the KVB asymptotics. We have shown by Monte Carlo simulations that KVB fixed-b asymptotics corrects the size distortion especially when a large trun-

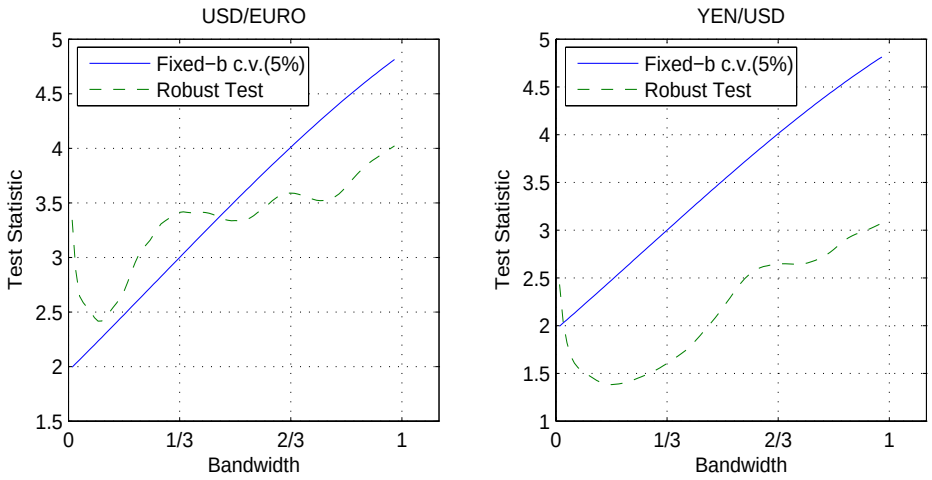

Figure 3.21: Values of the test statistic with various bandwidth and the Bartlett kernel. The solid line is two sided 5% level critical values from the fixed-b approximation. The fixed-b critical value at zero bandwidth is equal to the critical value from the standard normal approximation.

cation number M is used. A bootstrap method is compared with the normal and fixed-b approximations. It shows similar performance to the fixed-b asymptotics. The fixed-b approach showed reasonable local power in our examples especially when the Bartlett kernel is used. There is a trade-off between size and power in the bandwidth selection and the Bartlett kernel gave robust power and reasonably good size. The power is influenced by the correlation in the regressors and the errors and also by the degree of misspecification.

In an application to predicting future spot rates on currency exchanges (USD/EURO and YEN/USD), we find that a random walk model does slightly better in mean absolute prediction error than does a model based on the forward rate. The difference is not significant using the fixed-b asymptotic distribution, though it appears significant using the normal approximation. Thus the documented ten-

dency of the normal approximation to overreject could lead to overstatement of the data's ability to distinguish these two models in our sample.

Using the standard normal approximation for dynamic model selection test is not desirable unless the regressors are not correlated and small M is used for linear models. In general cases, the robust test should be used and the normal approximation should not. The KVB and bootstrap approximations are practical alternatives.

CHAPTER 4

DIFFERENTIAL GEOMETRY AND BIAS CORRECTION IN NONNESTED HYPOTHESIS TESTING

4.1 Introduction

nonnested 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 nonnested 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 nonnested 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

[0]Coauthored with Nicholas M. Kiefer.

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

$$KLIC_1 - KLIC_2 = E_\phi(l_\phi - l_1) - E_\phi(l_\phi - l_2) \tag{4.1.1}$$

$$= E_\phi(l_2 - l_1), \tag{4.1.2}$$

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 (4.1.2) for the test statistic.

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

4.2 Higher order bias correction of the test statistic

4.2.1 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 nonnested 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.

$$KLIC(p_0, p(\theta_1^*)) = KLIC(p_0, p(\theta_2^*)), \tag{4.2.1}$$

or equivalently,

$$KLIC(p_0, p(\theta_1^*)) - KLIC(p_0, p(\theta_2^*)) = E_0\left\{(l(\theta_2^*) - l_0) - (l(\theta_1^*) - l_0)\right\} \tag{4.2.2}$$

$$= E_0(l(\theta_2^*) - l(\theta_1^*)) = 0, \tag{4.2.3}$$

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

$$t_n = \frac{(l(\hat{\theta}_2) - l(\hat{\theta}_1))/\sqrt{n}}{\sqrt{\widehat{V}_n}}, \tag{4.2.4}$$

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

$$l(\theta_j) = \sum_{i=1}^{n} l_i(\theta_j), \tag{4.2.5}$$

$$\bar{l}(\theta_j) = \frac{1}{n}\sum_{i=1}^{n} l_i(\theta_j), \tag{4.2.6}$$

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

$$\widehat{V}_n = \frac{1}{n}\sum_{i=1}^{n}\left\{(l_i(\hat{\theta}_2) - \bar{l}(\hat{\theta}_2)) - (l_i(\hat{\theta}_1) - \bar{l}(\hat{\theta}_1))\right\}^2. \tag{4.2.7}$$

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 nonnested 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 (4.2.4) decomposing the term $l(\hat{\theta}_2) - l(\hat{\theta}_1)$ by

$$l(\hat{\theta}_2) - l(\hat{\theta}_1) = (l(\theta_2^*) - l(\theta_1^*)) + (l(\hat{\theta}_2) - l(\theta_2^*)) - (l(\hat{\theta}_1) - l(\theta_1^*)) \tag{4.2.8}$$

$$= S_1 + S_2, \tag{4.2.9}$$

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

$$E_0(S_1) = E_0(l(\theta_2^*) - l(\theta_1^*)) = 0. \tag{4.2.10}$$

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

$$l(\hat{\theta}_j) - l(\theta_j^*) = -\frac{1}{2} tr\{H(\theta_j^*)^{-1} s(\theta_j^*) s(\theta_j^*)^T\} + O_p(1/\sqrt{n}), \tag{4.2.11}$$

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

$$E_0(S_2) = E_0\left\{(l(\hat{\theta}_2) - l(\theta_2^*)) - (l(\hat{\theta}_1) - l(\theta_1^*))\right\} \tag{4.2.12}$$

$$= -\frac{1}{2} tr\{H(\theta_2^*)^{-1} J(\theta_2^*)\} + \frac{1}{2} tr\{H(\theta_1^*)^{-1} J(\theta_1^*)\} + O(1/\sqrt{n}), \tag{4.2.13}$$

$$= -\frac{1}{2} tr\{\bar{H}(\theta_2^*)^{-1} \bar{J}(\theta_2^*)\} + \frac{1}{2} tr\{\bar{H}(\theta_1^*)^{-1} \bar{J}(\theta_1^*)\} + O(1/\sqrt{n}), \tag{4.2.14}$$

where

$$J(\theta_j^*) = E_0(s(\theta_j^*)s(\theta_j^*)^T) \tag{4.2.15}$$

$$\bar{J}(\theta_j^*) = \frac{J(\theta_j^*)}{n}, \tag{4.2.16}$$

$$\bar{H}(\theta_j^*) = \frac{H(\theta_j^*)}{n}, \tag{4.2.17}$$

for $j = 1, 2$. We propose the correction from the first order term in equation (4.2.14),

$$b = -\frac{1}{2}tr\{\bar{H}(\theta_2^*)^{-1}\bar{J}(\theta_2^*)\} + \frac{1}{2}tr\{\bar{H}(\theta_1^*)^{-1}\bar{J}(\theta_1^*)\}. \tag{4.2.18}$$

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

$$\hat{b} = -\frac{1}{2}tr\{\bar{H}(\hat{\theta}_2)^{-1}\bar{J}(\hat{\theta}_2)\} + \frac{1}{2}tr\{\bar{H}(\hat{\theta}_1)^{-1}\bar{J}(\hat{\theta}_1)\}, \tag{4.2.19}$$

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

$$\tilde{t}_n = \frac{(l(\hat{\theta}_2) - l(\hat{\theta}_1) - \hat{b})/\sqrt{n}}{\sqrt{\widehat{V}_n}}, \tag{4.2.20}$$

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 (4.2.18) is reparameterization invariant and therefore a geometric object.

Theorem 4.2.2. *Let θ be the original parameterization and $\xi(\theta)$ be a locally one-to-one reparameterization of θ with $\xi^* = \xi(\theta^*)$. Then*

$$tr\{H(\theta^*)^{-1}J(\theta^*)\} \tag{4.2.21}$$

in equation (4.2.18) is invariant with respect to reparameterization $\xi(\theta)$, i.e.

$$tr\{H(\theta^*)^{-1}J(\theta^*)\} = tr\{H(\xi^*)^{-1}J(\xi^*)\}. \tag{4.2.22}$$

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

$$s(\xi) = D(\xi)s(\theta(\xi)), \tag{4.2.23}$$

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

$$J(\xi) = D(\xi)J(\theta(\xi))D(\xi)^T, \tag{4.2.24}$$

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

$$h_{ab}(\xi) = \sum_{k,l} D_{ak}(\xi)h_{kl}(\theta(\xi))D_{bl}(\xi) + \sum_k \partial D_{ak}(\xi)/\partial\xi_b \; s_k(\theta(\xi)), \tag{4.2.25}$$

and the second summation in the equation (4.2.25) above has zero expectation at ξ^* since $E_0\{s_k(\theta(\xi^*))\} = 0$ by definition of the pseudo-true value. Therefore we have

$$E_0(h_{ab}(\xi^*)) = H_{ab}(\xi^*) = \sum_{k,l} D_{ak}(\xi^*)H_{kl}(\theta(\xi^*))D_{bl}(\xi^*), \tag{4.2.26}$$

which also can be written as

$$H(\xi^*) = D(\xi^*)H(\theta(\xi^*))D(\xi^*)^T, \tag{4.2.27}$$

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

$$tr\{H(\xi^*)^{-1}J(\xi^*)\} = tr\left[\left\{D(\xi^*)H(\theta(\xi^*))D(\xi^*)^T\right\}^{-1} D(\xi^*)J(\theta(\xi^*))D(\xi^*)^T\right] \tag{4.2.28}$$

$$= tr\{H(\theta^*)^{-1}J(\theta^*)\}. \tag{4.2.29}$$

□

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.

4.2.2 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

$$p_0(y|\eta) = \exp\left[n\left\{\bar{y}^T\eta - \psi(\eta)\right\}\right] f(y), \tag{4.2.1}$$

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

$$p(y|\theta) = \exp\left[n\left\{\bar{y}^T\eta(\theta) - \psi(\eta(\theta))\right\}\right] f(y), \tag{4.2.2}$$

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

$$\kappa(\theta) = \|\eta'(\theta)\|^{-3}_{\eta(\theta)} \left[\|\eta'(\theta)\|^2_{\eta(\theta)} \|\eta''(\theta)\|^2_{\eta(\theta)} - \langle\eta'(\theta), \eta''(\theta)\rangle^2_{\eta(\theta)}\right]^{1/2}, \tag{4.2.3}$$

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 $\|x_1\|^2_{\eta(\theta)} = \langle x_1, x_1\rangle_{\eta(\theta)}$ 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 (4.2.2) in a k-dimensional full exponential family of equation (4.2.1). These two families are the candidates for the nonnested 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 $\mu = E_0(\bar{y})$ 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 (4.2.4) is given by

$$t_n = \frac{l(\hat{\theta}_2) - l(\hat{\theta}_1)/\sqrt{n}}{\sqrt{\widehat{V}_n}} \tag{4.2.4}$$

$$= \frac{\sqrt{n}\left[\bar{y}^T(\eta(\hat{\theta}_2) - \eta(\hat{\theta}_1)) - \left\{\psi(\eta(\hat{\theta}_2) - \psi(\eta(\hat{\theta}_1)))\right\}\right]}{\sqrt{\widehat{V}_n}}, \tag{4.2.5}$$

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

$$\widehat{V}_n = (\eta(\hat{\theta}_2) - \eta(\hat{\theta}_1))^T g(\eta(\bar{y}))(\eta(\hat{\theta}_2) - \eta(\hat{\theta}_1)), \tag{4.2.6}$$

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

$$g(\eta(\bar{y})) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T, \tag{4.2.7}$$

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

4.2.3 Bias correction for one-dimensional curved exponential models

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

$$s(\theta) = n\left\{\bar{y} - \psi'(\eta(\theta))\right\}^T \eta'(\theta), \tag{4.2.1}$$

$$h(\theta) = n\left[\left\{\bar{y} - \psi'(\eta(\theta))\right\}^T \eta''(\theta) - \eta'(\theta)^T\psi''(\eta(\theta))\eta'(\theta)\right] \tag{4.2.2}$$

$$= n\left[\left\{\bar{y} - \psi'(\eta(\theta))\right\}^T \eta''(\theta) - i(\theta)\right], \tag{4.2.3}$$

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

$$E\{s(\theta^*)\} = n(\mu - \psi'(\theta^*))^T \eta'(\theta^*) = 0, \tag{4.2.4}$$

$$\bar{H}(\theta^*) = \left[(\mu - \psi'(\theta^*))^T \eta''(\theta^*) - i(\theta^*)\right]. \tag{4.2.5}$$

Note that when the CEF contains the true model, we have $\mu = \psi'(\theta^*)$, and equation (4.2.5) becomes

$$\bar{H}(\theta^*) = -\eta'(\theta^*)^T \psi''(\theta^*)\eta'(\theta^*) = -i(\theta^*). \tag{4.2.6}$$

When the CEF is misspecified, we have $\mu - \psi'(\theta^*) \neq 0$, but by the orthogonality of $\mu - \psi'(\theta^*)$ and $\eta'(\theta^*)$, equation (4.2.4) 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

$$\bar{J}(\theta) = \eta'(\theta^*)^T g(\phi)\eta'(\theta^*), \tag{4.2.7}$$

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

When the parameter θ satisfies

$$i(\theta) = \|\eta'(\theta)\|^2_{\eta(\theta)} = 1, \text{ for all } \theta, \tag{4.2.8}$$

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.

$$\eta''(\theta) = (\eta''(\theta))_T + (\eta''(\theta))_N, \tag{4.2.9}$$

and

$$\langle \eta'(\theta), (\eta''(\theta))_N \rangle_{\eta(\theta)} = 0, \tag{4.2.10}$$

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

$$\kappa(\theta) = \| \ (\eta''(\theta))_N \ \|_{\eta(\theta)}, \tag{4.2.11}$$

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

Lemma 4.2.3. *Using the arclength parameterization, the bias in equation (4.2.18) can be calculated from*

$$tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\} = \frac{\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*)}{\left\langle (\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}, \left(\eta''(\theta_j^*)\right)_N \right\rangle_{\eta(\theta)} - 1}, \tag{4.2.12}$$

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

$$tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\} = -\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*). \tag{4.2.13}$$

Proof. Using equation (4.2.5), (4.2.7) and $i(\theta_j^*) = 1$, we have

$$tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\} = \frac{\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*)}{(\mu - \psi'(\theta_j^*))^T \eta''(\theta_j^*) - 1}, \tag{4.2.14}$$

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

$$(\mu - \psi'(\theta_j^*))^T \eta''(\theta_j^*) = \left\langle (\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}, \eta''(\theta_j^*) \right\rangle_{\eta(\theta)}. \tag{4.2.15}$$

Since the orthogonality condition in equation (4.2.4) implies $(\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}$ is orthogonal to $\eta'(\theta_j^*)$, i.e.

$$(\mu - \psi'(\theta^*))^T \eta'(\theta^*) = \left\langle (\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}, \eta'(\theta_j^*) \right\rangle_{\eta(\theta)} = 0, \tag{4.2.16}$$

we have

$$(\mu - \psi'(\theta_j^*))^T \eta''(\theta_j^*) = \left\langle (\mu - \psi'(\theta_j^*))^T g(\eta(\theta_j^*))^{-1}, \left(\eta''(\theta_j^*)\right)_N \right\rangle_{\eta(\theta)}, \tag{4.2.17}$$

from equation (4.2.9) and (4.2.15).

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 (4.2.12) 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

$$\eta'(\theta_j^*)^T g(\phi) \eta'(\theta_j^*) = 1, \tag{4.2.18}$$

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

Definition 4.2.4 (Critchley et al. (1994))**.** *For a fixed (true) model ϕ, define*

$$\mu^\phi(\eta) = E_\phi(s(\eta)), \tag{4.2.19}$$

$$g^\phi(\eta) = Var_\phi(s(\eta)), \tag{4.2.20}$$

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 4.2.5. *For one dimensional curved exponential family, the log of*

$$-tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\}$$

can be decomposed by

$$\ln\left(-tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\}\right) = P + K, \tag{4.2.21}$$

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 4.2.12 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 4.2.3. □

4.2.4 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

$$s(\theta) = n\eta'(\theta)^T(\bar{y} - \psi'(\theta)), \tag{4.2.1}$$

where $\eta'(\theta)$ is now the $k \times m$ matrix $\partial\eta(\theta)/\partial\theta' = [\partial\eta(\theta)/\partial\theta_1 \quad ... \quad \partial\eta(\theta)/\partial\theta_m]$, 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

$$\bar{J}(\theta_j^*) = \eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*). \tag{4.2.2}$$

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

$$h_{ab}(\theta) = n\left[(\bar{y} - \psi'(\eta(\theta)))^T\eta_{ab}(\theta) - i_{ab}(\theta)\right], \tag{4.2.3}$$

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

$$\bar{H}_{ab}(\theta_j^*) = (\mu - \psi'(\eta(\theta_j^*)))^T\eta_{ab}(\theta_j^*) - i_{ab}(\theta_j^*). \tag{4.2.4}$$

Using equation (4.2.2), (4.2.4), the bias term,

$$b = -\frac{1}{2}tr\{\bar{H}(\theta_2^*)^{-1}\bar{J}(\theta_2^*)\} + \frac{1}{2}tr\{\bar{H}(\theta_1^*)^{-1}\bar{J}(\theta_1^*)\}, \tag{4.2.5}$$

can be calculated from

$$tr\{\bar{H}(\theta_j^*)^{-1}\bar{J}(\theta_j^*)\} = tr(\left[\bar{H}_{ab}(\theta_j^*)\right]^{-1}\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*)) \tag{4.2.6}$$

$$= tr(\left[(\mu - \psi'(\eta(\theta_j^*)))^T\eta_{ab}(\theta_j^*) - i_{ab}(\theta_j^*)\right]^{-1}\eta'(\theta_j^*)^T g(\phi)\eta'(\theta_j^*)), \tag{4.2.7}$$

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 submanifold. We briefly summarize the differential geometrical approach of Amari (1982). We define the differential operator

$$\partial_a = \frac{\partial}{\partial\theta_a}, \tag{4.2.8}$$

$$\partial_a\partial_b = \frac{\partial^2}{\partial\theta_a\partial\theta_b}, \tag{4.2.9}$$

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

$$\langle\partial_a, \partial_b\rangle = Cov_\theta\left(\partial_a l(\theta), \partial_b l(\theta)\right) \tag{4.2.10}$$

$$= i_{ab}. \tag{4.2.11}$$

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 (4.2.10). 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

$$\partial_a = B_a^i\partial_i, \tag{4.2.12}$$

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.

$$\langle \partial_a, \partial_\gamma \rangle = 0 \quad \text{for } a = 1, ..., m \text{ and } \gamma = 1, ..., k-m, \tag{4.2.13}$$

$$\langle \partial_\gamma, \partial_\zeta \rangle = \delta_\gamma^\zeta \quad \text{for } \gamma = 1, ..., k-m, \text{ and } \zeta = 1, ..., k-m, \tag{4.2.14}$$

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

$$H_{ab\gamma}(\theta) = \langle \partial_a \partial_b, \partial_\gamma \rangle \tag{4.2.15}$$

$$= E\left\{ (\partial_a \partial_b - E\partial_a \partial_b)\partial_\gamma \right\}. \tag{4.2.16}$$

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 (4.2.15) is called a covariant derivative with respect to 1-connection, and from equation (4.2.3), we have

$$\partial_a \partial_b - E\partial_a \partial_b = n(\bar{y} - \psi'(\eta(\theta)))^T \eta_{ab}(\theta). \tag{4.2.17}$$

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

$$\partial_a\partial_b - E\partial_a\partial_b = n(\bar{y} - \psi'(\eta(\theta)))^T\eta_{ab}(\theta) \tag{4.2.18}$$

$$= \Gamma^c_{ab}\partial_c + H^\gamma_{ab}\partial_\gamma \tag{4.2.19}$$

$$= \Gamma^c_{ab}B^i_c\partial_i + H^\gamma_{ab}B^i_\gamma\partial_i, \tag{4.2.20}$$

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 (4.2.12). 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 4.2.6. *The term $(\mu - \psi'(\eta(\theta^*_j)))^T\eta_{ab}(\theta^*_j)$ in equation (4.2.4) is given by*

$$n(\mu - \psi'(\eta(\theta^*_j)))^T\eta_{ab}(\theta^*_j) = A_iB^i_\gamma H^\gamma_{ab}, \tag{4.2.21}$$

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 (4.2.12) and (4.2.19) respectively. If the model is 1-flat, or equivalently, has zero Euler-Schouten curvature (with respect to 1-connection) at θ^*_j, we have $n(\mu - \psi'(\eta(\theta^*_j)))^T\eta_{ab}(\theta^*_j) = 0$.

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

$$\partial^i = g^{ij}\partial_j, \tag{4.2.22}$$

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

$$n(\mu - \psi'(\eta(\theta_j^*)))^T \eta_{ab}(\theta_j^*) = E\left\{(\mu - \psi'(\eta(\theta_j^*)))^T g(\eta(\theta_j^*))^{-1} n(\bar{y} - \psi'(\eta(\theta_j^*)))\right\} \tag{4.2.23}$$

$$\left\{n(\bar{y} - \psi'(\eta(\theta_j^*)))^T \eta_{ab}(\theta_j^*)\right\} \tag{4.2.24}$$

$$= E\left\{A_i \partial^i\right\}\left\{\Gamma^c_{ab} B^i_c \partial_i + H^\gamma_{ab} B^i_\gamma \partial_i\right\} \tag{4.2.25}$$

$$= E\left\{A_i \partial^i\right\}\left(H^\gamma_{ab} B^i_\gamma \partial_i\right) \tag{4.2.26}$$

$$= A_i B^i_\gamma H^\gamma_{ab}, \tag{4.2.27}$$

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

$$E_0 \partial_c = (\mu - \psi'(\eta(\theta_j^*)))^T \eta'(\theta_j^*) \tag{4.2.28}$$

$$= \left\langle A_i \partial^i, B^i_c \partial_i \right\rangle \tag{4.2.29}$$

$$= E\left\{A_i \partial^i\right\}\left\{B^i_c \partial_i\right\} = 0. \tag{4.2.30}$$

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

$$tr\{H(\theta^*_j)^{-1}J(\theta^*_j)\} = tr(\left[A_i B^i_\gamma H^\gamma_{ab} - \delta^b_a\right]^{-1} \eta'(\theta^*_j)^T g(\phi)\eta'(\theta^*_j)), \tag{4.2.31}$$

for $j = 1, 2$ using Theorem 4.2.6. When the model j is exponential flat ($H^\gamma_{ab} = 0$), we have

$$tr\{H(\theta^*_j)^{-1}J(\theta^*_j)\} = -tr(\eta'(\theta^*_j)^T g(\phi)\eta'(\theta^*_j)). \tag{4.2.32}$$

Moreover if the full exponential family is totally flat ($g(\phi) = g(\eta(\theta_j))$), then the term $\eta'(\theta^*_j)^T g(\phi)\eta'(\theta^*_j)$ is also a $(m_j \times m_j)$ identity matrix since θ_j is 0-affine and we have

$$tr\{H(\theta^*_j)^{-1}J(\theta^*_j)\} = -m_j, \tag{4.2.33}$$

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

4.2.5 Summary and Extension

The term $tr\{H(\theta^*_j)^{-1}J(\theta^*_j)\}$ in equation (4.2.18) 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 4.2.5. 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

$$\tilde{l}(\eta) = l_j^* + \sum_{k=1}^{m} \eta_k l^*_{j/\theta_j^k} - \psi(\eta), \tag{4.2.1}$$

where

$$l_j^* = l_j(\theta_j^*), \tag{4.2.2}$$

$$l^*_{j/\theta_j^k} = \left. \frac{\partial^k}{\partial \theta_j^k} l_j(\theta_j) \right|_{\theta_j = \theta_j^*}, \tag{4.2.3}$$

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

$$\eta(\theta_j) = \left((\theta_j - \theta_j^*), \frac{1}{2}(\theta_j - \theta_j^*)^2, \cdots, \frac{1}{m!}(\theta_j - \theta_j^*)^m \right)^T. \tag{4.2.4}$$

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

$$\tilde{l}(\eta) = l_1^* + \sum_{k=1}^{m_1} \eta_k l^*_{1/\theta_1^k} \tag{4.2.5}$$

$$+ l_2^* + \sum_{k=1}^{m_2} \eta_{m_1+k} l^*_{2/\theta_2^k} - \psi(\eta). \tag{4.2.6}$$

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

$$\eta(\theta_1) = \left((\theta_1 - \theta_1^*), \frac{1}{2}(\theta_1 - \theta_1^*)^2, \cdots, \frac{1}{m_1!}(\theta_1 - \theta_1^*)^{m_1}, 0, 0, \cdots, 0 \right)^T, \tag{4.2.7}$$

and

$$\eta(\theta_2) = \left(0, 0, \cdots, 0, (\theta_2 - \theta_2^*), \frac{1}{2}(\theta_2 - \theta_2^*)^2, \cdots, \frac{1}{m_2!}(\theta_2 - \theta_2^*)^{m_2} \right)^T, \tag{4.2.8}$$

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

$$\mu(\phi) = E_0 \left(l^*_{1/\theta_1^1}, l^*_{1/\theta_1^2}, \cdots, l^*_{1/\theta_1^{m_1}}, l^*_{2/\theta_2^1}, l^*_{2/\theta_2^2}, \cdots, l^*_{2/\theta_2^{m_2}} \right)^T, \tag{4.2.9}$$

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.

4.3 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,

$$M_1 : (\eta_1 + 2)^2 + \eta_2^2 = 1, \tag{4.3.1}$$

$$M_2 : (\eta_1 + 0.5)^2 + \eta_2^2 = 1.5^2. \tag{4.3.2}$$

Figure 4.1 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.

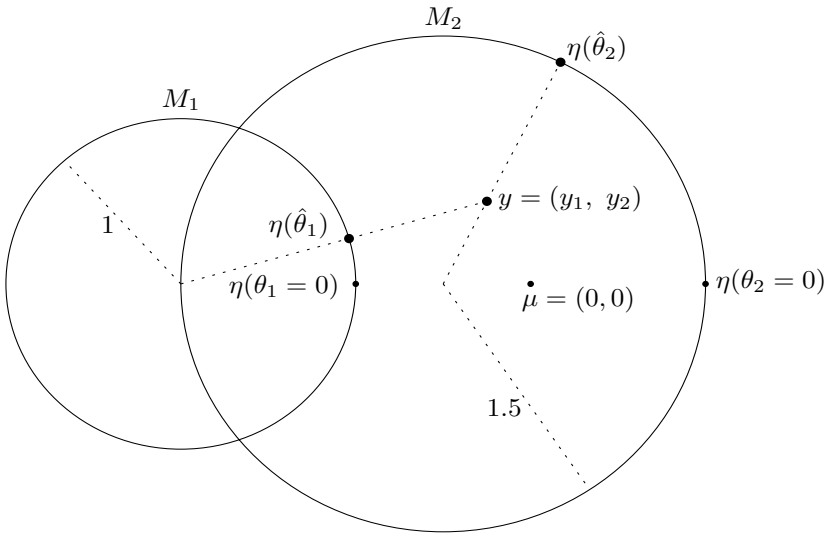

Figure 4.1: Two competing Fisher's circles

We compare the original Vuong test statistic (from equations (4.2.5) and (4.2.6))

$$t_1 = \frac{\{\eta(\hat{\theta}_2) - \eta(\hat{\theta}_1)\}^T y - \{\eta(\hat{\theta}_2)^T \eta(\hat{\theta}_2) - \eta(\hat{\theta}_2)^T \eta(\hat{\theta}_2)\}/2}{\parallel \eta(\hat{\theta}_2) - \eta(\hat{\theta}_1) \parallel} \tag{4.3.3}$$

and the bias corrected test statistic

$$t_2 = t_1 - \frac{\hat{b}}{\parallel \eta(\hat{\theta}_2) - \eta(\hat{\theta}_1) \parallel}, \tag{4.3.4}$$

where

$$\hat{b} = -\frac{1}{2}\left(\frac{1}{\eta''(\hat{\theta}_2)^T(y - \eta(\hat{\theta}_2)) - 1} - \frac{1}{\eta''(\hat{\theta}_1)^T(y - \eta(\hat{\theta}_1)) - 1}\right), \tag{4.3.5}$$

and

$$\kappa_1 = \parallel \eta''(\hat{\theta}_1) \parallel, \quad \kappa_2 = \parallel \eta''(\hat{\theta}_2) \parallel . \tag{4.3.6}$$

Since the embedding space is totally flat, the bias correction term is driven by the curvatures only. Figure 4.2 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 4.3 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.

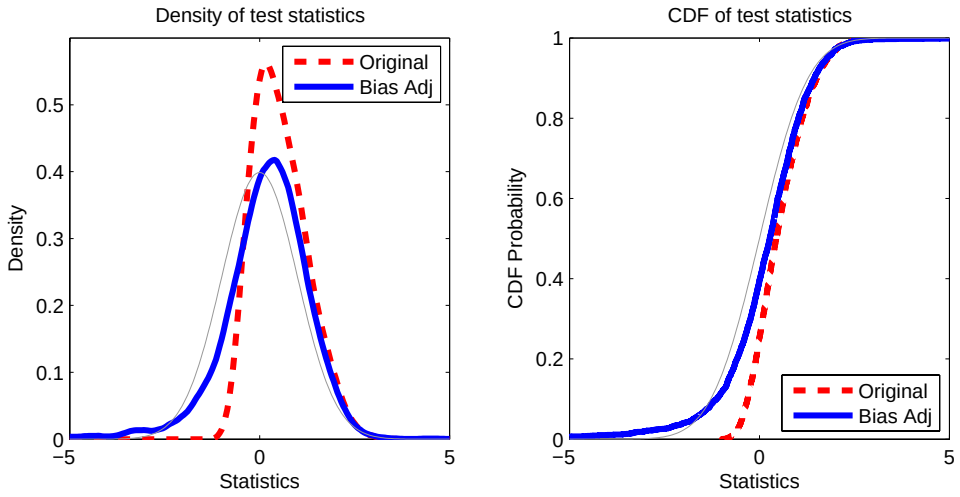

Figure 4.2: Comparison of the distributions of the original Vuong's and the bias corrected test statistics with the standard normal distribution. The thin lines are from $N(0,1)$

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 4.4 shows the CDF comparisons from the different radii of Model 2. As the

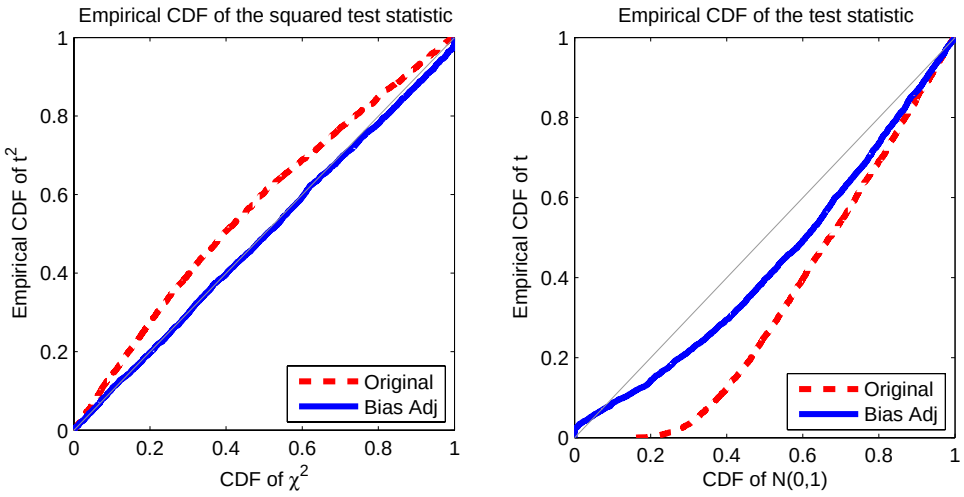

Figure 4.3: Empirical CDF of the squared test statistics with respect to the Chi-square CDF, and the empirical CDF of the test statistics with respect to the standard normal CDF. 45 degree lines imply exact match to the comparing CDF.

curvature of model 2 increases the improvement from the bias correction increases, as expected from our geometric analysis.

4.4 Conclusion

We showed that the numerator of the test statistic of the nonnested 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.

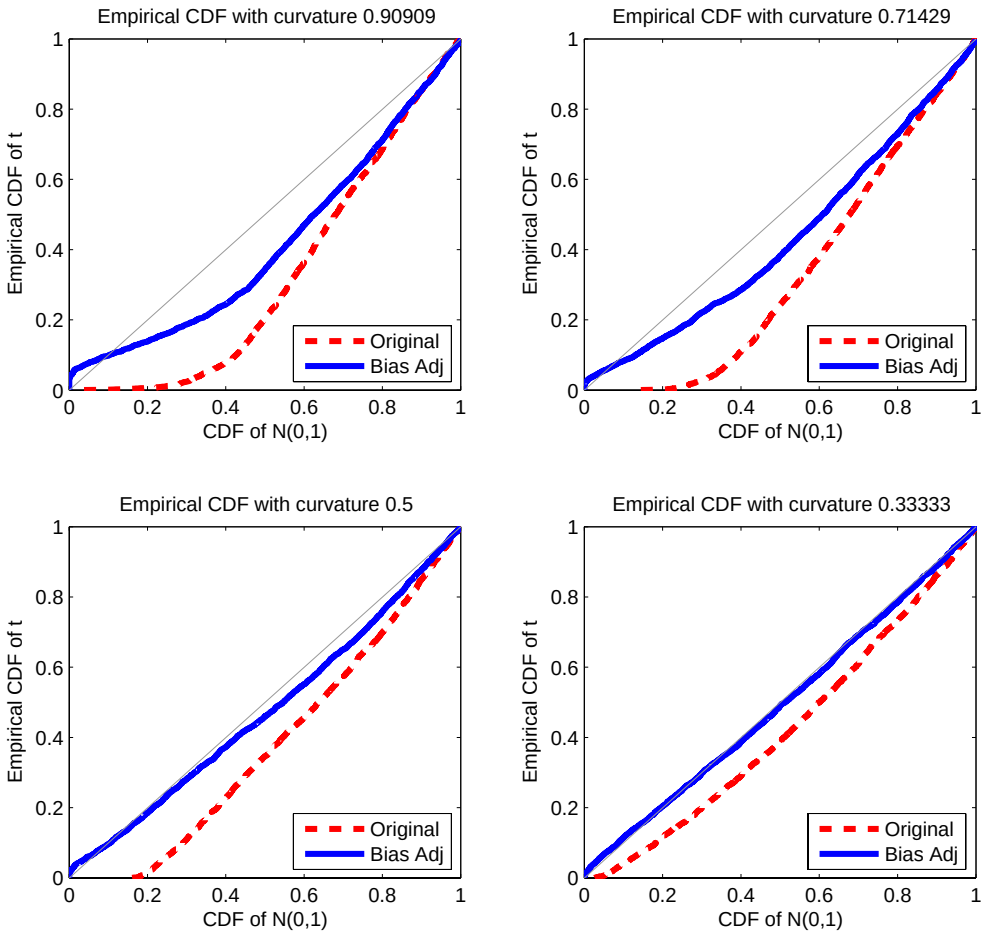

Figure 4.4: Comparison of CDFs of the test statistics from different curvatures for Model 2. The 45 degree line is the exact match of CDFs.

In a simulation, bias correction clearly improved the performance of the test statistic.

CHAPTER 5

CONCLUSION

I extended the applicability of the J test and Vuong's test to dynamic models and proposed the methods to improve the size properties of the tests. I also presented the differential geometry can play an important role in nonnested hypothesis testing by looking at the higher order asymptotic bias. I consider the differential geometry can be shown to be useful in understanding finite sample properties of other nonnested hypothesis testings and model selection procedures, and extensions to model selection issues in semiparametric models would be very important.

BIBLIOGRAPHY

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. *The 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.

Amari, S. I. and Nagaoka, H. (2000). *Methods of Information Geometry*, vol. 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI. Originally published in Japanese by Iwanami Shoten, Publishers, Tokyo, 1993.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59** 817–854.

Atkinson, A. C. (1970). A method for discriminating between models. *Journal of the Royal Statistical Society, Series B*, **32** 211–243.

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.

Bhattacharyya, A. (1943). On discrimination and divergence. In *29th Indian Sci. Cong. Part III*, vol. 13.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomical populations. *Sankhyā*, **7** 401–406.

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. (2005a). Robust model selection in dynamic models. *Working paper, Cornell University*.

Choi, H.-s. and Kiefer, N. M. (2005b). Robust nonnested testing and the demand for money. *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.

Dastoor, N. (1983). Some aspects of testing non-nested hypotheses. *Journal of Econometrics*, **21** 213–228.

Dastoor, N. K. and McAleer, M. (1989). Some power comparisons of joint and paired tests for non-nested models under local hypotheses. *Econometic Theory*, **5** 83–94.

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

Davidson, R. and MacKinnon, J. G. (1985). Heteroskedasticity-robust tests in regression directions. *Annales de l'INSEE*, **59/60** 183–218.

Davidson, R. and MacKinnon, J. G. (2002). Bootstrap j tests of nonnested linear regression models. *Journal of Econometrics*, **109** 167–193.

Davidson, R. and MacKinnon, J. G. (2004). Model selection based on information criteria. In *Econometric Theory and Methods*. Oxford University Press, New York, 675–676.

de Jong, R. M. and Davidson, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometica*, **68** 407–423.

Deaton, A. S. (1982). Model selection procedures, or, does the consumption function exist? In *Evaluating the reliability of macro-economic models* (G. Chow and P. Corsi, eds.). Wiley, New York, 43–65.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, **13** 253–263.

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.

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

Elyasiani, E. and Nasseh, A. (1994). The appropriate scale variable in the u.s. money demand: An application of nonnested tests of consumption versus income measures. *Journal of Business & Economic Statistics*, **12** 47–55.

Fan, Y. and Li, Q. (1995). Bootstrapping j-type tests for non-nested regression models. *Economics Letters*, **48** 107–112.

Fisher, G. and McAleer, M. (1981). Alternative procedures and associated tests of significance for non-nested hypotheses. *Journal of Econometrics*, **16** 103–119.

Godfrey, L. (1998). Tests of non-nested regression models: some results on small sample behaviour and the bootstrap. *Journal of Econometrics*, **84** 59–74.

Godfrey, L. and Pesaran, M. (1983). Tests of non-nested regression models: small sample adjustments and monte carlo evidence. *Journal of Econometrics*, **21** 133–154.

Goldfeld, S. M. and Quandt, R. E. (1972). *Nonlinear Methods in Econometrics*. North-Holland Publishing Company, Amsterdam-London.

Gonçalves, S. and Vogelsang, T. J. (2006). Block bootstrap HAC robust tests: The sophistication of the naive bootstrap. *Working paper, Université de Montréal and Cornell University*.

Gourieroux, C. and Monfort, A. (1999). Testing non-nested hypotheses. In *Handbook of Econometrics*, vol. 4. Elsevier Science Pub Co., North-Holland, 2583–2637.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Applications*. Academic Press, New York.

Hall, P. and Horowitz, J. L. (1996). Bootstrap critical values for tests based on generalized method of moments estimators. *Econometrica*, **64** 891–916.

Hansen, B. E. (1992). Consistent covariance matrix estimation for dependent heterogeneous processes. *Econometrica*, **60** 967–972.

Hansen, B. E. (2005a). Challenges for econometric model selection. *Econometric Theory*, **21** 60–68.

Hansen, P. R. (2005b). A test for superior predictive ability. *Journal of Business & Economic Statistics*, **23** 365–380.

Harvey, D., Leybourne, S. and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, **13** 281–291.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer Series in Statistics, Springer-Verlag, New York, NY.

Jansson, M. (2004). The error in rejection probability of simple autocorrelation robust tests. *Econometrica*, **72** 937–946.

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. (2002a). Heteroskedasticity-autocorrelation robust standard errors using the bartlett kernel without truncation. *Econometrica*, **70** 2093–2095.

Kiefer, N. M. and Vogelsang, T. J. (2002b). Heteroskedasticity-autocorrelation robust testing using bandwidth equal to sample size. *Econometric Theory*, **18** 1350–1366.

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

Kiefer, N. M., Vogelsang, T. J. and Bunzel, H. (2000). Simple robust testing of regression hypotheses. *Econometrica*, **68** 695–714.

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.

Ligeralde, A. V. and Brown, B. W. (1995). Band covariance matrix estimation using restricted residuals: A monte carlo analysis. *International Economic Review*, **36** 751–767.

Ling, S. (1999). On the probabilistic properties of a double threshold ARMA conditional heteroskedasticity model. *Journal of Applied Probability*, **36** 688–705.

Ling, S. and McAleer, M. (2002). Necessary and sufficient moment conditions for the garch(r, s) and asymmetric power garch(r, s) models. *Econometric Theory*, **18** 722–729.

Mankiw, N. G. and Summers, L. H. (1986). Money demand and the effects of fiscal policies. *Journal of Money, Credit and Banking*, **18** 415–429.

McAleer, M. (1995). The significance of testing empirical non-nested models. *Journal of Econometrics*, **67** 149–171.

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

Michelis, L. (1999). The distributions of the j and cox non-nested tests in regression models with weakly correlated regressors. *Journal of Econometrics*, **93** 369–401.

Pesaran, M. H. (1974). On the general problem of model selection. *Review of Economic Studies*, **41** 153–171.

Phillips, P. C. B. and Durlauf, S. N. (1986). Multiple time series regression with integrated processes. *The Review of Economic Studies*, **53** 473–495.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, **89** 1303–1313.

Pötscher, B. M. (1991). The effect of model selection on inference. *Econometric Theory*, **7** 163–185.

Quandt, R. E. (1974). A comparison of methods for testing non-nested hypotheses. *Review of Economics and Statistics*, **56** 92–99.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parametrers. *Bull. Calcutta Math. Soc.*, **37** 81–89.

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

Schouten, J. A. (1954). *Ricci-Calculus*. 2nd ed. Springer, Berlin.

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.

Čenčov, N. N. (1972). *Statistical Decision Rules and Optimal Inference*. Nauka, Moscow. In Russian. English translation: Chentsov (1982), Translation of Mathematical Monographs, Vol. 42. American Mathematical Society, Providence, Rhode Island.

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

White, H. (2000). A reality check for data snooping. *Econometrica*, **68** 1097–1126.