CAE Working Paper #06-09

Robust Model Selection in Dynamic Models with an Application to Comparing Predictive Accuracy

by

Hwan-sik Choi
and
Nicholas M. Kiefer

September 2006

Robust model selection in dynamic models with an application to comparing predictive accuracy

Hwan-sik Choi* Nicholas M. Kiefer†

Cornell University Cornell University

September, 2006

Abstract

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. We propose a robust test which generalizes Lien and Vuong's (1987) test with a Heteroscadasticity/Autocorrelation Consistent (HAC) variance estimator. We use the fixed-b asymptotics developed in Kiefer and Vogelsang (2005) to improve the asymptotic approximation to the sampling distribution of the test statistic. The fixed-b approach is compared with a bootstrap method and the standard normal approximation in Monte Carlo simulations. The fixed-b asymptotics and the bootstrap method are found to be markedly superior to the standard normal approximation. An empirical application for foreign exchange rate forecasting models is presented.

JEL classification: C12, C14, C15, C52

Keywords: Kullback-Leibler Information Criterion (KLIC), quasi-likelihood, dynamic models, fixed-b asymptotics, bootstrap method, Monte Carlo simulation.

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

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

1 Introduction

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, non-nested 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 non-nested models under the null hypothesis that one of the models is true. A test for non-nested 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 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 non-nested hypothesis testing (Davidson and MacKinnon (2004)). It has recently been extended for dynamic models using different criterion functions (see Rivers and Vuong (2002)).

In non-nested 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 (2005)). 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 Heteroscadasticity/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, Vogelsang, and Bunzel (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, Leybourne, and Newbold (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.

2 KVB fixed-b asymptotics

HAC variance estimators are used frequently in econometrics for test statistics involving serially correlated observations. The standard (normal) approximation to the sampling distributions of HAC estimators assumes that the long-run variance is known and equal to its estimated value. The resulting distribution does not depend on the kernel or bandwidth used in variance estimation and is known to give a poor approximation to the sampling distribution, especially for size calculation. Kiefer, Vogelsang, and Bunzel (2000), Kiefer and Vogelsang (2002a,b, 2005) proposed a new asymptotic approximation to the sampling distribution of HAC estimator (and test statistic). They proposed to generate the approximate distribution by fixing the *ratio* $M/T = b > 0$ as T goes to infinity. (In the conventional approach, b converges to zero.) Under this new set-up, the limit of HAC estimator does not converge to the long-run variance but to the long-run variance multiplied by a functional of a *Brownian bridge*. This approach is called the 'fixed-b' approach in comparison

to the conventional 'small-b' approach.

Let $\widehat{V}_T$ be a HAC estimator (used in the denominator of a test statistic) given by

$$\widehat{V}_T = \sum_{j=1-T}^{T-1} K\left(\frac{j}{M}\right)\hat{\gamma}(j), \tag{2.1}$$

where $K(x)$ is the kernel with support $[-1,1]$, T is the sample size, M is the number of lags or the truncation number, and

$$\hat{\gamma}(j) = \frac{1}{T}\sum_{t=j+1}^{T}(\hat{u}_t - \bar{u})(\hat{u}_{t-j} - \bar{u}), \tag{2.2}$$

where $\{\hat{u}_t\}$ is the estimated values of the stochastic process of interest, for example, residuals, scores, or other criteria used in a test statistic, and $\bar{u}$ is the sample mean of $\hat{u}_t$ (Often we have $\bar{u} = 0$ as in residuals from linear regression with a constant term or scores). The limiting distribution of $\widehat{V}_T$ under the fixed-b asymptotics assumption $M/T = b$ as $T \to \infty$ depends on the kernel function $K(x)$ and the bandwidth b. When the functional central limit theorem (FCLT) holds for the partial sum, i.e.

$$T^{-1/2}\sum_{t=1}^{[rT]} u_t \Rightarrow \lambda W(r), \tag{2.3}$$

where $\lambda^2 = \sum_{j=-\infty}^{\infty}\gamma(j)$ and $W(r)$ is the standard Brownian motion defined on $C[0,1]$, and when the Bartlett (triangular) kernel is used, we have

$$\widehat{V}_T \Rightarrow \frac{2\lambda^2}{b}\left[\int_0^1 \widetilde{W}(r)^2\, dr - \int_0^{1-b}\widetilde{W}(r+b)\widetilde{W}(r)\, dr\right], \tag{2.4}$$

where $\widetilde{W}(r)$ is a *Brownian bridge* defined as $\widetilde{W}(r) = W(r) - rW(1)$, under the assumption $M/T = b > 0$ as $T \to \infty$. In testing applications, λ^2 is cancelled out with the asymptotic variance of the numerator of a test statistic making the test statistic pivotal.

Conditions for the FCLT are slightly weaker than the assumptions required for the consistency of HAC estimator. A discussion of the assumptions imposed for the FCLT is given in Kiefer and Vogelsang (2005). Cases with different kernel functions are also in Kiefer and Vogelsang (2005). A simple example is in Choi and Kiefer (2005). We assume a FCLT applies to the related partial

sums in this paper.

3 Dynamic Model Selection Testing

3.1 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 3.1 *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 *For $i = 1, 2$, the estimator $\hat{\theta}_i$ of θ_i converges to a fixed vector θ_i^* in probability, i.e.*

$$\hat{\theta}_i \xrightarrow{p} \theta_i^*. \tag{3.1.1}$$

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 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 3.3 (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*

$$Q_i = \operatorname*{plim}_{T\to\infty} \widehat{Q}_{iT}. \tag{3.1.2}$$

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*

$$\left[\sqrt{T}(\widehat{Q}_{2T} - \widehat{Q}_{1T}) - \sum_{t=1}^{T}(\widehat{Q}_{2t}^{T} - \widehat{Q}_{1t}^{T})/\sqrt{T}\right] \xrightarrow{p} 0. \tag{3.1.3}$$

Assumption 3.3 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

$$\widehat{Q}_{it}^{T} = \left[\ln(\sigma_i^*)^2 + \frac{\hat{u}_{it}^2}{(\sigma_i^*)^2} - 1\right], \tag{3.1.4}$$

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

$$T^{-1/2} \sum_{t=1}^{[rT]} \hat{v}_t \Rightarrow \lambda W(r), \tag{3.1.5}$$

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 3.4 holds under a variety of regularity conditions and permits conditional heteroscadasticity in $\{\hat{v}_t\}$ but rules out most form of unconditional heteroscadasticity. 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, Vogelsang, and Bunzel (2000) for further discussion).

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

$$Q_1 - Q_2 = 0, \tag{3.1.6}$$

and the test statistic is given by

$$\tau_T = \frac{\sum_{t=1}^{T}(\widehat{Q}_{2t}^{T} - \widehat{Q}_{1t}^{T})/\sqrt{T}}{\sqrt{\widehat{V}_T}}, \tag{3.1.7}$$

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

$$\widehat{V}_T = \sum_{j=1-T}^{T-1} K\left(\frac{j}{M}\right)\hat{\gamma}(j), \tag{3.1.8}$$

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

$$\hat{\gamma}(j) = \frac{1}{T}\sum_{t=j+1}^{T}(\hat{v}_t - \bar{v})(\hat{v}_{t-j} - \bar{v}), \tag{3.1.9}$$

where

$$\bar{v} = \frac{1}{T}\sum_{t=1}^{T}\hat{v}_t. \tag{3.1.10}$$

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 $\tau_T \xrightarrow{a.s.} +\infty$ and vice versa, if we exclude the cases where Q_i is not defined under the alternative.

Theorem 3.5 *Under the assumption 3.1−3.4, the limiting distribution of the test statistic τ_T under $M/T \to b > 0$ is given by the KVB fixed-b asymptotics. For example, with the Bartlett kernel and for a bandwidth $M/T \to b \in (0, 1]$, we have*

$$\tau_T \xrightarrow{d} \frac{\mathcal{N}(0,1)}{\sqrt{\frac{2}{b}\left[\int_0^1 \widetilde{W}(r)^2 \, dr - \int_0^{1-b} \widetilde{W}(r+b)\widetilde{W}(r) \, dr\right]}}. \tag{3.1.11}$$

Proof. The result directly follows from Theorem 1 in Kiefer and Vogelsang (2005). ■

Different kernels give different denominators in the limiting distribution, thus our approximating distribution depends both on the kernel and bandwidths used.

3.2 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

$$\ln p_1(\hat{\theta}_1) - \ln p_2(\hat{\theta}_2) = \frac{T}{2} \ln(\hat{\sigma}_2^2/\hat{\sigma}_1^2), \tag{3.2.1}$$

and given by

$$\tau_T = \frac{\sqrt{T} \ln(\hat{\sigma}_2^2/\hat{\sigma}_1^2)}{\sqrt{\widehat{V}_T}}, \tag{3.2.2}$$

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

$$\hat{v}_t = \widehat{Q}_{2t}^T - \widehat{Q}_{1t}^T \tag{3.2.3}$$

$$= \left[\ln(\sigma_2^*)^2 - \ln(\sigma_1^*)^2 + \frac{\hat{u}_{2t}^2}{(\sigma_2^*)^2} - \frac{\hat{u}_{1t}^2}{(\sigma_1^*)^2} \right], \tag{3.2.4}$$

is given by plugging $\hat{v}_t$ into eq. (3.1.9) and using the estimated $\hat{\sigma}_i^2$ for $(\sigma_i^*)^2$ in eq. (3.2.4). 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 3.3 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.

3.3 The bootstrap method

Bootstrap methods are popular alternatives to the conventional asymptotic approximation in econometrics. In the non-nested 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 (2005).

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

$$t_b = \frac{\sqrt{T}\left(\ln(\tilde{\sigma}_2^2/\tilde{\sigma}_1^2) - C_0\right)}{\sqrt{\widetilde{V}_T}}, \tag{3.3.1}$$

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

$$C_0 = \ln \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2}, \tag{3.3.2}$$

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. (3.1.8) and (3.1.9) with

$$\tilde{v}_t = \left[\frac{\tilde{u}_{2t}^2}{\hat{\sigma}_2^2} - \frac{\tilde{u}_{1t}^2}{\hat{\sigma}_1^2} - D_1\right], \tag{3.3.3}$$

where

$$D_1 = \frac{\tilde{\sigma}_2^2}{\hat{\sigma}_2^2} - \frac{\tilde{\sigma}_1^2}{\hat{\sigma}_1^2}. \tag{3.3.4}$$

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.

3.4 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

$$y_t = w_t'\delta + x_t'\alpha_1 + z_t'\alpha_2 + u_t \ (t = 1, ..., T), \tag{3.4.1}$$

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

$$H_1 : y_t = w_t'\delta_1 + x_t'\beta_1 + u_{1t}, \tag{3.4.2}$$

$$H_2 : y_t = w_t'\delta_2 + z_t'\beta_2 + u_{2t}, \tag{3.4.3}$$

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

Assumption 3.7 *The regressors w_t, x_t, and z_t are serially uncorrelated but possibly correlated contemporaneously.*

Assumption 3.8 *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 3.9 *Under assumptions 3.6, 3.7 and 3.8, 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

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

Under assumption 3.8 we have

$$E(U_t) = 0,$$

therefore

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

If we put

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

we have

$$\begin{aligned}
E(U_t, U_{t-j}) &= E(2u_tA_t + B_t)(2u_{t-j}A_{t-j} + B_{t-j}) \\
&= 4E(u_tu_{t-j}A_tA_{t-j}) + E(B_tB_{t-j}) \\
&= 4\gamma(j)\gamma_A(j) + \gamma_B(j), \qquad (3.4.4)
\end{aligned}$$

from the independence between u_t and A_t by the assumption 3.6. Assumption 3.7 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 3.9 implies that under the assumptions 3.6 and 3.8, 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.

3.5 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.

4 Monte Carlo Study

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

4.1 Size Comparison

4.1.1 MA(2) model

Consider the following MA(2) true data generating process

$$y_t = \varepsilon_t + 0.5\varepsilon_{t-1} + \varepsilon_{t-2} \ (t = 1, ..., T), \tag{4.1.1}$$

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

$$H_1 : y_t = \alpha_1 + \beta y_{t-1} + \varepsilon_{1t}, \tag{4.1.2}$$

$$H_2 : y_t = \alpha_2 + \delta y_{t-2} + \varepsilon_{2t}, \tag{4.1.3}$$

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

$$\text{plim } \hat{\sigma}_1^2 = \text{plim } \hat{\sigma}_2^2, \tag{4.1.4}$$

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

$$\tau_T = \frac{\sqrt{T-2}\ln\left(\hat{\sigma}_2^2/\hat{\sigma}_1^2\right)}{\sqrt{\widehat{V}_T}}. \tag{4.1.5}$$

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 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.

4.1.2 Linear regression model

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

$$u_t = \alpha u_{t-1} + \varepsilon_t,\ \varepsilon_t \sim i.i.d.N\ (0,1), \tag{4.1.6}$$

$$w_t = \rho w_{t-1} + \zeta_{1t}, \tag{4.1.7}$$

$$x_t = \rho x_{t-1} + \zeta_{2t}, \tag{4.1.8}$$

$$z_t = \rho z_{t-1} + \zeta_{3t}, \tag{4.1.9}$$

and

$$\begin{pmatrix} \zeta_{1t} \\ \zeta_{2t} \\ \zeta_{3t} \end{pmatrix} \sim i.i.d.\ N\left(0, \begin{bmatrix} 1 & \kappa_1 & \kappa_1 \\ \kappa_1 & 1 & \kappa_2 \\ \kappa_1 & \kappa_2 & 1 \end{bmatrix}\right), \tag{4.1.10}$$

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

$$Case\ I : y_t = w_t + 0.5x_t + 0.5z_t + u_t, \tag{4.1.11}$$

$$Case\ II : y_t = w_t + 0.5x_t + 0.5z_t + 0.5y_{t-1} + u_t. \tag{4.1.12}$$

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

$$H_1 : y_t = \alpha_1 + w_t\delta_1 + x_t\beta_1 + u_{1t}, \tag{4.1.13}$$

$$H_2 : y_t = \alpha_2 + w_t\delta_2 + z_t\beta_2 + u_{2t}. \tag{4.1.14}$$

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 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 $2-5$. The results under *Case II* are in Figures $6-9$.

As shown in the theorem 3.9, 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.

4.2 Power Comparison

4.2.1 MA(2) model

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

$$y_t = \varepsilon_t + 0.5(1 + c)\varepsilon_{t-1} + \varepsilon_{t-2}, \tag{4.2.1}$$

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 $10 - 12$ are the power comparisons from 5,000 iterations for each of following

experiments.

Experiment 1 (Figure 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 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 12): For the five different kernel functions, Bartlett, Parzen, Quadratic spectral, Daniell, and Bohman, compare the local powers 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).

4.2.2 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

$$Case\ I : y_t = w_t + 0.5(1+c)x_t + 0.5(1-c)z_t + u_t, \tag{4.2.2}$$

$$Case\ II : y_t = w_t + 0.5(1+c)x_t + 0.5(1-c)z_t + 0.5y_{t-1} + u_t, \tag{4.2.3}$$

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 $13-15$ (CASE I) and Figures $16-18$ (CASE II) are the power comparisons from 5,000 iterations for the three experiments as in the MA(2) model power comparison.

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.

5 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,

$$E|e_{it}| = E|y_{t+3} - \hat{y}_{it}|, \text{ for } i = 1, 2, \tag{5.0.1}$$

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 statistics is the same as the DM test given by

$$\tau_T = \frac{\sqrt{T}\bar{d}}{\sqrt{\widehat{V}_T}}, \tag{5.0.2}$$

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. (3.1.9), but we use the fixed-b approximation.

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

Figure 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 to $(h-1)$ (Harvey, Leybourne, and Newbold (1997)). The fixed-b approximation properly addresses the size distortion problem by giving larger critical values for larger bandwidths.

6 Conclusion

For comparing non-nested 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 truncation 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.

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.

References

Atkinson, A. C. (1970): "A Method for Discriminating Between Models," *Journal of the Royal Statistical Society, Series B*, 32, 211–243.

Choi, H.-s., and N. M. Kiefer (2005): "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, pp. 105–123.

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

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

——— (2002): "Bootstrap J tests of nonnested linear regression models," *Journal of Econometrics*, 109, 167–193.

——— (2004): "Model selection based on information criteria," in *Econometric Theory and Methods*, pp. 675–676. Oxford University Press, New York.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–263.

Fan, Y., and Q. Li (1995): "Bootstrapping J-Type Tests for Non-Nested Regression Models," *Economics Letters*, 48, 107–112.

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 M. Pesaran (1983): "Tests of non-nested regression models: small sample adjustments and Monte Carlo evidence," *Journal of Econometrics*, 21, 133–154.

Gonçalves, S., and T. J. Vogelsang (2006): "Block Bootstrap HAC Robust Tests: The Sophistication of the Naive Bootstrap," *Working paper, Université de Montréal and Cornell University*.

Gourieroux, C., and A. Monfort (1999): "Testing Non-Nested Hypotheses," in *Handbook of Econometrics*, vol. 4, pp. 2583–2637. Elsevier Science Pub Co., North-Holland.

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

Hall, P., and J. L. Horowitz (1996): "Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators," *Econometrica*, 64, 891–916.

Hansen, B. E. (2005a): "Challenges for Econometric Model Selection," *Econometric Theory*, 21(1), 60–68.

Hansen, P. R. (2005b): "A Test for Superior Predictive Ability," *Journal of Business & Economic Statistics*, 23(4), 365–380.

Harvey, D., S. Leybourne, and P. Newbold (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(3), 937–946.

Kiefer, N. M., and T. J. Vogelsang (2002a): "Heteroskedasticity-Autocorrelation Robust Standard Errors Using the Bartlett Kernel Without Truncation," *Econometrica*, 70, 2093–2095.

Kiefer, N. M., and T. J. Vogelsang (2002b): "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size," *Econometric Theory*, 18, 1350–1366.

Kiefer, N. M., and T. J. Vogelsang (2005): "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 21, 1130–1164.

KIEFER, N. M., T. J. VOGELSANG, AND H. BUNZEL (2000): "Simple Robust Testing of Regression Hypotheses," *Econometrica*, 68, 695–714.

KULLBACK, S., AND R. LEIBLER (1951): "On information and sufficiency," *Annals of Mathematical Statistics*, 22(1), 79–86.

LIEN, D., AND H. Q. VUONG (1987): "Selecting the best linear regression model: A classical approach," *Journal of Econometrics*, 35, 3–23.

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

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

PHILLIPS, P. C. B., AND S. N. DURLAUF (1986): "Multiple Time Series Regression with Integrated Processes," *The Review of Economic Studies*, 53(4), 473–495.

POLITIS, D. N., AND J. P. ROMANO (1994): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89(428), 1303–1313.

PÖTSCHER, B. M. (1991): "The Effect of Model Selection on Inference," *Econometric Theory*, 7, 163–185.

RIVERS, D., AND H. Q. VUONG (2002): "Model selection tests for nonlinear dynamic models," *Econometrics Journal*, 5, 1–39.

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

WHITE, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68(5), 1097–1126.

A Figures

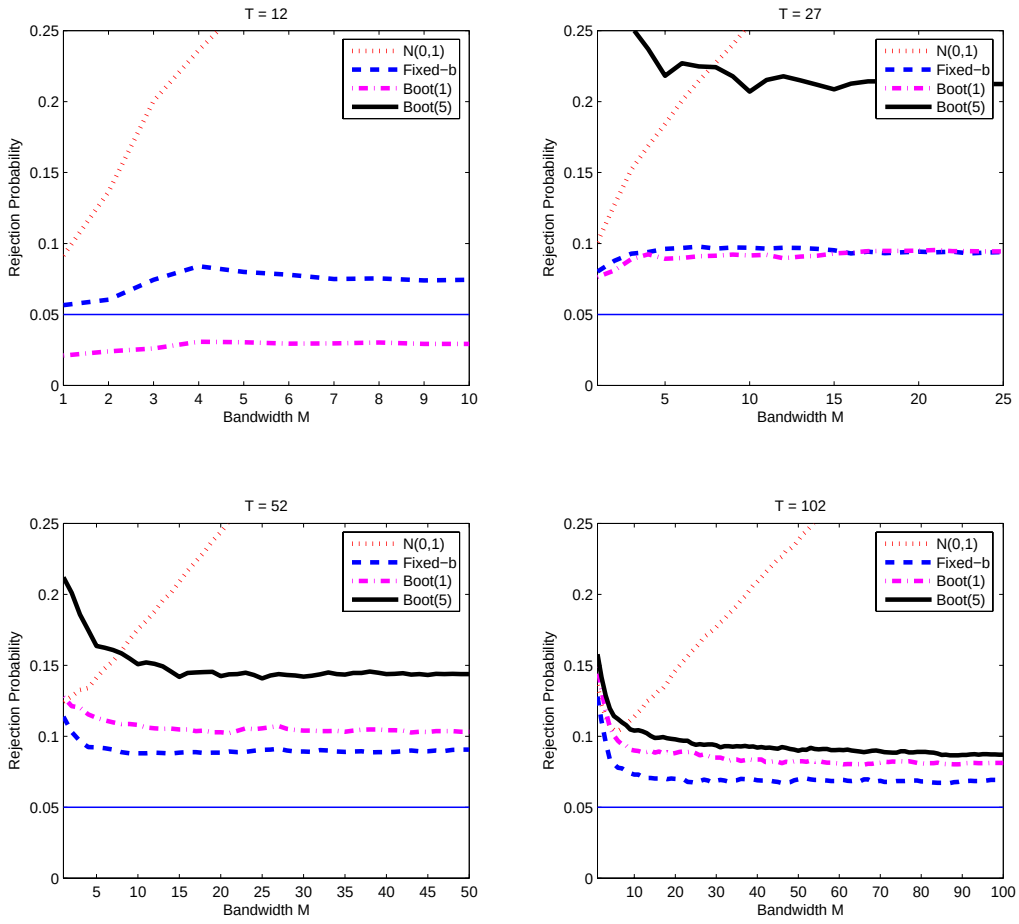

Figure 1: MA(2) DGP with two competing AR(1) models

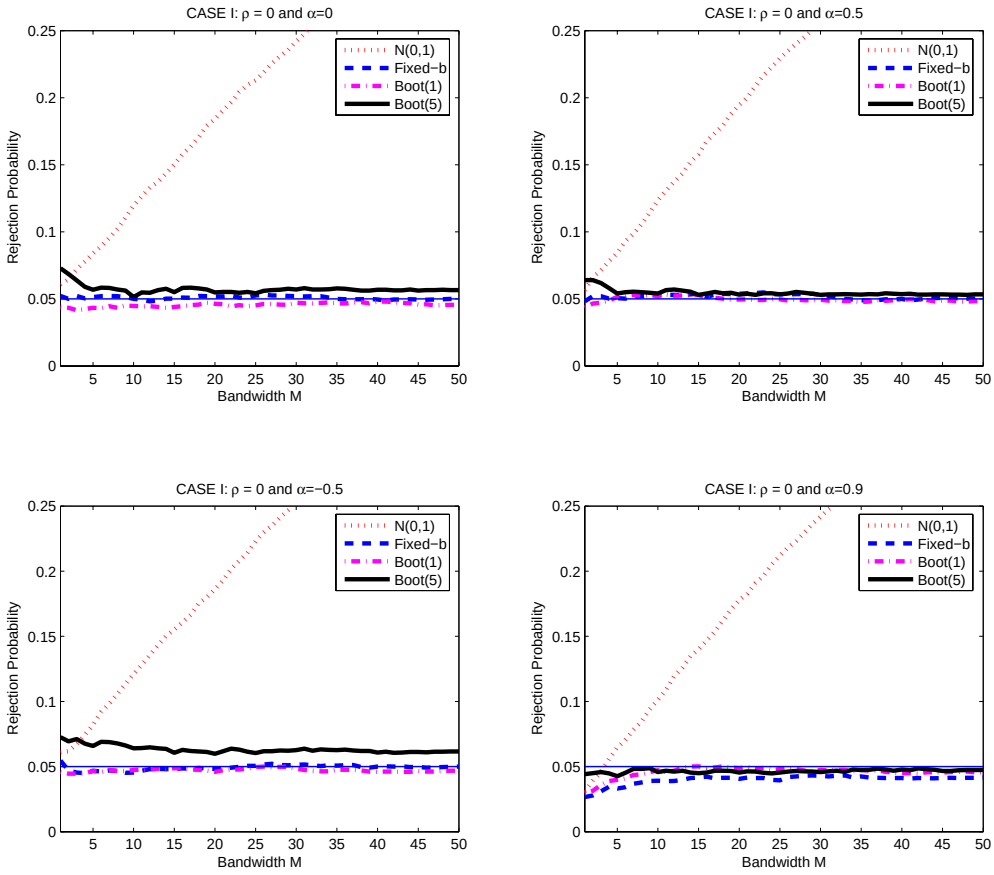

Figure 2: (CASE I) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0$, of the errors is α.

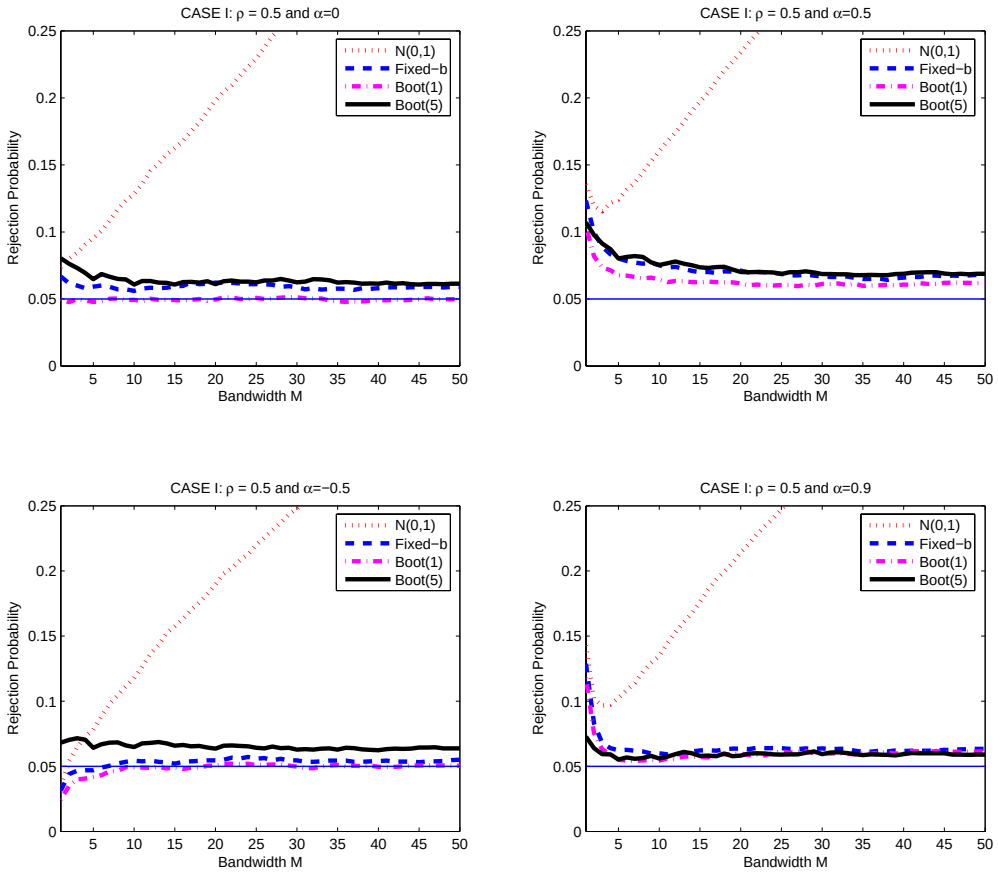

Figure 3: (CASE I) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0.5$, of the errors is α.

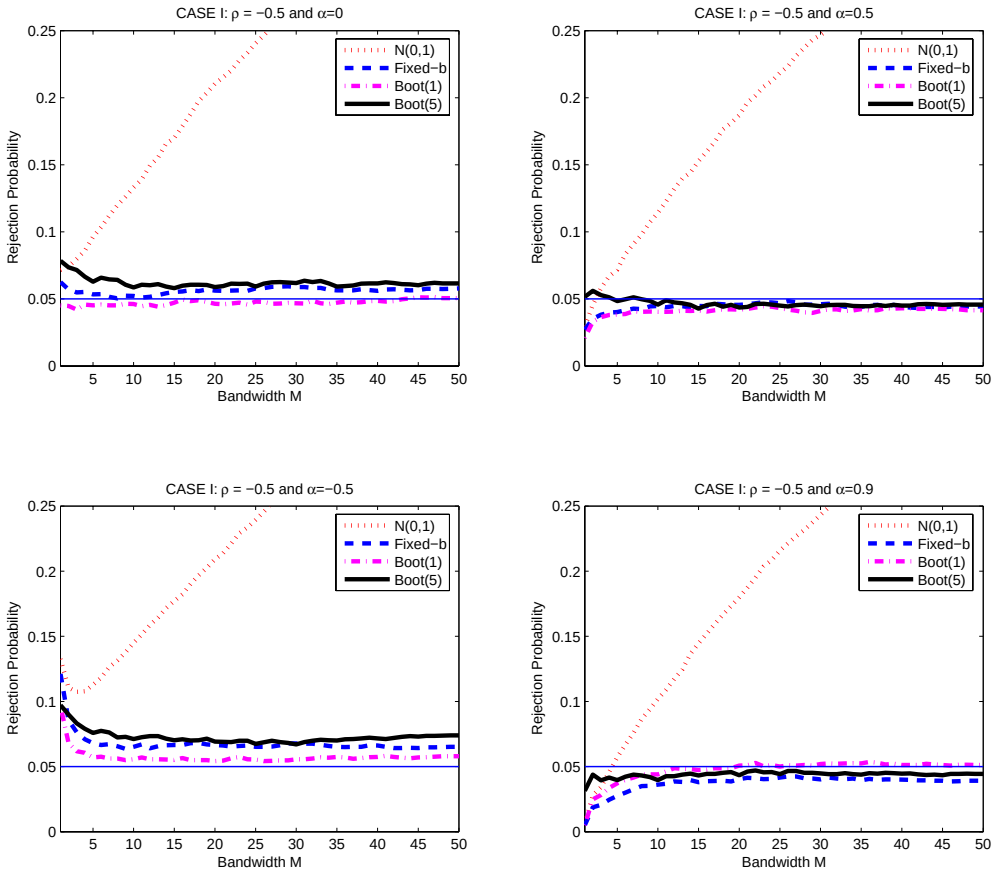

Figure 4: (CASE I) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = -0.5$, of the errors is α.

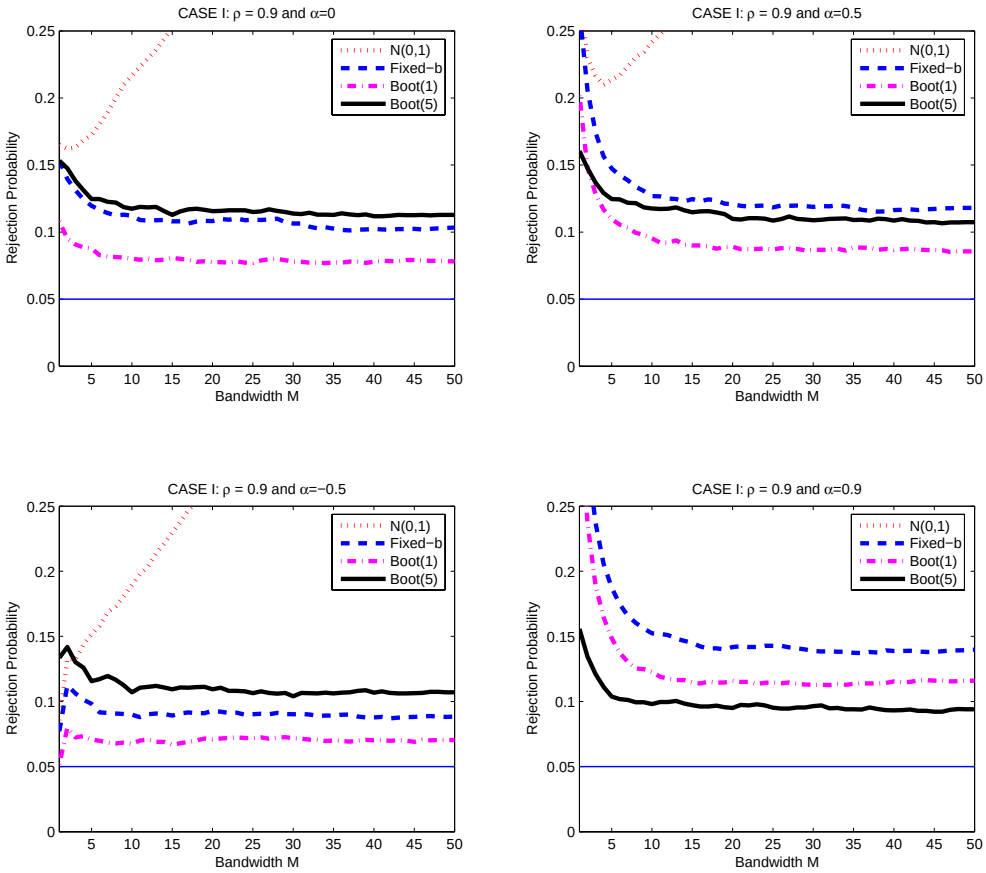

Figure 5: (CASE I) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0.9$, of the errors is α.

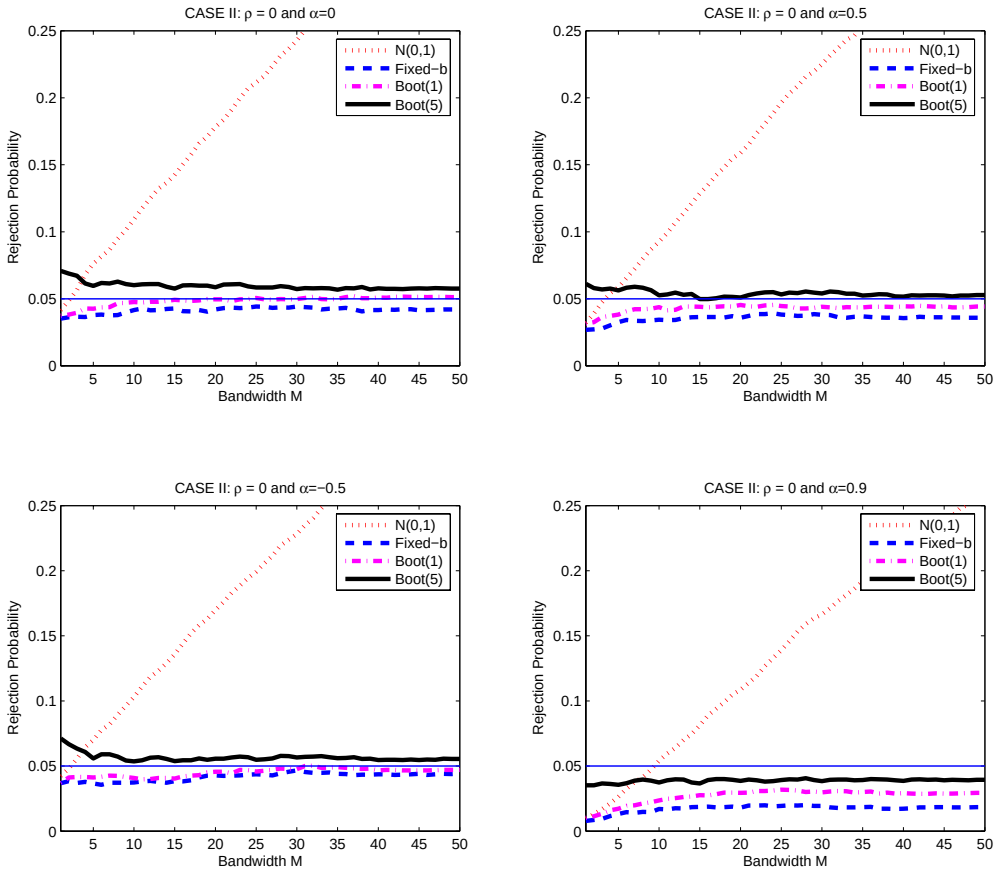

Figure 6: (CASE II) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0$, of the errors is α.

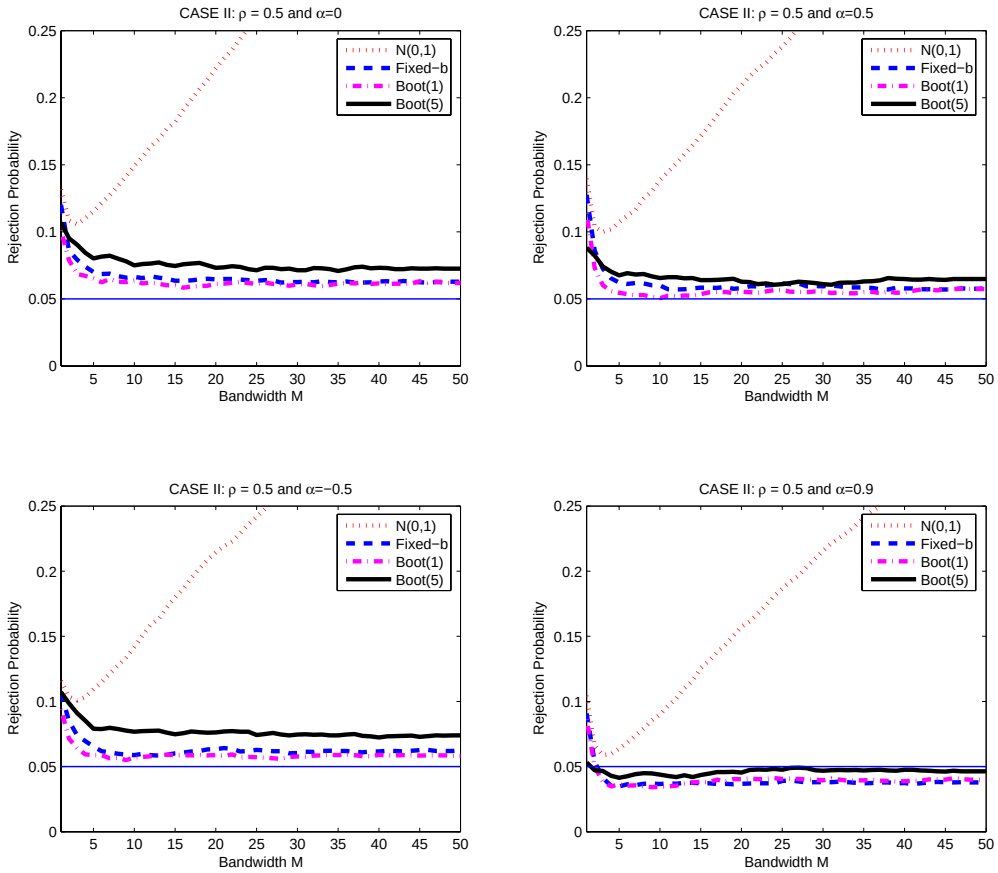

Figure 7: (CASE II) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0.5$, of the errors is α.

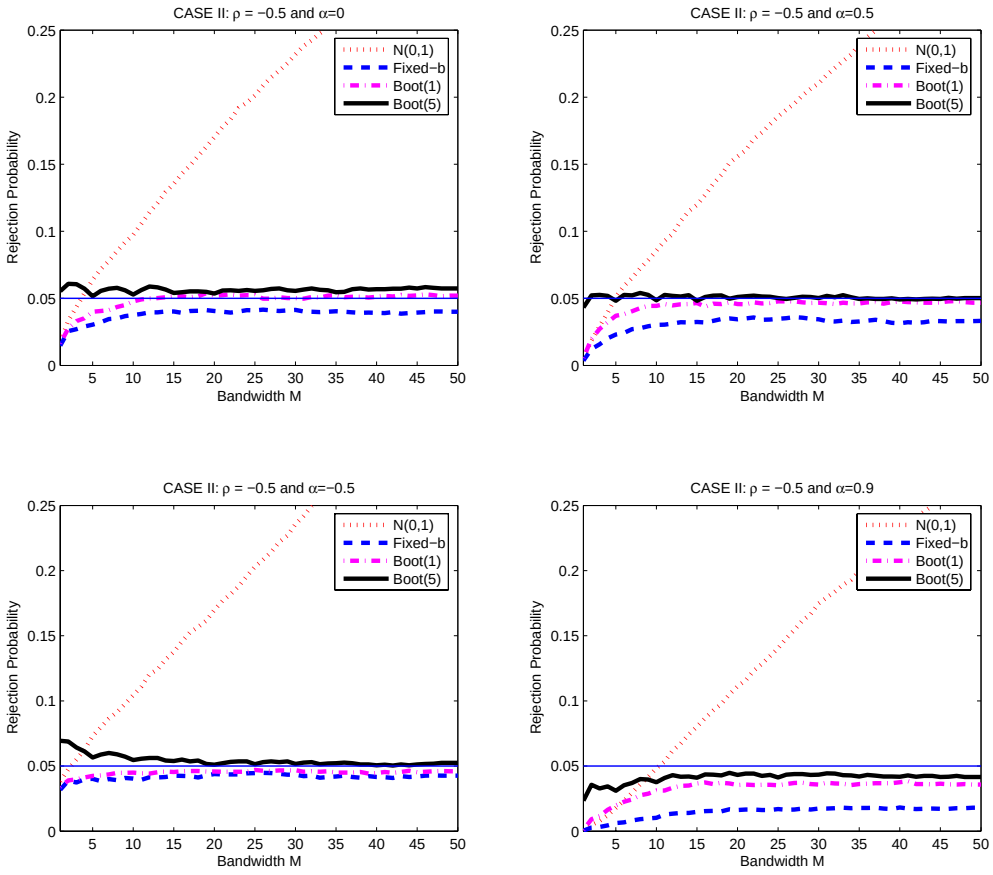

Figure 8: (CASE II) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = -0.5$, of the errors is α.

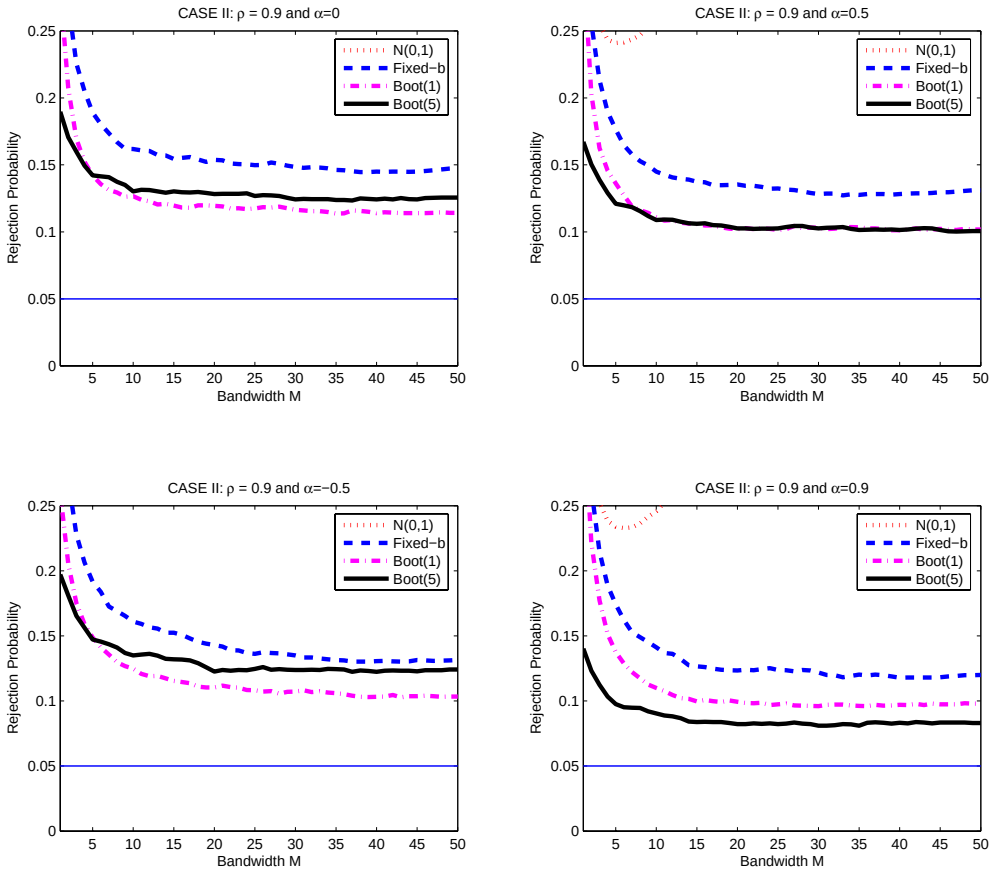

Figure 9: (CASE II) Two Competing Linear Models. AR(1) coefficient of the regressors is $\rho = 0.9$, of the errors is α.

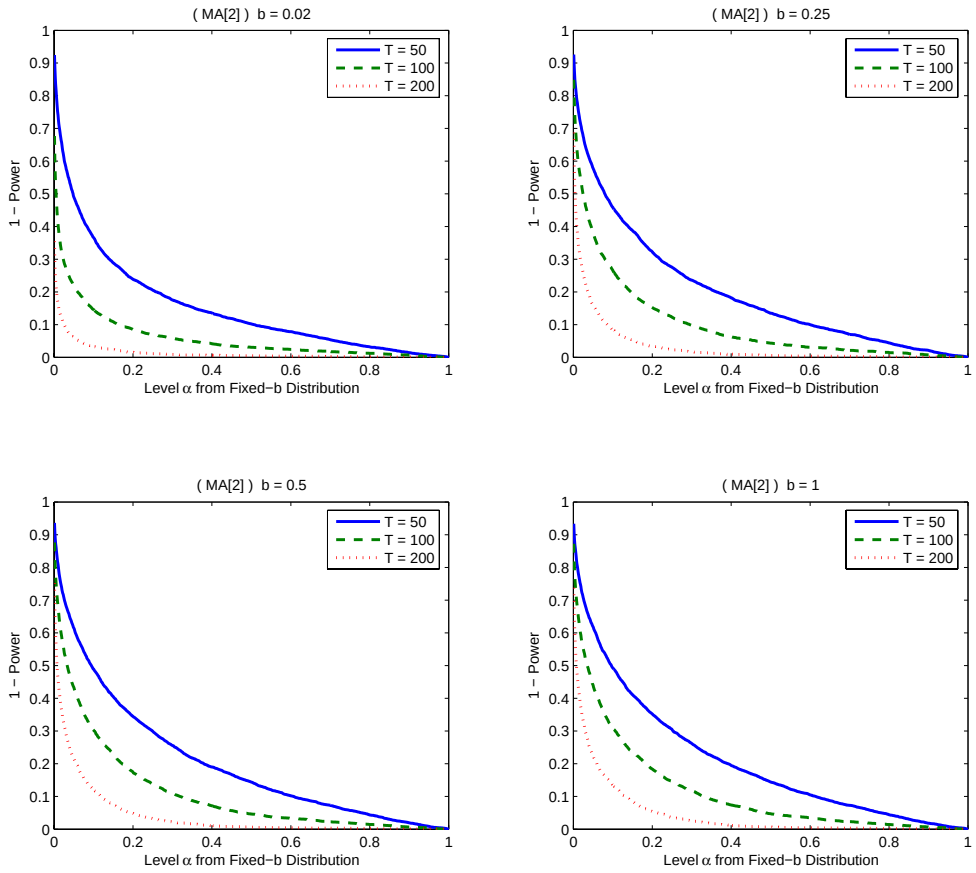

Figure 10: (MA(2)) Type II error $(1 - Power)$ as a function of the level α implied by the fixed-b approximating distribution. $b = M/T$ is the bandwidth and T is the sample size.

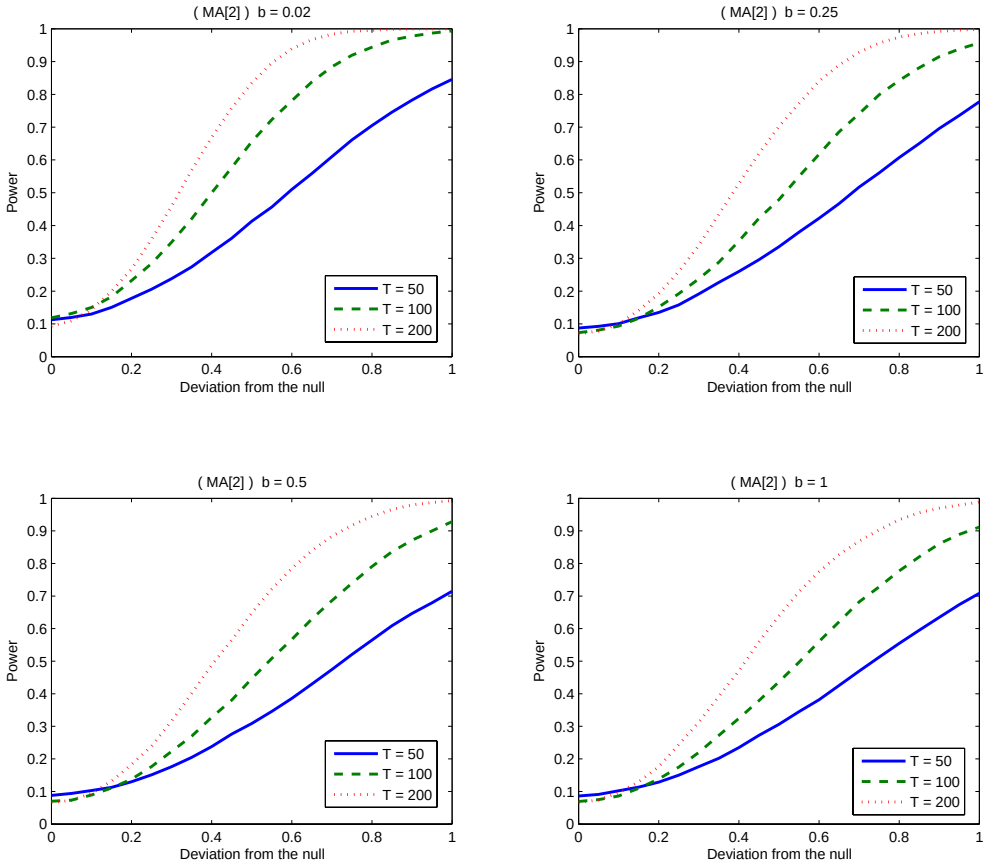

Figure 11: (MA(2)) Local power curves for different sample sizes T. $b = M/T$ is the bandwidth.

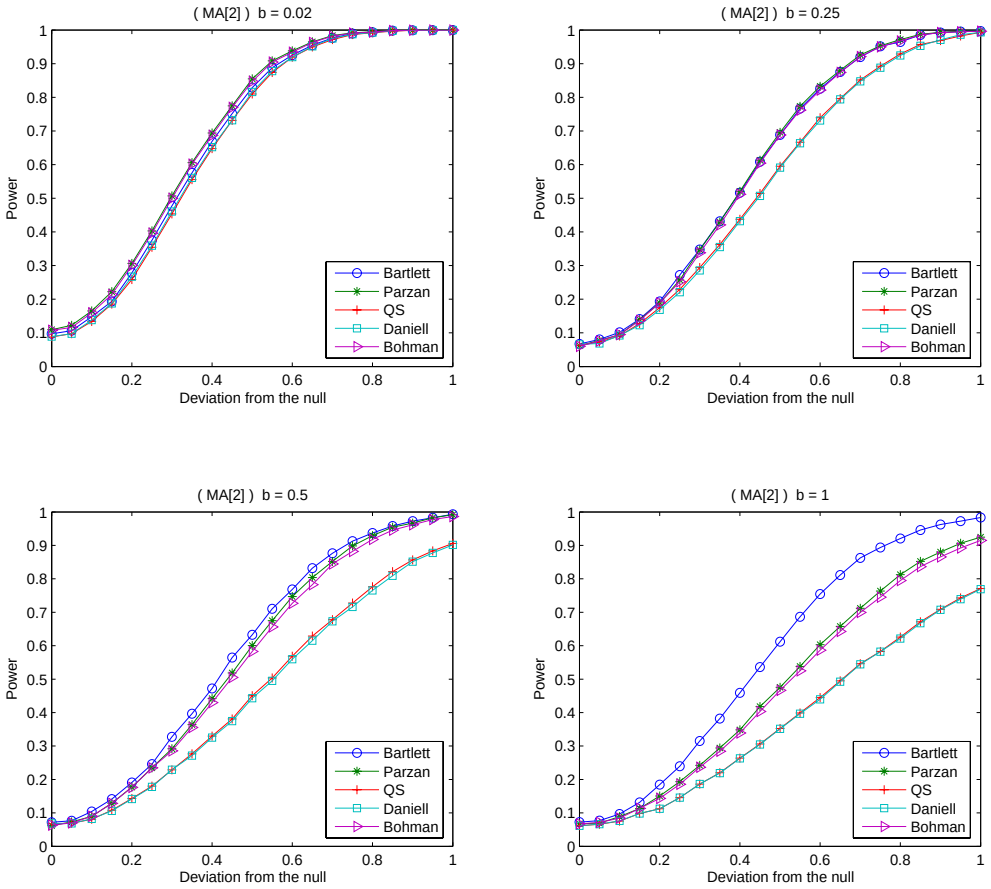

Figure 12: (MA(2)) Local power curves for different kernel functions. $b = M/T$ is the bandwidth.

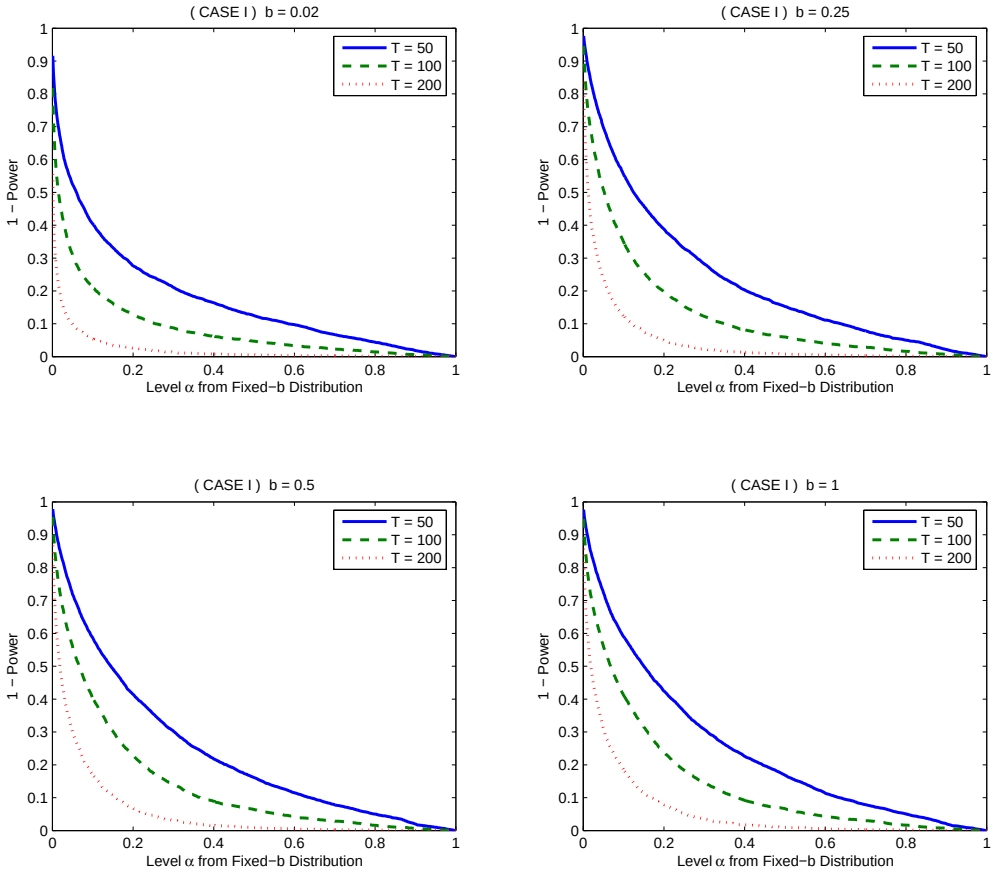

Figure 13: (CASE I) Type II error $(1 - Power)$ as a function of the level α implied by the fixed-b approximating distribution. $b = M/T$ is the bandwidth and T is the sample size.

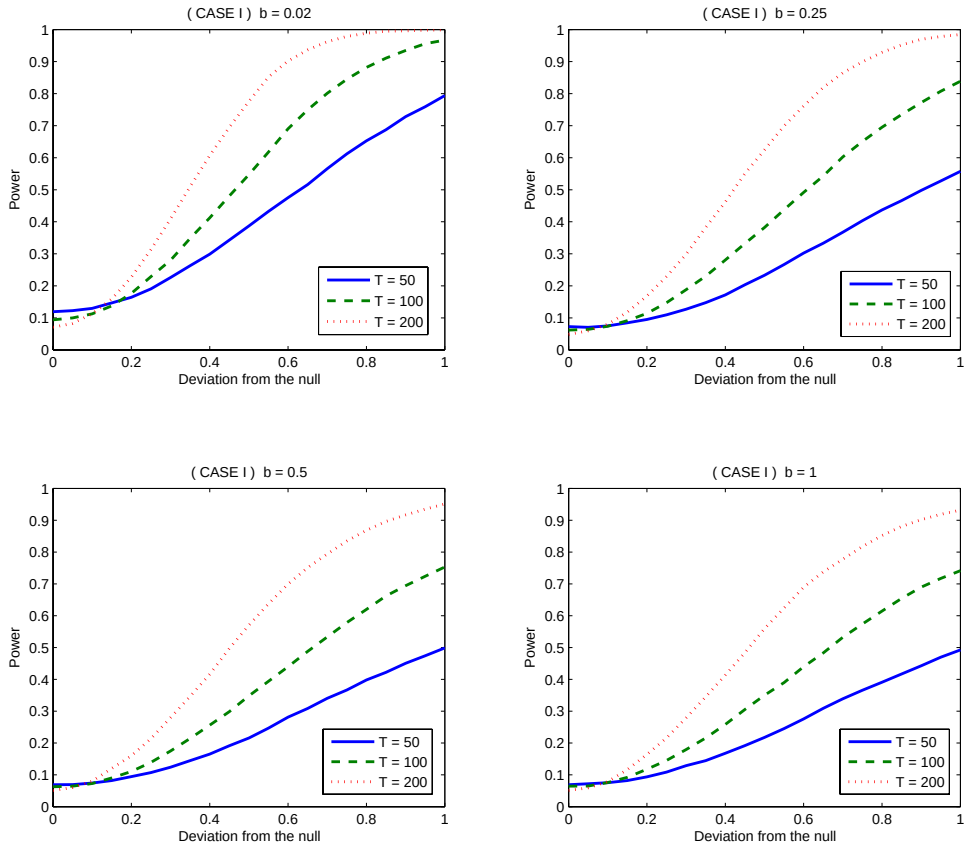

Figure 14: (CASE I) Local power curves for different sample sizes T. $b = M/T$ is the bandwidth.

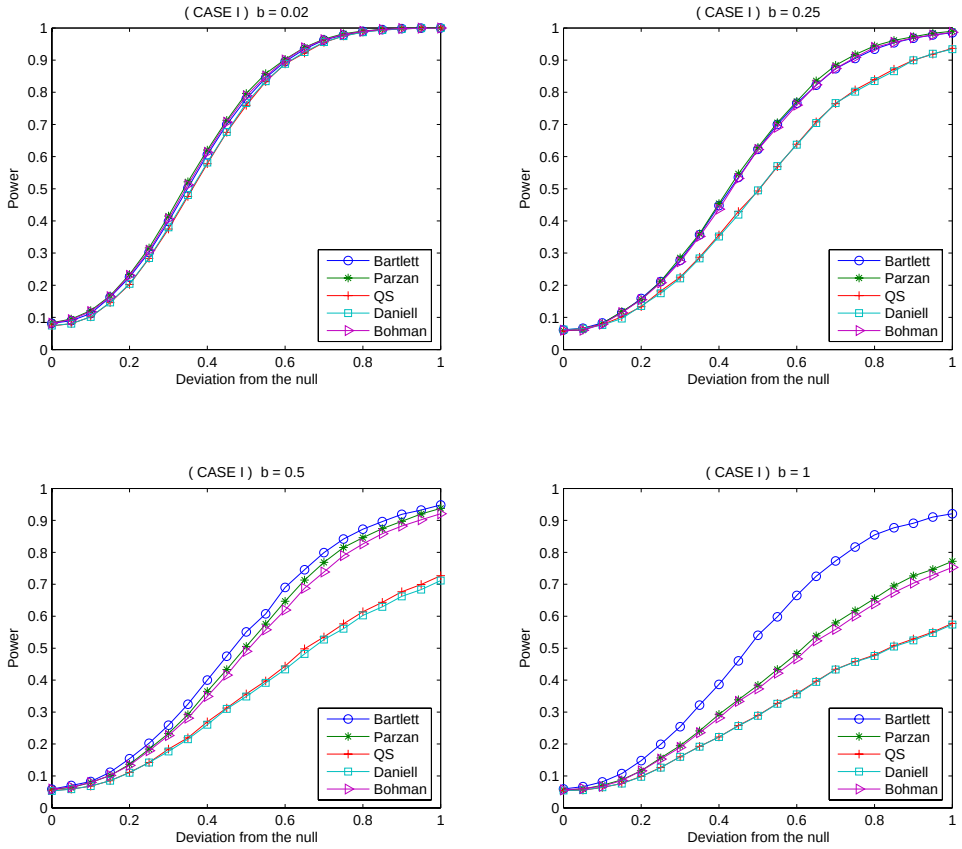

Figure 15: (CASE I) Local power curves for different kernel functions. $b = M/T$ is the bandwidth.

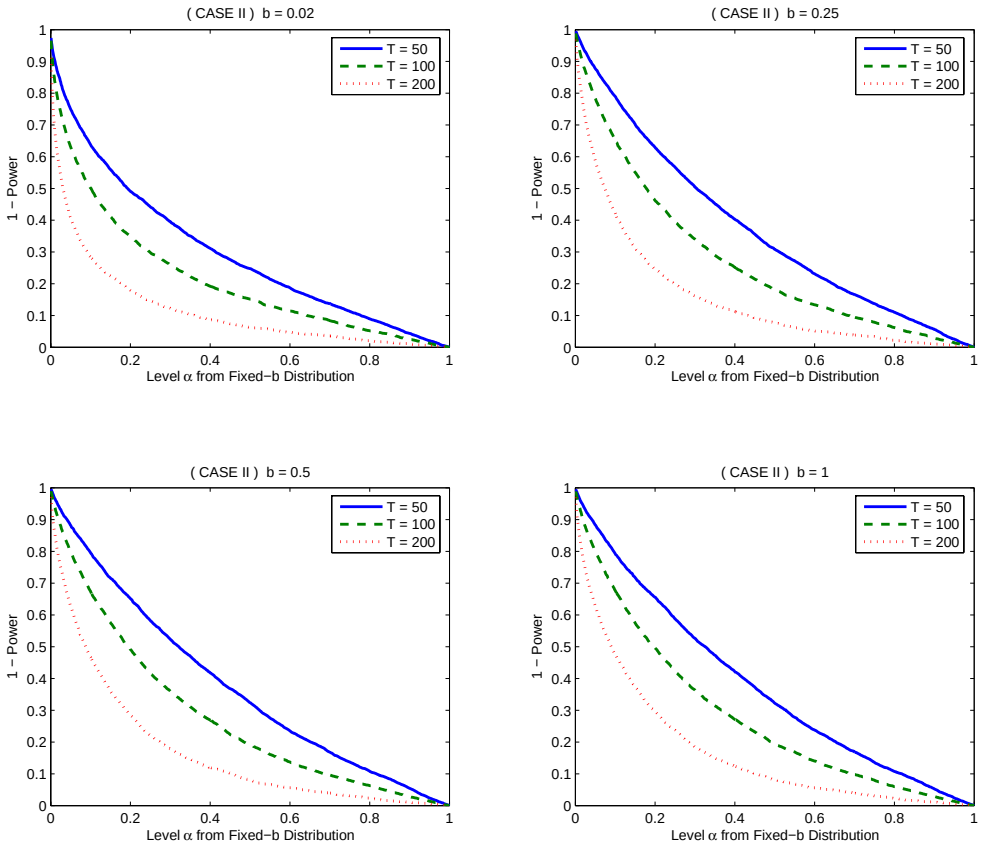

Figure 16: (CASE II) Type II error $(1 - Power)$ as a function of the level α implied by the fixed-b approximating distribution. $b = M/T$ is the bandwidth and T is the sample size.

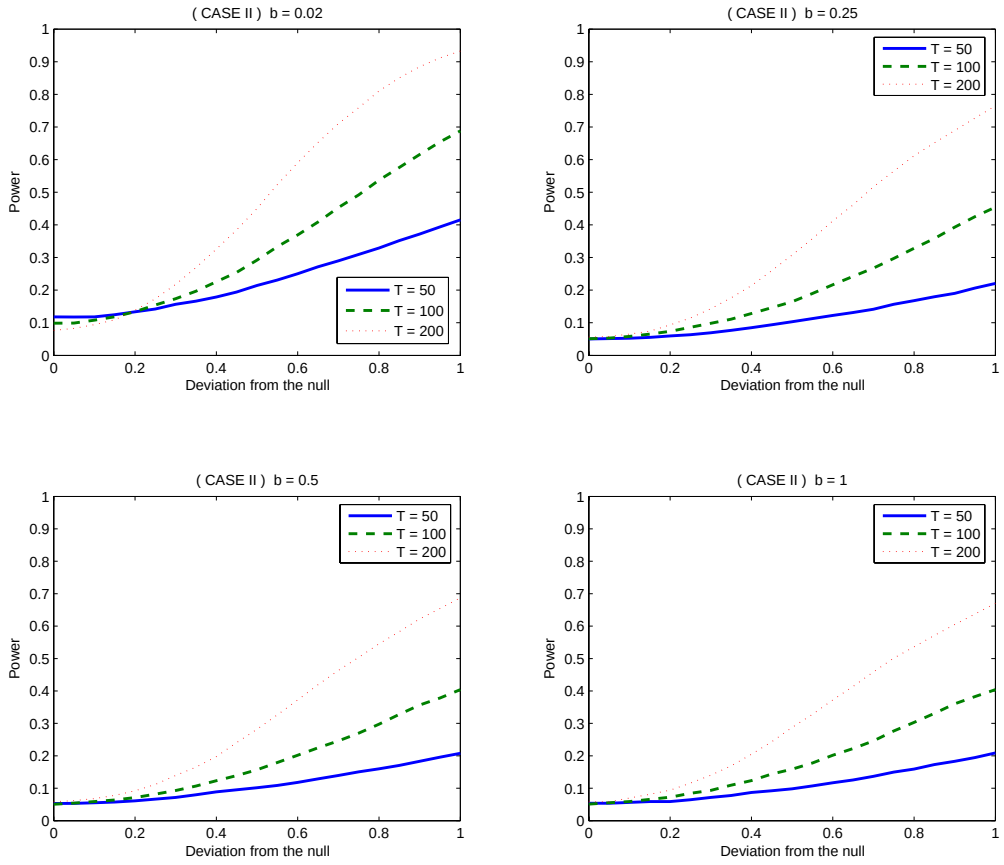

Figure 17: (CASE II) Local power curves for different sample sizes T. $b = M/T$ is the bandwidth.

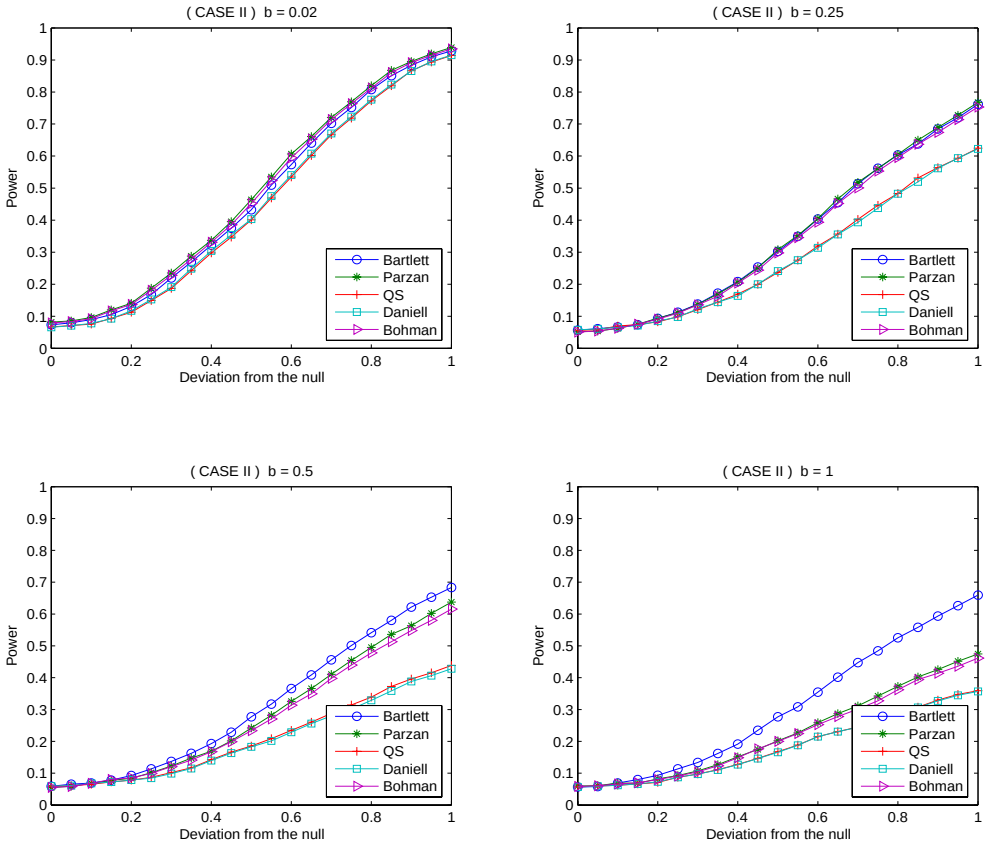

Figure 18: (CASE II) Local power curves for different kernel functions. $b = M/T$ is the bandwidth.

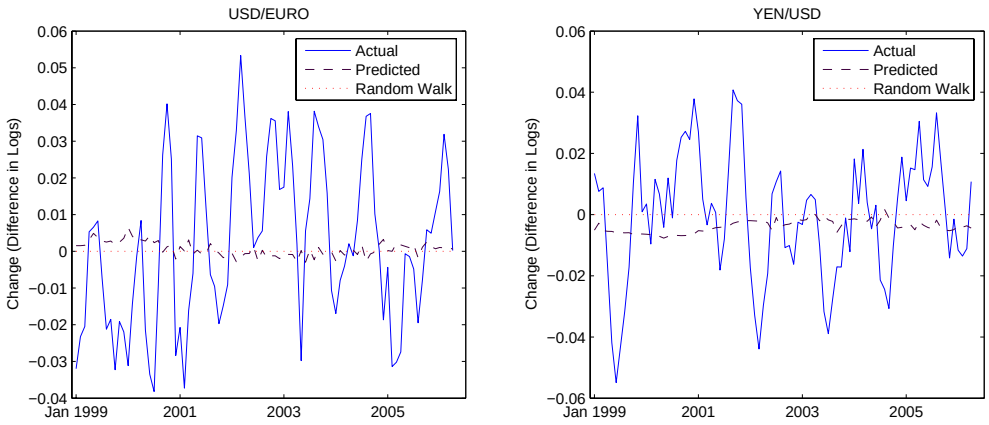

Figure 19: Three months change of exchange rates (monthly data). The solid line is actual changes, the dashed is from the 3-months forward rate model, and the dotted is from the random walk prediction (no change).

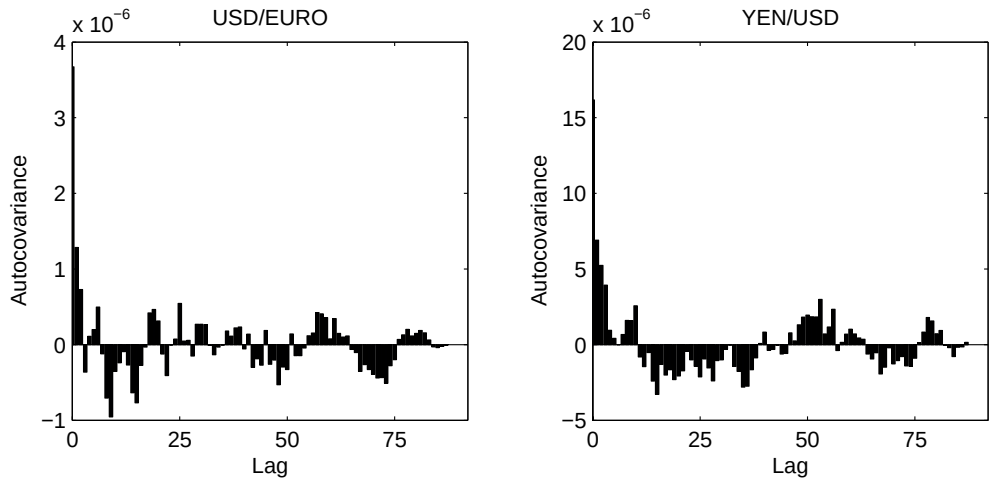

Figure 20: Autocovariance function of the difference in absolute prediction error, $\{|e_{1t}| - |e_{2t}|\}$, where e_{1t} =actual change−forward rate model and e_{2t} =actual change−random walk model.

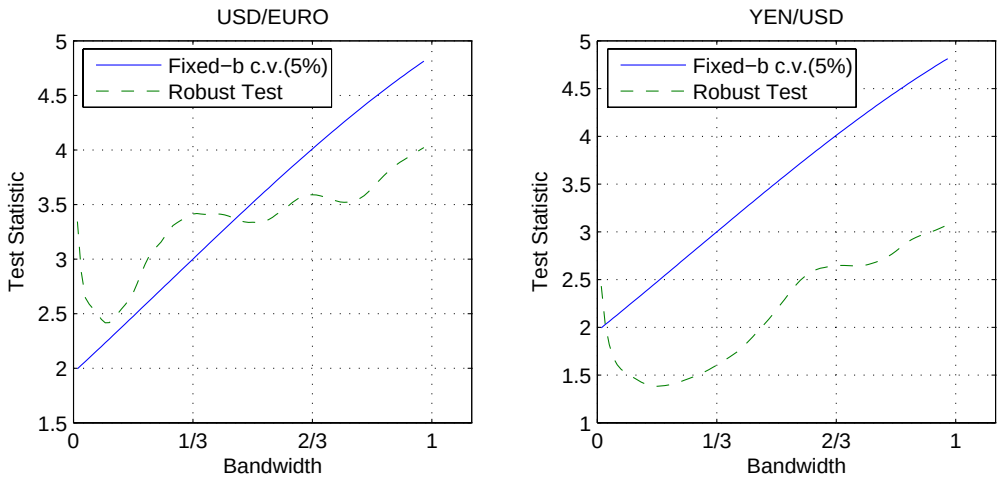

Figure 21: Values of the test statistic with various bandwidth and the Bartlett kernel. The solid line is two sided 5% level critical values from the fixed-b approximation. The fixed-b critical value at zero bandwidth is equal to the critical value from the standard normal approximation.