Bias Reduction and Likelihood Based Almost-Exactly Sized Hypothesis Testing in Predictive Regressions using the Restricted Likelihood

Willa W. Chen * Rohit S. Deo †

August 21, 2008

Abstract: Difficulties with inference in predictive regressions are generally attributed to strong persistence in the predictor series. We show that the major source of the problem is actually the nuisance intercept parameter and propose basing inference on the Restricted Likelihood, which is free of such nuisance location parameters and also possesses small curvature, making it suitable for inference. The bias of the Restricted Maximum Likelihood (REML) estimates is shown to be approximately 50% less than that of the OLS estimates near the unit root, without loss of efficiency. The error in the chi-square approximation to the distribution of the REML based Likelihood Ratio Test ($RLRT$) for no predictability is shown to be $\left(3/4-\rho^2\right) n^{-1}\left(G_3\left(\cdot\right)-G_1\left(\cdot\right)\right)+O\left(n^{-2}\right)$, where $|\rho|<1$ is the correlation of the innovation series and $G_s\left(\cdot\right)$ is the c.d.f. of a χ_s^2 random variable. This very small error, free of the AR parameter, suggests that the $RLRT$ for predictability has very good size properties even when the regressor has strong persistence. The Bartlett corrected $RLRT$ achieves an $O\left(n^{-2}\right)$ error. Power under local alternatives is obtained and extensions to more general univariate regressors and vector AR(1) regressors, where OLS may no longer be asymptotically efficient, are provided. In simulations the $RLRT$ maintains size well, is robust to non-normal errors and has uniformly higher power than the Jansson-Moreira test with gains that can be substantial. The Campbell-Yogo Bonferroni Q test is found to have size distortions and can be significantly oversized.

Keywords: Bartlett correction, likelihood ratio test, curvature

*Department of Statistics, Texas A&M University, College Station, Texas 77843, USA. Chens research was supported by NSF grant DMS-0605132.

†New York University. Part of this research was completed when Deo was at the University of Texas-Austin. We would like to thank four referees and the Editor for comments that led to a significantly improved paper. We would also like to thank Michael Jansson and Marcelo Moreira for providing their code and Yi Wang for his assistance with programming.

1 Introduction

A question of interest in financial econometrics is whether future values of one series $\{Y_t\}$ can be predicted from lagged values of another series $\{X_t\}$. The hypothesis of no predictability is commonly tested under the assumption that the two series $\{Y_t\}_{t=1}^n$ and $\{X_t\}_{t=0}^n$ obey the following model:

$$Y_t = \eta + \beta X_{t-1} + u_t, \tag{1}$$

$$X_t = \mu + \alpha X_{t-1} + v_t, \tag{2}$$

where $|\alpha| < 1$, $u_t = \phi v_t + e_t$, $(e_t, v_t) \sim N\left(\mathbf{0}, \text{diag}\left(\sigma_e^2, \sigma_v^2\right)\right)$ are an i.i.d. series and $X_0 \sim N\left(\mu\left(1-\alpha\right)^{-1}, \sigma_v^2\left(1-\alpha^2\right)^{-1}\right)$. Interest generally centers on the case where the regressor series $\{X_t\}$ possesses a strong degree of autocorrelation with the autoregressive parameter α lying close to the unit root. It is well known (Stambaugh, 1999) that the standard ordinary least squares (OLS) estimate of β is biased when the errors $\{u_t, v_t\}$ are contemporaneously correlated, with the amount of bias increasing as α gets closer to unity. This bias results in the corresponding t-statistic being biased with poor size properties. However, Stambaugh (1999) provided a simple estimable expression for the bias in the OLS estimate of β that allows the researcher to compute a bias-corrected OLS estimate as well as a t-statistic based on it. See, for example, Amihud and Hurvich (2004). There is currently, however, no known theoretical justification that such a t-statistic based on the bias-corrected OLS estimate will have improved size properties relative to the test based on the uncorrected OLS estimate. Indeed, Sprott and Viveros-Aguilera (1984) and Sprott (1990) point out that if inference is the goal of the researcher, computing pivotal $t-$statistics from bias corrected point estimates is not necessarily guaranteed to improve finite sample performance. Sprott and Viveros-Aguilera (Section 6, 1984) provide an example where even the use of the exact bias correction does not result in accurate finite sample inference since the resulting t-statistic can still be very far from normal. (One could perhaps use conservative bias corrections, such as in Lewellen (2004), which can yield tests that are under-sized but at the cost of significant power loss. See page 232 of Lewellen, 2004). Instead, following Fisher (1973), Sprott argued in a series of fundamental papers (1973, 1975, 1980, 1990) that from an inferential point of view, issues such as the bias of point estimates may be irrelevant in small

samples, and he stressed the importance of examining the likelihood.

At first glance, it may seem surprising that a likelihood ratio test may provide well-behaved hypothesis tests in situations when the t-statistic does not since the two test statistics are closely related. More specifically, under standard regularity conditions we can expand the Likelihood Ratio Test, $LRT(\hat{\theta}, \theta)$, for a parameter θ in terms of its t-statistic, $t(\hat{\theta}, \theta)$, as

$$LRT(\hat{\theta}, \theta) = t^2(\hat{\theta}, \theta) + R_n(\hat{\theta}),$$

where the remainder term $R_n(\hat{\theta})$ converges to zero in probability. This expansion suggests that t-statistics with poor finite sample size properties would correspond to LRT′s that are also poorly behaved. However, this intuition is misleading since it ignores a crucial property of a likelihood that is not shared by the t-statistic, viz. invariance under 1-1 parameter transformations $g(\theta)$. As we argue next based on some of the key ideas in Sprott (1975, 1980), this property can prove invaluable for the LRT. Since the LRT for θ is identical to the LRT for $g(\theta)$ due to invariance, we can as well expand the LRT in terms of the t-statistic for $g(\theta)$, yielding

$$LRT(\hat{\theta}, \theta) = LRT(g(\hat{\theta}), g(\theta)) = t^2(g(\hat{\theta}), g(\theta)) + R_n(g(\hat{\theta})). \tag{3}$$

Sprott argued that

(i) there may exist some transformation $g(\theta)$ such that the remainder term $R_n(g(\hat{\theta}))$ in the above expansion is close to zero, thus making the likelihood ratio approximately quadratic in the t-statistic based on that transformed parameter $g(\theta)$.

(ii) the conditions that allow for this quadratic approximation to be adequate also improve the normal approximation to the distribution of the t-statistic $t(g(\hat{\theta}), g(\theta))$ in the transformed parameter.

As a result, it follows from these two observations and the expansion in (3) that the likelihood ratio will be well approximated by a chi-square variable in finite samples as long as some parametrisation satisfying (i) exists, *even if one does not know what that parametrisation is.*

Sprott (1973, 1975, 1980) showed that such a parametrisation would exist if the "curvature" of the log-likelihood[1], as measured by a function of its higher order derivatives, was small. The use of such a likelihood would then result in a well behaved likelihood ratio test (LRT) in finite samples. See also Efron (1975) and McCullagh and Cox (1986) for a geometrical approach to curvature and likelihood ratio based hypothesis testing.

The approach we take in this paper is guided by the intuition given above. We find that in a univariate autoregressive (AR) process of order one, the likelihood has very small curvature and hence yields tests with good finite sample behaviour when there is no intercept in the model. However, the inclusion of an intercept in the model causes the likelihood ratio to lose this property, thus pointing to the intercept as the source of the problem. This motivates the use of the restricted likelihood, which is free of the nuisance intercept parameter and hence able to imitate the likelihood of the no-intercept univariate model with its attendant small curvature. This suggests that the restricted likelihood will also be useful in the related predictive regression problem, which is a bivariate $AR(1)$ with intercept. Indeed, we are able to obtain theoretical results that demonstrate that the LRT based on the restricted likelihood ($RLRT$) has good finite sample performance for both estimation and inference in this context. A curvature related approach to tackling the predictive regression problem for the model in (1) and (2) was also taken by Jansson and Moreira (2006, henceforth JM). We compare our results in more detail with those in JM, first on a theoretical basis in Section 3 and then through simulations in Section 5.

The layout of the paper is as follows. In Section 2, we provide motivation for considering the restricted likelihood by studying the related problem of inference in the univariate $AR(1)$ model. We then obtain the restricted likelihood for the bivariate predictive regression model and provide results on the bias of the REML estimates. In Section 3**,** we state our result on the finite sample behaviour of the $RLRT$ for β and compare its power under a sequence of Pitman alternatives to that of the restricted likelihood based Wald and Rao score test. A comparison with the results in JM is also provided. Section 4 provides results on extensions of the REML method to higher

[1]There are several formal measures of curvature based on higher order derivatives of the likelihood in the literature. See, for example, Kass and Slate (1994). We will not define any such measure explicitly since that is not the focus of our work here.

order AR processes for the regressor series, as well as multivariate AR(1) regressors. The finite sample performance of the REML estimates and the $RLRT$ is studied through simulations in Section 5 and is compared to the performance of the procedures developed by Jansson and Moreira (2006) and Campbell and Yogo (2006). All proofs are relegated to the Appendix at the end of the paper.

2 Restricted Maximum Likelihood Estimation

To understand why the restricted likelihood may yield well behaved LRT's in the predictive regression context, it is instructive to consider LRT's for α in the univariate AR(1) model given in (2). If $LRT_{\alpha,\mu}$ denotes the LRT for testing $H_0 : \alpha = \alpha_0$ versus $H_1 : \alpha \neq \alpha_0$ in that model for $|\alpha_0| < 1$, then Theorem 2 of van Giersbergen (2006) in conjunction[2] with the results of Hayakawa (1977, 1987), Cordeiro (1987), Barndorff-Nielsen and Hall (1988) and Chesher and Smith (1995) yields the following formal expansion[3]

$$P\left(LRT_{\alpha,\mu} \leq x\right) - G_1\left(x\right) = 0.25\left(1 + 7\alpha_0\right)\left(1 - \alpha_0\right)^{-1} n^{-1}\left(G_3\left(x\right) - G_1\left(x\right)\right) + O\left(n^{-2}\right), \quad (4)$$

where $G_s\left(x\right)$ is the c.d.f. of a χ^2_s random variable. On the other hand, if LRT_α denotes the LRT for testing $H_0 : \alpha = \alpha_0$ versus $H_1 : \alpha \neq \alpha_0$ in the univariate AR(1) model given in (2) *with* μ *known to be 0,* then it follows from Theorem 1 of van Giersbergen (2006) that we get the formal expansion

$$P\left(LRT_{\alpha} \leq x\right) - G_1\left(x\right) = -0.25n^{-1}\left(G_3\left(x\right) - G_1\left(x\right)\right) + O\left(n^{-2}\right), \quad (5)$$

It is obvious from (4) that $LRT_{\alpha,\mu}$ is very unstable when the autoregressive parameter α is close to the unit root with the leading error term in the expansion going to infinity. In stark contrast, we see from (5) that LRT_α is very well behaved with the leading term in the expansion of its distribution being both very small and free of α. Figure 1 shows the empirical densities

[2] van Giersbergen (2006) provides the expected value of the LRT, while the combined results from the remaining references show that the leading term in the expansion (4) is half that expected value.

[3] All the expansions that we provide in this paper are formal. Though we do not attempt to do so here, it may be possible to show that these expansions are valid by using the work of Chandra and Ghosh (1979).

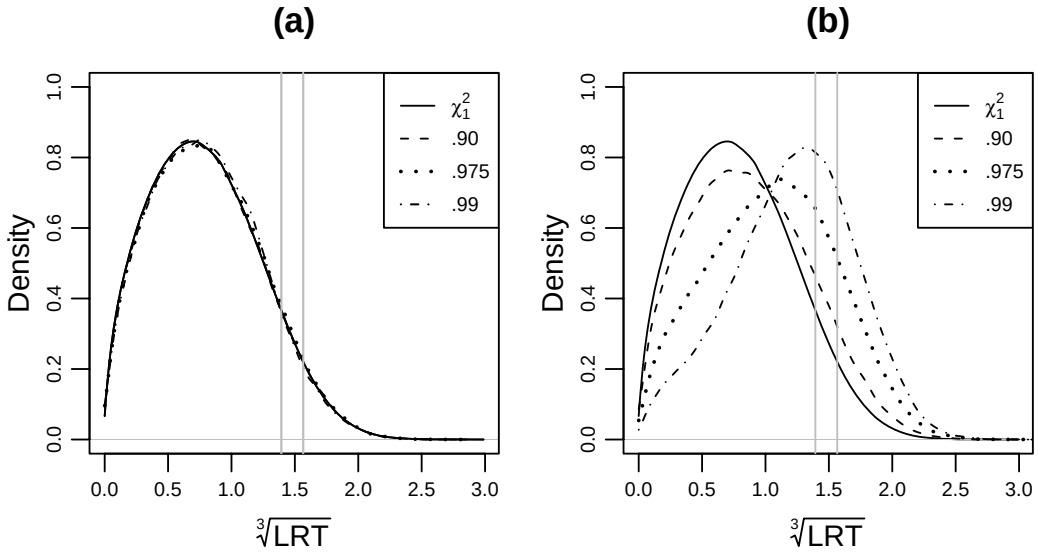

Figure 1: Empirical densities of cubic root transformed LRT statistics of AR(1) processes without intercept, plot (a) and with intercept, plot (b). The vertical lines are 90th and 95th percentiles. Both plots are based on 100,000 repetitions of an AR(1) with sample size $n = 100$ and AR coefficient $\alpha = .9, .975$.and 99.

of LRT_α and $LRT_{\alpha,\mu}$ based on samples of size $n = 100$ for various values of $\alpha = 0.9, 0.975$ and 0.99 plotted together with the limiting χ^2_1 density. (Since a χ^2_1 density is very right skewed, we plot the density of the cube root of the LRT to ensure a density that looks more symmetric in order to make the comparisons in the right tail clearer). The empirical densities of LRT_α (the zero intercept case) are seen to be remarkably well-approximated by the limiting χ^2_1 distribution, both when α is far from the unit root as well as when α is close to unity, while those of $LRT_{\alpha,\mu}$ (the intercept case) are far moved from that of the χ^2_1. Simulation results in Figure 1 of van Giersbergen (2006) further confirm the accuracy of the standard χ^2_1 approximation for LRT_α, even when α is close to unity. Not surprisingly, van Garderen (1999) has found that for the univariate AR(1) model in (2) *with μ known to be 0,* the Efron (1975) curvature (one of the standard measures of curvature of the likelihood) is very small[4], being of the order $O\left(n^{-2}\right)$ and

[4]Interestingly and somewhat surprisingly, van Garderen (1999) found that if the innovation variance σ_v^2 is known the Efron curvature of the model is $2n^{-1} + O\left(n^{-2}\right)$, which, though still small, is larger than when σ_v^2 is unknown and he provided a geometrical explanation for this phenomenon. Correspondingly, results in van Giersbergen (2006) imply that the coefficient of the leading term in (5) increases from $-0.25n^{-1}$ to $-1n^{-1}$ when σ_v^2 is known, indicating that the LRT is better approximated by the limiting chi-square distribution when the innovation variance is unknown than when it is known, though, of course, both approximations are very good by themselves.

converging to zero as $\alpha \to 1$ for every fixed n.

These results indicate that the culprit in the finite sample failure of the LRT in the univariate AR(1) model is the unknown intercept μ. Since the bivariate prediction model in (1) and (2) is a vector AR(1) (with the first column of the coefficient matrix restricted to zero) with an intercept, one is led to suspect from the discussion above that the LRT for β will also be poorly behaved in finite samples due to the nuisance intercept vector. This is indeed the case, as seen in the simulation study presented in Table I. Even when the $AR(1)$ coefficient is 0.9, and thus far from the unit root, the usual LRT for β has inflated size and the problem is exarcebated significantly as either the correlation between the innovations (u_t, v_t) or the AR coefficient increases. Thus, there is no advantage in using the usual LRT instead of the t-statistic in this case. However, the discussion above leads us to believe that the LRT for β may perhaps be well behaved if the intercept vector (η, μ) were known. Since the assumption that (η, μ) is known is extremely unrealistic, we are prompted to seek a likelihood that does not involve the location parameters and yet possesses small curvature properties similar to those of the model with known location parameters. The restricted likelihood turns out to be the one that has such properties and we turn next to defining it and stating some of its properties.

The idea of restricted likelihood was originally proposed by Kalbfleisch and Sprott (1970) precisely as a means of eliminating the effect of nuisance location (more generally, regression coefficient) parameters when estimating the parameters of the model covariance structure in a linear model. More specifically, assume that we observe data on the vector $\mathbf{Z}$ which follows the linear model

$$\mathbf{Z} = \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\mathbf{W}$ and $\boldsymbol{\theta}$ are the design matrix and coefficient parameter vector respectively and the error vector $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\psi}))$ with $\boldsymbol{\psi}$ being the parameters that describe the variance covariance matrix. Suppose that interest centers on the parameters $\boldsymbol{\psi}$ of the error covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\psi})$ and that the regression coefficients $\boldsymbol{\theta}$ are nuisance parameters. Kalbflesich and Sprott (1970) defined the restricted likelihood to be the exact likelihood of the linearly transformed data $\mathbf{TZ}$, where $\mathbf{T}$ is any matrix of full row-rank equal to $n - rank(\mathbf{W})$ such that $\mathbf{TW} = \mathbf{0}$. Thus,

the likelihood of the transformation $\mathbf{TZ}$ does not depend on the nuisance regression coefficient parameters $\boldsymbol{\theta}$. The particular choice of the matrix $\mathbf{T}$ is irrelevant since the likelihood of $\mathbf{TZ}$ will change only by a multiplicative constant for different choices of $\mathbf{T}$ (Harville, 1974) and hence will have no effect on either estimation or testing of hypothesis. Harville (1974) showed that the restricted likelihood for the process (6), up to a multiplicative constant, is given by

$$RL\left(\mathbf{Z},\boldsymbol{\psi}\right)=\left|\mathbf{W}'\mathbf{W}\right|^{1/2}\left|\boldsymbol{\Sigma}\left(\boldsymbol{\psi}\right)\right|^{-1/2}\left|\mathbf{W}'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\psi}\right)\mathbf{W}\right|^{-1/2}\exp\left(-\frac{1}{2}\widetilde{\mathbf{Z}}\left(\boldsymbol{\psi}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\psi}\right)\widetilde{\mathbf{Z}}\left(\boldsymbol{\psi}\right)\right), \quad (7)$$

where $\widetilde{\mathbf{Z}}\left(\boldsymbol{\psi}\right)=\mathbf{Z}-\mathbf{W}\left(\mathbf{W}'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\psi}\right)\mathbf{W}\right)^{-1}\mathbf{W}'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\psi}\right)\mathbf{Z}$. Harville (1977) showed that REML estimates of the parameters of the covariance structure do not suffer any loss of efficiency due to the linear transformation of the data. Harville (1974) also provided a Bayesian interpretation of the restricted likelihood, while Smyth and Verbyla (1996) showed that the restricted likelihood is also the exact conditional likelihood of the original data given the complete sufficient statistic for the regression coefficient parameters. Though the restricted likelihood has been studied primarily in the context of variance component models, there has also been some work on it in the context of time series models. See, for example, Tunnicliffe Wilson (1989) and Rahman and King (1997), among others. Francke and de Vos (2006) studied unit root tests for $AR(1)$ models based on the restricted likelihood, while Chen and Deo (2006, 2007) showed that confidence intervals for the sum of the autoregressive coefficients of univariate $AR(p)$ processes based on the restricted likelihood have good coverage properties, even when the series is close to a unit root process. The restricted maximum likelihood (REML) estimates also are less biased than regular ML estimates in nearly integrated univariate AR models with intercept (Cheang and Reinsel, 2000) and with trend (Kang, Shin and Lee, 2003).

The restricted likelihood can also be interpreted as the exact likelihood of the maximal invariant (see Section 6.2 of Lehmann and Romano, 2005) which is invariant under transformations of the form $\mathbf{Z}\rightarrow\mathbf{Z}+\mathbf{W}\varkappa$, where $\varkappa$ is some vector. There are references in the econometrics literature that use the Restricted Likelihood in some version of the form given in (7) while explicitly stating that it is also the likelihood of the maximal invariant (See, for example Rahman and King, 1997 and Francke and de Vos, 2006). The paper by JM also uses the Restricted Likelihood for the model in (1) and (2), though they refer to it only as the likelihood of the maximal

invariant. Since JM make the simplifying assumption that $\mu = 0$ in (2), they are able to write the Restricted Likelihood of the process, which becomes the exact likelihood of $(y_t - y_{t-1}, x_t)$ without having to exploit the form in (7). There is another strand of the literature in econometrics (eg., Dufour and King, 1991) that considers tests that are invariant under transformations of the form $\mathbf{Z} \rightarrow \mathbf{Z} + \mathbf{W}\varkappa$, but this literature focuses on point optimal tests and hence does not derive the likelihood of the maximal invariant of the form in (7). We now exploit the expression in (7) to obtain the Restricted Likelihood for an $AR(1)$ with intercept, which helps develop some intuition for why the RL can be of use in the bivariate predictive regression model.

When $\mathbf{X} = (X_0, ..., X_n)'$ follows the univariate AR(1) model in (2), we can express the vector $\mathbf{X}$ in the form (6), where now the design matrix $\mathbf{W}$ is a vector of ones, the regression coefficient parameter $\boldsymbol{\theta}$ is given by $\mu(1-\alpha)$ and the error vector $\boldsymbol{\varepsilon}$ is a vector following a zero-mean $AR(1)$ process. As a result, the restricted likelihood for this model is merely the exact likelihood of the first differences $\{X_t - X_{t-1}\}_{t=1}^n$ and from the expression (7) above, the restricted log-likelihood of $\mathbf{X}$ is given by

$$L_R\left(\sigma_v^2, \alpha, \mathbf{X}\right) = -\left(\frac{n}{2}\right)\log\sigma_v^2 + \frac{1}{2}\log\left(\frac{1+\alpha}{(n-1)(1-\alpha)+2}\right) - \frac{1}{2\sigma_v^2}Q(\alpha), \tag{8}$$

where

$$Q(\alpha) = \mathbf{X}'\boldsymbol{\Sigma}_X^{-1}\mathbf{X} - \frac{\left(\mathbf{X}'\boldsymbol{\Sigma}_X^{-1}\mathbf{1}\right)^2}{\mathbf{1}'\boldsymbol{\Sigma}_X^{-1}\mathbf{1}} \tag{9}$$

and $\boldsymbol{\Sigma}_X \equiv Var(\mathbf{X})$. On the other hand, the regular likelihood of $\mathbf{X}$ for model (2) with known μ (set, w.l.o.g., to zero) is

$$L\left(\sigma_v^2, \alpha, \mathbf{X}\right) = -\left(\frac{n+1}{2}\right)\log\sigma_v^2 + \frac{1}{2}\log\left(1-\alpha^2\right) - \frac{1}{2\sigma_v^2}\mathbf{X}'\boldsymbol{\Sigma}_X^{-1}\mathbf{X}. \tag{10}$$

On comparing (8) and (10) and noting that the second term in $Q(\alpha)$ is $O(1)$ whereas $\mathbf{X}'\boldsymbol{\Sigma}_X^{-1}\mathbf{X}$ is $O(n)$, it immediately becomes apparent that the restricted likelihood in this case differs on a relative scale by only an order $O\left(n^{-1}\right)$ from the likelihood of the AR(1) process with known intercept μ . As a matter of fact, Cheang and Reinsel (2000) show that the restricted likelihood for the AR(1) provides REML estimates of α whose bias is $-2\alpha n^{-1} + O\left(n^{-2}\right)$, which is identical, up to order $O\left(n^{-1}\right)$, to the bias of the maximum likelihood estimate when the intercept is known

(Marriott and Pope, 1954). Since the bias of the maximum likelihood estimate of α when the intercept is not known is $-(1+3\alpha)\,n^{-1}+O\left(n^{-2}\right)$, the REML estimate is able to achieve a bias reduction of approximately 50% when the autoregressive parameter is close to the unit root. The ability of the Restricted Likelihood to imitate the regular likelihood of a zero-mean process goes even further. A little simple algebra shows that the higher order derivatives ($w.r.t$ the parameters α, σ_v^2) of the restricted likelihood in (8) are identical, up to $O(n^{-1})$, to those of the zero-mean regular likelihood in (10). As a result, we can use the calculations in van Garderen (1999) obtained for the zero-mean $AR(1)$ model and establish that the Efron curvature properties of the restricted likelihood for the AR(1) are the same as those of the AR(1) model with known intercept, up to order $O\left(n^{-1}\right)$. In addition, we also get the following Theorem which provides a formal expansion for the distribution of the $RLRT$ in the model in (2) by arguments identical to those for Theorem 1 of van Giersbergen (2006), established for the zero-mean $AR(1)$ model (See also footnote 2) .

Theorem 1 *Let* $\mathbf{X} = (X_0, ..., X_n)'$ *follow the univariate AR(1) model in (2) and let* $RLRT_\alpha$ *denote the restricted likelihood ratio test based on the expression in (8) for testing* $H_0 : \alpha = \alpha_0$ *vs.* $H_1 : \alpha \neq \alpha_0$, $|\alpha_0| < 1$. *Then,*

$$P\left(RLRT_\alpha \leq x\right) - G_1\left(x\right) = -0.25n^{-1}\left(G_3\left(x\right) - G_1\left(x\right)\right) + O\left(n^{-2}\right).$$

Remark 1 *It is worth noting that the result of Theorem 1 continues to hold if the initial value* X_0 *in the* $AR(1)$ *model is assumed to follow* $N\left(\mu, \sigma_v^2\right)$ *instead of the stationary distribution. The reason for this is that the leading terms in the derivatives of the Restricted Likelihood are unaffected by the specification of the initial value.*

Comparing the result in Theorem 1 with the expressions in (4) and (5), we see that the distribution of the $RLRT$ for α in the model with intercept behaves like that of the regular LRT for the zero-mean model and should be well approximated by the χ_1^2 even when α is close to the unit root. This can also be seen in the empirical densities of the $RLRT$ for α in an intercept model shown in Figure 2, which are plotted together with the limiting χ_1^2 density.

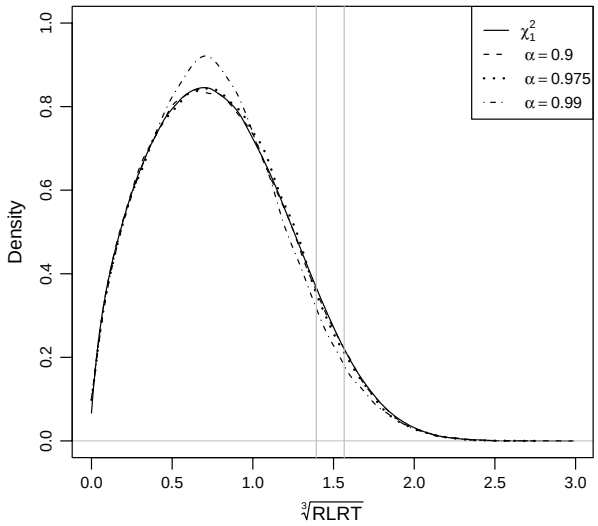

Figure 2: Empirical densities of cubic root transformed Restricted LRT statistics of AR(1) processes with intercept. The vertical lines are 90th and 95th percentiles. The plot is based on 100,000 repetitions of an AR(1) with sample size $n = 100$ and AR coefficient $\alpha = .9, .975$, and .99.

(As in Figure 1, we plot the density of the cube root of the $RLRT$). Furthermore, Table II shows that the simulation rejection percentages of the $RLRT$ based on χ^2_1 critical values are close to the nominal levels. The discussion above shows that the restricted likelihood provides a great advantage for both hypothesis testing and estimation for the univariate AR(1) model through the elimination of the nuisance intercept parameter. The general problem of $RLRT$ based inference in univariate $AR(p)$ models with intercept/trend is studied in Chen and Deo (2007). In this paper, we focus our attention on the use of the restricted likelihood for carrying out inference on β in the bivariate predictive regression model in (1) and (2). We first obtain a tractable expression for the restricted likelihood for this model.

We start by noting that the vector $(\mathbf{Y}', \mathbf{X}')' = (Y_1, \ldots, Y_n, X_0, \ldots, X_n)'$ defined by (1) and (2) can be expressed in the form (6), where the design matrix $\mathbf{W}$ is now of the form

$$\mathbf{W} = \begin{bmatrix} \mathbf{1}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n+1} \end{bmatrix}, \tag{11}$$

where $\mathbf{1}$ is a vector of ones and the regression coefficient vector is of the form $\boldsymbol{\theta} = (\eta + \beta\mu/(1-\alpha), \mu/(1-\alpha))'$.

However, the covariance matrix of the resulting error vector ε in this representation takes on an awkward form. As a result, though in principle we could obtain the restricted likelihood of $(\mathbf{Y}', \mathbf{X}')'$ by using the form (7) given above, the resulting expression is not simple to study. Hence, we derive the restricted likelihood by appealing to its basic definition and exploiting the structure of the model. We start by noting that since the design matrix $\mathbf{W}$ now has the form (11), the restricted likelihood of $(\mathbf{Y}', \mathbf{X}')'$ is the exact likelihood of $(\mathbf{Y}'\mathbf{T}_1', \mathbf{X}'\mathbf{T}_2')'$, where $\mathbf{T}_1$ and $\mathbf{T}_2$ are full row rank matrices of dimension $(n-1) \times n$ and $n \times (n+1)$ respectively, satisfying $\mathbf{T}_1\mathbf{1} = \mathbf{0}$ and $\mathbf{T}_2\mathbf{1} = \mathbf{0}$. (In other words, the RL in this context is the exact likelihood of $\{(Y_t - Y_{t-1})_{t=2}^n\,,\ (X_t - X_{t-1})_{t=1}^n\}$). We next note that the exact likelihood of $(\mathbf{Y}'\mathbf{T}_1', \mathbf{X}'\mathbf{T}_2')'$ can be factorised as

$$L\left(\left(\mathbf{Y}'\mathbf{T}_1', \mathbf{X}'\mathbf{T}_2'\right)'\right) = L\left(\mathbf{T}_1\mathbf{Y} \mid \mathbf{T}_2\mathbf{X}\right) L\left(\mathbf{T}_2\mathbf{X}\right).$$

By definition, the likelihood $L\left(\mathbf{T}_2\mathbf{X}\right)$ is just the restricted likelihood of $\mathbf{X}$ given in (8) above. In the Appendix, we obtain a simple expression for $L\left(\mathbf{T}_1\mathbf{Y} \mid \mathbf{T}_2\mathbf{X}\right)$ by using the model structure, and in conjunction with (8) thus obtain a simple expression for the restricted likelihood of the predictive regression model. This simple expression for the restricted likelihood that we state in Theorem 2 below allows both very easy calculation of the $REML$ estimates as well as a useful expression for their finite sample bias. We first define some quantities that will be useful to us in stating our results.

For the observed data $(\mathbf{Y}', \mathbf{X}')' = (Y_1, \ldots, Y_n, X_0, \ldots, X_n)'$ define $\mathbf{X}_1 = (X_1, \ldots, X_n)'$ and $\mathbf{X}_0 = (X_0, X_1, ..., X_{n-1})'$. Define the sample means $\bar{Y} = n^{-1}\mathbf{1}'\mathbf{Y}$, $\bar{X} = (n+1)^{-1}\mathbf{1}'\mathbf{X}$, $\bar{X}_1 = n^{-1}\mathbf{1}'\mathbf{X}_1$ and $\bar{X}_0 = n^{-1}\mathbf{1}'\mathbf{X}_0$ and the sample mean corrected data $\mathbf{Y}_c = \mathbf{Y} - \mathbf{1}\bar{Y}$, $\mathbf{X}_c = \left[\mathbf{X}_1 - \mathbf{1}\bar{X}_1,\ \mathbf{X}_0 - \mathbf{1}\bar{X}_0\right]$. Define

$$S\left(\phi, \beta, \alpha\right) = \left(\mathbf{Y}_c - \phi\mathbf{X}_{c,1} - \left(\beta - \phi\alpha\right)\mathbf{X}_{c,2}\right)'\left(\mathbf{Y}_c - \phi\mathbf{X}_{c,1} - \left(\beta - \phi\alpha\right)\mathbf{X}_{c,2}\right) \tag{12}$$

and note that for computational purposes $Q(\alpha)$ given in (9) can be written as

$$Q(\alpha) = \sum_{t=0}^{n} \left(X_t - \bar{X}\right)^2 + \alpha^2 \sum_{t=1}^{n-1} \left(X_t - \bar{X}\right)^2 - 2\alpha \sum_{t=0}^{n-1} \left(X_t - \bar{X}\right)\left(X_{t+1} - \bar{X}\right)$$
$$- \frac{(1-\alpha)\,\alpha^2}{(n-1)(1-\alpha)+2}\left[X_0 + X_n - 2\bar{X}\right]^2.$$

We now can state the following theorem.

Theorem 2 *For the model given by (1) and (2), the REML log-likelihood up to an additive constant is given by*

$$L\left(\beta, \alpha, \phi, \sigma_v^2, \sigma_e^2\right) = -\left(\frac{n-1}{2}\right)\log\sigma_e^2 - \frac{1}{2\sigma_e^2}S(\phi, \beta, \alpha) \tag{13}$$
$$- \left(\frac{n}{2}\right)\log\sigma_v^2 + \frac{1}{2}\log\left(\frac{1+\alpha}{(n-1)(1-\alpha)+2}\right) - \frac{1}{2\sigma_v^2}Q(\alpha).$$

The REML estimates $\hat{\boldsymbol{\psi}} = \left(\hat{\beta}, \hat{\alpha}, \hat{\phi}, \hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ *are given by*

$$\hat{\alpha} = \arg\min_{\alpha} \left\{ n\log Q(\alpha) - \log\left(\frac{1+\alpha}{(n-1)(1-\alpha)+2}\right)\right\},$$

$$\left(\hat{\phi}, \hat{\beta}\right)' = \begin{bmatrix} 1 & 0 \\ \hat{\alpha} & 1 \end{bmatrix} \left(\mathbf{X}_c'\mathbf{X}_c\right)^{-1}\mathbf{X}_c'\mathbf{Y}_c,$$

$$\hat{\sigma}_e^2 = \frac{S\left(\hat{\phi}, \hat{\beta}, \hat{\alpha}\right)}{n-1}$$

and

$$\hat{\sigma}_v^2 = \frac{Q(\hat{\alpha})}{n}.$$

The bias in $\left(\hat{\alpha}, \hat{\beta}\right)$ *is given by*

$$E(\hat{\alpha} - \alpha) = -\frac{2\alpha}{n-1} + o\left(n^{-1}\right)$$

and

$$E\left(\hat{\beta}-\beta\right)=\phi E\left(\hat{\alpha}-\alpha\right)$$
$$=-\phi\left(\frac{2\alpha}{n-1}\right)+o\left(n^{-1}\right),$$

and

$$E\left(\hat{\phi}-\phi\right)=0.$$

Remark 2 *It is obvious that obtaining the REML estimates is computationally easy since all the estimates are obtained in succession after the optimisation of a one-dimensional function which is almost quadratic.*

Remark 3 *Note that the restricted likelihood in (13) is well defined at the unit root* $\alpha=1$*, without having to assume that the initial value* X_0 *is fixed when* $|\alpha|<1$*.*

It is interesting to compare the bias in $\hat{\beta}$ with the bias in the OLS estimate $\hat{\beta}_{OLS}$, given by (see Stambaugh 1999)

$$E\left(\hat{\beta}_{OLS}-\beta\right)=\phi E\left(\hat{\alpha}_{OLS}-\alpha\right)=-\phi\left(\frac{1+3\alpha}{n-1}\right)+o\left(n^{-1}\right),$$

where $\hat{\alpha}_{OLS}$ is the OLS estimate of α. Thus, the bias in $\hat{\beta}$ depends upon the bias in $\hat{\alpha}$ in a manner identical to the way the bias in $\hat{\beta}_{OLS}$ depends on the bias in $\hat{\alpha}_{OLS}$. Consequently, the approximately 50% reduction in bias that $\hat{\alpha}$ achieves compared to $\hat{\alpha}_{OLS}$ close to the unit root is inherited by $\hat{\beta}$, relative to $\hat{\beta}_{OLS}$. The bias expression in Theorem 2 also suggests a bias corrected version of the REML estimate of β that may be computed as

$$\hat{\beta}_c=\hat{\beta}+\hat{\phi}\left(\frac{2\hat{\alpha}_c}{n-1}\right),$$

where

$$\hat{\alpha}_c=\hat{\alpha}\left(\frac{n+1}{n-1}\right)$$

is the bias corrected REML estimate of α. Since the bias correction term $\phi 2\alpha (n-1)^{-1}$ is smaller than $\phi (1+3\alpha)(n-1)^{-1}$, one would expect the bias corrected REML estimate $\hat{\beta}_c$ to have both less bias and a smaller variance than its bias corrected OLS counterpart,

$$\hat{\beta}_{OLS,c} = \hat{\beta}_{OLS} + \hat{\phi}\left(\frac{1+3\hat{\alpha}_c}{n-1}\right),$$

particularly since the parameter ϕ can take any value in $(-\infty, \infty)$ [5]. This is indeed the case, as we see in the simulations reported in Section 5 below.

The restricted likelihood given in (13) is derived for the situation where we assume that the initial value X_0 comes from the stationary distribution $N\left(\mu(1-\alpha)^{-1},\ \sigma_v^2\left(1-\alpha^2\right)^{-1}\right)$. A similar argument can be used to obtain the restricted likelihood for the model in (1) and (2) where the regressor series follows an asymptotically stationary process, given by $X_t = \mu + \widetilde{X}_t$, where $\widetilde{X}_t = \alpha\widetilde{X}_{t-1} + v_t$ for $t \geq 1$ and $\widetilde{X}_0 = v_0$. Under this assumption, the restricted likelihood is given by

$$\begin{aligned}\widetilde{L}\left(\beta, \alpha, \phi, \sigma_v^2, \sigma_e^2\right) = &-\left(\frac{n-1}{2}\right)\log\sigma_e^2 - \frac{1}{2\sigma_e^2}S(\phi, \beta, \alpha) \\ &-\left(\frac{n}{2}\right)\log\sigma_v^2 + \frac{1}{2}\log\left(\frac{1}{n(1-\alpha)^2+1}\right) - \frac{1}{2\sigma_v^2}\widetilde{Q}(\alpha),\end{aligned} \tag{14}$$

where

$$\widetilde{Q}(\alpha) = (X_0 - \hat{\mu}(\alpha))^2 + \sum_{t=1}^{n}(X_t - \hat{\mu}(\alpha) - \alpha(X_{t-1} - \hat{\mu}(\alpha)))^2$$

and

$$\hat{\mu}(\alpha) = \frac{X_0 + (1-\alpha)\sum_{t=1}^{n}(X_t - \alpha X_{t-1})}{1 + n(1-\alpha)^2}.$$

The bias results of Theorem 2 continue to hold for the REML estimates under this initial value condition.

In the next section, we provide a theorem that shows that the REML based LRT has very good finite sample properties, in that its finite sample distribution approaches the limiting one

[5] Based on monthly data with a sample size of $n = 379$, Amihud and Hurvich (Table 3, 2004) find the empirical estimate of ϕ to be approximately -92. The corresponding estimate of β is approximately 2 and that of α is 0.99. In their other empirical study based on annual data with $n = 45$, they find that $\phi \simeq -95$, $\beta \simeq 20$ and $\alpha \simeq 0.9$

very quickly and is practically unaffected by nuisance parameters, while also maintaining power against local alternatives when compared to the Wald and score test.

3 REML Likelihood Ratio Test

One standard method for testing the composite hypothesis $H_0 : \beta = 0$ vs. $H_a : \beta \neq 0$ is the likelihood ratio test (LRT) which compares the log-likelihood evaluated at the unrestricted estimates of the parameters to the log-likelihood evaluated at the parameter estimates obtained under the restriction that the null hypothesis $H_0 : \beta = 0$ is true. Using the quantities defined in (12) and just above it, it can be easily verified that under $H_0 : \beta = 0$ the restricted estimates $\hat{\boldsymbol{\psi}}_0 = \left(0, \hat{\alpha}_0, \hat{\phi}_0, \hat{\sigma}^2_{v0}, \hat{\sigma}^2_{e0}\right)$ are obtained as

$$\hat{\alpha}_0 = \arg\min_{\alpha} \; n \log Q(\alpha) - \log\left(\frac{1+\alpha}{(n-1)(1-\alpha)+2}\right) + (n-1)\log R(\alpha),$$

where

$$R(\alpha) = \mathbf{Y}_c' \left(\mathbf{I} - \mathbf{Z}_c(\alpha)\left(\mathbf{Z}_c'(\alpha)\mathbf{Z}_c(\alpha)\right)^{-1}\mathbf{Z}_c'(\alpha)\right)\mathbf{Y}_c,$$

$$\mathbf{Z}_c(\alpha) = \mathbf{X}_1 - \mathbf{1}\bar{X}_1 - \alpha\left(\mathbf{X}_0 - \mathbf{1}\bar{X}_0\right),$$

$$\hat{\phi}_0 = \left(\mathbf{Z}_c'(\hat{\alpha}_0)\mathbf{Z}_c(\hat{\alpha}_0)\right)^{-1}\mathbf{Z}_c'(\hat{\alpha}_0)\mathbf{Y}_c,$$

$$\hat{\sigma}^2_{v,0} = \frac{Q(\hat{\alpha}_0)}{n}, \qquad \hat{\sigma}^2_{e,0} = \frac{1}{n-1}R(\hat{\alpha}_0).$$

Just as in the unrestricted case, it is obvious that obtaining the restricted estimates requires only the optimisation of a nearly quadratic one dimensional function. The REML based likelihood ratio test ($RLRT$) for testing $H_0 : \beta = 0$ vs. $H_a : \beta \neq 0$ is now given by

$$R_T = -2L\left(\hat{\boldsymbol{\psi}}_0\right) + 2L\left(\hat{\boldsymbol{\psi}}\right), \tag{15}$$

where $L(\cdot)$ is the REML log-likelihood presented in (13). Under $H_0 : \beta = 0$, the asymptotic distribution of R_T is χ^2_1, the chi-square distribution with one degree of freedom. The following Theorem provides insight into the finite sample behaviour of the $RLRT$ through a formal

expansion of its distribution.

Theorem 3 *Under $H_0 : \beta = 0$ in the model given by (1) and (2), we have*

$$P\left(R_T \leq x\right) = P\left(\chi_1^2 \leq x\right) + \left(3/4 - \rho^2\right) n^{-1} \left[P\left(\chi_3^2 \leq x\right) - P\left(\chi_1^2 \leq x\right)\right] + O\left(n^{-2}\right), \qquad (16)$$

where $\rho = Corr\left(u_t, v_t\right)$.

Theorem 3 in conjunction with the result of Barndorff-Nielsen and Hall (1988) yields the following corollary.

Corollary 1 *If the Bartlett corrected RLRT is defined as*

$$R_{TB} = \left(1 + 2\left(3/4 - \hat{\rho}^2\right) n^{-1}\right)^{-1} R_T,$$

where

$$\hat{\rho}^2 = \left(\hat{\phi}^2 \hat{\sigma}_v^2 + \hat{\sigma}_e^2\right)^{-1} \hat{\phi}^2 \hat{\sigma}_v^2,$$

then

$$P\left(R_{TB} \leq x\right) = P\left(\chi_1^2 \leq x\right) + O\left(n^{-2}\right).$$

Remark 4 *Inspection of the proof of Theorem 3 shows that the expansion (16) depends entirely on the expected values of the higher order derivatives (w.r.t. the parameters) of the restricted log-likelihood. Hence, the results of Theorem 3 and Corollary 1 continue to hold for the restricted likelihood given in (14) when the initial condition X_0 is not from the stationary distribution.*

The results of Theorem 3 obviously imply that the χ_1^2 approximation to R_T is very good and almost unaffected by the nuisance parameters. Most importantly, the leading term in the error is free of the AR parameter, which most affects the finite sample performance of t-statistic based tests, particularly when it is close to unity. Theorem 3 also suggests two very simple ways in which one could adjust the p-value when carrying out a test of $H_0 : \beta = 0$. One would be

to use the first two terms on the right hand side of (16), with ρ^2 replaced by $\hat{\rho}^2$. The other is to use the Bartlett corrected statistic R_{TB}. However, (16) suggests that the original test R_T used in conjunction with the standard χ^2_1 distribution should be very well behaved and any improvements will be minimal at best. This belief is supported by the simulations that we provide in Section 5. Also, the correction factor $1 + 2\left(3/4 - \hat{\rho}^2\right) n^{-1}$ in Corollary 1 is almost unity for any reasonable sample size and hence the correction will be negligible. Hence, we do not pursue the use of the Bartlett corrected test in Corollary 1, merely noting that it can achieve an $O\left(n^{-2}\right)$ error rate. It is also worth noting that though the REML likelihood does not provide an unbiased estimate of β (indeed, the bias of $\hat{\beta}$ can be arbitrarily large due to the fact that ϕ is unbounded, as noted below Theorem 2), the REML likelihood yields a very well behaved test for β, irrespective of how large ϕ is. This result serves to illustrate the point that it may be more desirable at times to carry out tests of hypothesis using appropriate likelihoods rather than using parameter point estimates and supports the idea that bias can be irrelevant in inference, as described in the Introduction.

It is worthwhile to compare our results above with those provided in JM. We first note that the Restricted Likelihood approach provides a way of not only carrying out reliable inference but also yields point estimates with significantly reduced bias, as well as an estimable bias correction unlike the procedure in JM which is only a hypothesis test. JM assumed that the error covariance parameters $(\phi, \sigma_e, \sigma_v)$ are known, that the series $\{X_t\}$ has zero intercept (i.e. $\mu = 0$) and that the initial value X_0 is known to be 0. Under these assumptions, JM obtained a test with certain finite sample optimality properties. Such assumptions, however, would normally not be satisfied in practice. In their Theorem 6, JM allow for an unknown μ taking potentially non-zero values. However, in this Theorem 6 they provide asymptotic results on the size and power properties of their test statistic only along a sequence of local-to-unity parameter values for the autoregressive coefficient α, which is parametrised as $\alpha = 1 - cn^{-1}$ for some fixed $c \geq 0$. The limiting distribution for their test statistic $\hat{R}$, which is not "self-normalised" unlike, for example, the t-statistic for α in a univariate $AR(1)$ process, will be asymptotically degenerate even if the AR coefficient were approaching unity at a rate slower than n^{-1} (for example $\alpha = 1 - ck_n^{-1}$ for $k_n^{-1} + k_n n^{-1} \to 0$) based on the results of Phillips and Magdalinos (2007) and Giraitis

and Phillips (2006). The k_n in the above framework may be arbitrarily "close" to n, such as $k_n = n/\log(\log n)$.

The result provided in Theorem 3 for the $RLRT$ is for a fixed value α of the AR coefficient and is thus a point-wise result. Though we do not currently have any results on the behaviour of the $RLRT$ under a local-to-unity framework, our simulations reported below show that the $RLRT$ works very well even in such scenarios. At the boundary value $\alpha = 1$, the distribution of the $RLRT$ will not be chi-square and the chi-square approximation will fail. In practice however, the boundary value of $\alpha = 1$ is not relevant in most finance applications since the predictor series is stationary by construction. For example, Baker, Taliaferro and Wurgler (2006) assert on page 1715 that "The predictor variables we consider are theoretically stationary by construction (although in any given small sample, of course, one might not be able to reject a unit root)." A similar case for stationarity of the predictor series is made on page 213 of Lewellen (2004). Furthermore, the dependent series y_t in most finance applications is a returns series, which is unquestionably stationary and hence, as Lewellen (2004) states on page 213, "It also makes little sense to predict returns with a nonstationary variable". In light of this context in which the predictive regression model is most often used, we argue that the boundary case of the unit root, where the chi-square approximation to the $RLRT$ will fail, is not as relevant as it is in univariate AR modelling of economic series.

The fact that the leading error term in (16) is minimised at $\rho = \pm\sqrt{3/4}$ and not at $\rho = 0$ seems somewhat puzzling in light of the observation that the t-statistic for β provides exact inference when $\rho = 0$. This puzzling result can be explained by comparing the quality of the chi-square approximation in Theorem 3**,** where all four parameters $\left(\alpha, \phi, \sigma_v^2, \sigma_e^2\right)$ are nuisance parameters, to the approximation in the case where the parameters $\left(\phi, \sigma_v^2, \sigma_e^2\right)$ are known and α is the only nuisance parameter. From equation (36) and the discussion below it in the proof of Theorem 3, it can be seen that if $\left(\phi, \sigma_v^2, \sigma_e^2\right)$ are known the relevant $RLRT$ for testing $H_0 : \beta = 0$, denoted by $R_{T,1}$, would have a distribution that satisfies the formal expansion

$$P\left(R_{T,1} \leq x\right) = P\left(\chi_1^2 \leq x\right) - \rho^2 n^{-1}\left[P\left(\chi_3^2 \leq x\right) - P\left(\chi_1^2 \leq x\right)\right] + O\left(n^{-2}\right). \tag{17}$$

In this situation, the $RLRT$ will be best approximated by a χ_1^2 variable when $\rho = 0$. As a matter of fact, some trivial calculations show that under the assumption that $\left(\phi, \sigma_v^2, \sigma_e^2\right)$ are known the $RLRT$ when $\rho = 0$ is exactly a χ_1^2 variable. By comparing the result in (17) with that in Theorem 3, one sees that the quantity $3/4$ is a measure of the extent to which lack of knowledge of the innovation parameters (ϕ, τ_v, τ_e) affects the finite sample distribution of R_T. Since $\sup_\rho \rho^2 > \sup_\rho \left|\rho^2 - 3/4\right|$ for $\rho \in (-1, 1)$, the chi-square approximation to the $RLRT$ is better in a "minimax" sense over all possible values of ρ for the case when the innovation covariance parameters are unknown than when they are known[6]. Hence, though the $RLRT$ may not provide exact inference when $\rho = 0$, it is able to keep the error in check over the entire parameter space of ρ, whereas the t-statistic for β works perfectly at $\rho = 0$ but fails badly as ρ moves away from 0.

The result in Theorem 3 guarantees that the $RLRT$ will yield a test that is almost of exact size in finite samples. However, one would also like to ensure that this is not achieved at the expense of loss of power. The obvious tests that are competitors to the $RLRT$ are the Wald and Rao score test based on the restricted likelihood. It is a well known fact that just as these three tests share the identical limiting distribution under the null hypothesis, they also have the same power properties to first order. Hence, in order to distinguish between them one has to consider a sequence of local Pitman alternatives given by $H_a : \beta = \beta_0 + \xi n^{-1/2}$. The next Theorem obtains the power function of the $RLRT$, Wald and Rao score test against such local alternatives.

Theorem 4 *Let RLRT, W and RS denote the LRT, Wald test and Rao score test respectively of $H_0 : \beta = \beta_0$ based on the restricted likelihood in (13) of the model in (1) and (2). Assume that the true value of β is given by $\beta = \beta_0 + \xi n^{-1/2}$. Define*

$$\Delta = \frac{1}{1-\alpha^2}\frac{\sigma_v^2}{\sigma_u^2}\xi^2 + O\left(n^{-1}\right),$$

$$C_1 = \frac{-2\alpha\phi\sigma_v^4\xi^3}{\left(1-\alpha^2\right)^2\sigma_u^4}$$

[6]This finding is very much in keeping with the results of van Giersbergen (2006) for the LRT in the univariate AR(1) model that we described in the footnote at the beginning of section 2.

and

$$C_2 = \frac{-3\alpha\phi\sigma_v^2\xi}{(1-\alpha^2)\,\sigma_u^2}.$$

Let $\bar{G}_{s,\Delta}(x)$ *denote the survival function of a non-central* χ^2 *random variable with* s *degrees of freedom and non-centrality parameter* Δ*. Then,*

$$P(RLRT > x) = \bar{G}_{1,\Delta}(x) + \frac{C_1}{n^{1/2}}\left(\bar{G}_{3,\Delta}(x) - 0.5\bar{G}_{1,\Delta}(x) - 0.5\bar{G}_{5,\Delta}(x)\right) + O\left(n^{-1}\right),$$

$$P(W > x) = P(RLRT > x) + O\left(n^{-1}\right),$$

and

$$\begin{aligned} P(RS > x) = P(RLRT > x) &+ \frac{C_1}{n^{1/2}}\left(0.5\bar{G}_{5,\Delta}(x) - 0.5\bar{G}_{7,\Delta}(x)\right) \\ &+ \frac{C_2}{n^{1/2}}\left(0.5\bar{G}_{1,\Delta}(x) - 0.5\bar{G}_{5,\Delta}(x)\right) + O\left(n^{-1}\right). \end{aligned} \tag{18}$$

From Theorem 4 we see that the *RLRT* and the Wald test based on the restricted likelihood have identical power up to second order (i.e. up to $O\left(n^{-1}\right)$) against local Pitman alternatives. Since $\bar{G}_{s,\Delta}(x) - \bar{G}_{l,\Delta}(x) < 0$ for all $x > 0$ when $l > s$, it follows from (18) that *RLRT* will be guaranteed to be more powerful than the Rao score test against local alternatives if $C_1 > 0$ and $C_2 > 0$. This will be the case if $\phi < 0$ and $\xi > 0$, which is exactly the part of the parameter space which is of relevance in empirical applications in finance and economics. It is also interesting to note that the non-centrality parameter Δ, which will be the main source of power, increases as α gets closer to the unit root.

In the next Section we derive the REML likelihood under more general models for the regressor series as well as for multiple regressors and discuss some efficiency and computational issues.

4 REML for more general regressor models

It is easy to generalise the REML likelihood in two directions that are both of practical interest. One generalisation is to the case where the predictor series is a multivariate AR(1) process. Applications of such models can be found, for example, in Amihud and Hurvich (2004), who considered dividend yield and earnings to price ratio as bivariate predictors of market returns. The other generalisation is to the case where the univariate predictor follows a higher order AR process. We will state the REML log-likelihood for both these cases, starting with the multivariate AR(1) predictor model. The method of obtaining the log-likelihood is identical to that used in Theorem 2.

4.1 Multivariate regressors

Assume that the data $(Y_1, ..., Y_n, \mathbf{X}_0', ..., \mathbf{X}_n')$ follows

$$Y_t = \eta + \boldsymbol{\beta}' \mathbf{X}_{t-1} + u_t, \tag{19}$$

$$\mathbf{X}_t = \boldsymbol{\mu} + \mathbf{A}\mathbf{X}_{t-1} + \mathbf{v}_t, \tag{20}$$

where $u_t = \boldsymbol{\phi}' \mathbf{v}_t + e_t$, $(e_t, \mathbf{v}_t')' \sim N\left(\mathbf{0}, \text{diag}(\sigma_e^2, \boldsymbol{\Sigma}_v)\right)$ is an *i.i.d.* series and $\mathbf{A}$ is a $k \times k$ matrix with all eigenvalues less than unity in absolute value. Let $\boldsymbol{\Sigma}_v \equiv Var(\mathbf{v}_t)$ and $\boldsymbol{\Sigma}_X \equiv Var(\mathbf{X}_t)$, given by

$$\text{vec}(\boldsymbol{\Sigma}_X) = (I_{k^2} - \mathbf{A} \otimes \mathbf{A})^{-1} \text{vec}(\boldsymbol{\Sigma}_v)$$

and define

$$\hat{\boldsymbol{\tau}} = \left[\boldsymbol{\Sigma}_X^{-1} + n(\mathbf{I} - \mathbf{A})' \boldsymbol{\Sigma}_v^{-1} (\mathbf{I} - \mathbf{A})\right]^{-1} \left[\boldsymbol{\Sigma}_X^{-1} \mathbf{X}_0 + (\mathbf{I} - \mathbf{A})' \boldsymbol{\Sigma}_v^{-1} (\mathbf{I} - \mathbf{A}) \sum_{t=1}^{n} \mathbf{X}_t\right].$$

Lemma 1 *Then the REML log-likelihood up to an additive constant for the model in (19) and*

(20) is given by

$$L_M = -\left(\frac{n-1}{2}\right)\log\sigma_e^2 - \frac{1}{2\sigma_e^2}S\left(\boldsymbol{\phi},\boldsymbol{\beta},\mathbf{A}\right) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_X\right| - \frac{n}{2}\log\left|\boldsymbol{\Sigma}_v\right| \tag{21}$$
$$- \frac{1}{2}\log\left|\boldsymbol{\Sigma}_X^{-1} + n\left(\mathbf{I}-\mathbf{A}\right)'\boldsymbol{\Sigma}_v^{-1}\left(\mathbf{I}-\mathbf{A}\right)\right|$$
$$- \frac{1}{2}\left\{\left(\mathbf{X}_0-\hat{\boldsymbol{\tau}}\right)'\boldsymbol{\Sigma}_X^{-1}\left(\mathbf{X}_0-\hat{\boldsymbol{\tau}}\right) + \sum_{t=1}^{n}\left(\mathbf{X}_t-\hat{\boldsymbol{\tau}}-\mathbf{A}\left(\mathbf{X}_{t-1}-\hat{\boldsymbol{\tau}}\right)\right)'\boldsymbol{\Sigma}_v^{-1}\left(\mathbf{X}_t-\hat{\boldsymbol{\tau}}-\mathbf{A}\left(\mathbf{X}_{t-1}-\hat{\boldsymbol{\tau}}\right)\right)\right\}.$$

where

$$S\left(\boldsymbol{\phi},\boldsymbol{\beta},\mathbf{A}\right) = \sum_{t=1}^{n}\left(Y_{t,c} - \boldsymbol{\phi}'\mathbf{X}_{t,c} - \left(\boldsymbol{\beta}'-\boldsymbol{\phi}'\mathbf{A}\right)\mathbf{X}_{t-1,c}\right)^2,$$

$\mathbf{X}_{t,c} = \mathbf{X}_t - n^{-1}\sum_{t=1}^{n}\mathbf{X}_t$ *and* $\mathbf{X}_{t-1,c} = \mathbf{X}_{t-1} - n^{-1}\sum_{t=1}^{n}\mathbf{X}_{t-1}$.

To ease the computational burden during optimisation, the likelihood can be defined in terms of the re-parametrised set $\left(\boldsymbol{\Sigma}_v,\sigma_e^2,\mathbf{A},\boldsymbol{\phi},\boldsymbol{\gamma}\right)$, where $\boldsymbol{\gamma} = \boldsymbol{\beta} - \mathbf{A}\boldsymbol{\phi}$. This re-parametrisation allows us to concentrate $\left(\sigma_e^2,\boldsymbol{\phi},\boldsymbol{\gamma}\right)$ out of the likelihood, thus reducing the dimensionality of the optimisation problem. The likelihood can then be sequentially optimised, first over $(\boldsymbol{\Sigma}_v,\mathbf{A})$, with the REML estimates of $\left(\sigma_e^2,\boldsymbol{\phi},\boldsymbol{\gamma}\right)$ being then obtained by OLS through the minimisation of $S\left(\boldsymbol{\phi},\boldsymbol{\gamma}\right)$. A further simplification of the above likelihood occurs if the coefficient matrix $\mathbf{A}$ is diagonal, given by $\mathbf{A} = \mathrm{diag}\left(\alpha_1,...,\alpha_k\right)$ with $\max_i\left|\alpha_i\right| < 1$, in which case one gets

$$Var\left(\mathbf{X}_t\right) \equiv \boldsymbol{\Sigma}_X = \left(\left(\frac{\sigma_{v,ij}}{1-\alpha_i\alpha_j}\right)\right),$$

where $\boldsymbol{\Sigma}_v = \left(\left(\sigma_{v,ij}\right)\right)$. Amihud and Hurvich (2004) find evidence to support this model with a diagonal coefficient matrix in the empirical example that they consider. It should be noted that in the case where $\mathbf{A}$ can be assumed to be a diagonal matrix, the predictive regression model is no longer a SUR system and hence OLS will no longer be efficient. However, REML will clearly retain efficiency, no matter what the form of $\mathbf{A}$ is, thus giving it an advantage both in terms of asymptotic efficiency and power over any OLS based procedure.

Since the dimension of the parameter space is very large in the vector case, it is not feasible to obtain a result such as Theorem 3 in the most general case. However, in the case where $\mathbf{A}$ is a diagonal matrix and where $\left(\sigma_e^2,\boldsymbol{\phi},\boldsymbol{\Sigma}_v\right)$ are assumed known with $\boldsymbol{\Sigma}_v$ diagonal, we are able to

obtain the following result on the finite sample behaviour of the *RLRT* for testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$. The proof follows along lines similar to those for Theorem 3 and is omitted.

Theorem 5 *In the model given by (19) and (20), assume that* $\mathbf{A} = \text{diag}\,(\alpha_1, ..., \alpha_k)$, *with* $\max_i |\alpha_i| < 1$, *and that* $\left(\sigma_e^2, \boldsymbol{\phi}, \boldsymbol{\Sigma}_v\right)$ *are known with* $\boldsymbol{\Sigma}_v = \text{diag}\left(\sigma_{v,11}^2, ..., \sigma_{v,kk}^2\right)$. *Let* R_M *denote the RLRT based on the restricted likelihood in (21) for testing* $H_0 : \boldsymbol{\beta} = \mathbf{0}$. *Then,*

$$P\left(R_M \leq x\right) = P\left(\chi_k^2 \leq x\right) - n^{-1}\left(\sum_{i=1}^{k} \frac{\phi_i^2 \sigma_{v,ii}^2 \sigma_e^{-2}}{1 + \phi_i^2 \sigma_{v,ii}^2 \sigma_e^{-2}}\right)\left[P\left(\chi_{k+2}^2 \leq x\right) - P\left(\chi_k^2 \leq x\right)\right] + O\left(n^{-2}\right).$$

Since $0 < n^{-1} \sum_{i=1}^{k} \phi_i^2 \sigma_{v,ii}^2 \sigma_e^{-2} \left(1 + \phi_i^2 \sigma_{v,ii}^2 \sigma_e^{-2}\right)^{-1} < n^{-1} k$ trivially, the result shows that the χ^2 distribution once again provides a very good approximation to the *RLRT* in this situation. It is useful to note that Theorem 5 shows that the quality of the χ^2 approximation to the *RLRT* is affected only minimally by the dependence between u_t and $\mathbf{v}_t$, over which one has no control. However, once one X variable has been chosen, we can control which other X variables should be included in the model and it is preferable to use a group of X variables that have low (ideally zero) correlation among themselves to avoid unnecessary multicollinearity. Hence, the assumption in the Theorem that $\boldsymbol{\Sigma}_v$ is diagonal is not very unreasonable. Finally, from the discussion below equation (5) and below Theorem 3, one would expect the χ^2 approximation to continue to work well when $\left(\sigma_e^2, \boldsymbol{\phi}, \boldsymbol{\Sigma}_v\right)$ are unknown (with $\boldsymbol{\Sigma}_v$ diagonal) and, indeed, we find this to be the case in our simulations in Section 5 below. As a matter of fact, the simulations show that the *RLRT* behaves very well even when the cross-correlation in the $\mathbf{X}$ variables is as high as 0.9

The restricted likelihood for the $AR(p)$ regressor case can be derived in a manner analogous to that for the $AR(1)$ case and is given next.

4.2 Higher order autoregressive regressors

Let the observed data $(Y_1, ..., Y_n, X_{-p+1}, X_{-p+2}, ..., X_n)$ follow

$$Y_t = \eta + \beta X_{t-1} + u_t \tag{22}$$

and

$$X_t = \mu + \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + v_t, \tag{23}$$

where $u_t = \phi v_t + e_t$ and $(e_t, v_t)\tilde{}N\left(\mathbf{0}, \text{diag}\left(\sigma_e^2, \sigma_v^2\right)\right)$ are an *i.i.d.* series. Furthermore, assume that all the roots of the polynomial $z^p - \sum_{s=1}^{p} z^{p-s}\alpha_s$ lie within the unit circle. Define

$$\mathbf{Y}_c = \mathbf{Y} - \mathbf{1}\bar{Y}, \qquad \mathbf{X}_c = \left[\mathbf{X}_1 - \mathbf{1}\bar{X}_1, ..., \ \mathbf{X}_{-p+1} - \mathbf{1}\bar{X}_{-p+1}\right],$$

where $\mathbf{X}_i = (X_i, X_{i+1}, ..., X_{n-1-i})$ and $\bar{X}_i = n^{-1}\mathbf{1}'\mathbf{X}_i$.

Lemma 2 *The REML log-likelihood up to an additive constant for the model in (22) and (23) is given by*

$$\begin{aligned} L\left(\alpha, \beta, \phi, \sigma_e^2, \sigma_v^2\right) = &- \left(\frac{n-1}{2}\right)\log\sigma_e^2 - \frac{1}{2\sigma_e^2}S\left(\phi, \beta, \alpha_1, ..., \alpha_p\right) \\ &- \left(\frac{n+p-1}{2}\right)\log\sigma_v^2 - \frac{1}{2}\log\left|\mathbf{\Sigma}\right| - \frac{1}{2}\log\left|\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}\right| - \frac{1}{2\sigma_v^2}\left(\mathbf{X} - \mathbf{1}\hat{\tau}\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{X} - \mathbf{1}\hat{\tau}\right) \end{aligned} \tag{24}$$

where $Var\left(\mathbf{X}\right) = \sigma_v^{-2}\mathbf{\Sigma}$, $\hat{\tau} = \left(\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{X}$ *and* $S\left(\phi, \beta, \alpha_1, ..., \alpha_p\right) = \mathbf{Z}_c'\mathbf{Z}_c$ *where*

$$\mathbf{Z}_c = \left(\mathbf{Y}_c - \phi\mathbf{X}_{c,1} - \left(\beta - \phi\alpha_1\right)\mathbf{X}_{c,2} + \sum_{i=2}^{p}\alpha_{c,i}\mathbf{X}_{c,i+1}\right).$$

Though at first glance the expression in (24) looks formidable, it is actually very easy to compute. The quantity $S\left(\phi, \beta, \alpha_1, ..., \alpha_p\right)$ is, of course, just a quadratic form. It is also well known that both the determinant $|\mathbf{\Sigma}|$ and bilinear/quadratic forms of the type $y'\mathbf{\Sigma}^{-1}x$ are very easy to compute for $AR(p)$ models, since for such models $\mathbf{\Sigma}^{-1}$ can be easily expressed as $\mathbf{B}'\mathbf{B}$ for some lower triangular matrix $\mathbf{B}$. (See equation 30 below). After a re-parametrisation

$\gamma = \beta - \phi\alpha_1$, the parameters $\left(\sigma_e^2, \sigma_v^2, \phi, \gamma\right)$ can be concentrated out of (24) and the concentrated log-likelihood optimised over the remaining parameters $(\alpha_2, ..., \alpha_p)$.

In the next Section we report the results of our simulations.

5 Simulations

We first study the performance of four estimators of β: (i) The OLS estimate $\hat{\beta}_{OLS}$ (ii) The bias corrected OLS estimate $\hat{\beta}_{OLS,c}$ (iii) The REML estimate $\hat{\beta}$ and (iv) The bias corrected REML estimate $\hat{\beta}_c$. The bias corrected estimators are defined below Theorem 2. The data was simulated from the model given by equations (1) and (2) with sample sizes, $n = 50, 100$, 200 and 400. For each sample size, the predictive regression slope coefficient β was set to 0 while the autoregressive coefficient α was set to $\alpha = 1 - cn^{-1}$ for $c = 1, 5$ and the initial value X_0 was drawn from its stationary distribution. The innovation variances were set as $\sigma_u^2 = \sigma_v^2 = 1$ and $\phi = -0.98$. When $\sigma_u^2 = \sigma_v^2 = 1$, the parameter ϕ reduces to the correlation between (u_t, v_t). This normalisation of the innovation variances was also used in Campbell and Yogo (2006) and Jansson and Moreira (2006). The high negative value of ϕ was chosen to reflect the kind of values seen empirically (See Campbell and Yogo, 2006).

Tables III and IV report the simulation means and standard deviations of these four estimates based on 10,000 replications for each of the values of α. As predicted by the theory, the bias of $\hat{\beta}$ is uniformly less than that of $\hat{\beta}_{OLS}$, and approximately half its value when α is close to unity. At the same time, $\hat{\beta}$ does not suffer any loss of efficiency, indeed having a uniformly smaller standard deviation than that of $\hat{\beta}_{OLS}$. The bias of $\hat{\beta}_c$ is also always substantially less than that of $\hat{\beta}_{OLS,c}$, at times by as much as 80% and its standard deviation is uniformly lower too. To get an idea of the magnitude of the bias and the impact of the bias correction, we note that Campbell and Yogo (Table 5, 2006) report empirical estimates of β (after normalisation to ensure unit innovation variances, as above) that range between 0.01 and 0.02 for monthly data and between 0.03 and 0.3 for annual data. Similarly, from page 826 of Amihud and Hurvich (2004), we get an estimated (normalised) β of approximately 0.17 for annual data and 0.02 for monthly data. The

simulation results clearly demonstrate the advantage enjoyed by the REML estimate as well as its bias-corrected version, in terms of both bias and standard deviation over the corresponding OLS counterparts. To test the robustness of the procedure to non-normal thick tailed errors, we also generated data from the same model and parameter configurations, but letting (e_t, v_t) be independent t_5 errors. The results are presented in the second set of columns in Tables II and III and it is seen that once again, $\hat{\beta}$ and $\hat{\beta}_c$ are consistently better than $\hat{\beta}_{OLS}$ and $\hat{\beta}_{OLS,c}$ respectively in terms of both bias and standard deviation.

We next turn to studying the quality of the χ^2_1 approximation to the distribution of the $RLRT$ for the above model and parameter configurations with both normal and t_5 innovations. The χ^2_1 approximation was assessed by two measures (i) QQ plots of the $RLRT$ against the theoretical quantiles of a χ^2_1 distribution and (ii) simulation sizes at the 5% and 1% level. The simulation sizes are reported in Table V while the QQ plots are shown in Figure 3. (We present the QQ plots only for the normal innovations since the plots for t_5 errors are qualitatively similar in nature). The $RLRT$ is seen to be very well approximated by the χ^2_1 distribution according to each of the two measures. Comparing the results of Table V with the performance of the usual LRT shown in Table II demonstrates the significant advantage that the $RLRT$ provides over the usual LRT. It is also worth stressing again that the performance of the $RLRT$ is not affected by the bias of $\hat{\beta}$ at all.

We also carried out a simulation study to compare the size and power of our procedure with that of Jansson and Moreira (2006) and Campbell and Yogo (2006). We note that the $RLRT$ procedure is very straighforward to use while the Jansson-Moreira procedure is much more complex. The sample size was set to $n = 100, 200$ and 400 while $\alpha = 1 - cn^{-1}$ with $c = 0, 1, 5, 10$ and 20. The innovation variances were set as $\sigma_u^2 = \sigma_v^2 = 1$ and $\phi = Corr(u_t, v_t) = -0.98$. The initial value X_0 was drawn from $N\left(0, \sigma_v^2\right)$, since the JM (2006) and Campbell and Yogo (2006) procedures require the initial value to be $o_p\left(n^{1/2}\right)$. As a result, the RL used was of the form in (14). Following JM, the parameter β was defined as $\beta = n^{-1}b\sqrt{1-\phi^2}$, where $b = 0, 25$ and 50. Thus, $b = 0$ corresponds to the size simulations while $b = 25, 50$ corresponds to the power of the procedure against local alternatives.

Table VI shows the size of the Campbell and Yogo (2006) Bonferroni Q procedure for a variety of parameter values across 10,000 replications. It is seen that their procedure can produce significant size distortions even when local-to-unity (which is the framework for which their test is designed), with the tests being significantly over-sized under the null hypothesis. Since these tests are based on a Bonferroni style inequality, it is not clear as to how one should compute size-adjusted power for them and as a result we do not provide a power comparison with the Campbell and Yogo (2006) procedure, simply noting that it can be significantly over-sized. It is somewhat puzzling that a test which is supposed to be based on a Bonferroni style inequality yields sizes that are significantly larger than the declared nominal size. However, we point out that in the simulation results that Campbell and Yogo (2006) themselves present in their Table 3, the nominal 5% Bonferroni Q-test has rejection percentages as large as 0.117 and 0.09 in 10,000 replications, which is well above what is to be expected even after accounting for simulation error. One potential cause for this poor performance may be the fact that the Campbell Yogo procedure mean corrects the predictor series by subtracting a GLS estimate of the mean that is computed under the assumption that $\alpha = 1 - 7n^{-1}$ (See Campbell and Yogo, 2005). If the true value of α deviates substantially from this assumption, the mean adjustment will poor and may potentially degrade the size performance of the test.

The results of the comparison between the $RLRT$ and the JM procedures are reported in Table VII for 5,000 replications. As is to be expected from the simulation reported above in Table V, the $RLRT$ maintains its size very well at all the stationary parameter configurations, while JM shows a little size distortion when $c = 20$. The $RLRT$ is oversized when the predictor series is non-stationary and $Corr(u_t, v_t) \simeq -1$, which is not surprising since the chi-square limit distribution will no longer hold at the boundary value of $\alpha = 1$. As we see in Table VIII however, the $RLRT$ is not oversized at the unit root when $Corr(u_t, v_t) = -0.5$, suggesting that the chi-square approximation is degraded at the boundary only when the innovation correlation is extremely high. Furthermore, as argued in the discussion after Theorem 3, we note again that the case of a non-stationary predictor series is not relevant in financial applications where the predictive regression model is most commonly used.

The power comparison between the two procedures in Table VII shows that the $RLRT$

provides uniformly higher power than JM at all the alternatives considered, with significant power gains ranging from twice as much to seven times as much. These simulations provide strong evidence that the RL procedure performs significantly better than the JM procedure in terms of both size and power even when the predictor series is nearly integrated, while also yielding estimates with low (and estimable) bias. The $RLRT$ also does not impose any restrictions on the specification of the initial value and extends easily to the multivariate regressor case.

We also include a comparison between the $RLRT$ and the JM procedure using the parameter configuration in Table II on page 702 of JM. This configuration sets the innovation correlation at -0.5 and 0.5 and allows us to compare the $RLRT$ against the various procedures reported in Table II in JM. The results of this comparison, based on $n = 1000$ and 500 replications to be consistent with the design in JM, are reported in Table VIII. We first note that the $RLRT$ is now no longer oversized at the unit root, suggesting that the chi-square approximation works even at the boundary value $\alpha = 1$ if the innovation correlation is not very high. Furthermore, the $RLRT$ maintains nominal size at all the configurations while once again uniformly dominating the JM procedure in terms of power.

We finally generate data from a model in which the regressors are a bivariate AR(1) model. More specifically, we use the model $Y_t = \boldsymbol{\beta}'\mathbf{X}_{t-1} + u_t$ and $\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{v}_t$, where $u_t = \boldsymbol{\phi}'\mathbf{v}_t + e_t$. The sample size was set at $n = 200$, the slope vector $\boldsymbol{\beta}$ was set to zero, while $\boldsymbol{\phi} = (-80, -80)$. Two configurations of the autoregressive coefficient matrix $\mathbf{A}$ were considered, $\text{diag}\,(0.95, 0.8)$ and $\text{diag}\,(0.95, 0.95)$. The innovation matrix $\boldsymbol{\Sigma}_v$ was set to one of the following three configurations: (a) the identity matrix (b) variances equal to 2 and correlation $\rho_v = 0.5$ (c) variances equal to 10 and correlation $\rho_v = 0.9$. In all cases, the innovation variance σ_e^2 was set to unity. This design (except for $\boldsymbol{\Sigma}_v = \mathbf{I}$) matches that used in Amihud and Hurvich (2004). The number of replications for each parameter configuration was 5,000 in the vector case. As before, the quality of the χ_2^2 approximation to the distribution of the $RLRT$ was assessed via two measures (i) QQ plots of the $RLRT$ against the theoretical quantiles of a χ_2^2 distribution(Figure 4) and (ii) simulation sizes at the 5% and 1% level. The results are provided in Tables VII and VIII. As noted in sub-section 4.1 above, OLS is no longer asymptotically efficient if $\mathbf{A}$ is

a diagonal matrix since the system is no longer a SUR. Hence, not only does REML afford a dramatic reduction in bias over OLS, it also provides a great reduction in the standard deviation, as can be seen in Table VII with the reduction being as much as 80%. Furthermore, for the inference problem it is once again seen that the $RLRT$ is very well approximated by the χ^2 distribution in all the cases we consider.

The overall conclusion to be had from the simulations is that the REML procedure yields point estimates that are much less biased than their OLS counterparts and also an $RLRT$ that is very well behaved, even when the regressors are close to being integrated.

6 Appendix

Proof of Theorem 2:

As noted at the start of Section 2) above, the REML likelihood corresponds to the likelihood of $\mathbf{T}(\mathbf{Y}', \mathbf{X}')'$, where $\mathbf{T}$ is any full row rank matrix such that $\mathbf{T1} = \mathbf{0}$. We will obtain this likelihood by choosing $\mathbf{T}$ to have the form

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 \end{bmatrix},$$

where $\mathbf{T}_1$ and $\mathbf{T}_2$ are full row rank matrices of dimension $(n-1)\times n$ and $n\times(n+1)$ respectively, satisfying $\mathbf{T}_1\mathbf{1} = \mathbf{0}$, $\mathbf{T}_2\mathbf{1} = \mathbf{0}$, $\mathbf{T}_1\mathbf{T}_1' = \mathbf{I}$ and $\mathbf{T}_2\mathbf{T}_2' = \mathbf{I}$ and using the fact that

$$L\left(\mathbf{T}(\mathbf{Y}', \mathbf{X}')'\right) = L\left(\mathbf{T}_1\mathbf{Y} \mid \mathbf{T}_2\mathbf{X}\right) L\left(\mathbf{T}_2\mathbf{X}\right).$$

We first obtain $L\left(\mathbf{T}_1\mathbf{Y} \mid \mathbf{T}_2\mathbf{X}\right)$. Since $u_t = \phi v_t + e_t$, where $\phi = \sigma_{uv}/\sigma_v^2$ and $e_t \tilde{} N\left(0, \sigma_e^2\right)$ is a series independent of $\{v_t\}$, we get

$$\begin{aligned} Y_t &= \eta + \beta X_{t-1} + \phi\left(X_t - \mu - \alpha X_{t-1}\right) + e_t \\ &= \eta + \phi X_t + \left(\beta - \phi\alpha\right) X_{t-1} + e_t. \end{aligned} \tag{25}$$

Let $\tilde{\mathbf{Y}} = \mathbf{T}_1\mathbf{Y}$, $\tilde{\mathbf{X}}_1 = \mathbf{T}_1\mathbf{X}_1$ and $\tilde{\mathbf{X}}_0 = \mathbf{T}_1\mathbf{X}_0$. From (25) it then follows that

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\theta} + \tilde{\mathbf{e}}, \tag{26}$$

where $\tilde{\mathbf{X}} = \left[\tilde{\mathbf{X}}_1,\ \tilde{\mathbf{X}}_0\right]$, $\tilde{\mathbf{e}} = \mathbf{T}_1\mathbf{e}$, $\mathbf{e} = (e_1, ..., e_n)'$ and $\boldsymbol{\theta} = (\phi,\ \beta - \phi\alpha)'$. Since X_t is a function only of $\{v_t, v_{t-1}, ...\}$, the series $\{X_t\}$ is independent of $\{e_t\}$. Furthermore, knowledge of $\mathbf{T}_2\mathbf{X}$, where $\mathbf{T}_2$ is any full row rank matrix such that $\mathbf{T}_2\mathbf{1} = \mathbf{0}$, implies knowledge of $\tilde{\mathbf{X}}$. Hence, from (26) the conditional distribution of $\tilde{\mathbf{Y}}$ given $\mathbf{T}_2\mathbf{X}$ is $N\left(\tilde{\mathbf{X}}\boldsymbol{\theta},\ \sigma_e^2\mathbf{I}\right)$, since $\mathbf{T}_1\mathbf{T}_1' = \mathbf{I}$. It follows that the conditional log-likelihood of $\left(\tilde{\mathbf{Y}} \mid \mathbf{T}_2\mathbf{X}\right)$ up to an additive constant is given by

$$l_1\left(\tilde{\mathbf{Y}} \mid \mathbf{T}_2\mathbf{X},\ \boldsymbol{\theta}, \sigma_e^2\right) = -\left(\frac{n-1}{2}\right)\log\sigma_e^2 - \frac{1}{2\sigma_e^2}\left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right)^*\left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right).$$

Note, however, that $\tilde{\mathbf{X}}\boldsymbol{\theta} = \mathbf{T}_1\left[\mathbf{X}_1,\ \mathbf{X}_0\right]\boldsymbol{\theta}$. Thus,

$$\left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right)^*\left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right) = \left(\mathbf{Y} - \left[\mathbf{X}_1,\ \mathbf{X}_0\right]\boldsymbol{\theta}\right)'\mathbf{T}_1^*\mathbf{T}_1\left(\mathbf{Y} - \left[\mathbf{X}_1,\ \mathbf{X}_0\right]\boldsymbol{\theta}\right).$$

Since the matrix $\mathbf{T}_1$ when augmented by the row $n^{-1/2}\mathbf{1}'$ is an orthogonal matrix, it follows that $\mathbf{T}_1^*\mathbf{T}_1 = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$ and hence

$$\begin{aligned}\left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right)^*\left(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\theta}\right) &= \left(\mathbf{Y} - \left[\mathbf{X}_1,\ \mathbf{X}_0\right]\boldsymbol{\theta}\right)'\left[\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'\right]\left(\mathbf{Y} - \left[\mathbf{X}_1,\ \mathbf{X}_0\right]\boldsymbol{\theta}\right) \\ &= S\left(\phi, \beta, \alpha\right).\end{aligned}$$

Thus, we get

$$l_1\left(\tilde{\mathbf{Y}} \mid \tilde{\mathbf{X}}, \boldsymbol{\theta}, \sigma_e^2\right) = -\left(\frac{n-1}{2}\right)\log\sigma_e^2 - \frac{1}{2\sigma_e^2}S\left(\phi, \beta, \alpha\right). \tag{27}$$

The log-likelihood of $\mathbf{T}_2\mathbf{X}$ is obtained from Harville (1974) and up to an additive constant is given by

$$l_2\left(\mathbf{T}_2\mathbf{X},\ \alpha, \sigma_v^2\right) = -\left(\frac{n}{2}\right)\log\sigma_v^2 - \frac{1}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\log\left|\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\right| - \frac{1}{2\sigma_v^2}\left(\mathbf{X} - \mathbf{1}\hat{\tau}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - \mathbf{1}\hat{\tau}\right), \tag{28}$$

where $\hat{\tau} = \left(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ and $\boldsymbol{\Sigma} = \sigma_v^{-2}Var\left(\mathbf{X}\right)$. From (27) and (28), we obtain the log-

likelihood of $(\mathbf{T}_1\mathbf{Y},\ \mathbf{T}_2\mathbf{X})$ up to an additive constant to be

$$L\left(\mathbf{T}_1\mathbf{Y},\ \mathbf{T}_2\mathbf{X},\ \alpha,\beta,\phi,\sigma_v^2,\sigma_e^2\right) = -\left(\frac{n-1}{2}\right)\log\sigma_e^2 - \frac{1}{2\sigma_e^2}S\left(\phi,\beta,\alpha\right) \\ -\left(\frac{n}{2}\right)\log\sigma_v^2 - \frac{1}{2}\log\left|\mathbf{\Sigma}\right| - \frac{1}{2}\log\left|\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}\right| - \frac{1}{2\sigma_v^2}Q\left(\alpha\right). \tag{29}$$

The final form, as stated in (13) is obtained by algebraic simplification, using the fact that $\mathbf{\Sigma}^{-1} = \mathbf{B}'\mathbf{B}$, where $\mathbf{B}$ is the $(n+1)\times(n+1)$ matrix given by

$$\mathbf{B} = \begin{bmatrix} \sqrt{1-\alpha^2} & 0 & 0 & \cdots & 0 & 0 \\ -\alpha & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\alpha & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & & -\alpha & 1 \end{bmatrix} \tag{30}$$

and noting that

$$\hat{\tau} = \frac{(1-\alpha)\sum_{i=0}^{n-2}X_i + X_{-1} + X_{n-1}}{(n-1)(1-\alpha)+2}.$$

To obtain the REML estimates of $\left(\alpha,\beta,\phi,\sigma_e^2,\sigma_v^2\right)$, it helps to consider the re-parametrised set of parameters $\left(\alpha,\gamma,\phi,\sigma_e^2,\sigma_v^2\right)$, where $\gamma = \beta - \phi\alpha$. It is then immediately obvious from inspecting (29) that the REML estimates of $\left(\alpha,\sigma_v^2\right)$ are obtained by simply maximising just

$$l_2\left(\mathbf{T}_2\mathbf{X},\ \alpha,\sigma_v^2\right) = -\left(\frac{n}{2}\right)\log\sigma_v^2 - \frac{1}{2}\log\left|\mathbf{\Sigma}\right| - \frac{1}{2}\log\left|\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}\right| - \frac{1}{2\sigma_v^2}\left(\mathbf{X}-\mathbf{1}\hat{\tau}\right)'\mathbf{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{1}\hat{\tau}\right).$$

In other words, the REML estimates of $\left(\alpha,\sigma_v^2\right)$ are just those estimates that would have been obtained by maximising the REML likelihood function of $(X_1,...,X_n)$. It thus follows immediately from Cheang and Reinsel (2000) that the bias of $\hat{\alpha}$ is

$$E\left(\hat{\alpha}-\alpha\right) = \frac{2\alpha}{n-1} + o\left(n^{-1}\right).$$

The REML estimates of (ϕ,γ) are obtained as the least squares estimates

$$\left(\hat{\phi},\hat{\gamma}\right)' = \left(\mathbf{X}_c{}'\mathbf{X}_c\right)^{-1}\mathbf{X}_c{}'\mathbf{Y}_c,$$

which then yields the REML estimate $\hat{\beta} = \hat{\gamma} + \hat{\phi}\hat{\alpha}$. Since

$$\begin{pmatrix} \hat{\phi} \\ \hat{\beta} \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ \hat{\alpha} & 1 \end{bmatrix} \begin{pmatrix} \hat{\phi} \\ \hat{\gamma} \end{pmatrix},$$

it follows that the REML estimates of the original parameters (ϕ, β) can also be obtained in a direct regression of $\mathbf{Y}_c$ on $\left(\mathbf{X}_1 - \mathbf{1}\bar{X}_1 - \hat{\alpha}(\mathbf{X}_0 - \mathbf{1}\bar{X}_0),\ \mathbf{X}_0 - \mathbf{1}\bar{X}_0\right)$. Thus, $\hat{\beta}$ is identical to the ARM estimate considered by Amihud and Hurvich (2004) using the REML estimate $\hat{\alpha}$ as a proxy for α. Hence, the bias of $\hat{\beta}$ can be obtained from Theorem 2 of Amihud and Hurvich (2004) and is

$$\begin{aligned} E\left(\hat{\beta} - \beta\right) &= \phi E\left(\hat{\alpha} - \alpha\right) \\ &= -\frac{2\alpha\phi}{n-1} + o\left(n^{-1}\right). \end{aligned}$$

Finally, Lemma 1 of Amihud and Hurvich (2004) implies that $E\left(\hat{\phi}\right) = \phi$.

Proof of Theorem 3:

Since the LRT is invariant to re-parametrisation we choose to work with the re-parametrisation $\lambda = (\beta, \alpha, \phi, \tau_v, \tau_e) \equiv (\beta, \lambda_2)$, where $\tau_v = \sigma_v^{-2}$ and $\tau_e = \sigma_e^{-2}$ since this greatly reduces the burden of our computations. For the REML log-likelihood given in (13), we will denote expectations of the log-likelihood derivatives as

$$\kappa_{rs} = n^{-1}E\left(\frac{\partial^2 L}{\partial\lambda_r\partial\lambda_s}\right), \qquad \kappa_{rst} = n^{-1}E\left(\frac{\partial^3 L}{\partial\lambda_r\partial\lambda_s\partial\lambda_t}\right), \qquad \kappa_{rs}^{(t)} = \frac{\partial\kappa_{rs}}{\partial\lambda_t}.$$

Letting $\delta = \tau_e/\tau_v$, it is easily seen that the information matrix $\mathbf{K} = ((-\kappa_{rs}))$ is given by

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{\beta,\alpha} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{\phi,\tau_v,\tau_e} \end{bmatrix} + O\left(n^{-1}\right), \tag{31}$$

where

$$\mathbf{K}_{\beta,\alpha} = \frac{\delta}{1-\alpha^2}\begin{bmatrix} 1 & -\phi \\ -\phi & \left(\phi^2 + \delta^{-1}\right) \end{bmatrix},$$

and

$$\mathbf{K}_{\phi,\tau_v,\tau_e} = \text{diag}\left\{\delta, \frac{1}{2\tau_v^2}, \frac{1}{2\tau_e^2}\right\}.$$

Let κ^{rs} denote the entries of $-\mathbf{K}^{-1}$ and $\tilde{\kappa}^{rs}$ denote the entries of $-\mathbf{K}_{22}^{-1}$, where $\mathbf{K}_{22}$ is the lower right 4×4 sub-matrix of $\mathbf{K}$.

Theorem 1 of Hayakawa (1977) provides an expression for the formal expansion of the distribution of R_T. This expression is in notation that is not easy to use and also contains a term, A_2, which was subsequently shown (Chesher and Smith, 1995) to be identically zero in regular models. As a result, we use the notation of Cordeiro (1987) and from the discussion on page 342 of Cribari-Neto and Cordeiro (1996) obtain the expansion

$$P\left(R_T \leq x\right) = P\left(\chi_1^2 \leq x\right) + An^{-1}\left[P\left(\chi_3^2 \leq x\right) - P\left(\chi_1^2 \leq x\right)\right] + O\left(n^{-2}\right), \tag{32}$$

where

$$A = \frac{1}{2}\left\{\sum_{\lambda}\left(l_{rstu} - l_{rstuvw}\right) - \sum_{\lambda_2}\left(\tilde{l}_{rstu} - \tilde{l}_{rstuvw}\right)\right\}, \tag{33}$$

$$l_{rstu} = \kappa^{rs}\kappa^{tu}\left(\frac{1}{4}\kappa_{rstu} - \kappa_{rst}^{(u)} + \kappa_{rt}^{(su)}\right), \qquad \tilde{l}_{rstu} = \tilde{\kappa}^{rs}\tilde{\kappa}^{tu}\left(\frac{1}{4}\kappa_{rstu} - \kappa_{rst}^{(u)} + \kappa_{rt}^{(su)}\right),$$

$$l_{rstuvw} = \kappa^{rs}\kappa^{tu}\kappa^{vw}\left(\frac{1}{6}\kappa_{rtv}\kappa_{suw} + \frac{1}{4}\kappa_{rtu}\kappa_{svw} - \kappa_{rtv}\kappa_{sw}^{(u)} - \kappa_{rtu}\kappa_{sw}^{(v)} + \kappa_{rt}^{(v)}\kappa_{sw}^{(u)} + \kappa_{rt}^{(u)}\kappa_{sw}^{(v)}\right)$$

and

$$\tilde{l}_{rstuvw} = \tilde{\kappa}^{rs}\tilde{\kappa}^{tu}\tilde{\kappa}^{vw}\left(\frac{1}{6}\kappa_{rtv}\kappa_{suw} + \frac{1}{4}\kappa_{rtu}\kappa_{svw} - \kappa_{rtv}\kappa_{sw}^{(u)} - \kappa_{rtu}\kappa_{sw}^{(v)} + \kappa_{rt}^{(v)}\kappa_{sw}^{(u)} + \kappa_{rt}^{(u)}\kappa_{sw}^{(v)}\right).$$

Exploiting the near diagonal (up to $O\left(n^{-1}\right)$) structure of $\mathbf{K}$, we can simplify the term $\sum_{\lambda}\left(l_{rstu} - l_{rstuvw}\right)$ in A as

$$\begin{aligned}\sum_{\lambda}\left(l_{rstu} - l_{rstuvw}\right) = &\sum_{(\beta,\alpha)}\left(l_{rstu} - l_{rstuvw}\right) + \sum_{(\phi,\tau_v,\tau_e)}\left(l_{rstu} - l_{rstuvw}\right) \\ &+ \sum_{((\beta,\alpha),(\phi,\tau_v,\tau_e))}\left(l_{rstu} - l_{rstuvw}\right) + O\left(n^{-1}\right),\end{aligned} \tag{34}$$

where $\sum_{((\beta,\alpha),(\phi,\tau_v,\tau_e))}$ denotes that at least one index in the summand must come from (β,α) and at least one index from (ϕ,τ_v,τ_e) . By the same logic, we can simplify the term $\sum_{\lambda_2}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)$ in A as

$$\begin{aligned}\sum_{\lambda_2}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right) &= \sum_{\alpha}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)+\sum_{(\phi,\tau_v,\tau_e)}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)\\ &+\sum_{(\alpha,(\phi,\tau_v,\tau_e))}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)+O\left(n^{-1}\right)\\ &=\sum_{\alpha}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)+\sum_{(\phi,\tau_v,\tau_e)}\left(l_{rstu}-l_{rstuvw}\right)\\ &+\sum_{(\alpha,(\phi,\tau_v,\tau_e))}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)+O\left(n^{-1}\right), \qquad (35)\end{aligned}$$

where the last step in (35) follows from the fact that the entries of $\mathbf{K}$ for (ϕ,τ_v,τ_e) are diagonal up to $O\left(n^{-1}\right)$. It thus follows from (33), (34) and (35) that

$$\begin{aligned}A &= \frac{1}{2}\left\{\sum_{(\beta,\alpha)}\left(l_{rstu}-l_{rstuvw}\right)-\sum_{\alpha}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)\right\}\\ &+\frac{1}{2}\left\{\sum_{((\beta,\alpha),(\phi,\tau_v,\tau_e))}\left(l_{rstu}-l_{rstuvw}\right)-\sum_{(\alpha,(\phi,\tau_v,\tau_e))}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)\right\}+O\left(n^{-1}\right)\\ &\equiv A_{(\beta,\alpha)}+C_{\lambda}+O\left(n^{-1}\right), \qquad (36)\end{aligned}$$

It is obvious from the structure of $A_{(\beta,\alpha)}$ in (36) that $A_{(\beta,\alpha)}$ would be the leading remainder term in the expansion of the distribution of R_T of the form in (32) if (ϕ,τ_v,τ_e) were known and α were the only nuisance parameter. The term C_{λ}, thus, is a measure of the extent to which lack of knowledge of the parameters (ϕ,τ_v,τ_e) affects the finite sample distribution of R_T. We now compute the two terms $A_{(\beta,\alpha)}$ and C_{λ}, beginning with $A_{(\beta,\alpha)}$.

Though (β,α) are not orthogonal, the computation of $\sum_{(\beta,\alpha)}\left(l_{rstu}-l_{rstuvw}\right)$ can be simplified by working with the transformed parameters (γ,α) , where $\gamma=\beta-\phi\alpha$, since γ is orthogonal to α (Note that as stated above, when computing $A_{(\beta,\alpha)}$ the remaining parameters (ϕ,τ_v,τ_e) are fixed). Since (γ,α) is an affine transformation of (β,α) , we can exploit the fact that the

first term of $A_{(\beta,\alpha)}$ is invariant under such transformations (see page 371 of Hayakawa 1977) and thus get

$$\frac{1}{2}\sum_{(\beta,\alpha)}\left(l_{rstu}-l_{rstuvw}\right)=\frac{1}{2}\sum_{(\gamma,\alpha)}\left(l_{rstu,p}-l_{rstuvw,p}\right), \tag{37}$$

where the extra subscript p will mean that the computation is being carried out for the re-parameterised form of (13) with $\gamma=\beta-\phi\alpha$. The right hand side of (37) is much simpler to compute due to the fact that $\kappa_{\gamma\alpha,p}=\kappa^{(\gamma)}_{\gamma\gamma,p}=\kappa^{(\gamma\gamma)}_{\gamma\gamma,p}=0$ and all the terms $\kappa_{rst,p}$ and $\kappa_{rstu,p}$ are at most $O\left(n^{-1}\right)$, for all permutations of the subscripts. Since $\kappa_{\gamma\gamma,p}=-\delta\left(1-\alpha^2\right)^{-1}$ and $\kappa_{\alpha\alpha,p}=-\left(1-\alpha^2\right)^{-1}$, it follows that

$$\frac{1}{2}\sum_{(\gamma,\alpha)}l_{rstu,p}=\frac{1}{2}\kappa_p^{\alpha\alpha}\kappa_p^{\alpha\alpha}\left(\kappa^{(\alpha\alpha)}_{\alpha\alpha,p}\right)+O\left(n^{-1}\right)=-\left(1+3\alpha^2\right)\left(1-\alpha^2\right)^{-1}+O\left(n^{-1}\right) \tag{38}$$

and

$$\frac{1}{2}\sum_{(\gamma,\alpha)}l_{rstuvw,p}=\frac{1}{2}\kappa_p^{\alpha\alpha}\kappa_p^{\alpha\alpha}\kappa_p^{\alpha\alpha}2\left(\kappa^{(\alpha)}_{\alpha\alpha}\right)^2+O\left(n^{-1}\right)=-4\alpha^2\left(1-\alpha^2\right)^{-1}+O\left(n^{-1}\right). \tag{39}$$

¿From (37), (38) and (39), we get

$$\frac{1}{2}\sum_{(\beta,\alpha)}\left(l_{rstu}-l_{rstuvw}\right)=-1+O\left(n^{-1}\right). \tag{40}$$

The second term in $A_{(\beta,\alpha)}$ is not invariant under affine transformations and we revert to the original log-likelihood (13) to compute it. Noting that $\kappa_{\alpha\alpha}=-\left(\phi^2\delta+1\right)\left(1-\alpha^2\right)^{-1}$ and $\tilde{\kappa}^{\alpha\alpha}=-\left(\phi^2\delta+1\right)^{-1}\left(1-\alpha^2\right)$, we have

$$\begin{aligned}\frac{1}{2}\sum_{\alpha}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)&=\frac{1}{2}\tilde{\kappa}^{\alpha\alpha}\tilde{\kappa}^{\alpha\alpha}\left(\kappa^{(\alpha\alpha)}_{\alpha\alpha}\right)-\frac{1}{2}\tilde{\kappa}^{\alpha\alpha}\tilde{\kappa}^{\alpha\alpha}\tilde{\kappa}^{\alpha\alpha}2\left(\kappa^{(\alpha)}_{\alpha\alpha}\right)^2+O\left(n^{-1}\right)\\&=-\frac{1}{\phi^2\delta+1}+O\left(n^{-1}\right).\end{aligned} \tag{41}$$

¿From (40) and (41) we conclude that

$$A_{(\beta,\alpha)}=-\frac{\phi^2\delta}{\phi^2\delta+1}+O\left(n^{-1}\right)=-\rho^2+O\left(n^{-1}\right). \tag{42}$$

We now turn our attention to computing the second term, C_λ, in (36). We first note from the near diagonal structure of $\mathbf{K}$ that

$$\begin{aligned}\frac{1}{2}\sum_{((\beta,\alpha),(\phi,\tau_v,\tau_e))}(l_{rstu}-l_{rstuvw}) &= \frac{1}{2}\sum_{(\beta,(\phi,\tau_v,\tau_e))} l_{rrtt}+\frac{1}{2}\sum_{(\alpha,(\phi,\tau_v,\tau_e))} l_{rrtt} \\ &\quad -\frac{1}{2}\sum_{(\beta,(\phi,\tau_v,\tau_e))} l_{rrttvv}-\frac{1}{2}\sum_{(a,(\phi,\tau_v,\tau_e))} l_{rrttvv} \\ &\quad -\frac{1}{2}\sum_{(\beta,\alpha,(\phi,\tau_v,\tau_e))} l_{rstuvw}.\end{aligned} \tag{43}$$

Each of the terms in (43) are now computed. The details are not provided here, both to save space and also because the computation does not afford any special insight into the problem. The detailed calculations, however, are available from the authors. The terms in $\sum_{(\alpha,(\phi,\tau_v,\tau_e))}\left(\tilde{l}_{rstu}-\tilde{l}_{rstuvw}\right)$ can be decomposed in a manner similar to that in (43), except that in this case there are no terms in β. When all these terms are put together, one gets

$$C_\lambda = \frac{3}{4}+O\left(n^{-1}\right). \tag{44}$$

The theorem now follows from (32), (36), (42) and (44).

Proof of Theorem 4:

The expansion of the distribution of the LRT and the Wald test under local Pitman alternatives is given in Hayakawa (1975), while that of the distribution of the Rao score test is given by Harris and Peers (1980). These results are consolidated using simpler notation in Cordeiro, Botter and Ferrari (1994) and we follow the notation used in their work. To obtain the results of Theorem 4, we calculate the quantities in equations (1) - (5b) on page 711 of Cordeiro et al. (1994). Letting $\lambda=(\beta,\alpha,\phi,\tau_v,\tau_e)$, where $\tau_v=\sigma_v^{-2}$ and $\tau_e=\sigma_e^{-2}$, we note that the quantities $\kappa_{ij}=n^{-1}E\left(\partial^2 L/\partial\lambda_i\partial\lambda_j\right)$ have been already obtained in (31). To obtain the quantities $\kappa_{i,jk}=n^{-1}E\left[\left(\partial L/\partial\lambda_i\right)\left(\partial^2 L/\partial\lambda_j\partial\lambda_k\right)\right]$ and $\kappa_{i,j,k}=n^{-1}E\left[\left(\partial L/\partial\lambda_i\right)\left(\partial L/\partial\lambda_j\right)\left(\partial L/\partial\lambda_k\right)\right]$, we exploit the Bartlett identities (Bartlett, 1953) to obtain

$$\kappa_{i,j,k}=2\kappa_{ijk}-\frac{\partial\kappa_{jk}}{\partial\lambda_i}-\frac{\partial\kappa_{ik}}{\partial\lambda_j}-\frac{\partial\kappa_{ij}}{\partial\lambda_k}$$

and

$$\kappa_{i,jk} = -\kappa_{ijk} + \frac{\partial \kappa_{jk}}{\partial \lambda_i}.$$

The quantities $\partial \kappa_{jk}/\partial \lambda_i$ are then calculated using (31). Using the same notation (note, however, that we define all our cumulants κ to be $O(1)$ whereas Cordeiro et al. (1994) define them to be $O(n)$) as in equations (3a) - (5b) of Cordeiro et al. (1994), we get

$$b_{10} = -0.5n^{-1/2}C_1 + O\left(n^{-1}\right),\ b_{11} = n^{-1/2}C_1 + O\left(n^{-1}\right),\ b_{12} = -0.5n^{-1/2}C_1 + O\left(n^{-1}\right),\ b_{13} = 0,$$

$$b_{20} = -0.5n^{-1/2}C_1 + O\left(n^{-1}\right),\ b_{21} = n^{-1/2}C_1 + O\left(n^{-1}\right),\ b_{22} = -0.5n^{-1/2}C_1 + O\left(n^{-1}\right),\ b_{23} = O\left(n^{-1}\right)$$

and

$$b_{30} = -0.5n^{-1/2}C_1 + O\left(n^{-1}\right),\ b_{31} = n^{-1/2}\left(C_1 + C_2\right) + O\left(n^{-1}\right)$$
$$b_{32} = -n^{-1/2}C_1 + O\left(n^{-1}\right),\ b_{33} = -0.5n^{-1/2}C_1 + O\left(n^{-1}\right).$$

The result now follows from equation 2 of Cordeiro et al. (1994).

References

Amihud, Y. and Hurvich, C. (2004): "Predictive Regressions: A Reduced-Bias Estimation Method," *Journal of Financial and Quantitative Analysis,* 39, 813-841.

Baker, M., Taliaferro, R. and Wurgler, J. (2006): "Predicting Returns with Managerial Decision Variables: Is there a Small Sample Bias?," *Journal of Finance,* LXI, No. 4, 1711-1730.

Barndorff-Nielsen, O.E. and Hall, P. (1988): "On the Level-Error after Bartlett Adjustment of the Likelihood Ratio Statistic," *Biometrika,* 75, 374-378.

Bartlett, M. (1937): "Properties of Sufficiency and Statistical Tests," *Proc. Roy. Soc. Ser. A,* 160, 268-282.

Bartlett, M. (1953): "Approximate Confidence Intervals III. More than One Unknown Parameter," *Biometrika,* 40, 306-317.

Campbell, J. and Yogo, M. (2005): "Implementing the Econometric Methods in "Efficient Tests of Stock Return Predictability"," http://finance.wharton.upenn.edu/~yogo/

Campbell, J. and Yogo, M. (2006): "Efficient Tests of Stock Return Predictability," *Journal of Financial Economics*, 81, 27-60.

Chandra, Tapas K.; Ghosh, J. K. Valid asymptotic expansions for the likelihood ratio statistic and other perturbed chi-square variables. *Sankhya, Ser. A*. 41 (1979), no. 1-2, 22–47

Chen, W. and Deo, R. (2006): "A Smooth Transition to the Unit Root Distribution via the Chi-Square Distribution with Interval Estimation for Nearly Integrated Autoregressive Processes," Working Paper.

Chen, W. and Deo, R. (2007): "The Chi-Square Approximation of the Restricted Likelihood Ratio Test for the Sum of Autoregressive Coefficients with Interval Estimation," Working Paper.

Cheang, W.K. and Reinsel, G. (2000): "Bias Reduction of Autoregressive Estimates in Time Series Regression Model through Restricted Maximum Likelihood," *Journal of the American Statistical Association*, 95**,** 1173-1184.

Chesher, A. and Smith, R. (1995): "Bartlett Corrections to Likelihood Ratio Tests," *Biometrika*, 82, 433-436.

Cordeiro, G. (1987): "On the Corrections to the Likelihood Ratio Statistics," *Biometrika,* 74, 265-274.

Cordeiro, G., Botter, D. and Ferrari, S. L. (1994): "Nonnull Asymptotic Distributions of Three Classic Criteria in Generalised Linear Models," *Biometrika,* 81, 709-720.

Cribari-Neto, F. and Cordeiro, G. (1996): "On Bartlett and Bartlett-Type Corrections," *Econometric Reviews,* 15, 339-367.

Dufour, J-M. and King, M. L. (1991): "Optimal Invariant Tests for the Autocorrelation Coefficient in Linear Regressions with Stationary or Non-stationary AR(1) Errors," *Journal of Econometrics*, 47, 115-143.

Fisher, R. (1973): "*Statistical Methods and Scientific Inference.*" Haffner, New York.

Francke, M, and de Vos, A. (2006): "Marginal Likelihood and Unit Roots," *Journal of Econometrics,* 137, 708-728.

Efron, B. (1975): "Defining the curvature of a statistical problem (with applications to second order efficiency)," *Annals of Statistics*, 3, 1189–1242.

van Garderen, K. (1999): "Exact Geometry of Autoregressive Models," *Journal of Time Series Analysis,* 20, 1-21.

van Giersbergen, N (2006): "Bartlett Correction in the Stable AR(1) Model with Intercept and Trend," Working paper, Universiteit van Amsterdam.

Giraitis, L. and Phillips, P. C. B. (2006): Uniform Limit Theory for Stationary Autoregression," *Journal of Time Series Analysis,* 27, 51-60.

Harris, P. and Peers, H. W. (1980): "The Local Power of the Efficient Scores Test Statistic," *Biometrika,* 67, 525-529.

Harville, D. (1974): "Bayesian Inference of Variance Components Using Only Error Contrasts," *Biometrika,* 61, 383-385.

Harville, D. (1977): "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of the American Statistical Association*, 72, 320-338.

Hayakawa, T. (1975): "The Likelihood Ratio Criterion for a Composite Hypothesis under a Local Alternative," *Biometrika,* 62, 451-460.

Hayakawa, T. (1977): "The Likelihood Ratio Criterion and the Asymptotic Expansion of its Distribution," *Ann. Inst. Statist. Math.,* 29, 359-378.

Hayakawa, T. (1987): Correction, *Ann. Inst. Statist. Math.,* 39, 681.

Jansson, M. and Moreira, M. (2006): "Optimal Inference in Regression Models with Nearly Integrated Regressors," *Econometrica*, 74, 681-714.

Kalbfleisch, J. and Sprott, D. (1970): "Application of Likelihood Methods to Models Involving Large Numbers of Parameters," *Journal of the Royal Statistical Society B,* 32, 175-194.

Kang, W., Shin, D. and Lee, Y. (2003): "Baises of the Restricted Maximum Likelihood Estimators for ARMA Processes with Polynomial Time Trend," *Journal of Statistical Planning and Inference,* 116, 163-176.

Kass, R. E. and Slate, E. H. (1994): "Some diagnostics of maximum likelihood and posterior nonnormality," *Annals of Statistics*, 22, 668–695.

Lawley, D. (1956): "A General Method for Approximating to the Distribution of the Likelihood Ratio Criteria," *Biometrika,* 43, 295-303.

Lewellen, J. (2004): "Predicting Returns with Financial Ratios," *Journal of Financial Economics*, 74, 209-235.

Lehmann, E. and Romano, J. (2005): *Testing Statistical Hypotheses, Third Edition.* Springer

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

Marriott, F. and Pope, J. (1954): "Bias in the Estimation of Autocorrelations," *Biometrika,* 41, 390-402.

Phillips, P.C.B. (1987): "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika,*74, 535-547.

Phillips, P. C. B. and Magdalinos, T. (2007): "Limit Theory for Moderate Deviations from a Unit Root," *Journal of Econometrics,* 136, 115-130.

Rahman, S. and King, M. (1997): "Marginal-Likelihood Score-Based Tests of Regression Disturbances in the Presence of Nuisance Parameters," *Journal of Econometrics,* 82, 81-106.

Smyth, G. and Verbyla, A. (1996): "A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models," *Journal of the Royal Statistical Society Ser. B*, 58, 565-572.

Sprott, D. (1973): "Normal Likelihoods and their Relation to Large Sample Theory of Estimation," *Biometrika,* 60, 457-465.

Sprott, D. (1975): "Application of Maximum Likelihood Methods to Finite Samples," *Sankhya, Ser. B.*, 37, 259-270.

Sprott, D. (1980): "Maximum Likelihood in Small Samples: Estimation in the Presence of Nuisance Parameters," *Biometrika,* 67, 515-523.

Sprott, D. and Viveros-Aguilera, R. (1984): "The Interpretation of Maximum-Likelihood Estimation," *Canadian Journal of Statistics,* 12, 27-38.

Sprott, D. (1990): "Inferential Estimation, Likelihood, and Linear Pivotals," *Canadian Journal of Statistics,* 18, 1-10.

Stambaugh, R. (1999): "Predictive Regressions," *Journal of Financial Economics,* 54, 375-421.

Tunnicliffe Wilson, G. (1989): "On the Use of Marginal Likelihood in Time Series Model Estimation," *Journal of the Royal Statistical Society Ser. B*, 51, 15-27.

Table I. Rejection Rates of LRT for Predictive Regressions with 5% Nominal Rate

$\beta = 0$, $\sigma_v^2 = \sigma_u^2 = 1$

	ϕ	-.5				-.98			
α	n	50	100	200	400	50	100	200	400
.90		.0680	.0620	.0532	.0528	.1127	.0838	.0690	.0604
.95		.0784	.0696	.0580	.0546	.1570	.1149	.0843	.0696
.99		.0904	.0882	.0791	.0639	.3114	.2392	.1748	.1260
.995		.0946	.0936	.0897	.0713	.3629	.3095	.2370	.1700

Table II. Rejection Rates of RLRT for AR(1) Processes

	intercept			intercept & trend		
α	10%	5%	1%	10%	5%	1%
.9	.1025	.0504	.0112	.1026	.0499	.0106
.95	.1006	.0505	.0112	.0903	.0435	.0091
.99	.0847	.0404	.0095	.0927	.0442	.0088
.995	.0827	.0403	.0091	.0927	.0437	.0089

Table III. Mean and Standard Deviation of $\hat{\beta}$: $c = 1$

$\beta = 0$, $\alpha = 1 - c/n$, $\phi = -.98$, $\sigma_v^2 = \sigma_u^2 = 1$

innovations			Gaussian errors				t_5 errors			
n	α		$\hat{\beta}_{OLS}$	$\hat{\beta}_{REML}$	$\hat{\beta}_{OLS,c}$	$\hat{\beta}_{REML,c}$	$\hat{\beta}_{OLS}$	$\hat{\beta}_{REML}$	$\hat{\beta}_{OLS,c}$	$\hat{\beta}_{REML,c}$
50	.98	bias	.0945	.0484	.0164	.0097	.0953	.0491	.0173	.0105
		s.d.	.0841	.0789	.0884	.0821	.0839	.0785	.0883	.0817
100	.99	bias	.0486	.0247	.0094	.0052	.0491	.0254	.0010	.0059
		s.d.	.0436	.0401	.0447	.0409	.0442	.0408	.0453	.0416
200	.995	bias	.0245	.0124	.0049	.0027	.0247	.0125	.0051	.0028
		s.d.	.0226	.0204	.0228	.0207	.0225	.0206	.0227	.0208
400	.9975	bias	.0123	.0061	.0025	.0012	.0126	.0065	.0028	.0016
		s.d.	.0114	.0104	.0114	.0104	.0114	.0104	.0114	.0105

[7]Tables I, II and III are based on 10,000 replications.

Table IV. Mean and Standard Deviation of $\hat{\beta}$: $c = 5$

$\beta = 0$, $\alpha = 1 - c/n$, $\phi = -.98$, $\sigma_v^2 = \sigma_u^2 = 1$

innovations			Gaussian errors				t_5 errors			
n	α		$\hat{\beta}_{OLS}$	$\hat{\beta}_{REML}$	$\hat{\beta}_{OLS,c}$	$\hat{\beta}_{REML,c}$	$\hat{\beta}_{OLS}$	$\hat{\beta}_{REML}$	$\hat{\beta}_{OLS,c}$	$\hat{\beta}_{REML,c}$
50	.90	bias	.0796	.0370	.0058	.0011	.0811	.0389	.0074	.0031
		s.d.	.0927	.0927	.0981	.0965	.0926	.0933	.0980	.0971
100	.95	bias	.0424	.0199	.0043	.0011	.0427	.0206	.0047	.0018
		s.d.	.0493	.0482	.0506	.0492	.0492	.0489	.0506	.0499
200	.975	bias	.0219	.0103	.0026	.0007	.0222	.0109	.0029	.0014
		s.d.	.0255	.0247	.0258	.0249	.0256	.0252	.0260	.0255
400	.9875	bias	.0110	.0051	.0013	.0002	.0113	.0056	.0015	.0007
		s.d.	.0132	.0128	.0133	.0128	.0129	.0127	.0129	.0128

Table V. Rejection Rates, Simulation Mean and Variance of RLRT – univariate regressor

$\beta = 0$, $\alpha = 1 - c/n$, $\phi = -.98$, $\sigma_v^2 = \sigma_u^2 = 1$

	$c = 1$					$c = 5$				
		Gaussian		t_5			Gaussian		t_5	
n	α	5%	1%	5%	1%	α	5%	1%	5%	1%
50	.98	.0445	.0084	.0449	.0092	.90	.0506	.0100	.0519	.0131
100	.99	.0444	.0095	.0504	.0101	.95	.0472	.0093	.0548	.0123
200	.995	.0429	.0093	.0456	.0104	.975	.0460	.0103	.0533	.0134
400	.9975	.0443	.0108	.0473	.0097	.9875	.0497	.0103	.0550	.0136

Table VI. Sizes of Campbell Yogo procedure

$\beta = 0$, $\alpha = 1 - c/n$, $\phi = -.98$, $\sigma_v^2 = \sigma_u^2 = 1$

	$c = 1$		$c = 5$		$c = 20$	
n	5% 2-side	2.5% 1-side	5% 2-side	2.5% 1-side	5% 2-side	2.5% 1-side
100	.0647	.0604	.0724	.0590	.0999	.0541
200	.0572	.0519	.0664	.0496	.0758	.0325
400	.0490	.0433	.0586	.0406	.0670	.0302

[8]Tables IV, V and VI are based on 10,000 replications.

Table VII. Rejection Rates for Testing $H_0 : \beta = 0$

True $\beta = b/n\sqrt{1-\phi^2}$, $\phi = -.98$, $\alpha = 1 - c/n$, 5,000 replications

		$b = 0$		$b = 25$		$b = 50$	
c	n	RLRT	J&M	RLRT	J&M	RLRT	J&M
0	100	.0852	.0462	1.000	.7718	1.0000	.9852
	200	.0880	.0544	1.000	.7846	1.0000	.9940
	400	.0860	.0448	1.000	.7906	1.0000	.9992
1	100	.0556	.0506	.9982	.4796	1.0000	.9990
	200	.0530	.0522	.9984	.4688	1.0000	.9990
	400	.0496	.0500	.9980	.4680	1.0000	.9996
5	100	.0552	.0492	.3300	.0782	1.0000	.5454
	200	.0500	.0500	.3386	.0828	1.0000	.5352
	400	.0486	.0568	.3328	.0886	.9988	.5124
10	100	.0566	.0560	.1612	.0724	.6508	.1110
	200	.0488	.0500	.1598	.0650	.6932	.1016
	400	.0456	.0570	.1690	.0728	.7082	.1044
20	100	.0530	.0610	.1162	.0646	.3180	.0784
	200	.0496	.0570	.1072	.0634	.3066	.0770
	400	.0474	.0622	.1088	.0684	.3196	.0804

Table VIII. Rejection Rates for Testing $H_0 : \beta = 0$

True $\beta = b/n\sqrt{1-\phi^2}$, $\alpha = 1 - c/n$, $n = 1,000$, 500 replications

	c	0		5		10		15	
ϕ	b	RLRT	J&M	RLRT	J&M	RLRT	J&M	RLRT	J&M
$-.5$	0	.058	.054	.046	.058	.042	.042	.048	.066
	5	.516	.424	.240	.102	.170	.076	.130	.080
	10	.898	.826	.644	.338	.450	.180	.340	.154
	15	.984	.946	.914	.702	.782	.392	.650	.230
.5	0	.052	.046	.050	.036	.044	.018	.048	.022
	5	.368	.412	.195	.132	.146	.054	.110	.040
	10	.754	.602	.526	.220	.402	.122	.306	.088
	15	.922	.720	.780	.274	.634	.196	.546	.082

Table IX. Mean and Variance of $\hat{\beta}$ – bivariate regressor

$\beta = 0$, $\phi = c(-80, -80)$, $\sigma_e^2 = 1$.

diag **A**		(.95, .80)				(.95, .95)			
ρ_v		$\hat{\beta}_{1,OLS}$	$\hat{\beta}_{2,OLS}$	$\hat{\beta}_{1,REML}$	$\hat{\beta}_{2,REML}$	$\hat{\beta}_{1,OLS}$	$\hat{\beta}_{2,OLS}$	$\hat{\beta}_{1,REML}$	$\hat{\beta}_{2,REML}$
0	bias	1.8375	2.0377	0.7480	0.6828	2.1131	2.1399	0.7579	0.7653
	s.d.	3.3518	5.4121	2.3268	3.6798	3.5761	3.6521	2.3144	2.4042
.5	bias	2.4394	1.2751	0.6125	0.7532	2.1007	2.1479	0.7012	0.7411
	s.d.	4.3870	7.1407	2.0053	3.2747	5.0234	5.1596	2.0516	2.1255
.9	bias	3.9979	-1.2468	0.3587	0.5306	2.0745	2.1766	0.6291	0.6384
	s.d.	6.5038	10.1373	1.2865	2.1064	11.4054	11.5401	1.6685	1.6961

Table X. Rejection Rates, Simulation Mean and Variance of RLRT– bivariate regressor

$\beta = 0$, $\phi = c(-80, -80)$, $\sigma_e^2 = 1$.

diag **A**	(.95, .80)		(.95, .95)	
ρ_v	5%	1%	5%	1%
0	.0494	.0106	.0554	.0124
.5	.0536	.0112	.0558	.0132
.9	.0572	.0122	.0548	.0104

[9]Tables IX and X are based on 10,000 replications.

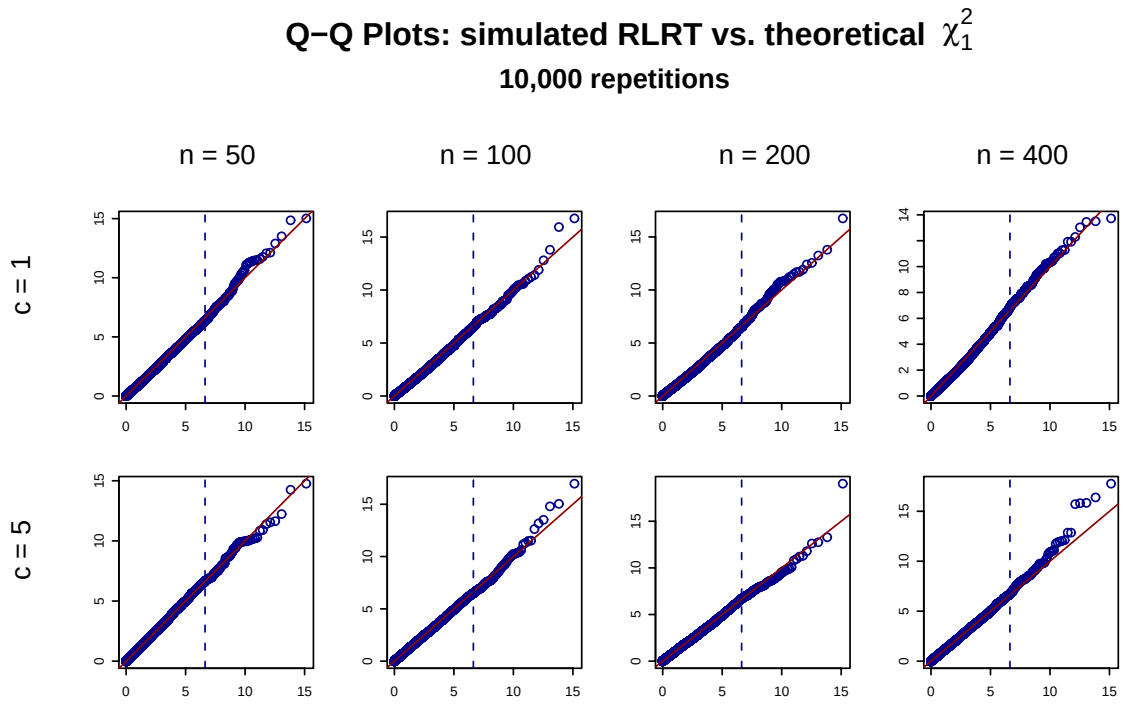

Figure 3: QQ plots of RLRT from simulations of 10,000 repetitions with $n = 50, 100, 200$ and 400. The vertical dashed lines are the 99th percentile of χ^2_1. Data are generated from $Y_t = \beta X_{t-1} + u_t$, where $\beta = 0$, $X_t = \alpha X_{t-1} + v_t$ with $\alpha = 1 - c/n$ with $c = 1,\ 5$, and $corr(v_t, u_t) = -.98$.

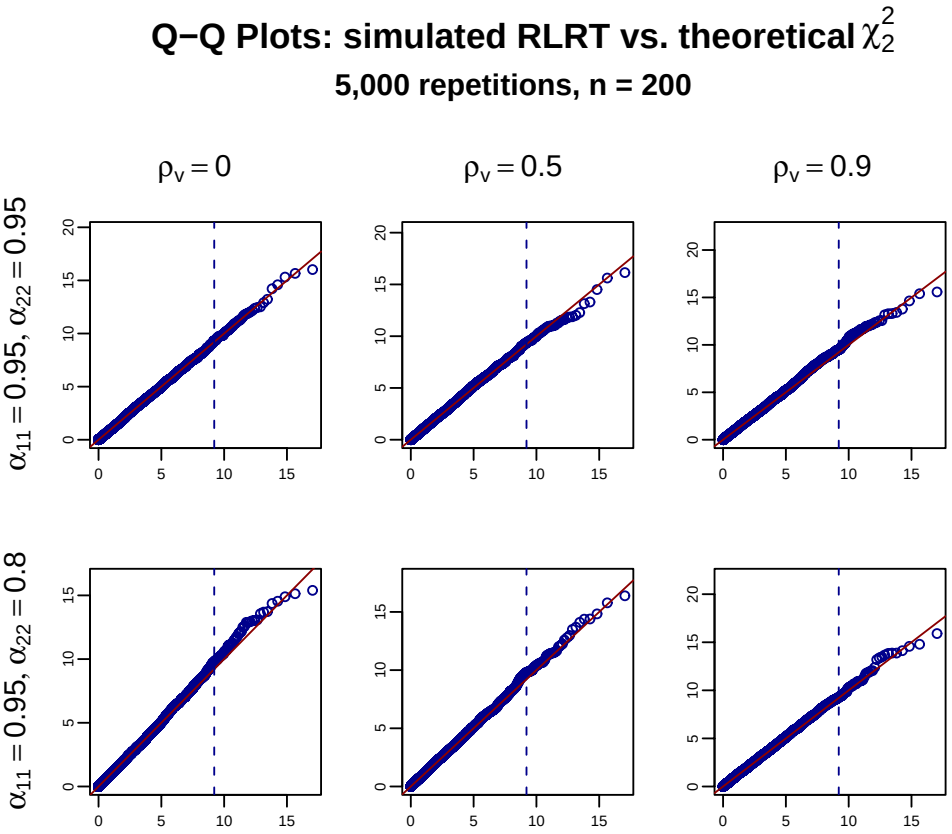

Figure 4: QQ plots of LRT from simulations of 5,000 repetitions with $n = 200$. The vertical dashed lines are the 99th percentile of χ^2_2. Data are generated from $Y_t = \boldsymbol{\beta}'\mathbf{X}_{t-1} + u_t$, where $\boldsymbol{\beta} = 0$, $\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{v}_t$ with $\mathbf{A} = \text{diag}(\alpha_{11}, \alpha_{22})'$ and $\rho_v = 0, .5$ and $.9$.