*Biometrika* (2016), **103**, 4, *pp.* 905–914
*Printed in Great Britain*
doi: 10.1093/biomet/asw048

# Nonnested model comparisons for time series

By T. S. McELROY

*Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, District of Columbia 20233-9100, U.S.A.*
tucker.s.mcelroy@census.gov

Summary

This paper addresses the topic of nonnested time series model comparisons. The main result is a central limit theorem for the likelihood ratio statistic when the models are nonnested and non-equivalent. The concepts of model equivalence and forecast equivalence, which are important for determining the parameter subset corresponding to the null hypothesis, are developed. The method is validated through a simulation study and illustrated on a retail time series.

*Some key words*: Akaike information criterion; Autoregressive integrated moving average model; Gaussian quadratic form; Generalized likelihood ratio; Goodness-of-fit.

## 1. Introduction

In using time series software it is common to be provided with a number of competing models that are identified from empirical criteria, or may be members of a default list of contenders. When several such models have adequate goodness-of-fit diagnostics, practitioners may need to make a final decision about the best model. Model comparison measures, such as the Akaike information criterion, can be utilized; if one is examining only two nested models, the likelihood ratio statistic can be used instead. This paper determines the limit theory for the likelihood ratio statistic for both nested and nonnested models and introduces the notions of forecast equivalence and model equivalence.

There is a substantial literature on model goodness-of-fit testing for integrated autoregressive moving average processes models, as well as on nested model comparisons, but less treatment of nonnested model comparisons. Li (2004) gives an overview of classical diagnostic tests of goodness-of-fit, while Paparoditis (2000) and Chen & Deo (2004) discuss frequency-domain tests of fit. Nested model comparisons are discussed in Taniguchi & Kakizawa (2000), while recent work on the nonnested case includes McElroy & Findley (2010) and Clark & McCracken (2014). The nonnested case is important in time series modelling; although model building typically proceeds by forward addition or backward deletion of lagged variables, competing models may ultimately be nonnested, and a disciplined approach to model selection requires a distribution theory for this case. For example, any moving average model is nested within a moving average model of greater order, whereas a nonnested comparison arises when assessing a moving average model against an autoregressive model. This latter type of comparison arises frequently when modelling time series.

Model selection among nested models is commonly achieved via either likelihood ratio statistics or information criteria; see Findley et al. (1998) for discussion. Both approaches take into

account the complexity of the models under consideration, in order to protect against overfitting. The likelihood ratio test is an example of a model comparison statistic and hence can provide Type I error rate control and $p$-values, whereas information criteria typically do not. However, the likelihood ratio test compares two models only, and the classical theory requires the models to be nested; information criteria, in contrast, can be used to compare several models at once. When only two models are being compared, a likelihood ratio statistic is preferable to information criteria, because spurious rejections can then be avoided. When comparing more than two models, a theory for multiple comparisons is needed, requiring a joint asymptotic result for various pairs of likelihood ratios. Our main result is that the likelihood ratio statistic for nonnested model comparisons obeys a central limit theorem under a null hypothesis that both models forecast equally well asymptotically.

Our limit theory for nonnested likelihood ratio statistics is related to the approach of Vuong (1989) and Rivers & Vuong (2002), who provide a central limit theory for a broad class of model comparison statistics. This generalizes to time series the comparison of Cox (1961, 1962). However, the limiting variance for the nonnested likelihood ratio statistic can be zero, which happens if and only if the two models are indistinguishable with respect to the true process. We elaborate on these points in the paper; we refer to the preceding scenario as model equivalence, and illustrate that this is a much narrower concept than the null hypothesis of equal forecast capability, or forecast equivalence.

## 2. Nesting, pseudo-true values, and equivalence

The concept of pseudo-true value is discussed in Gourieroux et al. (1984), but we mainly follow the treatment of Taniguchi & Kakizawa (2000). Consider a sequence $X = \{X_1, \ldots, X_n\}^{\mathrm{T}}$ of size $n$ from a stationary time series $\{X_t\}$ with autocovariance function $\gamma_h$ and mean $\tilde{\mu}$. We compare stationary models; if the data are nonstationary, we assume they have already been correctly differenced to stationarity. The chief context is the fitting of autoregressive moving average models to $X$ utilizing a Gaussian likelihood. However, one may apply our results to stationary nonlinear time series models that satisfy certain cumulant conditions.

Time series models for stationary data are often formulated by specifying a parametric family of spectral densities $f_\theta$, where the parameter $\theta$ is to be estimated from the data. We formulate models in terms of spectra, because this broad framework encompasses nonlinear and long-memory processes. For any $g : [-\pi, \pi] \to \mathbb{R}$ let $\langle g \rangle = (2\pi)^{-1} \int_{-\pi}^{\pi} g(\lambda)\, \mathrm{d}\lambda$ and $\gamma_h(g) = \langle g e^{ih\lambda} \rangle$. The notation $\Sigma(g)$ denotes the Toeplitz covariance matrix corresponding to the function $g$, having $(j,k)$th entry $\gamma_{j-k}(g)$. Let

$$\mathcal{D}(\theta, g) = \langle \log f_\theta \rangle + \langle g / f_\theta \rangle \tag{1}$$

denote the Kullback–Leibler divergence, which measures the discrepancy between a given model spectral density $f_\theta$ and a target spectrum $g$. Likelihood ratio testing is based upon the Whittle likelihood (Taniguchi & Kakizawa, 2000) $\mathcal{D}(\theta, I)$, where for $\lambda \in [-\pi, \pi]$ the mean-centred periodogram is $I(\lambda) = n^{-1} \left| \sum_{t=1}^{n} (X_t - \tilde{\mu})\, e^{-it\lambda} \right|^2$. Let $\theta(g) = \arg\min_\theta \mathcal{D}(\theta, g)$; such minimizers need not be unique, but we will focus on situations where a unique minimizer exists. Quasilikelihood estimation yields $\theta(I)$ as the minimizer of $\mathcal{D}(\cdot, I)$, and these are called the quasilikelihood estimators; they are asymptotically efficient for a broad class of processes when the model is correctly specified.

If $\tilde{f}$ denotes the spectral density of the true process, then $\mathcal{D}(\theta, \tilde{f})$ measures the Kullback–Leibler divergence between the model and the true spectrum. Minimizing this criterion with respect to $\theta$ yields $\theta(\tilde{f})$, the so-called pseudo-true value. When a model is correctly specified, pseudo-true values correspond to the true parameters; more generally, they correspond to the probability limits of the quasilikelihood estimators. When the innovation variance $\sigma^2$ is a separate parameter, the model is said to be separable, and we write $f_\theta(\lambda) = f_\vartheta(\lambda)\,\sigma^2$ for $\theta = (\vartheta^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$ consisting of $r+1$ components. The innovation-free spectral density $f_\vartheta$ is obtained from $f_\theta$ by setting $\sigma^2 = 1$. Henceforth we focus on separable models; the Supplementary Material provides extensions to the nonseparable case.

In goodness-of-fit testing, we are interested in whether the data belong to a model $\mathcal{F} = \{(f_\vartheta, \sigma^2) : \vartheta \in \Theta, \sigma^2 \in (0, \infty)\}$, with $\Theta$ an $r$-dimensional space. The true spectrum of the process is some unknown $\tilde{f}$, and we seek to discern whether $\tilde{f} \in \mathcal{F}$ or not; if $\tilde{f} \notin \mathcal{F}$ then we can compute the pseudo-true values $\{\vartheta(\tilde{f}), \sigma^2(\tilde{f})\}$, which have the property that $f_{\vartheta}(\tilde{f})\,\sigma^2(\tilde{f})$ is as close as possible to $\tilde{f}$ in the Kullback–Leibler sense. By profiling the Kullback–Leibler divergence with respect to $\sigma^2$ (McElroy & Findley, 2015), we obtain the asymptotic one-step-ahead forecast mean square error arising from forecasting with the model $f_\theta$:

$$\mathcal{J}(\vartheta, \tilde{f}) = \langle \tilde{f}/f_\vartheta \rangle.$$

The minimizer is $\vartheta(\tilde{f})$, and the minimal value is $\sigma^2(\tilde{f}) = \langle \tilde{f}/f_{\vartheta(\tilde{f})} \rangle$.

With model comparisons we wish to know which of two models performs better according to some criterion such as asymptotic mean square error of one-step-ahead forecasting. Set $\phi = (\varphi^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$ and let the second model be denoted by $\mathcal{G} = \{(f_\varphi, \sigma^2) : \varphi \in \Phi, \sigma^2 \in (0, \infty)\}$ with $\Phi$ some $s$-dimensional space. While it is possible that $\tilde{f} \in \mathcal{F} \cap \mathcal{G}$, the true process may belong to neither model. The pseudo-true values for both models are denoted by $\{\vartheta(\tilde{f}), \sigma^2(\tilde{f})\}$ and $\{\varphi(\tilde{f}), \sigma^2(\tilde{f})\}$.

Given two models $\mathcal{F}$ and $\mathcal{G}$, we say that $\mathcal{F}$ is nested in $\mathcal{G}$ if for any $\theta$ there exists some $\phi$ such that $f_\phi \equiv f_\theta$. In this case we write $\mathcal{F} \subset \mathcal{G}$. If neither model is nested in the other, they are said to be nonnested. The model spectral densities evaluated at their pseudo-true values need not be equal, i.e., $f_{\theta(\tilde{f})}$ need not equal $f_{\phi(\tilde{f})}$. The definition of model equivalence is that $f_{\theta(\tilde{f})} = f_{\phi(\tilde{f})}$, which is a strong property. The overlapping property, by definition, holds if $\tilde{f} \in \mathcal{F} \cap \mathcal{G}$, which implies model equivalence. The definition of forecast equivalence is $\mathcal{J}\{\vartheta(\tilde{f}), \tilde{f}\} = \mathcal{J}\{\varphi(\tilde{f}), \tilde{f}\}$. For separable models forecast equivalence is the same as divergence equivalence, which is defined via $\mathcal{D}\{\theta(\tilde{f}), \tilde{f}\} = \mathcal{D}\{\phi(\tilde{f}), \tilde{f}\}$. Model equivalence implies forecast equivalence, but examples given in the Supplementary Material show that the converse is false; the following result gives a sufficient condition for forecast equivalence to imply model equivalence.

PROPOSITION 1. *Suppose that two nested models are separable and their pseudo-true values exist uniquely. If the models are forecast equivalent, then they are also model equivalent.*

The condition of Proposition 1 is hard to verify in some cases, but for all autoregressive models the pseudo-true values are unique, being given as the solution to the Yule–Walker equations. The result simplifies the discussion of model comparisons in the nested case, because we can focus on model equivalence.

## 3. Methodology for likelihood ratio statistics

The likelihood ratio statistic can be used to test forecast equivalence in the overlapping case, assuming the models are separable; the appeal, from a practical standpoint, is that no additional computation is necessary, once the practitioner has fitted both models and has recorded the values of their maximized loglikelihoods. Many other tests of forecast equivalence have been considered: Vuong (1989) developed a general approach to equivalence, which has been followed by other papers, e.g., West (1996), West & McCracken (1998), McCracken (2000, 2007), and Clark & McCracken (2001, 2005). Much of this literature considers an out-of-sample evaluation period, which is distinct from the span of data utilized to fit the model; the likelihood ratio statistic, in contrast, is an in-sample device, in that the full span of data is utilized to both fit the model and evaluate one-step-ahead forecast performance. Literature on the in-sample approach includes Findley (1991), Diebold & Mariano (1995), Rivers & Vuong (2002), McElroy & Holan (2009), and McElroy & Findley (2010).

As discussed in Taniguchi & Kakizawa (2000), the null hypothesis typically considered for a likelihood ratio test is that the nested model is correct; because the models are nested by assumption, this is the same as saying that the true process is overlapped by both models. This is stronger than simply assuming model equivalence. Suppose that we wish to compare separable models according to asymptotic mean square error of one-step-ahead forecasting. This is equivalent to comparing the Kullback–Leibler divergences of each model to the true spectrum. It is natural to test this by substituting empirical estimators for unknown quantities and computing the differences. This yields the definition

$$\text{LR} = \mathcal{D}\{\theta(I), I\} - \mathcal{D}\{\phi(I), I\}, \tag{2}$$

with the convention that the divergence of the second model $\mathcal{G}$ is subtracted from the divergence of the first model $\mathcal{F}$. When the models are nested, we always write the second as the nesting model and the first as the nested model; then LR must be positive, because the forecast performance of the nesting model will always be superior. When the models are nonnested, no such superiority is required and LR can be either positive or negative.

Theorem 1 below generalizes likelihood ratio theory for nested models in two main ways: the result discriminates between the hypotheses of forecast equivalence and model equivalence, and also considers nonnested models. The theory of likelihood ratio statistics discussed in Taniguchi & Kakizawa (2000) is derived under the Hosoya–Taniguchi conditions (Hosoya & Taniguchi, 1982), which we employ as well. These conditions require that the process $\{X_t\}$ be a causal filter of a higher-order martingale difference, which allows nonlinearity in the data process. We also require that the fourth-order cumulant function be identically zero, as otherwise the variances in some of the underlying central limit theory will have altered expressions.

We define the Hessian of $\mathcal{D}(\theta, g)$ as $M(\theta, g) = \nabla\nabla^{\mathrm{T}}\mathcal{D}(\theta, g)$, where the gradient is taken with respect to $\theta$. The generalized Fisher information matrix is defined as $H(\theta, g) = \langle \nabla \log f_\theta \, \nabla^{\mathrm{T}} \log f_\theta \, g^2 f_\theta^{-2} \rangle$, and equals the usual Fisher information matrix when $g = f_\theta$. Similarly, these matrices are defined for model $\mathcal{G}$, by substituting $\phi$ for $\theta$. We assume that the spectral density is sufficiently smooth, so that these matrices are well-defined. Elementary calculation yields

$$M(\theta, g) = \left\langle \nabla\nabla^{\mathrm{T}} f_\theta \, \left(1 - g f_\theta^{-1}\right) f_\theta^{-1} \right\rangle + \left\langle \nabla f_\theta \nabla^{\mathrm{T}} f_\theta \, \left(2 g f_\theta^{-1} - 1\right) f_\theta^{-2} \right\rangle,$$

so $M(\theta, f_\theta) = H(\theta, f_\theta)$, though in general this is false for choices of $g$ other than $g = f_\theta$. A third type of matrix is also needed to describe the nonnested case:

$$K(\theta, \phi, g) = \begin{bmatrix} \langle \nabla \log f_\theta \, \nabla^{\mathrm{T}} \log f_\theta \, g^2 f_\theta^{-1} f_\phi^{-1} \rangle & \langle \nabla \log f_\theta \, \nabla^{\mathrm{T}} \log f_\phi \, g^2 f_\theta^{-1} f_\phi^{-1} \rangle \\ \langle \nabla \log f_\phi \, \nabla^{\mathrm{T}} \log f_\theta \, g^2 f_\theta^{-1} f_\phi^{-1} \rangle & \langle \nabla \log f_\phi \, \nabla^{\mathrm{T}} \log f_\phi \, g^2 f_\theta^{-1} f_\phi^{-1} \rangle \end{bmatrix}.$$

The upper left and lower right submatrices of $K(\theta, \phi, g)$ are just the generalized Fisher information matrices for each model when they are equivalent.

THEOREM 1. *Suppose that the process $\{X_t\}$ has finite fourth moments, that conditions* (HT1)–(HT6) *of Taniguchi & Kakizawa (2000) hold, and that the fourth-order cumulants are zero. Also suppose that the pseudo-true values $\phi(\tilde{f})$ and $\theta(\tilde{f})$ exist uniquely in the interior of their respective parameter spaces, and that $M\{\phi(\tilde{f})\}$ and $M\{\theta(\tilde{f})\}$ are invertible.*

(i) *Assume that the models are nested and forecast equivalent. Let the nesting model be $\mathcal{G}$, so that the likelihood ratio statistic is always nonnegative. Then for $Z$ a zero-mean normal random vector with identity covariance matrix $I_s$ and likelihood ratio statistic given by* (2), *$n$* LR *converges in distribution as $n \to \infty$ to $Z^{\mathrm{T}} Q Z$, where*

$$Q = H\{\phi(\tilde{f}), \tilde{f}\}^{1/2} \left( M\{\phi(\tilde{f}), \tilde{f}\}^{-1} - \begin{bmatrix} M\{\theta(\tilde{f}), \tilde{f}\}^{-1} & 0 \\ 0 & 0 \end{bmatrix} \right) H\{\phi(\tilde{f}), \tilde{f}\}^{1/2}.$$

*In the case of overlapping models,*

$$Q = I_s - H\{\phi(\tilde{f}), \tilde{f}\}^{1/2} \begin{bmatrix} H\{\theta(\tilde{f}), \tilde{f}\}^{-1} & 0 \\ 0 & 0 \end{bmatrix} H\{\phi(\tilde{f}), \tilde{f}\}^{1/2},$$

*and the limiting distribution is $\chi^2_{s-r}$.*

(ii) *Assume that the models are nonnested and forecast equivalent. If they are not model equivalent, then $n^{1/2}$* LR *converges in distribution to a zero-mean random variable with variance $V = 2\langle (f^{-1}_{\theta(\tilde{f})} - f^{-1}_{\phi(\tilde{f})})^2 \tilde{f}^2 \rangle$, which is consistently estimated by $\hat{V} = \langle (f^{-1}_{\theta(I)} - f^{-1}_{\phi(I)})^2 I^2 \rangle$. Otherwise, if model equivalence holds, then $n$* LR *converges in distribution as $n \to \infty$ to $Z^{\mathrm{T}} Q Z$, where $Z$ is a zero-mean normal random vector with identity covariance matrix $I_{r+s}$ and*

$$Q = K\{\theta(\tilde{f}), \phi(\tilde{f}), \tilde{f}\}^{1/2} \begin{bmatrix} -M\{\theta(\tilde{f}), \tilde{f}\}^{-1} & 0 \\ 0 & M\{\phi(\tilde{f}), \tilde{f}\}^{-1} \end{bmatrix} K\{\theta(\tilde{f}), \phi(\tilde{f}), \tilde{f}\}^{1/2}. \quad (3)$$

*Remark* 1. The overlapping case of (i) can be found in Taniguchi & Kakizawa (2000), while (ii), assuming the models are not equivalent, can be found in McElroy & Findley (2010). The other cases of Theorem 1 are new. The case where nonnested models are forecast equivalent but not model equivalent can arise in practice, and is therefore pertinent; see the Supplementary Material for examples. Likewise, the case where nested models are model equivalent but not overlapping is important but has not been considered previously.

In the nonnested case, $V = 0$ if and only if model equivalence holds. Thus $\hat{V}$ could be used as a separate test statistic for model equivalence. While $\hat{V}$ converges in probability to $V$ when only forecast equivalence holds, the statistic is bounded in probability of order $n^{-1}$ when model equivalence is true as well, and the limiting distribution is fairly complicated. In this case, $n\hat{V}$

converges in distribution to a Gaussian quadratic form, so that $n^{1/2}\,\mathrm{LR}\,\hat{V}^{-1/2}$ has a nondegenerate limit whether or not model equivalence holds.

Corollary 1. *Under the same general assumptions as in Theorem 1, if the models are nonnested and forecast equivalent, then $n^{1/2}\,\mathrm{LR}\,\hat{V}^{-1/2}$ converges in distribution, where the limit is either a standard normal random variable or $Z^{\mathrm{T}}QZ\ (2\,Z^{\mathrm{T}}Q^2Z)^{-1/2}$, depending on whether the models are not equivalent or equivalent respectively, where $Z$ is a zero-mean normal random vector with identity covariance matrix $I_{r+s}$ and $Q$ is defined in (3).*

The Gaussian quadratic forms $Z^{\mathrm{T}}QZ$ and $Z^{\mathrm{T}}Q^2Z$ have distributions characterized by the eigenvalues of $Q$, and $\chi^2$ or $\gamma$ distributions can result when certain conditions on the trace of powers of $Q$ hold, as described in Tziritas (1987). In practice such conditions are difficult to verify, and one might instead use subsampling techniques (Politis et al., 1999) to approximate the sampling distribution; this would be somewhat expensive, because the models would have to be refitted on each subspan of data.

In most applications, nested models will be compared under the overlapping assumption, so that the classic $\chi^2$ distribution can be used. However, it is a fallacy to conclude that the models are not model equivalent when the $\chi^2$ test rejects; we only know that the overlapping condition is rejected, whereas the models may still be equivalent. In order to reject the weaker hypothesis of model equivalence, more evidence is needed and the critical values of $Z^{\mathrm{T}}QZ$ with $Q$ given in the first part of Theorem 1 are required. Thus, the likelihood ratio test for nested models with $\chi^2$ critical values is useful for testing model goodness-of-fit with respect to the particular alternative given by the nesting model, but is not informative as a model comparison test statistic unless we have some prior reason to believe that model equivalence is impossible.

In the special case that the innovation variance is the final parameter we have $\theta = (\vartheta^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$, and the optimizer for the innovation variance is $\sigma^2(g) = \langle f_{\vartheta(g)}^{-1} g\rangle$. Then the divergence becomes

$$\mathcal{D}\{\theta(g), g\} = \log\sigma^2(g) + \langle f_{\vartheta(g)}^{-1} g\rangle/\sigma^2(g) = \log\langle f_{\vartheta(g)}^{-1} g\rangle + 1,$$

and the likelihood ratio statistic is just the difference in the logarithms of the innovation variance estimators from the two models. In the nonnested case, the variance estimate is computed via

$$\hat{V} = \left\langle \left(f_{\theta(I)}^{-1} - f_{\phi(I)}^{-1}\right)^2 I^2 \right\rangle = R^{\mathrm{T}}\,\Sigma(f_{\theta(I)}^{-2})\,R - 2R^{\mathrm{T}}\,\Sigma(f_{\theta(I)}^{-1} f_{\phi(I)}^{-1})\,R + R^{\mathrm{T}}\,\Sigma(f_{\phi(I)}^{-2})\,R.$$

Here $R$ is the vector of sample autocovariances $R^{\mathrm{T}} = [\hat{\gamma}_{1-n}, \ldots, \hat{\gamma}_0, \ldots, \hat{\gamma}_{n-1}]$, where $\hat{\gamma}(h) = \langle I e^{ih\cdot}\rangle$, and the Toeplitz matrices $\Sigma$ have dimension $2n-1$. We only need to compute inverse Fourier transforms for $f_{\theta(I)}^{-2}, f_{\theta(I)}^{-1} f_{\phi(I)}^{-1}$, and $f_{\phi(I)}^{-2}$.

A final point is that the objective function commonly used is not the Whittle likelihood $\mathcal{D}(\theta, I)$ of (1), but rather $-2$ times the log Gaussian likelihood, typically stripped of normalizing constants. This objective function is written

$$\mathcal{L}(\theta; X) = X^{\mathrm{T}}\,\Sigma(f_\theta)^{-1}\,X + \log|\Sigma(f_\theta)|. \tag{4}$$

As discussed in Taniguchi & Kakizawa (2000), $\mathcal{L}(\theta, X)/n \approx \mathcal{D}(\theta, I)$, and parameter estimators from either (1) and (4) behave similarly in large Gaussian samples according to results of Dahlhaus & Wefelmeyer (1996). In empirical work, the assumptions of the theory should be checked by

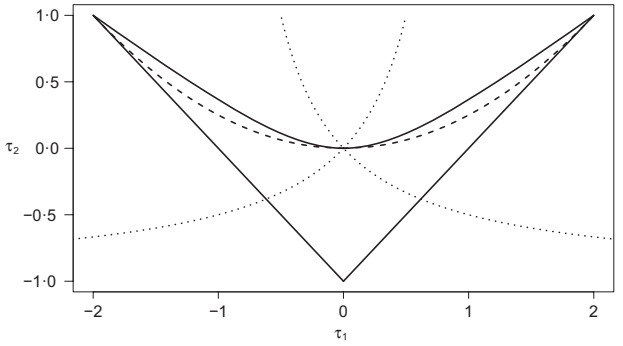


Fig. 1. The invertibility region for the order-two moving average process is within the solid lines, with dashed lines demarcating real roots (below) from complex roots (above). The forecast equivalence set is denoted by dotted lines. The origin is the model equivalence set.

applying normality and stationarity tests to the model residuals; if using the Gaussian likelihood (4) it is important to check that the data is Gaussian, while for nonlinear time series the Whittle likelihood can be used.

## 4. Simulation study

We consider fitting an order-one autoregressive model and a gap order-two autoregressive model, which means that the first autoregressive coefficient is constrained to be zero. The pseudo-true values are $\vartheta(\tilde{f}) = \rho_1$ for the first model and $\varphi(\tilde{f}) = \rho_2$ for the second model, where $\rho_1$ and $\rho_2$ are the lag-one and lag-two autocorrelations of the true process. These two models are nonnested, and the smallest model nesting both of them is the unconstrained order-two autoregressive model. The asymptotic mean square one-step-ahead forecast error for these models is

$$\mathcal{J}\{\vartheta(\tilde{f}),\tilde{f}\} = \gamma_0\left(1-\rho_1^2\right) \qquad \mathcal{J}\{\varphi(\tilde{f}),\tilde{f}\} = \gamma_0\left(1-\rho_2^2\right),$$

with $\gamma_0$ the variance of the process. Forecast equivalence holds if and only if $\rho_1^2 = \rho_2^2$, whereas model equivalence holds if and only if $\rho_1 = 0 = \rho_2$.

To study size and power, let the true process be an order-two moving average with moving average polynomial $1+\tau_1 B+\tau_2 B^2$, and unit innovation variance. This polynomial has real roots if $\tau_1^2 \geqslant 4\tau_2$ and the roots have magnitude greater than one, implying that the process is invertible, so long as $|\tau_1| - 1 \leqslant \tau_2 \leqslant \{(1+2\tau_1^2)^{1/2} - 1\}/2$. This region is depicted in Fig. 1. In terms of the moving average parameters, forecast equivalence means that $\tau_1 = \pm\tau_2/(1+\tau_2)$, whereas model equivalence means that $\tau_1 = \tau_2 = 0$.

To assess size, we simulated 1000 replications of a Gaussian moving average process from the forecast equivalence set of sample sizes $n = 50$ through $n = 300$, fitting both models and normalizing via $\hat{V}$, as described in Corollary 1. We took a range of $\tau_2$ values, from 0 down to $(-3+5^{1/2})/2$, and $\tau_1 = \tau_2/(1+\tau_2)$. We then measured the proportion of rejections of the null hypothesis, utilizing a two-sided test and normal quantiles at the 5% and 10% levels. This procedure is not asymptotically correct on the model equivalence set, and in Table 1 we see deterioration in size for processes that are forecast equivalent and close to being model equivalent. The results agree with the theory, as size improves with increasing $n$, and also improves as the distance from the origin is increased.

Table 1. *Finite-sample size* (%) *of likelihood ratio tests, at nominal* 5% *(left cell) and* 10% *(right cell)*

| | Sample size | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau_2$ | 50 | | 100 | | 150 | | 200 | | 250 | | 300 | |
| −0·032 | 0·0 | 0·6 | 0·2 | 0·6 | 0·1 | 0·3 | 0·0 | 0·4 | 0·0 | 0·5 | 0·0 | 0·3 |
| −0·082 | 0·6 | 2·0 | 0·3 | 1·6 | 0·1 | 1·3 | 0·3 | 1·8 | 0·1 | 2·4 | 0·6 | 3·6 |
| −0·132 | 1·0 | 2·6 | 1·2 | 4·0 | 1·6 | 4·2 | 1·5 | 5·6 | 1·8 | 6·3 | 2·8 | 7·4 |
| −0·182 | 1·4 | 3·5 | 2·8 | 6·0 | 2·0 | 7·1 | 3·0 | 8·5 | 3·3 | 8·9 | 4·1 | 8·4 |
| −0·232 | 2·6 | 6·8 | 4·5 | 8·6 | 3·5 | 9·7 | 5·0 | 9·3 | 4·4 | 9·8 | 4·9 | 9·6 |
| −0·282 | 4·0 | 9·3 | 4·4 | 9·2 | 4·1 | 9·1 | 4·5 | 10·3 | 4·8 | 10·8 | 5·4 | 9·9 |
| −0·332 | 5·6 | 11·5 | 4·2 | 9·1 | 4·6 | 9·4 | 4·4 | 9·5 | 4·8 | 9·6 | 5·5 | 10·8 |
| −0·382 | 5·6 | 10·4 | 4·3 | 11·0 | 5·3 | 11·1 | 4·4 | 9·8 | 4·4 | 8·8 | 4·7 | 10·0 |

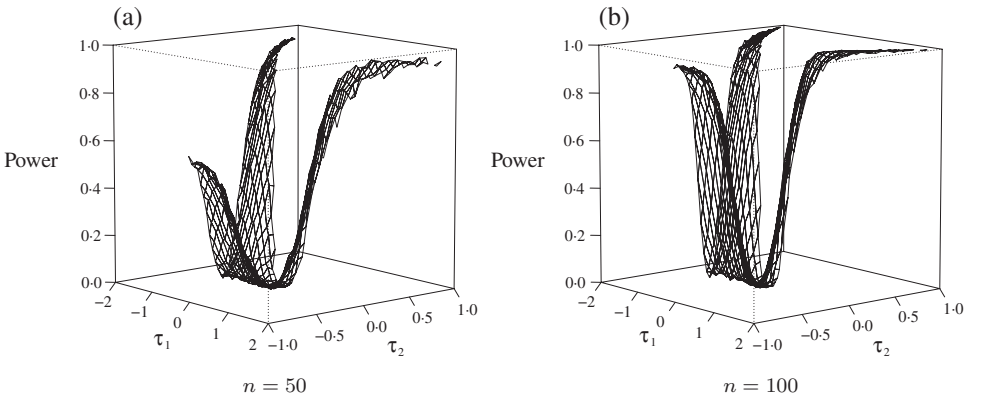


Fig. 2. Power surfaces for the likelihood ratio test statistic, comparing an order-one autoregressive model and a gap order-two autoregressive model over the space of invertible order-two moving average processes. Type I error rate is 5%, with sample sizes $n = 50$ and $n = 100$.

To investigate power we simulated the same moving average process over the entire invertibility region, incrementing the moving average parameters in steps of 0·05, and measured the incidence of rejections using normal quantiles at the 5% and 10% levels. This mimics how the procedure would be used in practice. Figure 2 displays the power surfaces for sample sizes $n = 50$ and $n = 100$ and $\alpha = 5\%$. The bird-like shape indicates higher power at the three corners/cusps of the invertibility region in Fig. 1. Again, on the model equivalence set this procedure is not consistent, and we may expect power to deteriorate in a neighbourhood of the origin. Along the forecast equivalence set the percent of rejections is quite low, as we expect, and this also holds for the model equivalence set, i.e., the origin. Results based upon simulations for higher sample sizes and $\alpha = 10\%$ were qualitatively similar to those in Fig. 2.

## 5. Analysis of electricity series

We study the series of Electronics and Appliance Stores over the pre-Recession span 1992–2007, which is published by the U.S. Census Bureau from the Monthly Retail Sales Survey. The series required a log transformation, and all types of fixed effects, such as outliers, Easter and trading day, were removed before further analysis. We selected a seasonal autoregressive moving average model, for which normality and stationarity of the model residuals was not rejected.

However, the Ljung–Box statistics indicate some potential residual correlation, as discussed in McElroy & Monsell (2014). In that paper a gap model was proposed, whereby the differenced series is modelled with an order-13 moving average, where all coefficients but the last one are constrained to be zero. Letting $\{X_t\}$ correspond to seasonal and nonseasonal differencing of the data process, the scaled likelihood (4) is $-1189{\cdot}51$ for the first model

$$(1 - 0{\cdot}37B + 0{\cdot}14B^2)X_t = (1 - 0{\cdot}46B)(1 - 0{\cdot}24B^{12})\epsilon_t,$$

where the log innovation variance is $-7{\cdot}65$. On the other hand, the more parsimonious order-13 moving average model has scaled likelihood $-1188{\cdot}94$, with a significant 13th moving average parameter of $0{\cdot}26$ and log innovation variance $-7{\cdot}65$. The model results were also adequate. McElroy & Monsell (2014) indicate that this model is superior to the identified seasonal autoregressive moving average via analysis of each model's residuals, but here we directly compare their likelihoods, which are extremely similar.

Now we can determine the exact conditions of model equivalence by equating the two spectral densities and solving. Using the uniqueness of spectral factorization, we find that the two autoregressive parameters of the first model must be zero if model equivalence is to hold, and deduce that the models are not equivalent unless they are both white noise. If we constrain the 13th moving average coefficient of the second model to be in the interval $(0, 1)$, then we ensure that the second model cannot be a white noise, implying that the models must not be equivalent. Hence we can safely use the standard normal critical values. The normalized likelihood ratio statistic is $n^{1/2}\,\mathrm{LR}\,\hat{V}^{-1/2} = -0{\cdot}017$, which is an insignificant result. Although the first model is favoured, it is not significantly superior.

## Acknowledgement

The author is grateful for helpful comments from three referees and the editor, as well as William Bell, Chris Blakely, Neil Ericsson, David Findley, Thomas Trimbur, and participants in the 2013 National Bureau of Economic Research/National Science Foundation Time Series Conference.

## Supplementary material

Supplementary material available at *Biometrika* online includes background, extensions, proofs, and a discussion of encompassing.

## References

Chen, W. W. & Deo, R. S. (2004). A generalized portmanteau goodness-of-fit test for time series models. *Economet. Theory* **20**, 382–416.

Clark, T. E. & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *J. Economet.* **105**, 85–110.

Clark, T. E. & McCracken, M. W. (2005). Evaluating long horizon forecasts. *Economet. Rev.* **24**, 369–404.

Clark, T. E. & McCracken, M. W. (2014). Tests of equal forecast accuracy for overlapping models. *J. Appl. Economet.* **29**, 415–30.

Cox, D. R. (1961). Test of separate families of hypotheses. In *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, vol. 1, J. Neyman, ed. Berkeley: University of California Press, pp. 105–23.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B **24**, 406–24.

Dahlhaus, R. & Wefelmeyer, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Ann. Statist.* **24**, 952–74.
Diebold, F. X. & Mariano, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econ. Statist.* **13**, 253–63.
Findley, D. F. (1991). Convergence of finite multistep predictors from incorrect models and its role in model selection. *Note di Mat. XI*, 145–55.
Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C. & Chen, B. C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program (with Discussion). *J. Bus. Econ. Statist.* **16**, 127–77
Gourieroux, C., Monfort, A. & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* **52**, 681–700.
Hosoya, Y. & Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *Ann. Statist.* **10**, 132–53.
Li, W. K. (2004). *Diagnostic Checks in Time Series*. New York: CRC Press.
McCracken, M. W. (2000). Robust out of sample inference. *J. Economet.* **99**, 195–223.
McCracken, M. W. (2007). Asymptotics for out of sample tests of Granger causality. *J. Economet.* **140**, 719–52.
McElroy, T. S. & Findley, D. F. (2010). Discerning between models through multi-step ahead forecasting errors. *J. Statist. Plan. Infer.* **140**, 3655–75.
McElroy, T. S. & Findley, D. F. (2015). Fitting constrained vector autoregression models. In *Empirical Economic and Financial Research – Theory, Methods, and Practice*, J. Beran, Y. Feng & H. Hebbel, eds. New York: Springer, pp. 451–70.
McElroy, T. S. & Holan, S. H. (2009). A local spectral approach for assessing time series model misspecification. *J. Mult. Anal.* **100**, 604–21.
McElroy, T. S. & Monsell, B. C. (2014). The multiple testing problem for Box–Pierce statistics. *Electron. J. Statist.* **8**, 497–522.
Paparoditis, E. (2000). Spectral density based goodness-of-fit tests for time series models. *Scand. J. Statist.* **27**, 143–76.
Politis, D. N., Romano, J. P. & Wolf, M. (1999). *Subsampling*. New York: Springer.
Rivers, D. & Vuong, Q. H. (2002). Model selection tests for nonlinear dynamic models. *Economet. J.* **5**, 1–39.
Taniguchi, M. & Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series.* New York: Springer.
Tziritas, G. G. (1987). On the distribution of positive-definite Gaussian quadratic forms. *IEEE Trans. Info. Theory* **33**, 895–906.
Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–33.
West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica* **64**, 1067–84.
West, K. D. & McCracken, M. W. (1998). Regression-based tests of predictive ability. *Int. Econ. Rev.* **39**, 817–40.

[*Received May* 2015. *Revised August* 2016]