Electronic Journal of Statistics
Vol. 8 (2014) 497–522
ISSN: 1935-7524
DOI: 10.1214/14-EJS892

The multiple testing problem for Box-Pierce statistics*

Tucker McElroy and Brian Monsell

Center for Statistical Research and Methodology
U.S. Census Bureau, 4600 Silver Hill Road
Washington, D.C. 20233-9100
e-mail: tucker.s.mcelroy@census.gov

Abstract: We derive the exact joint asymptotic distribution for multiple Box-Pierce statistics, and use these results to determine appropriate critical values in joint testing problems of time series goodness-of-fit. By sequentially testing at various lags, we can identify specific problems with a model, and identify superior models. A novel α-rationing scheme, motivated by the sequence of conditional probabilities for the statistical tests, is developed and implemented. This method can be used to produce critical values and p-values both for each step of the sequential testing procedure, and for the procedure as a whole. Efficient computational algorithms are discussed. Simulation studies assess the impact of finite samples on the real Type I error. It is also demonstrated empirically that the conventional χ^2 critical values for the Box-Pierce statistics are too small, with a Type I error rate greater than the nominal; the new method does not suffer from this defect, and allows for more rigorous model-building. We illustrate on several time series how model defects can be identified and ameliorated.

Keywords and phrases: ARIMA models, Ljung-Box statistic, time series residuals.

Received November 2013.

1. Introduction

The most popular time series goodness-of-fit diagnostic test statistics are the portmanteau Q statistics introduced by [3] and extended by [17]. The original idea is based on ascertaining model goodness-of-fit via examination of the correlation structure of time series model residuals. The presence of residual autocorrelation can be measured through the sample autocorrelation function – denoted by $\widehat{\rho}_k$ for $k \geq 0$ – of these residuals, and the Box-Pierce (BP) and Ljung-Box (LB) statistics are constructed from a cumulation of the square of this function over various time lags:

$$Q_{BP} = n \sum_{k=1}^{m} \widehat{\rho}_k^2 \qquad Q_{LB} = n(n+2) \sum_{k=1}^{m} \widehat{\rho}_k^2/(n-k) \tag{1}$$

*This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

defines the BP and LB Q statistics. Here m is chosen by the practitioner, and may be viewed as a fixed integer in the asymptotic analysis.

Despite many variants (see the treatment in [16, 24, 14, 15, 25, 11, 12] and [10]), the original statistics remain quite popular, being featured in commonly used statistical software packages such as X-12-ARIMA [6], TRAMO-SEATS [18], and STAMP [13]. The reason for this popularity is more than just cultural inertia; these statistics are intuitive, easy to calculate, and simple to interpret. Various improvements to the initial proposed test statistics have focused, over four decades, on improving finite sample size and power. However, an important but neglected topic is the issue of multiple testing, since typically many of these test statistics are used concurrently.

For example, in X-12-ARIMA the default diagnostic output produces Ljung-Box Q statistics (hereafter LBs) at a maximum lag up to 24, so that typically over 20 test statistics are presented simultaneously. Now for 20 independent tests of the same null hypothesis, with a nominal Type I error rate of .05, one is extremely likely to obtain at least one significant statistic by chance alone. (But if all the tests were fully dependent, then the rejection rate would be identical with the nominal of .05.) The primary aim of this paper is to use the joint asymptotic distribution of the sample autocorrelations of time series residuals to derive the joint cumulative distribution function (cdf) of BPs and LBs[1]. This mathematical result, along with a practical, fast technique for computing the joint cdf, allows one to determine the asymptotic p-values and critical values associated with a sequence of test statistics. Application of the methodology should be helpful in mitigating an over-abundance of Type I errors, along with the unfortunate behavior of "data-snooping" with ARIMA models.

The reason that many time series software packages provide multiple BP/LB statistics is that there is understood, among practitioners, to be some benefit to examining several such statistics simultaneously (or sequentially); also see the discussion in [12]. If a BP/LB statistic at lag m is significant, and all statistics at lower lags are *not* significant, then the problem with the model manifests at the mth lag of the residuals – this point is established rigorously in this paper. Hence, examination of all BP/LB statistics at a variety of lags, in a sequential manner, can be helpful towards ascertaining specific model defects, providing the practitioner with guidance about how to improve the model. We show that statistical power chiefly arises from alternatives corresponding to time series residuals with significant autocorrelation at lag m.

Prior results on the joint asymptotic limit of the sample autocovariances and autocorrelations of time series residuals, accounting for parameter estimation uncertainty, include [23, 4, 24, 25], and [12]. Our formulation extends beyond the ARMA and SARMA classes of models treated by the above authors, and our results include the important case of model misspecification. Also the derivations in [3] assume that the moving average coefficients decay as sample size increases, which effectively ensures that the limiting covariance matrix of the residual

[1] A similar result for the sample autocorrelations was used in [12] to derive joint results for alternative portmanteau statistics, that are essentially modified LBs.

autocorrelations is idempotent; our results dispense with this assumption, which we have found to be untenable for small samples and low LB lags.

The second major contribution of this paper (Section 3) is a novel scheme for multiple testing apportionment of Type I error (called α-rationing; cf. [28] and later literature) appropriate for sequences of test statistics. Then we assess (in Section 4) the actual Type I error rate in several ways. We conduct simulations from finite length Gaussian time series, and look at empirical Type I error rates based on conventional χ^2 critical values, as well as the new joint critical values discussed in Section 3. All proofs are in the Appendix, which also contains some supplementary material on time series residual processes.

2. Asymptotic theory for BP and LB statistics

2.1. Models and residuals

This section presents the main theoretical results of the paper. Theorem 1, along with its corollaries, provides joint asymptotic convergence results for the sample autocorrelations of time series model residuals, as well as for Q statistics, assuming a very broad class of time series models are utilized. Previous results, such as [3], have focused on ARIMA or SARIMA models, relying on the causal representation to define time series residuals. Here we instead suppose that a model is formulated via parametrizing a family of spectral densities; this generalizes ARMA models, and allows us to include unobserved components models as well.

Consider a sample of size n from a stationary time series, denoted $X = \{X_1, X_2, \ldots, X_n\}'$. (If the raw data is nonstationary, we assume it has been correctly differenced to stationarity already.) The series may have a nonzero mean $\widetilde{\mu}$. Time series models for stationary data are formulated by specifying a family of spectral densities f_θ that depend on a parameter θ, which is to be estimated from the data. We formulate models in terms of spectra, rather than using a Wold decomposition or State Space Form (SSF), because this is the most general treatment possible – including non-linear processes and long memory processes that cannot be represented in SSF.

We use the notation $\Sigma(f_\theta)$ to denote the Toeplitz covariance matrix corresponding to the model spectrum f_θ, i.e., with jkth entry given by $\gamma_{j-k}(f_\theta)$, the lag $j-k$ autocovariance (acv). More generally, we have the inverse Fourier Transform (FT) of any real-valued function of frequency g defined via $\gamma_h(g) = (2\pi)^{-1}\int_{-\pi}^{\pi} g(\lambda)e^{i\lambda h}\,d\lambda$. Such weighted integrals will be abbreviated with a $\langle\cdot\rangle$ notation, i.e., $\langle g\rangle = (2\pi)^{-1}\int_{-\pi}^{\pi} g(\lambda)d\lambda$. We will say that g is the FT associated with the Toeplitz matrix $\Sigma(g)$. The Gaussian log likelihood function multiplied by -2, i.e., the "deviance", is (dispensing with constants) simply

$$\mathcal{D}(\theta) = \log|\Sigma(f_\theta)| + (X-\mu\iota)'\Sigma^{-1}(f_\theta)\,(X-\mu\iota)\,, \tag{2}$$

where $|\Sigma(f_\theta)|$ denotes the determinant of $\Sigma(f_\theta)$. Here ι is a vector of n ones, and the mean μ is essentially a nuisance parameter. Maximum likelihood estimation yields $\widehat{\theta}$ as the minimizer of $\mathcal{D}$; we may very well wish to use (2) even

when our data is not Gaussian (or is not known to be Gaussian), since Maximum Likelihood Estimates (MLEs) are consistent for a fairly broad class of processes (cf., treatment in [29, Chap. 3]). Similar results can be formulated for parameter estimates that minimize the Whittle likelihood, which is discussed in [29, Chap. 3]; our Theorem 1 below relies on theoretical results derived in [20], which are established for both MLEs and Whittle estimates.

We focus upon separable models, where the innovation variance σ^2 is the final parameter of θ, and we write $\theta = \{[\theta]', \sigma^2\}'$. Also, $\overline{f}_{[\theta]}$ is defined by setting the innovation variance to unity, i.e., $\overline{f}_{[\theta]} = f_\theta/\sigma^2$ by definition. We are interested in testing the goodness-of-fit of the data to a model-class $\mathcal{F} = \{(\overline{f}_{[\theta]}, \sigma^2) : [\theta] \in \Theta, \sigma^2 \in (0, \infty)\}$, with Θ an r-dimensional space. The true spectrum of the process is some unknown $\widetilde{f}$, and we seek to discern whether $\widetilde{f} \in \mathcal{F}$ or not; if it is, there is some $[\widetilde{\theta}] \in \Theta$ and $\widetilde{\sigma}^2$ such that $\widetilde{f} = \overline{f}_{[\widetilde{\theta}]}\widetilde{\sigma}^2$. When $\widetilde{f} \notin \mathcal{F}$, the tilde notation on parameters then refers to Pseudo-True Values (PTVs), i.e., the minimizers of the Kullback-Leibler (KL) discrepancy between $\widetilde{f}$ and $\mathcal{F}$; see [29, Chap. 3] and [22] for background. To be precise, $[\widetilde{\theta}]$ is the minimizer of $[\theta] \mapsto \langle \widetilde{f}/\overline{f}_{[\theta]} \rangle$, and $\widetilde{\sigma}^2 = \langle \widetilde{f}/\overline{f}_{[\widetilde{\theta}]} \rangle$. The PTVs need not be unique, but in many cases of interest (e.g., AR models) they are.

Now the MLE for $[\theta]$ can be determined by minimizing (with respect to $[\theta]$) the sum of squared residuals, which we take to be the vector $\Sigma^{-1/2}(\overline{f}_{[\theta]})(X - \widetilde{\mu}\iota)$. The square root refers to the matrix square root described in [7, Chap. 4]. This vector of residuals exactly corresponds to a Gaussian white noise vector when the model is correctly specified and $([\theta], \mu)$ are replaced by the true parameters $([\widetilde{\theta}], \widetilde{\mu})$. Estimated residuals are obtained by substituting the MLE $[\widehat{\theta}]$ and $\overline{X} = n^{-1}\,\iota' X$ for $([\theta], \mu)$; although a GLS estimate for μ could also be used, the nature of the mean estimate won't be relevant for our asymptotic treatment of Q statistics. The result is a vector R of estimated residuals:

$$R = \Sigma^{-1/2}(\overline{f}_{[\widehat{\theta}]})(X - \overline{X}\iota). \tag{3}$$

This definition of residual corresponds to the innovations algorithm [4, Chap. 8], and is quite general[2]. This minimization produces the MLE $[\widehat{\theta}]$; the innovation variance is estimated by the average of squared centered residuals. The sample acvs of the residuals are

$$\widehat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k}(R_t - \overline{R})(R_{t+k} - \overline{R})$$

for $k \geq 0$, where we have taken the biased definition (this won't matter asymptotically, since we will always have $k = o(n)$ in our treatment). Because the innovation variance is defined via $\widetilde{\sigma}^2 = \langle \widetilde{f}/f_{[\widetilde{\theta}]} \rangle$, we can estimate it via $\widehat{\gamma}_0$.

[2] [1] offered this definition, where $\Sigma^{1/2}$ was the lower Cholesky factor of Σ, and demonstrated that statistics based upon such residuals had superior size properties than other definitions.

The BP and LB statistics are computed from weighted linear combinations of squared sample acvs $\widehat{\gamma}_k$. Let the sample autocorrelations (acs) be defined via $\widehat{\rho}_k = \widehat{\gamma}_k/\widehat{\gamma}_0$; then (1) gives the definition of the Q statistics.

The centering by $\overline{R}$ will not affect the asymptotic distributions, and so it is valid to approximate the sample acv by $n^{-1}R'L^{(k)}R$, where R is given in (3) and the matrix $L^{(k)}$ is the symmetrization of the kth power of the lag matrix, namely a matrix of all zeroes except the value 1/2 on the bands given by all entries i, j such that $|i - j| = k$. When $k = 0$ this is just the identity matrix, and in all cases it is a Toeplitz matrix with associated Fourier Transform given by $\cos(k\lambda)$ – which will be denoted by $c_k(\lambda)$. See [27] for background on lag matrices. Also, centering R by $\overline{X}$ is irrelevant asymptotically, so that the sample acv of the residuals can be approximately written as

$$\overline{\gamma}_k = \frac{1}{n}\left[X - \widetilde{\mu}\right]' \Sigma(\overline{f}_{[\widehat{\theta}]}^{-1} c_k)\ \left[X - \widetilde{\mu}\right].$$

Likewise, let $\overline{\rho}_k = \overline{\gamma}_k/\overline{\gamma}_0$.

2.2. Asymptotic results

The theoretical contribution of the paper is the following theorem and two corollaries, which generalize previous results of [25] to non-ARMA/SARMA models. The asymptotic covariance matrix of [3] was idempotent, though later derivations in [23] corrected this approximation, with which our result agrees. Moreover, we explicitly derive the limit theory for the case of a misspecified model, and consider multiple Q statistics jointly, both of which are new facets. We first focus on the joint asymptotic distribution of $\overline{\gamma}_k$ for $k = 0, 1, \ldots, m$. The Gaussian assumption can be relaxed if working with QMLEs instead of MLEs, so long as we include the extra conditions found in [8]; see the discussion at the end of the proof of Theorem 2 in [20]. Also, the asymptotic variance for non-Gaussian data depends on the fourth cumulants, and its estimation is not straight-forward (see Theorem 3.1.2 of [29, Chap. 3]. Below, we consider the gradient ∇ with respect to the full parameter vector θ, so that the final derivative is with respect to σ^2; hence the final component of $\nabla\overline{f}_{[\theta]}$ is zero.

Theorem 1. *Suppose that the PTVs $\widetilde{\theta}$ exist uniquely in the interior of the parameter space, which is compact and convex, and that the Hessian of the KL discrepancy is invertible at the PTVs. Suppose that the process $\{X_t\}$ is mean zero Gaussian and stationary, and that the model spectrum f_θ is twice continuously differentiable in θ and continuous in λ; also that the derivatives with respect to θ are uniformly bounded in λ away from zero and infinity. Then the following weak convergence holds as $n \to \infty$:*

$$\left\{\sqrt{n}\overline{\gamma}_k + \sqrt{n}\,\langle c_k\left(\widetilde{\sigma}^2 - \widetilde{f}/\overline{f}_{[\widetilde{\theta}]}\right)\rangle\right\}_{k=1}^{m} \stackrel{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, V(\widetilde{\theta})\right). \tag{4}$$

The asymptotic variance is given by

$$V_{k\ell}(\theta) = 2\,\langle c_k c_\ell \widetilde{f}^2 f_{[\theta]}^{-2}\rangle - 2b_k'(\theta)M_f^{-1}(\theta)\langle c_\ell \widetilde{f}^2 \nabla f_\theta f_\theta^{-2}\overline{f}_{[\theta]}^{-1}\rangle$$

$$
\begin{aligned}
&- 2b'_\ell(\theta) M_f^{-1}(\theta) \langle c_k \widetilde{f}^2 \nabla f_\theta f_\theta^{-2} \overline{f}_{[\theta]}^{-1} \rangle \\
&+ 2b'_k(\theta) M_f^{-1}(\theta) \langle \widetilde{f}^2 \nabla f_\theta \nabla' f_\theta f_\theta^{-4} \rangle M_f^{-1}(\theta) b_\ell(\theta) \\
b_k(\theta) = \; &\langle c_k \overline{f}_{[\theta]}^{-2} \nabla_\theta \overline{f}_{[\theta]} \widetilde{f} \rangle,
\end{aligned}
$$

where $M_f(\theta)$ is the Hessian of the KL discrepancy between $\widetilde{f}$ and f_θ. Here the indexing is $k, \ell = 1, 2, \ldots, m$. Moreover, the same results are true with $\widehat{\gamma}_k$ substituted for $\overline{\gamma}_k$. Under the Null Hypothesis of a correct model, the asymptotic variance simplifies to

$$V_{k\ell}(\theta) = 2\sigma^4 1_{\{k=\ell\}} - 2b'_k(\theta) M_f^{-1}(\theta) b_\ell(\theta)$$

and $\overline{\gamma}_0 \xrightarrow{P} \widetilde{\sigma}^2$.

Remark 1. The asymptotic bias term $\langle c_k(\widetilde{\sigma}^2 - \widetilde{f}/\overline{f}_{[\widetilde{\theta}]}) \rangle$ is identically zero when the model is correctly specified, but otherwise may be nonzero (note that $\widetilde{\sigma}^2 = \langle \widetilde{f}/\overline{f}_{[\widetilde{\theta}]} \rangle$). However, it is possible for this bias to be zero even when the model is incorrect, which has ramifications for the Q statistics considered below; in such a case the LB and BP test statistics will be inconsistent. Under the Null Hypothesis, the first term of V is just proportional to the identity matrix; the second term arises solely from the uncertainty in the MLEs. In particular, if we fix a particular parameter rather than estimating it, we can zero out the corresponding entries of b_k and b_ℓ. When σ^2 is a parameter, then for any $k = 1, 2, \ldots, m$ the vector $b_k(\theta)$ has $r+1$ components with the last component equal to $\langle c_k \rangle = 0$, whereas the first r components are given by $\sigma^2 \langle c_k \nabla_{[\theta]} \log \overline{f}_{[\theta]} \rangle$. $M_f(\theta)$ is the Hessian of the Kullback-Leibler discrepancy, which under the Null Hypothesis of a correct model is equal to twice the Fisher information matrix.

From Theorem 1 we can derive as a corollary the joint asymptotic limit of the sample acs. We first define the notion of residual autocorrelations: the lag k autocovariance of the asymptotic residual process is defined to be $\widetilde{\gamma}_k = \langle c_k \widetilde{f} \overline{f}_{[\widetilde{\theta}]}^{-1} \rangle$, and hence the lag k autocorrelation is $\widetilde{\rho}_k = \widetilde{\sigma}^{-2} \langle c_k \widetilde{f} \overline{f}_{[\widetilde{\theta}]}^{-1} \rangle$, using the fact that $\langle \widetilde{f} \overline{f}_{[\widetilde{\theta}]}^{-1} \rangle = \widetilde{\sigma}^2$. The sample residual autocorrelations are consistent for these asymptotic residual autocorrelations, as shown next.

Corollary 1. *Suppose the same conditions as Theorem 1. Whether or not the model is true,*

$$\{\sqrt{n}\,(\overline{\rho}_k - \widetilde{\rho}_k)\}_{k=1}^m \xRightarrow{\mathcal{L}} \mathcal{N}\left(0, V(\widetilde{\theta})/\widetilde{\sigma}^4\right). \tag{5}$$

Moreover, the same results are true with $\widehat{\rho}_k$ substituted for $\overline{\rho}_k$ in (5).

Remark 2. When the model is correctly specified, the vector of asymptotic residual autocorrelations – denoted by $\widetilde{\rho}$ – is zero. Denoting the limiting covariance matrix in (5) by $\overline{V}(\theta) = V(\theta)\sigma^{-4}$, we can write

$$\overline{V}_m(\theta) = 1_m - 2C' N^{-1} C$$

with N the upper $r \times r$ sub-matrix of M_f. Here 1_m denotes an identity matrix of dimension m; the subscript on $\overline{V}$ highlights the dimension of the matrix. The $r \times m$ dimensional matrix C is defined to have entry jk given by

$$C_{jk} = [b_k(\theta)]_j/\sigma^2 = \left\langle c_k \frac{\partial}{\partial \theta_j} \log \overline{f}_{[\theta]} \right\rangle.$$

The matrix $\overline{V}_m$ has a nested structure, which means that $\overline{V}_k$ is the upper $k \times k$ block of $\overline{V}_m$ whenever $k \leq m$ – this follows from the structure of C[3].

From Corollary 1 and the above Remark, we at once obtain distributional results for the BP and LB statistics, generalizing previously published results beyond the ARMA class. Let Q_m denote either Q_{BP} or Q_{LB}, as defined in (1) with lag equal to m, being based on m autocorrelations.

Corollary 2. *Suppose the same conditions as Theorem 1. For some integer M let Y_M be a mean zero Gaussian random vector of dimension M and covariance matrix $\overline{V}_M(\widetilde{\theta})$, and let Y_m for any $1 \leq m \leq M$ denote the first m components of Y_M. Then when the model is correctly specified,*

$$\{Q_m\}_{m=1}^M \stackrel{\mathcal{L}}{\Longrightarrow} \{Y_m' Y_m\}_{m=1}^M$$

as $n \to \infty$. When the model is incorrectly specified such that $\widetilde{\rho} \neq 0$, then

$$Q_m = o_P(1) + \sqrt{n}\,(\overline{\rho} - \widetilde{\rho})' \sqrt{n}\,(\overline{\rho} - \widetilde{\rho}) + 2\sqrt{n}\,\widetilde{\rho}' \sqrt{n}\,(\overline{\rho} - \widetilde{\rho}) + n\,\widetilde{\rho}'\widetilde{\rho}$$

for each $1 \leq m \leq M$.

In Corollary 2, the expansion for Q_m under the alternative hypothesis (i.e., for $\widetilde{\rho} \neq 0$) shows three terms, the first of which converges to $Y_m' Y_m$. The second term, when divided by $\sqrt{n}$, converges to $\widetilde{\rho}' Y_m$, and the third term is the positive constant $\widetilde{\rho}'\widetilde{\rho}$ times n. This indicates that the power of the test is asymptotically dominated by the quantity $\widetilde{\rho}'\widetilde{\rho}$; of course this can still be zero when a model is misspecified (e.g., an EXP model can produce this behavior), but taking a sufficiently high number of lags should guard against this occurrence.

2.3. Discussion of asymptotic theory

The theoretical results of this have ramifications for sequential testing. Sequential testing refers to examination of the various Q_m in the order $Q_1, Q_2, \ldots, Q_M$. If by the mth test we have so far failed to reject the null hypothesis, then we have failed to reject the hypothesis that $\widetilde{\rho}_k^2 = 0$ for $1 \leq k \leq m$. Now proceeding to test statistic Q_{m+1}, we seek to test whether $\sum_{k=1}^{m+1} \widetilde{\rho}_k^2 > 0$; given our previously retained null hypotheses, we essentially are testing whether $\widetilde{\rho}_{m+1} \neq 0$. If the $m+1$th test statistic is significant, we can conclude that there is residual autocorrelation at lag $m+1$, *and not at the other previous lags.* On the other hand if Q_{m+1} is not significant, we can conclude that $\widetilde{\rho}_{m+1} = 0$, and proceed.

[3] In comparison with [23], our matrix C' equals his X, and our matrix N equals his $I/2$ (half the Fisher information matrix).

Note that this sequential procedure gives more insight into potential modeling problems than just testing the portmanteau Q_M – a significant Q_M only indicates that $\sum_{k=1}^{M} \widetilde{\rho}_k^2 > 0$, but we don't know at which lags problems are really arising. But if we have an idea about which lags are causing difficulties, then we may be able to improve the model; this is further illustrated in Section 4.3.

We now provide some additional discussion of the limit distribution of Corollary 2. Let us focus on the case of just a single Q_m statistic. Since Y_m is normal with variance $\overline{V}_m$, we can simply characterize the limiting distribution of the Q_m statistic via the Laplace Transform (LT) as follows [30]:

$$\mathbb{E} \exp\{-\phi Y_m' Y_m\} = |1_M + 2\phi \overline{V}_m|^{-1/2}.$$

This formula, however, is of little use in determining quantiles except in special cases, discussed below. Also see [9] for background on this distribution.

Note that the first term of $\overline{V}_m$ (see Remark 2) is the identity matrix, which indicates that when we ignore parameter estimation error (i.e., set the b_k vectors identically to zero) the above LT reduces to $(1+2\phi)^{-m/2}$. This is recognizable as the LT for the sum of m *iid* χ^2 variables on one degree of freedom, i.e., in this special case $Y_m' Y_m$ is χ^2 on m degrees of freedom. But when parameter uncertainty is present, $\overline{V}_m$ will not be diagonal and the LT might *not* be the product of LTs for χ^2 variables.

In [3] the authors propose a formula for $\overline{V}_m$ that is somewhat different. In fact, accounting for differences in notation, they essentially propose

$$V_m^\sharp(\theta) = 1_m - C'(CC')^{-1}C,$$

which means replacing the upper left portion of the Fisher information matrix with CC'. Of course $N \neq 2CC'$, but [3] argue that it is a suitable approximation especially when m is large and assuming that the coefficients in the Wold decomposition decay at a suitable rate as the sample size increases. Moreover, it is a convenient substitution because then $1_m - V_m^\sharp$ is idempotent with rank r, so that $m-r$ eigenvalues of $V_m^\sharp$ are equal to unity, and the remaining r eigenvalues are zero. We can apply Proposition 2 of [30] to conclude that the limiting distribution would then be a χ^2_{m-r}. We know of one case (discussed below) where this approximation is actually valid without assuming that the Wold coefficients depend on sample size n; i.e., we discuss below a case where $\overline{V}_m$ has all its eigenvalues equal to either zero or one.

Consider the EXP(r) model of [2], which has spectral density given by $\overline{f}_{[\theta]}(\lambda) = \exp\{[\theta]' \overrightarrow{c}(\lambda)\}$, with $\overrightarrow{c}$ denoting the column vector of functions c_k. For this model, b_k is one half the kth unit vector, and $N = .5\, 1_r$. Hence $\overline{V}$ is diagonal with the first r entries equal to zero and the remaining $m-r$ entries equal to one. This is quite a special structure, and is not true for ARMA models. To digress briefly – since it is pertinent to the time series fitting problem in general – seems appropriate here.

Let the periodogram be defined as $I(\lambda) = n^{-1} \sum_{|k|<n} \overline{\gamma}_k e^{-i\lambda k}$. Gaussian maximum likelihood estimation – or asymptotically equivalently, fitting via minimization of KL discrepancy between model and periodogram – essentially works

to minimize the sample variance $\langle I/\overline{f}_{[\theta]}\rangle$ of time series residuals, whereas Q statistics test whether the residual spectrum $I/\overline{f}_{[\widehat{\theta}]}$ behaves like white noise. Because the gradient of KL for the EXP(r) model equals $\langle \overrightarrow{c}\, I/\overline{f}_{[\theta]}\rangle$, minimization necessarily entails that the first r sample autocorrelations of the residual process are zero – which is what the Q statistics are attempting to verify. It is this strong property of EXP models that is responsible for the simple asymptotic structure of the Q statistics.

As for the exact asymptotic distribution in the general case with fixed coefficients – leaving aside the useful approximation of [3] for the moment – we note that C has a null space of dimension at least $m-r$ by the rank-nullity theorem, and hence $m-r$ eigenvalues of $\overline{V}_m$ are equal to unity. Then the limiting LT is that of a χ^2_{m-r} variable plus an independent variable with LT

$$\exp\left\{-\frac{1}{2}\sum_{\ell=1}^{r}\log(1+2\phi[1-\lambda_\ell])\right\},$$

with $\lambda_1,\ldots,\lambda_r$ the r nonzero eigenvalues of $2C'N^{-1}C$. In practice, these r eigenvalues can be quite close to zero for values of m only a little larger than r, depending upon the model and the underlying process. This makes the inversion of $\overline{V}_m$ infeasible and the approximation of [3] quite useful. On the other hand, the degrees of freedom in the [3] approach is $m-r$, so that no distributional result can be used when $m \le r$; in these cases, $\overline{V}_m$ may be quite different from an idempotent matrix, and is *a fortiori* invertible. Moreover, in the sequential approach to testing advocated below, the joint behavior of the LBs for small m is indispensable. In applications one evaluates $\overline{V}_m$ at parameter estimates, such as MLEs.

As an example of the above claim, we have found that when fitting simulated data with an Airline model ($r=2$), the χ^2 approximation to the distribution is decent for $m \ge 5$ (this also depends upon the parameter values of the simulated true Airline process). In contrast, for $1 \le m \le 4$ the χ^2 approximation can be quite poor and $\overline{V}_m$ is easily invertible. In this case, the exact distribution should be used, it being superior to the χ^2_{m-2} – which moreover is not even defined for $m = 1, 2$.

3. Sequential testing of Q statistics

The joint testing problem is to determine a sensible sequence of critical values for a given overall Type I error rate α. The challenge is that there are so many ways to divide up the mass of a multivariate probability density. However, when the statistics have a sequential relationship, we can describe a methodical procedure to obtain critical values.

The basic idea of sequential testing is related to ideas in the biostatistics literature [28]. Consider a sequence of test statistics $\{T_m\}$ for $m = 1, 2, \ldots, M$ (where $M = \infty$ is allowed, although in our LB application $M < \infty$), where rejection occurs when $T_m > x_m$. Quantities of considerable interest are the

conditional probabilities

$$\alpha_{m+1} = \mathbb{P}\left[T_{m+1} > x_{m+1} | T_m \leq x_m, \ldots, T_1 \leq x_1\right]. \tag{6}$$

This represents, for a given sequence of critical values x_ℓ, the probability of rejection now, given that we have not rejected up to now. Also set $\alpha_1 = \mathbb{P}[T_1 > x_1]$. A closely related quantity is the joint probability

$$p_{m+1} = \mathbb{P}\left[T_{m+1} \leq x_{m+1}, T_m \leq x_m, \ldots, T_1 \leq x_1\right]. \tag{7}$$

Of course we have the relation $\alpha_{m+1} = 1 - p_{m+1}/p_m$. The overall assessment of the procedure involves computing the probability that at least one test rejects, i.e., the event $\cup_m \{T_m > x_m\}$. Now while [28] focus on computing the joint probabilities p_m (7), we emphasize the α_ms, although there is a ready equivalence between the two approaches. We first note a few elementary facts.

Let $\alpha = \mathbb{P}[\cup_m \{T_m > x_m\}]$, which is called the Type I error rate of the sequential procedure – whereas the sequential Type I error rates are the quantities α_m. This is nomenclature. By induction, $p_{m+1} = \Pi_{\ell=1}^{m+1}(1-\alpha_\ell)$. In addition $\alpha = 1 - p_M$, which also equals $\sum_{m=1}^{M} \alpha_m p_{m-1}$ (with $p_0 = 1$), as shown in [28].

Consider two extreme cases: first, if all the tests are independent, then α_m is the mth marginal probability. If the tests are also identically distributed, then the sequential critical values are all the same, and equal to $1 - (1-\alpha)^{1/M}$. If instead all the tests are fully dependent (say, actually identical), then there is really no multiple testing problem, and α must equal the sequential conditional probability. More generally, the test statistics are somewhat dependent and the relationship of marginal to joint distribution is more complicated.

The first challenge is to determine, for a given α, the critical values x_m such that the corresponding sequential conditional or joint probabilities aggregate appropriately to α. As an initial step, one must choose the numbers p_m or α_m to satisfy the appropriate constraint; then we may determine the x_m sequentially given certain information about the distribution functions. Whereas [28] choose to work with the p_m, partitioning them such that they sum to α, we in contrast work with the α_m. The reason is that we find these to be a more intuitive quantity, given their interpretation in the sequence of tests. In fact, it seems reasonable to impose that all the α_m numbers be identically equal, say to a common value α_0. Although this decision is arbitrary, it imposes an equitable restriction – each step of our testing procedure is treated equally. However, this approach need not generate the maximal possible power; determining a most powerful sequence of Type I error rates α_m for Q statistics is difficult to discover *a priori*, because power depends on the nature of the alternative hypothesis through the unknown $\widetilde{\rho}$, the residual asymptotic autocorrelations.

Our setting of equal sequential Type I error rates implies that $p_m = (1-\alpha_0)^m$ and we must choose $\alpha_0 = 1 - (1-\alpha)^{1/M}$ (this approach does not work when $M = \infty$, and in fact the sequential conditional probabilities must decay, being non-constant, in this case). This is equivalent to taking a geometrically decaying sequence of joint probabilities in the [28] rationing of α. Note that if the

tests happened to be independent, order would be unimportant and we should set all probabilities equal; setting all the conditional probabilities to be equal generalizes this concept to potentially dependent statistics.

A second challenge is to compute, for a given sequence x_m, the joint and conditional probabilities. In practice, this is much easier than finding critical values, so we describe it first. Note that when the various x_m correspond to the observed values of actual test statistics, the corresponding probabilities can be interpreted as sequential (conditional or joint) p-values. The algorithm is to compute $p_1, p_2, \ldots, p_M$ in sequence – potentially using Monte Carlo simulation if analytic formulas are unavailable – and then determine each α_m, which depend on knowing p_m and p_{m-1}. The p-value of the sequential procedure would then be just p_M. Explicitly, if we have J Monte Carlo draws of the various statistics, with the jth draw denoted $T_\ell^{(j)}$ for $1 \le \ell \le m$ and $1 \le j \le J$, then

$$p_m \approx J^{-1} \sum_{j=1}^{J} 1_{\{T_m^{(j)} \le x_m, \ldots, T_1^{(j)} \le x_1\}} = J^{-1} \# \left\{ j : \max_{1 \le \ell \le m} \left[T_\ell^{(j)} - x_\ell \right] \le 0 \right\}.$$

The second expression is useful for encoding the method, e.g., in R [26].

Now consider computation of critical values, using the α-rationing scheme described above. First compute x_1 such that $1-\alpha_0 = 1-\alpha_1 = \mathbb{P}[T_1 \le x_1]$, possibly by inverting the marginal distribution, otherwise by using the approximation to p_1 above, noting that $1-p_1 = \alpha_1$. Given a knowledge of $x_1, x_2, \ldots, x_m$, we wish to compute x_{m+1} such that $1-\alpha_{m+1}$ equals $\mathbb{P}[T_{m+1} \le x_{m+1} | T_m \le x_m, \ldots, T_1 \le x_1]$. So consider a subset $L(x_m, x_{m-1}, \ldots, x_1) \subset \{1, 2, \ldots, J\}$ consisting of only those Monte Carlo draws j such that $\max_{1 \le \ell \le m}[T_\ell^{(j)} - x_\ell]$ is less than or equal to zero. Then

$$1 - \alpha_0 = 1 - \alpha_{m+1} \approx |L(x_m, x_{m-1}, \ldots, x_1)|^{-1} \sum_{j \in L(x_m, x_{m-1}, \ldots, x_1)} 1_{\{T_{m+1}^{(j)} \le x_{m+1}\}}.$$

Hence, one only needs the $(1-\alpha_0)$th largest order statistic of the collection $T_{m+1}^{(j)} - x_{m+1}$ such that $j \in L(x_m, x_{m-1}, \ldots, x_1)$, and this will be the approximation to x_{m+1}.

We have found such a procedure to be effective, written in R, for determining critical values of Q statistics (more details given below). For simulation studies we lowered J to be 10^4 in order for the computations to finish in a reasonable amount of time, but for real data analysis $J = 10^5$ provided increased accuracy – while the calculations of all critical values and p-values completed in a few seconds.

It may arise that we desire some subset of the full sequence of test statistics. Of course, one could just relabel the subsequence and start the analysis over again. An alternative way of thinking about it is to view certain test statistics T_j as "missing": then set $p_j = p_{j-1}$ and $\alpha_j = 0$, and essentially declare $x_j = \infty$. Then when p-values are computed, the jth statistic offers no restrictions on the probabilities, as all Monte Carlo draws will be less than ∞. For critical values, use the same trick. Of course, if K of a sequence of M statistics are missing in this manner, then we should compute $\alpha_0 = 1 - (1-\alpha)^{1/(M-K)}$, since we only

really have $M - K$ statistics to consider. Our R code is adapted to handle the calculations for any subset of Q statistics that is desired.

Consider computation of the joint probabilities of asymptotic Q statistics, using the theoretical results of Section 2. Suppose that we consider a sequence of M asymptotic Q-statistics, denoted by the random vector $Q = [Q_1, Q_2, \ldots, Q_M]$, and we want to know the joint cdf for all the Q_m with $1 \leq m \leq M$, evaluated at non-negative numbers q_m. Clearly

$$\mathbb{P}\left[Q_1 \leq q_1, \ldots, Q_M \leq q_M\right] = \int_0^{q_M} \cdots \int_0^{q_1} p_Q(u_1, \ldots, u_M)\, du_1 \cdots du_M,$$

and we will denote this function by $F_Q(q_1, \ldots, q_M)$. Let Y_M denote a Gaussian vector of length M with covariance matrix $\overline{V}_M$; this is simple to simulate once $\overline{V}_M^{1/2}$ is computed, which is valid even when $\overline{V}_M$ is non-invertible (e.g., idempotent) or close to singular. Let Y_M^2 denote the component-wise squaring of Y_M, and let A be an $M \times M$ aggregation matrix with $A_{jk} = 1$ whenever $j \geq k$, and zero otherwise. Then Q is equal in distribution to AY_M^2. So by Monte Carlo methods we can easily approximate F_Q. Note that direct simulation of the entire process to get the finite-sample distribution is not feasible; each draw of a Gaussian series, taking the model and MLE as truth, would need to be fitted with residuals computed – this is too expensive to be practical.

$\overline{V}_M$ can be computed from the theoretical results, utilizing the null hypothesis, with MLEs substituted for PTVs. Details on the computation of $\overline{V}_M$ for the case of a SARIMA model can be found in supplementary material for this paper [21]. Expressions for the ARMA and SARMA cases can be found in [23].

4. Numerical studies and data analysis

In order to evaluate the practical importance of these ideas, it is helpful to do a simulation study. This will assess the impact of having a finite sample on the use of asymptotic critical values, under the rather idealized scenario of Gaussian data. Secondly, the new method should be compared to the standard Box-Pierce method on real time series, in order to form an idea of how much the proposed methodology really matters in practice. We first consider a simulation study of finite sample size impact, and then consider analysis of 9 U.S. Census Bureau time series.

4.1. Simulation study

Here we are interested in drawing samples from a monthly Gaussian Airline model with parameters .6 and .6 for the nonseasonal and seasonal moving average parameters, and unit innovation variance, with sample sizes of 10, 15, and 20 years. Since the data is seasonal, there may be considerable interest in residual autocorrelations at lags 12 and 24. By a "full" set of Q statistics, we mean the sequential procedure involving Q_m for $1 \leq m \leq 24$. But we might also consider certain subsets of the full $M = 24$ collection of Q statistics, as alluded to in Section 3.

TABLE 1
Simulation Results for LB Statistics

	n=120			n=180			n=240		
Type	.01	.05	.10	.01	.05	.10	.01	.05	.10
Full	.0166	.0564	.1098	.0168	.0540	.0990	.0170	.0582	.1038
Partial	.0162	.0588	.1034	.0132	.0536	.0988	.0140	.0588	.1056
Restricted	.0162	.0552	.1100	.0160	.0514	.1014	.0164	.0590	.1012
Maximal	.0166	.0578	.1060	.0156	.0524	.0980	.0176	.0608	.1024
Classical	.0848	.2766	.4408	.0790	.2846	.4598	.0838	.2820	.4560

Empirical Type I error for the sequential procedure, based on a nominal $\alpha = .01, .05, .10$, for each of the five methods (see text). Results are based on 5000 simulations of a Gaussian airline model with parameters .6, .6, with sample sizes $n = 120, 180, 240$.

TABLE 2
Simulation Results for BP Statistics

	n=120			n=180			n=240		
Type	.01	.05	.10	.01	.05	.10	.01	.05	.10
Full	.0058	.0298	.0616	.0082	.0342	.0676	.0120	.0418	.0784
Partial	.0094	.0412	.0772	.0096	.0440	.0812	.0096	.0476	.0896
Restricted	.0054	.0274	.0542	.0068	.0314	.0630	.0092	.0404	.0770
Maximal	.0056	.0232	.0444	.0082	.0306	.0570	.0110	.0402	.0728
Classical	.0600	.2252	.3844	.0650	.2494	.4226	.0722	.2572	.4262

Empirical Type I error for the sequential procedure, based on a nominal $\alpha = .01, .05, .10$, for each of the five methods (see text). Results are based on 5000 simulations of a Gaussian airline model with parameters .6, .6, with sample sizes $n = 120, 180, 240$.

In particular, we might only be interested in those lags of the Q statistics deemed to be important to the model. One such subset – henceforth referred to as the "partial" set – consists of lags 1,2,3,4,12, and 24. Or we might just take the seasonal lags Q_{12} and Q_{24}; this choice will be called the "restricted" set. Finally, one might just consider Q_{24}, which being a marginal distribution has no multiple testing issue – this will be called the "maximal" set. Then for any of the four sets – full, partial, restricted, or maximal – we can determine the sequential Type I error rate appropriately, given a selection of the α for the sequential procedure described in Section 3. Of course, final results are contingent upon the set of statistics that is utilized.

For each simulation, we fit the airline model and construct critical values using the sequential procedure, for each of the four sets of Q statistics, for $\alpha = .01, .05, .10$. We also compute the critical values for the classical method, which is defined by utilizing χ^2_{m-r} quantiles when the lag m exceeds r (this method does not use the sequential procedure, because it does not assume anything about the joint distribution of the Q statistics). By determining empirical model rejection rates over many simulations, we can evaluate the competing methods in terms of their Type I error, taking finite-sample effects into account.

The simulations were 5000 draws from a Gaussian airline model, with parameters .6, .6, with sample sizes $n = 120, 180, 240$. Tables 1 and 2 summarize the empirical size results for the four subsets of statistics, both LB and BP, as well as the classical method. The coverage of the new methods is somewhat rough when $n = 120$, and yet far superior to the classical method, which rejects far

TABLE 3
Descriptions of Monthly Retail Sales Series, covering the period 1992 through 2007 (Source: U.S. Census Bureau)

Series	Description of Retail Sales Series
Elect	Electronics and Appliance Stores
Food	Food Services and Drinking Places
Furn	Furniture and Home Furnishing Stores
Gas	Gasoline Stations
GenMerch	General Merchandise Stores
Groc	Grocery
MenCloth	Men's Clothing
Motor	Motor Vehicle and Parts Dealers
WomCloth	Women's Clothing

too often (as expected). The LB statistics were over-sized, with only marginal improvement as sample size increased. The BP statistics were under-sized, but actually improved quite a bit by sample size $n = 240$. Overall, the LB statistics have better size than the BP statistics, which is not surprising given the motivation for their definition [17]. Although the finite sample distribution is slow to converge to the asymptotic, the coverage for these subsets of Q statistics is adequate for practical applications, and is greatly superior to the classical coverage. The coverage for the partial and restricted cases is quite similar to that of the maximal case, which is encouraging.

4.2. Census bureau time series

We consider nine seasonal time series published by the Census Bureau from the Monthly Retail Sales Survey. Table 3 gives the names and descriptions of these series[4]. All series cover the period 1992 through 2007 inclusive (truncated to avoid the Great Recession, for simplicity). In each case we have performed the following analysis: fitted a SARIMA model (identified as best according to the *automdl* spec of X-12-ARIMA), with fixed effects handled appropriately; computed $\overline{V}(\widehat{\theta})$ at the MLEs, as well as the LB statistics for lags 1 through 24; evaluated our proposed methodology with a sequential procedure α of $.01, .05, .10$ using either the full, partial, restricted, or maximal sets of Q statistics, along with the classical procedure. The competing sets of critical values are plotted along with the actual LB statistics.

In each graph (Figure 1 through 9), for a fixed value of α, we see the actual LB statistics plotted as a function of lag, with the critical values plotted in other colors. If the former curve crosses above any of the critical values, it indicates rejection of the specified model according to that particular criterion. It is apparent that the results are sensitive to whether we adopt the full, partial, restricted, or maximal sets of Q statistics, as well as what the given α is set to be.

[4]Descriptions of data sources and reliability are available from the Census Bureau web site http://www.census.gov/retail/mrts/how_surveys_are_collected.html. Program overviews and current data are available from the site http://www.census.gov/cgi-bin/briefroom/BriefRm.

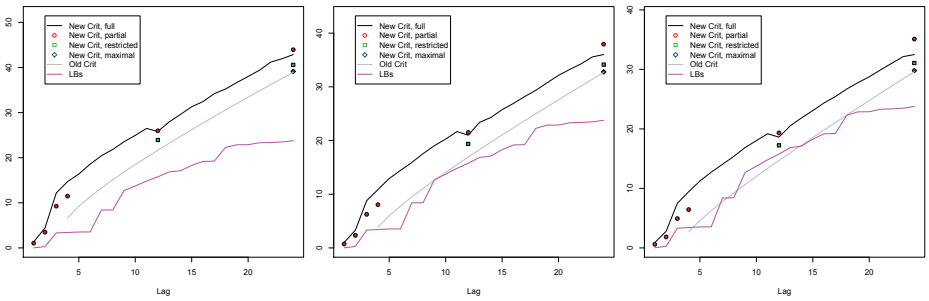

FIG 1. *Critical values and LB statistics for Motor series. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

In general, the modified critical values increase as a function of lag, but less smoothly than with the classical method. In most cases, the classical critical values are lower than the proposed full critical values, so that fewer models are rejected with the proposed method. However, this story changes when we move to the partial or restricted sets of Q statistics. Note that there is not so much discrepancy between the classical and proposed critical values as might be thought initially, which is due to the fact that the sequence of Q statistics are cross-correlated; recall that when the test statistics are fully correlated, there is no multiple testing problem. Also, since the critical values of the classical method ignore multiple testing, they should approximately agree (at lag 24) with the maximal critical value of our new method. This is because the only discrepancies between them would be due to our use of Corollary 1 to compute the critical values, as opposed to a χ^2_{24-r}. Since $24-r$ is fairly large, the idempotent approximation of $\overline{V}$ is reasonably accurate, so that the exact asymptotic distribution differs very little from the χ^2_{24-r}, as discussed in Section 2.

Another general feature is that the first r critical values for the default method are not available, since the degrees of freedom would not be positive in this case. Also, critical values decrease as we move from the upper $\alpha = .01$ panel to the bottom $\alpha = .10$ panel; note that the y-axes on the three panels have not been standardized, since it is not our primary intention to make comparisons across α.

Let us now discuss the individual results. All series required a log transformation, and were linearized (i.e., all types of fixed effects, such as outliers, Easter and trading day, were removed) before further analysis. For the Motor series (Figure 1) a (012)(011) model was identified, and there seem to be no problems with it according to any of the four proposed sets of Q statistics, though according to the classical method rejection at the .05 and .10 levels is warranted. The Food series (Figure 2) has an airline model, and there is rejection – according to classical criteria – at $\alpha = .05, .10$ at a few distinct lags. But accounting for multiple testing indicates this model would not be rejected at all. For the Elect series (Figure 3) a (211)(011) model was identified, so that $r = 4$ (fairly high for a SARIMA model). For $\alpha = .05$ we have rejection of the model based on the

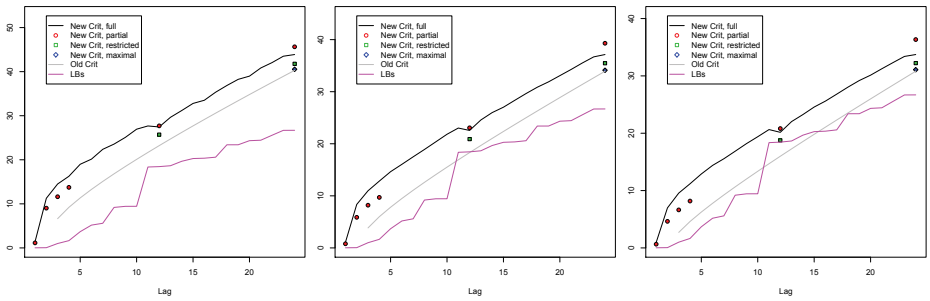

FIG 2. *Critical values and LB statistics for Food series. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

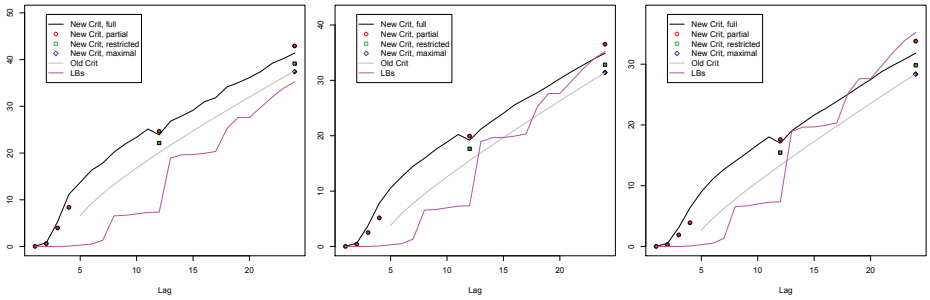

FIG 3. *Critical values and LB statistics for Elect series. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

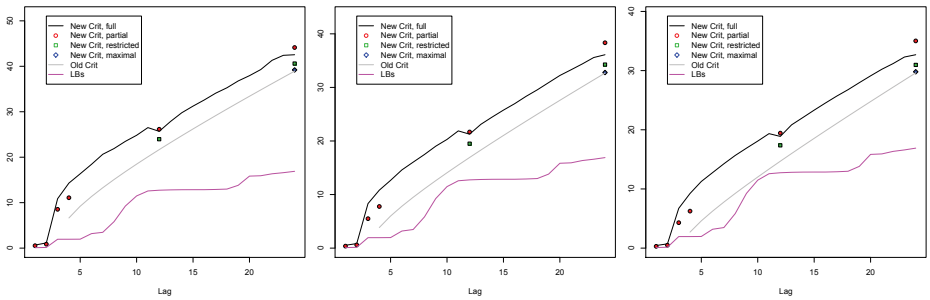

FIG 4. *Critical values and LB statistics for Furn series. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

classical, full, restricted, and maximal schemes, but not for the partial scheme. Rejection at rate $\alpha = .10$ occurs for all the schemes, and the problems seem to arise from the higher lags; no rejections occur at $\alpha = .01$.

The Furn series was identified as a (210)(011) model, and Figure 4 gives no reason to reject it (by any of the methods). The Gas series in Figure 5 follows a (012)(011) model, and at the $\alpha = .05$ level is rejected under the classical,

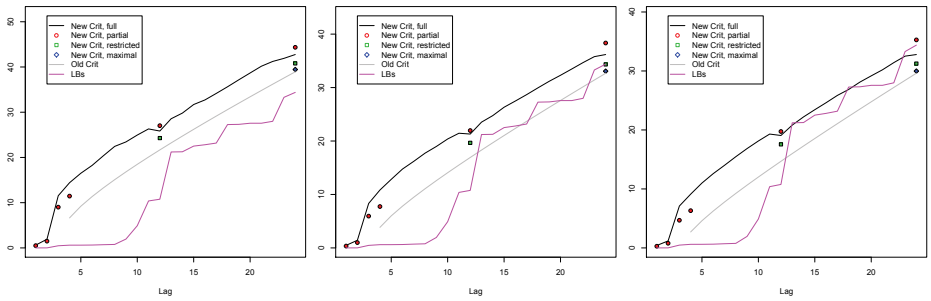

FIG 5. *Critical values and LB statistics for Gas series. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

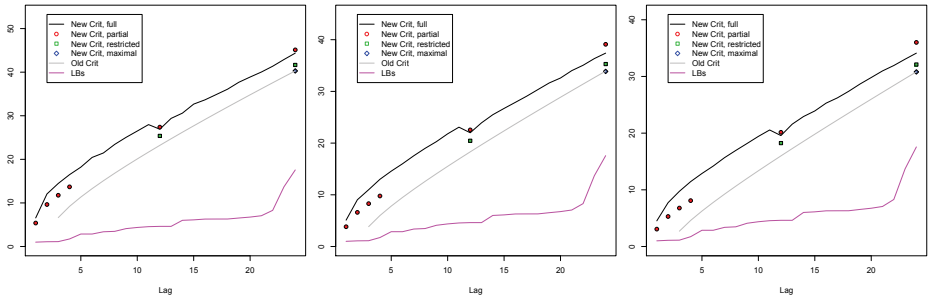

FIG 6. *Critical values and LB statistics for GenMerch series. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

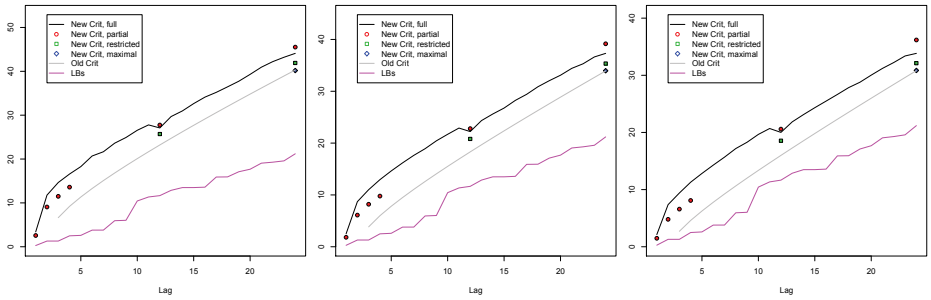

FIG 7. *Critical values and LB statistics for Groc series. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

maximal, and restricted schemes. The problem lags occur at lags 11 and 13 in this case. At $\alpha = .10$ the model would be rejected by the full scheme as well, while there would be no rejections for $\alpha = .01$. The GenMerch series of Figure 6 follows an (011)(110) model, and there is no evidence whatsoever to reject it. The story is the same for the Groc series (Figure 7), which was identified with a (110)(011) model. Likewise, the MenCloth and WomCloth series

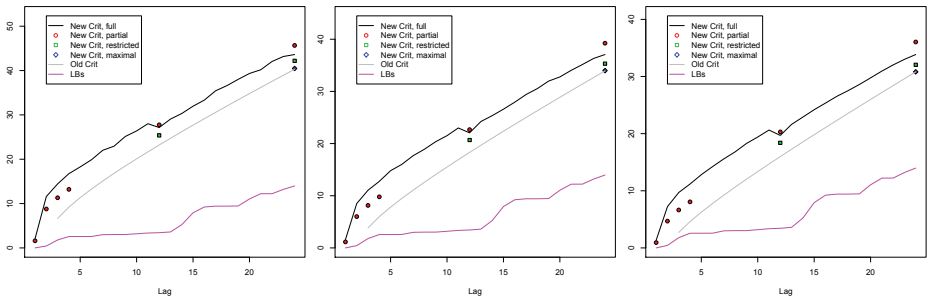

FIG 8. *Critical values and LB statistics for MenCloth series. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

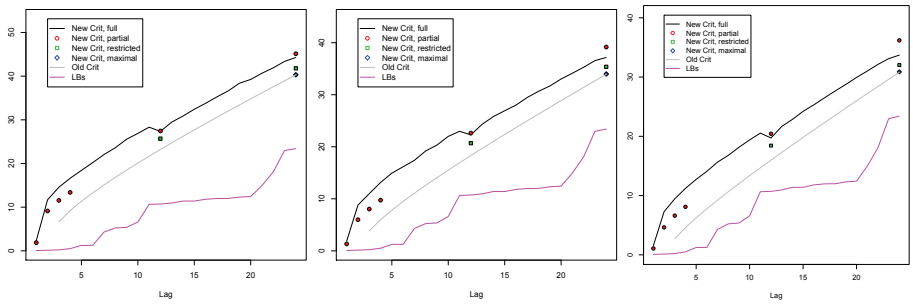

FIG 9. *Critical values and LB statistics for WomCloth series. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

(Figures 8 and 9) were both identified with (011)(011) models, which cannot be rejected.

In summary, five of the nine series (Furn, GenMerch, Groc, MenCloth, WomCloth) provide no evidence of model misspecification. Two of the series (Elect and Gas) yield model rejection results both for the classical and the proposed methods, so there is an agreement of decisions. Finally, two of the series (Motor and Food) would be rejected by the classical method, while not being rejected by any of the proposed methods. We know that critical values are increased by accounting for multiple testing, so it is not surprising that sometimes we will incorrectly reject some models when using conventional χ^2 critical values.

4.3. A refined model for elect series

We now investigate the Elect series further. Figure 3 indicates that lag 13 can be an issue in the residuals, because the Q statistics are not significant for lags 1 through 12. Examination of the sample autocorrelation function for the differenced series indicates there is negative correlation at lag 13, but little correlation at other lags – this structure is difficult to capture with a SARIMA

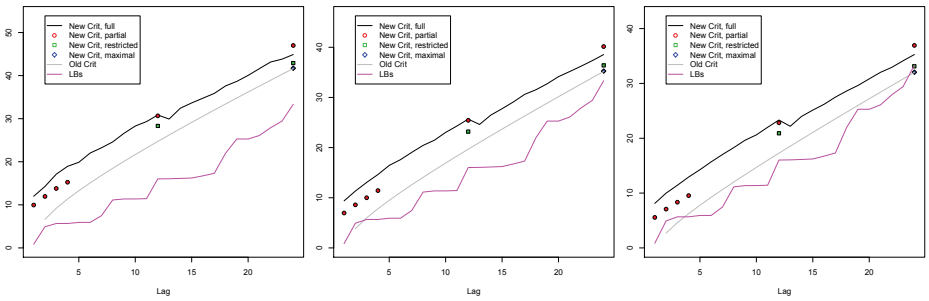

FIG 10. *Critical values and LB statistics for Elect series, using gap moving average model. From left to right, the procedure's Type I error rates are* .01, .05, *and* .10.

model. In particular, a moving average process with nonzero coefficient only at lag 13 cannot be described by an airline model (for the differenced series), because the lag 13 coefficient of an airline model is the product of the model's two parameters, each of which in turn equals the lag 1 and the lag 12 coefficients of the moving average. So instead of the identified (211)(011) SARIMA model, we will consider an order 13 moving average model where the first 12 coefficients are forced to be zero.

For this model, the one parameter is $\widehat{\theta}_{13} = .261$ and the residuals appear to be white upon visual inspection. The LB statistics were then computed, and the revised results are plotted in Figure 10; this can be compared with Figure 3. (Note that in the calculation of $\overline{V}_M$, having fixed coefficients in an MA(13) model means that the first 12 components of each b_k vector in Theorem 1 is zero, and $\overline{V}_M$ must be adjusted accordingly.) According to the full or partial criteria, the model is adequate, although the classical method flags some problems, and with $\alpha = .10$ the restricted and maximal criteria still flag a problem at lag 24. However, the model seems to be a great improvement from the standpoint of time series residuals.

We might also evaluate the new model against the old model according to other criteria. The models are non-nested, so a Generalized Likelihood Ratio test cannot be applied. We can assess out-of-sample forecast performance by forecasting to horizons 1 and 12 based upon the two models being fitted to reduced spans of data; we utilized the revisions history diagnostic of X-12-ARIMA and X-13ARIMA-SEATS. The revisions history of the accumulated forecast error is generated with a start date of January 1992 and terminal dates of January 2000 up through December 2007.

Both the SARIMA (211)(011) – denoted as Model 1 – and the new model – denoted as Model 2 – are fitted to each span, their forecasts generated to either 1 or 12 steps ahead, and cumulative sums of squared forecast errors are generated. The differences of the accumulating sums of squared forecast errors between the two competing models are plotted in the top panel of Figure 11. As each new data point is added, the span expands by one month and both models are re-

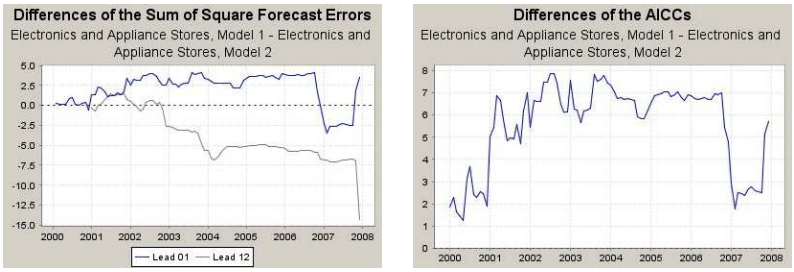

FIG 11. *Forecasting revisions history (left panel) and AICC difference history (right panel) plots for comparison of Model 1 (the SARIMA model) and Model 2 (the new moving average model).*

fitted, and new forecasts at horizons 1 and 12 are computed and a new difference in the accumulated forecast error is generated. (The dates on the x-axis of Figure 11 are not the span terminal dates, but the date of the corresponding forecast.)

If the direction of the accumulating differences is predominantly downward, this means that the forecast errors are persistently smaller for the first model, which indicates a preference for the second model. For more examples of forecast error history plots, see [5].

We see that Model 2 performs better than Model 1 at 1-step ahead, but is inferior at 12 step-ahead forecasting. [22] argues that when working with misspecified models – so that one has not effectively "whitened the data's spectrum" – parameter estimates attuned to 1-step ahead or 12-step ahead loss can be quite different, and the 12-step ahead performance of models fitted according to a 1-step ahead criterion can be disappointing. In this case, the new model is useful for 1-step ahead forecasting, but for horizon 12 the default SARIMA (211)(011) is preferable.

There is a sizable dip in the 1-step ahead performance of Model 2, relative to Model 1, in the final year. Additional analysis reveals that the structure of the series changes slightly near the beginning of 2007 (leading into the Great Recession). Examining the revision history of the differences of the AICC diagnostic for the two models (bottom panel of Figure 11) shows the same pattern in the final years. This is not surprising, given that AICC is calculated from the Gaussian likelihood, which is directly related to the sum of squared forecast errors.

In summary, the sequential examination of Q statistics leads us to consider the 13th lag, and an alternative model that better whitens the spectrum. The resulting model has better 1-step ahead forecast performance, and superior AIC, as compared to the SARIMA (211)(011) contender, although the 12-step ahead forecasting performance is inferior. This demonstrates chiefly that the sequence of Q statistics – as opposed to sole consideration of one Q statistics with a high lag m – can be useful for determining alternative models. This is because the power of a Q statistic at lag $m+1$ – given that all others at lags less than $m+1$ are not significant – is chiefly generated by residual autocorrelation at lag $m+1$.

This reasoning, without formal justification until now, has served to justify the widespread use of multiple Q statistics in time series software used around the world.

5. Conclusion

This paper makes several novel contributions to an important problem in time series analysis. The use of Q statistics is widespread, has a long legacy (more than four decades), and despite recent alternatives seems likely to continue to occupy a central place among time series model diagnostics. Two outstanding issues with the conventional use of BP and LB statistics are that the asymptotic theory currently in common use is flawed, and secondly the use of multiple Q statistics suffers from the ubiquitous multiple testing problem.

The first issue is shown in this paper to be of primary concern when the number of lags m in a Q statistic are small; for larger m the χ^2 approximate asymptotic distribution of [3] is highly accurate. But given that small lags are of key interest in practice – and that the BP method furnishes no critical values at all when m is exceeded by the model order, since the degrees of freedom would be in essence negative – our correct asymptotic distribution is compelling. Our analysis generalizes and extends previous treatments of the topic, and also furnishes additional insight by allowing examination of the eigenvalues, so that one can understand the real differences between the χ^2 heuristic and the actual limit.

The second issue is resolved in the paper through a sequential testing paradigm, which has precedent in [28], but is developed somewhat differently here. When a series of test statistics is fully dependent, one need not worry about multiple testing, but for independent or partially dependent statistics, getting the Type I error rate is a serious issue. Our approach is effective and practical, as illustrated through our numerical studies.

In particular, our procedure involves an initial specification of a Type I error rate for the entire testing procedure, which is split equally into sequential error rates for conditional probabilities of rejection given that rejection has not yet occurred. The computations of critical values require software to compute the crucial asymptotic covariance matrix $\overline{V}$, which is only approximately idempotent for large lags. Equipped with the MLEs and a knowledge of the fitted SARIMA model, R software can rapidly produce this matrix and determine the corresponding asymptotic distributions of sample autocorrelations of time series residuals[5]. We propose a Monte Carlo method for determining joint and conditional probabilities of test statistics, and for the corresponding critical values. In our implementation this process requires only a few seconds (this time depends on the number of Monte Carlo draws) for each series, no matter its length, and therefore is not onerous. Given the grossly inadequate inferences that can arise from using the classical method, i.e., by ignoring the multiple testing problem, our proposed method is both important and viable.

[5] R code for fitting SARIMA models, computing their time series residuals, computing $\overline{V}$, and calculating the critical values is available upon request.

The crucial defect of the classical method is its inadequate handling of the multiple testing problem; the use of χ^2 critical values is a secondary, and lesser problem. Our derivation of the exact distribution is important chiefly because it allows treatment of the joint distribution – if one were only concerned with a single Q statistic at high lag, then the methodology proposed here would grant little improvement in return for a slight delay in computing time, and we would not advocate it. The key is that typical users of time series software do indeed examine multiple Q statistics simultaneously. Our method is able to address this case, as well as the case where a single Q statistic for a low lag is of interest. The main tradeoff is the additional computational time required.

Appendix A: Time series residual processes

A key concept in time series model fitting is to "whiten the spectrum." A time series sample decorrelated by a given model (of the types considered in Section 2) has asymptotic spectrum given by $\widetilde{f}/\overline{f}_{[\widetilde{\theta}]}$. The Whittle likelihood seeks to fit models by minimizing the integral of an empirical version of the time series residual spectrum, namely the periodogram divided by model spectrum. But with either MLE or Whittle estimation, under typical regularity assumptions the asymptotic time series residual spectrum is $\widetilde{f}/\overline{f}_{[\widetilde{\theta}]}$.

This quantity is featured prominently in the asymptotic analysis of Q statistics – see Remark 1. One might hope that the autocovariances of this residual process would be small (or zero); when the model is correctly specified, $\overline{f}_{[\widetilde{\theta}]} = \widetilde{f}/\widetilde{\sigma}^2$, and the residual process is white noise. Typically, when the model is misspecified, the positive lag autocovariances of the residual process will be nonzero, but this need not always be true. As mentioned in Section 2, the EXP(r) model will have the first r positive lag autocovariances of the residual process equal to zero; this is because the PTVs (by definition) minimize the KL discrepancy, and therefore are zeroes of the gradient function, which has kth component given by the integral of c_k times the residual spectrum. In other words, when fitting a potentially misspecified EXP(r) model such that $\widetilde{\theta}$ is the PTV, then necessarily the first r autocovariances of the residual spectrum are zero, which ensures that the BP and LB test statistics are inconsistent when $m \leq r$.

However, this intriguing property need not be true for misspecified ARMA models. As an example, consider fitting an AR(1) to an MA(1). Say the MA process is written $X_t = (1 + \theta B)\epsilon_t$, so that the PTV for the AR(1) parameter is the lag one autocorrelation, or $\theta/(1+\theta^2)$. Plugging this value into the AR(1) spectrum, the residual spectrum becomes

$$\left| 1 + \frac{\theta^3}{1+\theta^2} z - \frac{\theta^2}{1+\theta^2} z^2 \right|^2,$$

where $z = e^{-i\lambda}$. So the residual process is an MA(2). For an invertible MA, $\theta \in (-1, 1)$, which we henceforth assume. The autocovariances of the residual

process then are given by

$$\gamma_0 = (1+\theta^2)^{-2}\left(1+2\theta^2+2\theta^4+\theta^6\right)$$
$$\gamma_1 = (1+\theta^2)^{-2}\theta^3$$
$$\gamma_2 = (1+\theta^2)^{-1}(-\theta^2).$$

At higher lags the autocovariances are zero. The maximal values of the lag one and two autocorrelations is $\pm 1/6$, and occur for $\theta = \pm 1$. These are clearly the cases of worst model misspecification, whereas $\theta = 0$ implies no model misspecification, and the residual spectrum is white noise.

More generally, these types of calculations are extremely difficult to perform analytically – although the limiting MLEs for fitted AR models are simple enough (they are just solutions of Yule-Walker equations, expressible through the true autocovariances of the DGP), for more general ARMA models the solution requires nonlinear optimization. The point is that residual autocorrelations may be fairly large when there is misspecification present; these are the quantities driving the power of the Q statistics, as is clear from the asymptotic bias in Theorem 1.

Appendix B: Proofs

We first introduce a concept from [19]: we say that $A \sim B$ for two matrices A and B if $Z'AZ - Z'AZ = O_P(1)$ for all vectors Z with uniformly bounded second moments, as the dimension $n \to \infty$.

Lemma 1. *Suppose that the model spectral density is continuously differentiable in λ and is positive. Then*

$$\Sigma^{-1/2}(f_{\widehat{\theta}})L^{(k)}\Sigma^{-1/2}(f_{\widehat{\theta}}) \sim \Sigma(f_{\widehat{\theta}}^{-1}c_k).$$

The result follows from Lemmas 2, 3, and 4 of [19], which can be easily extended to handle spectra depending on random coefficients, so long as the spectra are continuously differentiable.

Proof of Theorem 1. Let the periodogram of the uncentered data be defined as $I(\lambda) = n^{-1}|\sum_{t=1}^{n} X_t e^{-i\lambda t}|^2$ for $\lambda \in [-\pi, \pi]$ (the previous definition above inserted a centering by the sample mean), and note that $\overline{\gamma}_k = \langle \overline{f}_{[\widehat{\theta}]}^{-1} c_k I\rangle$. Since we have a linear functional of the periodogram, it makes no difference whether we consider an integral or a discrete sum over Fourier frequencies, and we may apply Theorem 2 of [20] with the weighting functions $g_{\theta,k} = c_k/\overline{f}_{[\theta]}$. Then $\nabla g_{\theta,k} = -c_k \nabla \overline{f}_{[\theta]}/\overline{f}_{[\theta]}^2$, where the last component is zero (since the derivative is with respect to σ^2). It also follows that

$$\nabla f_\theta = \begin{bmatrix} \sigma^2\, \overline{\nabla}\, \overline{f}_{[\theta]} \\ \overline{f}_{[\theta]} \end{bmatrix},$$

where $\overline{\nabla}$ refers to the gradient with respect to $[\theta]$. Then the expression $b_k(\theta)$ from Theorem 2 of [20] is

$$\begin{aligned} b_k(\theta) &= \langle (f_\theta - \widetilde{f})\nabla g_{\theta,k} + g_{\theta,k}\nabla f_\theta \rangle \\ &= -\langle c_k(f_\theta - \widetilde{f})\nabla \overline{f}_{[\theta]}\overline{f}_{[\theta]}^{-2} - c_k \overline{f}_{[\theta]}^{-1}\nabla f_\theta \rangle \\ &= \langle c_k \overline{f}_{[\theta]}^{-2}\Big(\nabla f_\theta \overline{f}_{[\theta]} - \nabla \overline{f}_{[\theta]}(f_\theta - \widetilde{f})\Big)\rangle \\ &= \langle c_k \overline{f}_{[\theta]}^{-2}\Big(\overline{\nabla}' \overline{f}_{[\theta]}\widetilde{f}, \overline{f}_{[\theta]}^2\Big)'\rangle \\ &= \langle c_k \widetilde{f}\overline{f}_{[\theta]}^{-2}\nabla \overline{f}_{[\theta]}\rangle. \end{aligned}$$

The function $p_{\theta,k}$ of [20] is then given by

$$p_{\theta,k} = -f_\theta^{-2}\, b_k'(\theta)\, M_f^{-1}(\theta)\, \nabla f_\theta.$$

Then the stated formula for $V_{k\ell}(\theta)$ follows from the formula in Theorem 2 of [20]:

$$V_{k\ell}(\theta) = 2\,\langle g_{\theta,k} g_{\theta,\ell}\widetilde{f}^2\rangle + 2\,\langle p_{\theta,k} g_{\theta,\ell}\widetilde{f}^2\rangle + 2\,\langle g_{\theta,k} p_{\theta,\ell}\widetilde{f}^2\rangle + 2\,\langle p_{\theta,k} p_{\theta,\ell}\widetilde{f}^2\rangle.$$

Furthermore, under the Null Hypothesis – evaluating at the PTVs – we obtain $\widetilde{f} = \widetilde{\sigma}^2\,\overline{f}_{[\widetilde{\theta}]}$, and

$$b_k(\widetilde{\theta}) = \widetilde{\sigma}^2\,\langle c_k\, \overline{f}_{[\widetilde{\theta}]}^{-1}\,\nabla \overline{f}_{[\widetilde{\theta}]}\rangle.$$

Moreover, we have

$$\langle \widetilde{f}^2 \nabla f_\theta \nabla' f_\theta f_\theta^{-4}\rangle = \langle \nabla \log f_{\widetilde{\theta}} \nabla' \log f_{\widetilde{\theta}}\rangle,$$

which equals $M_f(\widetilde{\theta})$ (the Hessian of KL discrepancy equals twice the Fisher information when the Null Hypothesis holds). Finally,

$$\langle c_k \widetilde{f}^2 \nabla f_{\widetilde{\theta}} f_{\widetilde{\theta}}^{-2}\overline{f}_{[\theta]}^{-1}\rangle = \langle c_k \nabla f_{\widetilde{\theta}}\overline{f}_{[\widetilde{\theta}]}^{-1}\rangle = \widetilde{\sigma}^2\,\langle c_k \nabla \overline{f}_{[\widetilde{\theta}]}\overline{f}_{[\widetilde{\theta}]}^{-1}\rangle,$$

where the last equality follows from $\langle c_k\rangle = 0$. Since this quantity equals $b_k(\widetilde{\theta})$, the formula for $V_{k\ell}$ simplifies as stated. The convergence in probability of $\overline{\gamma}_0$ follows similarly:

$$\langle I\overline{f}_{[\widehat{\theta}]}^{-1}\rangle \xrightarrow{P} \langle \widetilde{f}\overline{f}_{[\widetilde{\theta}]}^{-1}\rangle = \widetilde{\sigma}^2,$$

so long as the Null Hypothesis holds.

The conditions of the theorem guarantee that Lemma 1 holds. Also since $\overline{X} = O_P(n^{-1/2})$, we can show using Lemmas 2, 3, and 4 of [19], along with the Cauchy-Schwarz inequality, that $\overline{R} = O_P(1/\sqrt{n})$. (We use the fact that $\iota'\Sigma^{-1/2}(\overline{f}_{[\theta]})\iota = O(n)$.) Then $\sqrt{n}\widehat{\gamma}_k = n^{-1/2}R'L^{(k)}R + O_P(n^{-1/2})$. Expanding again using the same techniques,

$$n^{-1/2}R'L^{(k)}R = n^{-1/2}Y'\Sigma^{-1/2}(\overline{f})L^{(k)}\Sigma^{-1/2}(\overline{f})Y + O_P(n^{-1/2}),$$

where Y is the demeaned X vector. In these calculations, the spectral density can be evaluated at any parameter, even $\widehat{\theta}$. Finally, we can apply Lemma 1 to conclude that $\sqrt{n}\widehat{\gamma}_k = O_P(n^{-1/2}) + \sqrt{n}\overline{\gamma}_k$. □

Proof of Corollary 1. The result is immediate from Slutsky's theorem. Also since $\widehat{\gamma}_k$ and $\overline{\gamma}_k$ are asymptotically equivalent, the same follows for the autocorrelations. □

Supplementary Material

Supplement to "The multiple testing problem for Box-Pierce statistics"
(doi: 10.1214/14-EJS892SUPP; .pdf).

References

[1] Ansley, C. and Newbold, P. (1979). On the finite sample distribution of residual autocorrelations in autoregressive-moving average models. *Biometrika* **66** 547–553.

[2] Bloomfield, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika* **60** 217–226. MR0323048

[3] Box, G. and Pierce, D. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association* **65** 1509–1526. MR0273762

[4] Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods*, 2nd ed. Springer, New York. MR1093459

[5] Findley, D. F. and Hood, C. C. H. (1999). X-12-ARIMA and its application to some Italian indicator Series. In *Seasonal Adjustment Procedures – Experiences and Perspectives*, pp. 231–251. Rome: Istituto Nazionale di Statistica (ISTAT).

[6] Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., and Chen, B. C. (1998). New capabilities and methods of the X-12-ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics* **16** 127–177 (with discussion).

[7] Golub, G. and Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore. MR1417720

[8] Hosoya, Y. and Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *Ann. Statist.* **10** 132–153. MR0642725

[9] Imhof, J. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48** 419–426. MR0137199

[10] Kan, R. and Wang, X. (2010). On the distribution of sample autocorrelation coefficients. *Journal of Econometrics* **154** 101–121. MR2558954

[11] Katayama, N. (2008). An improvement of the portmanteau statistic. *Journal of Time Series Analysis* **29** 359–370. MR2392777

[12] Katayama, N. (2009). On multiple portmanteau tests. *Journal of Time Series Analysis* **30** 487–504. MR2560415

[13] Koopman, S., Harvey, A., and Doornik, J. (2000). *STAMP 6.0: Structural Time Series Analyser, Modeller, and Predictor*. Timberlake Consultants, London.

[14] Kwan, A. and Sim, A. (1996). On the finite-sample distribution of modified portmanteau tests for randomness of a Gaussian time series. *Biometrika* **83** 938–943. MR1440057
[15] Kwan, A. and Wu, Y. (1997). Further results on the finite-sample distribution of Monti's portmanteau test for the adequacy of an $ARMA(p,q)$ model. *Biometrika* **84** 733–736. MR1603964
[16] Ljung, G. (1986). Diagnostic testing of univariate time series models. *Biometrika* **73** 725–730. MR0897866
[17] Ljung, G. and Box, G. (1978). On a measure of lack of fit in time series models. *Biometrika* **65** 297–303.
[18] Maravall, A. and Caporello, G. (2004). Program TSW: Revised Reference Manual. *Working Paper 2004, Research Department, Bank of Spain.* http://www.bde.es.
[19] McElroy, T. (2008). Statistical properties of model-based signal extraction diagnostic tests. *Communications in Statistics, Theory and Methods* **37** 591–616. MR2392345
[20] McElroy, T. and Holan, S. (2009). A local spectral approach for assessing time series model misspecification. *Journal of Multivariate Analysis* **100** 604–621. MR2478185
[21] McElroy, T. and Monsell, B. (2014). Supplement to "The multiple testing problem for Box-Pierce statistics". DOI:10.1214/14-EJS892SUPP.
[22] McElroy, T. and Wildi, M. (2013). Multi-step ahead estimation of time series models. *International Journal of Forecasting* **29** 378–394.
[23] McLeod, A.I. (1978). On the distribution of residual autocorrelations in Box-Jenkins models. *Journal of the Royal Statistical Society, Series B* **40** 296–302. MR0522212
[24] Monti, A. (1994). A proposal for residual autocorrelation test in linear models. *Biometrika* **81** 776–780. MR1326425
[25] Peña, D. and Rodriguez, J. (2002). A powerful portmanteau test of lack of fit for time series. *Journal of the American Statistical Association* **97** 601–610. MR1941476
[26] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
[27] Pollock, D. (1999). *A Handbook of Time-Series Analysis, Signal Processing and Dynamics*. Academic Press, New York. MR1737528
[28] Slud, E. and Wei, L. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* **77** 862–868. MR0686410
[29] Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer-Verlag, New York. MR1785484
[30] Tziritas, G. (1987). On the distribution of positive-definite Gaussian quadratic forms. *IEEE Transactions on Information Theory* **33** 895–906. MR0923244