RESEARCH REPORT SERIES
(Statistics #2012-15)

The Multiple Testing Problem for Box-Pierce Statistics

Tucker McElroy
Brian Monsell

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: September 24, 2012

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

The Multiple Testing Problem for Box-Pierce Statistics

Tucker McElroy*and Brian Monsell[†]
U.S. Census Bureau

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. 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.

Keywords. ARIMA models, Ljung-Box statistic, Time series residuals.

Disclaimer 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.

1 Introduction

The most popular time series goodness-of-fit diagnostic test statistics are the portmanteau statistics introduced by Box and Pierce (1970) and extended by Ljung and Box (1978). 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 of these residuals, and the Box-Pierce statistic is constructed from a cumulation of the square of this function over various time lags. Despite many variants (see the

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

[†] Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100, brian.c.monsell@census.gov

treatment in Ljung (1986), Monti (1994), Kwan and Wu (1997), Peña and Rodriguez (2002), and Kan and Wang (2010)), the original statistics remain quite popular, being featured in commonly used statistical software packages such as X-12-ARIMA (Findley et al., 1998), TRAMO-SEATS (Marvall and Caparello, 2004), and STAMP (Koopman et al., 2000). 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.) This general multiple testing phenomena is well-known in the field of statistics, but has received little serious attention in time series analysis[1], where typically many iterative phases of exploratory analysis, model estimation, and model goodness-of-fit testing are involved. 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 Box-Pierce statistics (BPs) and LBs. 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.

Statistics based on sample autocorrelations are generally given a distribution (under the Null Hypothesis that the model was correctly specified) derived from certain asymptotic calculations, and typically two sources of error creep into this substitution for the exact finite sample distribution. Firstly, there is the limiting effect of sample size tending to infinity. Secondly, there is the effect of parameter uncertainty entering into the structure of the residuals. Our first result (given in Section 2) is the joint asymptotic limit of the sample autocovariances and autocorrelations of time series residuals, accounting for parameter uncertainty in the Maximum Likelihood Estimates (MLEs), and this is applied to joint asymptotics for BPs and LBs. This result was first derived in Box and Pierce (1970), and re-capitulated and extended in other works such as Brockwell and Davis (1991), Monti (1994) and Peña and Rodriguez (2002). Our formulation extends beyond the ARMA and SARMA classes of models treated by the above authors, and our results include the important case

[1] Literature such as Box and Jenkins (1976) and Brockwell and Davis (1991) provide cautions against the multiple testing fallacy – i.e., applying the Type I error rates for a single test to a collection of tests – but a procedure is not set forth to integrate the various dependent statistics.

of misspecification. Also the derivations in Box and Pierce (1970) assume that the moving average coefficients decay as sample size increases, which effectively ensures that the limiting covariance matrix of the residual autocorrelations is idempotent; our results dispense with this assumption, which we have found to be untenable for small samples and low LB lags.

Size distortions for BPs observed in finite samples have led to various modifications, such as those in Ljung and Box (1978), Ljung (1986), and later literature. A recent proposal for improving the asymptotic distribution is that of Kan and Wang (2010): the authors base their entire analysis on a finite sample of data drawn from an elliptical distribution, which obviates the first source of approximation error mentioned above. Unfortunately, this elliptical distribution presumes an iid structure, which is manifestly false for time series residuals – as Box and Pierce (1970) correctly derived four decades earlier. Thus, at least from a theoretical standpoint, the Kan and Wang (2010) analysis is not satisfactory, although it performs reasonably well in simulation studies.

Although the joint structure of residual autocorrelations has been considered in the prior literature, to our knowledge this has not been utilized to provide a joint distribution for the suite of BPs and LBs, which is essential for the multiple testing problem. While one might hope that the asymptotic distribution might behave as a random walk of χ^2 variables on 1 degree of freedom – so that conditional on previous BP/LBs, the current statistic is just χ^2 – this is not true in general, as our results show; actually, this result does hold true for the exponential (EXP) model of Bloomfield (1973), but not for ARMA models.

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. Slud and Wei (1982) and later literature) appropriate for sequences of test statistics. We argue that setting the sequential conditional probabilities to a common error rate is intuitive, allowing for a common inferential basis at each step of the sequence of tests, and facilitates computation of p-values and critical values. This general technique is applied to sequences of BP/LBs, which are asymptotically given as Gaussian quadratic forms; details of the algorithm (which depends on computation of the asymptotic covariance matrix of the sample autocorrelations) are given in Section 3.

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. As sample size increases, we expect the new critical values to provide close to nominal rejection rates. For any real time series, we also can plug the χ^2 critical values into the sequential probabilities to determine an asymptotic assessment of the Type I error rate accruing from using the (wrong) marginal critical values; we compare these results to the correct asymptotic joint critical values, by examining many dozens of time series. When a rejection does occur, we can also identify the exact lags of the BP/LBs for which these rejections occur, and use this information to further refine the model. Although the classical approach can also be used to do this, mistakes will occur due to the method being badly mis-sized.

In practice – as demonstrated through our examples – the understatement of the true Type I error rate arising from using classical χ^2 critical values is not so bad as might be feared at first (and in some cases the critical values are actually larger than the appropriate ones). The main reason is that actual BP/LBs tend to be highly correlated with one another, and this dependence offsets the erroneous conclusions of the multiple testing fallacy (recall that if 20 tests are completely dependent, the multiple testing problem evaporates). A secondary reason is that the χ^2 distribution is indeed a decent approximation to the marginal distribution of the BP/LBs for high lags. However, the low lag behavior (where the idempotent approximation of Box and Pierce (1970) for the autocorrelation covariance matrix tends to be invalid) can be quite different from χ^2; these topics are summarized in Section 5, and all proofs are in the Appendix, which also contains some supplementary material on time series residual processes.

2 Asymptotic Theory for Q Statistics

We will refer to either BP or LB statistics as Q statistics. Here we present the main theoretical results of the paper. Consider a sample of size n from a stationary time series, denoted $X = \{X_1, X_2, \cdots, X_n\}'$. (If the raw data is nonstationary, we assume it has been correctly differenced to stationarity already.) The series may have a nonzero mean $\widetilde{\mu}$. We are interested in studying the asymptotic distribution for Q statistics, which are based on linear combinations of squared sample autocorrelations of model residuals. The chief context is the fitting of ARMA and SARMA models to the data X utilizing a Gaussian likelihood function; however, our results also extend to structural models (Harvey, 1989), the chief requirement being a spectral density that is sufficiently smooth with respect to its parameters.

We develop the asymptotic theory from the fundamental results of McElroy and Holan (2009), which reveal the joint distribution of multiple Q statistics. This framework is convenient because it does not restrict us to a certain model class, such as ARMA, and permits us to describe the asymptotics under misspecification as well. It will be seen directly in the asymptotic results how parameter estimation error prevents the limiting marginal distributions from corresponding exactly to a χ^2 variable; the true result is instead a nontrivial Gaussian quadratic form.

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. Given the spectral density f_θ, it is typically a simple matter to compute autocovariances (see below), which together with the mean is sufficient to determine a Gaussian time series' distribution. We proceed to 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. It will allow us to formulate the time series residuals in a general way; by reducing to a discussion of the primitive concept (the spectrum), we automatically treat all cases of practical interest (i.e., stationary time series with finite variance).

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 $< \cdot >$ notation, i.e., $< g > = (2\pi)^{-1} \int_{-\pi}^{\pi} g(\lambda) d\lambda$. 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 simply

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

where $|\Sigma(f_\theta)|$ denotes the determinant of $\Sigma(f_\theta)$. Here ι is a vector of n ones. Maximum likelihood estimation yields $\widehat{\theta}$ as the minimizer of $\mathcal{D}$; we may very well wish to use (1) even when our data is not Gaussian (or is not known to be Gaussian), since Maximum Likelihood Estimates (MLEs) are asymptotically efficient for a broad class of processes (cf. treatment in Taniguchi and Kakizawa (2000)). The mean μ is essentially a nuisance parameter for purposes of the Q statistics.

Now it is convenient to distinguish within our notation the innovation variance from the other parameters, when it is a separate parameter (e.g., in an ARMA model). We will write $f_\theta(\lambda) = \overline{f}_{[\theta]}(\lambda)\sigma^2$, where $\theta = \{[\theta]', \sigma^2\}'$ consists of $r+1$ components, the last component being the innovation variance σ^2. The other parameter components are denoted by $[\theta]$. The "innovation-free" spectral density $\overline{f}_{[\theta]}$ is obtained from f_θ by setting $\sigma^2 = 1$. Note that for structural models the innovation variance is not a parameter (i.e., it is a nonlinear function of the model parameters); in such a case we can identify $\overline{f}_{[\theta]}$ with f_θ in the subsequent treatment; in our exposition we focus on the case that σ^2 is a parameter.

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 Taniguchi and Kakizawa (2000) and McElroy and Wildi (2012) for background. To be precise, $[\widetilde{\theta}]$ is the minimizer of $[\theta] \mapsto < \widetilde{f}/\overline{f}_{[\theta]} >$, and $\widetilde{\sigma}^2 = < \widetilde{f}/f_{[\widetilde{\theta}]} >$. More intuitively, a PTV is the vector (or one of a set of vectors) to which the MLE converges when the model is misspecified (i.e., the spectral density is misspecified). 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 Golub and Van Loan (1996). 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:

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

This is not the only way to define time series residuals; see alternative treatments in Box and Jenkins (1976) and Brockwell and Davis (1991). Although our definition is not the most computationally efficient, it seems quite natural (given the form of the likelihood) and also has the merit of being well-defined for any parametric model (whereas the residuals of the above authors require an ARMA or SARMA specification).

This minimization produces the MLEs $[\widehat{\theta}]$; the innovation variance is estimated by the average of squared residuals, i.e., $\widehat{\sigma}^2 = n^{-1}R'R$. The sample acvs of the residuals are

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

where we have taken the biased definition (this won't matter asymptotically, since we will always have $k = o(n)$ in our treatment).

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

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

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. A third Q statistic, which we won't pursue further here, is $Q_{KS} = 4^{-1}\sum_{k=1}^{m}(n-k-3)\log^2[(1+\widehat{\rho}_k)/(1-\widehat{\rho}_k)]$, discussed in Kwan and Sim (1996).

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 (2) 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 Pollock (1999) 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

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

Likewise, let $\overline{\rho}_k = \overline{\gamma}_k/\overline{\gamma}_0$. Our main theoretical results focus on the joint asymptotic distribution of $\overline{\gamma}_k$ for $k = 0,1,\cdots,m$.

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 $n \to \infty$:*

$$\left\{ \sqrt{n}\overline{\gamma}_k + \sqrt{n} < c_k \left(\widetilde{\sigma}^2 - \widetilde{f}/\overline{f}_{[\widetilde{\theta}]} \right) > \right\}_{k=1}^{m} \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}\left(0, V(\widetilde{\theta})\right). \tag{3}$$

Also $\overline{\gamma}_0 \xrightarrow{P} \widetilde{\sigma}^2$. The asymptotic variance under the null hypothesis of a correct model is given by

$$\begin{aligned} V_{k\ell}(\theta) &= 2\sigma^4 < c_k c_\ell > -2b_k'(\theta) M_f^{-1}(\theta) b_\ell(\theta) \\ b_k(\theta) &=< c_k \overline{f}_{[\theta]}^{-1} \nabla_\theta f_\theta > \\ M_f(\theta) &=< \nabla_\theta \log f_\theta \nabla_\theta' \log f_\theta > . \end{aligned}$$

Here the indexing is $k, \ell = 1, 2, \cdots, m$. Moreover, the same results are true with $\widehat{\gamma}_k$ substituted for $\overline{\gamma}_k$.

Remark 1 The asymptotic bias term $< c_k \left(\widetilde{\sigma}^2 - \widetilde{f}/\overline{f}_{[\widetilde{\theta}]} \right) >$ is identically zero when the model is correctly specified, but otherwise may be nonzero. 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. Clearly, 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 $k = 1, 2, \cdots, m$ the vector $b_k(\theta)$ has $r + 1$ components with the last component equal to $< c_k >= 0$, whereas the first r components are given by $\sigma^2 < c_k \nabla_{[\theta]} \log \overline{f}_{[\theta]} >$. $M_f(\theta)$ is the Hessian of the Kullback-Leibler discrepancy, which under the null hypothesis of a correct model is equal to the Fisher information matrix.

Remark 2 The asymptotic distribution, when the model is correctly specified, is identical with the derivations of Box and Pierce (1970) and later authors, except that their limiting covariance matrix is idempotent (see below). This may assure the reader that our particular definition of time series residual does not alter the asymptotics from the formulation peculiar to ARMA and SARMA models (see Brockwell and Davis (1991) for full treatment).

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 $\widetilde{\gamma}_k =< c_k \widetilde{f} \overline{f}_{[\widetilde{\theta}]}^{-1} >$, and hence the lag k autocorrelation is $\widetilde{\rho}_k =$

$\widetilde{\sigma}^{-2} < c_k \widetilde{f} \overline{f}^{-1}_{\widetilde{[\theta]}} >$, using the fact that $< \widetilde{f} \overline{f}^{-1}_{\widetilde{[\theta]}} >= \widetilde{\sigma}^2$. 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,*

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

for $k = 1, 2, \cdots, m$. Moreover, the same results are true with $\widehat{\rho}_k$ substituted for $\overline{\rho}_k$ in (4).

Remark 3 When the model is correctly specified, the vector of asymptotic residual autocorrelations – denoted by $\widetilde{\rho}$ – is zero. Denoting the limiting covariance matrix in (4) by $\overline{V}(\theta)$, we can write

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

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

$$C_{jk} = [b_k(\theta)]_j/\sigma^2 = < c_k \nabla_{[\theta]} \log \overline{f}_{[\theta]} > .$$

From Corollary 1 and the above Remark, we at once obtain distributional results for the BP and LB statistics. Denote the limiting multivariate normal vector in (4) by Y. The Null Hypothesis states that the model is correctly specified, and hence guarantees hat $\widetilde{\rho} = 0$. We derive the general case (where the Alternative Hypothesis holds, so that perhaps $\widetilde{\rho} \neq 0$) for the Box-Pierce statistic:

$$\begin{aligned} Q_{BP} &= (\sqrt{n}\widehat{\rho})'(\sqrt{n}\widehat{\rho}) \\ &= \sqrt{n}\,(\overline{\rho} - \widetilde{\rho})'\,\sqrt{n}\,(\overline{\rho} - \widetilde{\rho}) + 2\sqrt{n}\,\widetilde{\rho}'\,\sqrt{n}\,(\overline{\rho} - \widetilde{\rho}) + n\,\widetilde{\rho}'\widetilde{\rho} \\ &\overset{\mathcal{L}}{\Longrightarrow} Y'Y + 2\sqrt{n}\,\widetilde{\rho}'Y + n\,\widetilde{\rho}'\widetilde{\rho}. \end{aligned}$$

Formally the result in the last line only makes sense when $\widetilde{\rho} = 0$; otherwise, it is interpreted as giving an order of divergence of the statistic to ∞, such that $n^{-1}Q_{BP} \xrightarrow{P} \widetilde{\rho}'\widetilde{\rho}$. The same result holds for the Ljung-Box statistic, since m is held constant in the asymptotics. Note that the same derivation (under the Null Hypothesis) can also be found in Brockwell and Davis (1991) and Peña and Rodriguez (2002) for ARMA and SARMA models, though the method of analysis is different (being based on time domain techniques, rather than frequency domain techniques).

More generally, the joint asymptotic distribution of several Q_{BP} or Q_{LB} statistics can be described via the Cramer-Wold device by considering $\sum_{m=1}^M \alpha_m Q_m$ (where Q_m denotes either Q statistic, computed using the first m sample autocorrelations), and has limit $Y'AY$ with A an M-dimensional diagonal matrix with mth entry $\sum_{\ell=m}^M \alpha_\ell$. Now since Y is normal with variance $\overline{V}$, we can simply characterize the limiting distribution of the Q statistic via the Laplace Transform (LT) as follows (Tziritas, 1987):

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

This formula, however, is of little use in determining quantiles except in special cases, discussed below. Also see Imhof (1961) for background on this distribution.

Let us focus on the case of just a single Q_M statistic by taking $A = 1_M$. Note that the first term of $\overline{V}$ 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 $\sum_{k=1}^{M} Y_k^2$ is χ^2 on M degrees of freedom. But when parameter uncertainty is present, $\overline{V}$ will not be diagonal and the LT might *not* be the product of LTs for χ^2 variables.

In Box and Pierce (1970) the authors propose a formula for $\overline{V}$ that is somewhat different. In fact, accounting for differences in notation, they essentially propose

$$V^\sharp(\theta) = 1_M - C'\big(CC'\big)^{-1}C,$$

which means replacing the upper left portion of the Fisher information matrix with $2CC'$. Of course $N \neq 2CC'$, but Box and Pierce (1970) 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 $1_M - V^\sharp$ is idempotent with rank r, so that $M - r$ eigenvalues of $V^\sharp$ are equal to unity, and the remaining r eigenvalues are zero. We can apply Proposition 2 of Tziritas (1987) 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 V has all its eigenvalues equal to either zero or one.

Consider the EXP(r) model of Bloomfield (1973), 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 $< I/\overline{f}_{[\theta]} >$ 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 $< \overrightarrow{c} I/\overline{f}_{[\theta]} >$, 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 exponential 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 Box and Pierce (1970) 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 $\overline{V}$ are equal to unity. Then the limiting LT is that of a χ^2_{M-r} variable plus an independent variable with LT

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

with $\lambda_1, \cdots, \lambda_r$ the r nonzero eigenvalues of $2C'N^{-1}C$. In practice, these r eigenvalues can be quite close to zero for values of M as small as 5 or 6, depending upon the model and the underlying process. This makes the inversion of $\overline{V}$ infeasible and the approximation of Box and Pierce (1970) quite useful. On the other hand, the degrees of freedom in the Box and Pierce (1970) approach is $M-r$, so that no distributional result can be used when $M \leq r$; in these cases, $\overline{V}$ may be quite different from idempotency, and is moreover 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 $\overline{V}$ at parameter estimates, such as MLEs.

3 Joint 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 idea is to use conditional probabilities of prior statistics, for which critical values have already been obtained, such that the Type I error at each step is controlled as desired.

In this section we describe a novel sequential testing procedure, which is related to ideas first published in Slud and Wei (1982). We first develop the ideas somewhat generally, and then specialize to the case of BPs and LBs.

3.1 Sequential Testing

The basic idea of sequential testing is related to ideas in the biostatistics literature (Slud and Wei, 1982). Consider a sequence of test statistics $\{T_k\}$ for $k = 1, 2, \cdots, M$ (where $M = \infty$ is allowed, although in our LB application $M < \infty$). Quantities of considerable interest are the conditional probabilities

$$\alpha_{k+1} = \mathbb{P}\left[T_{k+1} > x_{k+1} | T_k \leq x_k, \cdots, T_1 \leq x_1\right]. \tag{5}$$

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

$$p_{k+1} = \mathbb{P}\left[T_{k+1} \leq x_{k+1}, T_k \leq x_k, \cdots, T_1 \leq x_1\right]. \tag{6}$$

Of course we have the relation $\alpha_{k+1} = 1 - p_{k+1}/p_k$. The overall assessment of the procedure involves computing the probability that at least one test rejects, i.e., the event $\cup_k\{T_k > x_k\}$. Now while Slud and Wei (1982) focus on computing the joint probabilities p_k (6), we emphasize the α_ks, although there is a ready equivalence between the two approaches. We first note a few elementary facts.

Let $\alpha = \mathbb{P}[\cup_k\{T_k > x_k\}]$, which is called the Type I error rate of the sequential procedure – whereas the sequential Type I error rates are the quantities α_k. This is nomenclature. By induction, $p_{k+1} = \Pi_{\ell=1}^{k+1}(1 - \alpha_\ell)$. In addition $\alpha = 1 - p_M$, which also equals $\sum_{k=1}^{M} \alpha_k p_{k-1}$ (with $p_0 = 1$), as shown in Slud and Wei (1982).

Consider two extreme cases: first, if all the tests are independent, then α_k is the kth 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_k such that the corresponding sequential conditional or joint probabilities aggregate appropriately to α. As an initial step, one must choose the numbers p_k or α_k to satisfy the appropriate constraint; then we may determine the x_k sequentially given certain information about the distribution functions. Whereas Slud and Wei (1982) choose to work with the p_k, partitioning them such that they sum to α, we in contrast work with the α_k. 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_k 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_k 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 $p_k = (1 - \alpha_0)^k$ 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 Slud and Wei (1982) 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_k, 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_k 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 $p_1, p_2, \cdots, p_M$ in sequence – potentially using Monte Carlo simulation if analytic formulas are unavailable – and then determine each α_k, which depend on knowing p_k and p_{k-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 $1 \leq \ell \leq k$ and $1 \leq j \leq J$, then

$$p_k \approx J^{-1} \sum_{j=1}^{J} 1_{\{T_k^{(j)} \leq x_k, \cdots, T_1^{(j)} \leq x_1\}} = J^{-1} \#\{j : \max_{1 \leq \ell \leq k} \left[T_\ell^{(j)} - x_\ell \right] \leq 0\}.$$

The second expression is useful for encoding the method, e.g., in R (R Development Core Team, 2009).

Now consider computation of critical values, using the α-rationing scheme described above. First compute x_1 such that $1 - \alpha_0 = 1 - \alpha_1 = \mathbb{P}[T_1 \leq x_1]$, 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 $x_1, x_2, \cdots, x_k$, we wish to compute x_{k+1} such that $1 - \alpha_{k+1}$ equals $\mathbb{P}[T_{k+1} \leq x_{k+1} | T_k \leq x_k, \cdots, T_1 \leq x_1]$. So consider a subset $L(x_k, x_{k-1}, \cdots, x_1) \subset \{1, 2, \cdots, J\}$ consisting of only those Monte Carlo draws j such that $\max_{1 \leq \ell \leq k} [T_\ell^{(j)} - x_\ell]$ is less than or equal to zero. Then

$$1 - \alpha_0 = 1 - \alpha_{k+1} \approx L^{-1}(x_k, x_{k-1}, \cdots, x_1) \sum_{j \in L(x_k, x_{k-1}, \cdots, x_1)} 1_{\{T_{k+1}^{(j)} \leq x_{k+1}\}}.$$

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

We have found such a procedure to be effective, written in R, for determining critical values of Q statistics (more details given below). Determination of p-values can be more problematic in practice, since the formula $\alpha_{k+1} = 1 - p_{k+1}/p_k$ produces negative conditional probabilities whenever $p_{k+1} < p_k$ – of course, such an occurrence is impossible in theory, but due to Monte Carlo error our estimate of p_{k+1} may be less than p_k. However, there seems to be little difficulty with the critical values that are produced; additional accuracy can be obtained by increasing J. 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.

3.2 Q Statistics

We here 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 $Q = [Q_1, Q_2, \cdots, Q_M]$, 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

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

and we will denote this function by $F_Q\left(q_1, \cdots, q_M\right)$. Let Y denote a Gaussian vector of length M with covariance matrix $\overline{V}$; this is simple to simulate once $\overline{V}^{1/2}$ is computed, which is valid even when $\overline{V}$ is non-invertible (e.g., idempotent) or close to singular. Let Y^2 denote the component-wise squaring of Y, 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^2. So by Monte Carlo methods we can easily approximate F_Q.

$\overline{V}$ can be computed from the theoretical results, utilizing the null hypothesis, with MLEs substituted for PTVs. Details on the computation of $\overline{V}$ for the case of a SARIMA model follow, utilizing standard Box and Jenkins (1976) notation. Although the ARMA case is treated in Brockwell and Davis (1991), we are unaware of published details on the SARMA case. We are not concerned with nonstationary differencing here, so we instead consider just the SARMA(p,q,P,Q) of period s given by the difference equation

$$\Phi(B)\Xi(B^s)X_t = \Psi(B)\Omega(B^s)\epsilon_t.$$

The nonseasonal AR polynomial is $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$; the nonseasonal MA polynomial is $\Psi(B) = 1 - \psi_1 B - \cdots - \psi_q B^q$; the seasonal AR polynomial is $\Xi(B^s) = 1 - \xi_1 B^s - \cdots - \xi_P B^{Ps}$; the seasonal MA polynomial is $\Omega(B^s) = 1 - \omega_1 B^s - \cdots - \omega_Q B^{Qs}$. These parameters are entered in this order in the vector $[\theta]$, i.e., $[\theta]' = \{\phi_1, \cdots, \phi_p, \psi_1, \cdots, \psi_q, \xi_1, \cdots, \xi_P, \omega_1, \cdots, \omega_Q\}$. Of course $r = p + q + P + Q$.

In order to present the results it is useful to introduce a notation for reciprocal polynomials. We consider the inverses of $\Phi(B)$, $\Psi(B)$, $\Xi(B^s)$, and $\Omega(B^s)$ as causal power series in B, and denote the resulting coefficients by $\phi_t^\sharp$, $\psi_t^\sharp$, $\xi_t^\sharp$, and $\omega_t^\sharp$ respectively, for $t \geq 0$. Note that in the case of the seasonal AR and seasonal MA the coefficients are for the power series in B, not B^s. Then by direct calculation the jkth entry of C is given by

$$
\begin{aligned}
C_{jk} &= \begin{cases}
\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos(k\lambda)\left(\frac{z^j}{\Phi(z)}+\frac{z^{-j}}{\Phi(\overline{z})}\right)\,d\lambda & 1\leq j\leq p\\
\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos(k\lambda)\left(\frac{-z^j}{\Psi(z)}+\frac{-z^{-j}}{\Psi(\overline{z})}\right)\,d\lambda & 1\leq j-p\leq q\\
\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos(k\lambda)\left(\frac{z^{sj}}{\Xi(z^s)}+\frac{z^{-sj}}{\Xi(\overline{z}^s)}\right)\,d\lambda & 1\leq j-p-q\leq P\\
\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos(k\lambda)\left(\frac{-z^{sj}}{\Omega(z^s)}+\frac{-z^{-sj}}{\Omega(\overline{z}^s)}\right)\,d\lambda & 1\leq j-p-q-P\leq Q
\end{cases}\\
&= \begin{cases}
\phi^{\sharp}_{k-j} & 1\leq j\leq p\\
-\psi^{\sharp}_{k-j} & 1\leq j-p\leq q\\
\xi^{\sharp}_{k-sj} & 1\leq j-p-q\leq P\\
-\omega^{\sharp}_{k-sj} & 1\leq j-p-q-P\leq Q
\end{cases}
\end{aligned}
$$

Here $z=e^{-i\lambda}$. The calculations for N are more complicated, but there are essentially 16 blocks to compute. Recall that our notation is $\gamma_h(f)=(2\pi)^{-1}\int_{-\pi}^{\pi}\overline{z}^h f(\lambda)\,d\lambda$. Then

$$
N_{jk} = \begin{cases}
\gamma_{j-k}\left(|\Phi(z)|^{-2}\right)+\gamma_{k-j}\left(|\Phi(z)|^{-2}\right) & 1\leq j,k\leq p\\
-\gamma_{j-k}\left(\Phi^{-1}(\overline{z})\Psi^{-1}(z)\right)-\gamma_{k-j}\left(\Phi^{-1}(z)\Psi^{-1}(\overline{z})\right) & 1\leq j\leq p,\ 1\leq k-p\leq q\\
\gamma_{j-k}\left(\Phi^{-1}(\overline{z})\Xi^{-1}(z^s)\right)+\gamma_{k-j}\left(\Phi^{-1}(z)\Xi^{-1}(\overline{z}^s)\right) & 1\leq j\leq p,\ 1\leq k-p-q\leq P\\
-\gamma_{j-k}\left(\Phi^{-1}(\overline{z})\Omega^{-1}(z^s)\right)-\gamma_{k-j}\left(\Phi^{-1}(z)\Omega^{-1}(\overline{z}^s)\right) & 1\leq j\leq p,\ 1\leq k-p-q-P\leq Q\\
-\gamma_{j-k}\left(\Psi^{-1}(\overline{z})\Phi^{-1}(z)\right)-\gamma_{k-j}\left(\Psi^{-1}(z)\Phi^{-1}(\overline{z})\right) & 1\leq j-p\leq q,\ 1\leq k\leq p\\
\gamma_{j-k}\left(|\Psi(z)|^{-2}\right)+\gamma_{k-j}\left(|\Psi(z)|^{-2}\right) & 1\leq j-p\leq q,\ 1\leq k-p\leq q\\
-\gamma_{j-k}\left(\Psi^{-1}(\overline{z})\Xi^{-1}(z^s)\right)-\gamma_{k-j}\left(\Psi^{-1}(z)\Xi^{-1}(\overline{z}^s)\right) & 1\leq j-p\leq q,\ 1\leq k-p-q\leq P\\
\gamma_{j-k}\left(\Psi^{-1}(\overline{z})\Omega^{-1}(z^s)\right)+\gamma_{k-j}\left(\Psi^{-1}(z)\Omega^{-1}(\overline{z}^s)\right) & 1\leq j-p\leq q,\ 1\leq k-p-q-P\leq Q\\
\gamma_{j-k}\left(\Xi^{-1}(\overline{z}^s)\Phi^{-1}(z)\right)+\gamma_{k-j}\left(\Xi^{-1}(z^s)\Phi^{-1}(\overline{z})\right) & 1\leq j-p-q\leq P,\ 1\leq k\leq p\\
-\gamma_{j-k}\left(\Xi^{-1}(\overline{z}^s)\Psi^{-1}(z)\right)-\gamma_{k-j}\left(\Xi^{-1}(z^s)\Psi^{-1}(\overline{z})\right) & 1\leq j-p-q\leq P,\ 1\leq k-p\leq q\\
\gamma_{j-k}\left(|\Xi(z^s)|^{-2}\right)+\gamma_{k-j}\left(|\Xi(z^s)|^{-2}\right) & 1\leq j-p-q\leq P,\ 1\leq k-p-q\leq P\\
-\gamma_{j-k}\left(\Xi^{-1}(\overline{z}^s)\Omega^{-1}(z^s)\right)-\gamma_{k-j}\left(\Xi^{-1}(z^s)\Omega^{-1}(\overline{z}^s)\right) & 1\leq j-p-q\leq P,\ 1\leq k-p-q-P\leq Q\\
-\gamma_{j-k}\left(\Omega^{-1}(\overline{z}^s)\Phi^{-1}(z)\right)-\gamma_{k-j}\left(\Omega^{-1}(z^s)\Phi^{-1}(\overline{z})\right) & 1\leq j-p-q-P\leq Q,\ 1\leq k\leq p\\
\gamma_{j-k}\left(\Omega^{-1}(\overline{z}^s)\Psi^{-1}(z)\right)+\gamma_{k-j}\left(\Omega^{-1}(z^s)\Psi^{-1}(\overline{z})\right) & 1\leq j-p-q-P\leq Q,\ 1\leq k-p\leq q\\
-\gamma_{j-k}\left(\Omega^{-1}(\overline{z}^s)\Xi^{-1}(z^s)\right)-\gamma_{k-j}\left(\Omega^{-1}(z^s)\Xi^{-1}(\overline{z}^s)\right) & 1\leq j-p-q-P\leq Q,\ 1\leq k-p-q\leq P\\
\gamma_{j-k}\left(|\Omega(z^s)|^{-2}\right)+\gamma_{k-j}\left(|\Omega(z^s)|^{-2}\right) & 1\leq j-p-q-P\leq Q,\ 1\leq k-p-q-P\leq Q
\end{cases}
$$

The recommended method is first to obtain the causal power series representation for each of the four reciprocal polynomials – the coefficients can be taken out to a thousand terms or so with little effort[2]. Then for the block diagonal terms in N one can utilize a procedure to compute autocovariances from the resulting moving average – the advantage of doing the calculation in two steps is that problems arising from the direct approach when auto-regressive roots are close to unity can be circumvented. For the off-block diagonal terms in N, one should rationalize by multiplying top and bottom by the conjugate; then the terms are calculated as certain linear combinations of autocovariances. Code developed using version 2.7.1 of R (R Development Core Team, 2009) for the entire computation of $\overline{V}$ is available from the authors.

Example: Consider the Airline model, which has spectral density $\overline{f}_{[\theta]}(\lambda) = |1-\theta_1 z|^2 |1-\theta_2 z^{12}|^2$. Then we obtain

$$N = \begin{bmatrix} \frac{2}{1-\theta_1^2} & 0 \\ 0 & \frac{2}{1-\theta_2^2} \end{bmatrix}.$$

Taking $M = 2$ for illustration, we obtain

$$C = \begin{bmatrix} -1 & -\theta_1 \\ 0 & 0 \end{bmatrix} \qquad \overline{V} = \begin{bmatrix} \theta_1^2 & -\theta_1(1-\theta_1^2) \\ -\theta_1(1-\theta_1^2) & 1-\theta_1^2(1-\theta_1^2) \end{bmatrix}.$$

The determinant of $\overline{V}$ is θ_1^4, which is zero iff $\theta_1 = 0$, in which case the eigenvalues are zero and one. Hence if the true DGP has $\widetilde{\theta}_1 = 0$, then the first two LB statistics will be asymptotic to point mass at zero (at lag 1) and χ^2 on one degree of freedom (at lag 2). When $\widetilde{\theta}_1 \neq 0$ this degeneracy is *not* the case. Note that when adopting our method there is no problem when degeneracy happens to arise (since the square root of $\overline{V}$ is well-defined even when the matrix is singular).

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

[2] In R one can use the function *ARMAtoMA*.

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 at the end of section 3.1.

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.

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, 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 2 and 3 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 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 (Ljung and Box, 1978). 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.

4.2 Census Bureau Time Series

We consider nine seasonal time series published by the Census Bureau from the Monthly Retail Sales Survey. Table 1 gives the names and descriptions of these series[3].

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;

[3] 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`.

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

Table 1: Descriptions of Monthly Retail Sales Series, covering the period 1992 through 2007 (Source: U.S. Census Bureau).

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 default 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.

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 Elec 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 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, maximal, and restricted schemes. The problem lags occur at lag 12 and 24 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 (Figure 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 (Elec 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.

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 Box and Pierce (1970) 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 our analysis 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 Slud and Wei (1982), 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[4]. 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.

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.

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

Appendix

A.1 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

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

At higher lags the autocovariances are zero. The maximal values of the lag one and two autocorrelations is ±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.

A.2 Proofs of Technical Results

We first introduce a concept from McElroy (2008): we say that $A \sim B$ for two matrices A and B if $Z'AZ - Z'AZ = O(1)$ for all vectors Z with uniformly bounded second moments, as the dimension $n \rightarrow \infty$.

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 McElroy (2008), 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 $\overline{\gamma}_k =< \overline{f}_{[\widehat{\theta}]}^{-1} c_k I >$. 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 McElroy and Holan (2009) with the weighting functions $g_{\theta,k}(\lambda) = \cos(k\lambda)/f_\theta(\lambda)$. The formulas for the asymptotic bias and null variance matrix follow from the same results. The variance matrix is complicated, but typically only needs to be computed under the Null Hypothesis, which greatly simplifies things. The convergence in probability of $\overline{\gamma}_0$ follows as well from the weak convergence.

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 McElroy (2008), along with the Cauchy-Schwarz inequality, that $\overline{R} = O_P(1/\sqrt{n})$. (We use the fact that $\iota'\Sigma^{-1/2}(\overline{f})\iota = O(n)$.) 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. □

References

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

[2] Box, G. and Jenkins, G. (1976) *Time Seris Analysis, Forecasting and Control.* San Francisco: Holden-Day.

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

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

[5] 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).

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

[7] Harvey, A. (1989) *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

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

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

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

[11] 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.

[12] 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.

[13] Ljung, G. (1986) Diagnostic testing of univariate time series models. *Biometrika* **73**, 725–730.

[14] Ljung, G. and Box, G. (1978) On a measure of lack of fit in time series models. *Biometrika* **65**, 297–303.

[15] Maravall, A. and Caporello, G. (2004) Program TSW: Revised Reference Manual. *Working Paper 2004, Research Department, Bank of Spain.* http://www.bde.es

[16] McElroy, T. (2008) Statistical properties of model-based signal extraction diagnostic tests. *Communications in Statistics, Theory and Methods* **37**, 591–616.

[17] McElroy, T. and Holan, S. (2009) A local spectral approach for assessing time series model misspecification. *Journal of Multivariate Analysis* **100**, 604–621.

[18] McElroy, T. and Wildi, M. (2012) Multi-Step Ahead Estimation of Time Series Models. Forthcoming, *International Journal of Forecasting.*

[19] Monti, A. (1994) A proposal for residual autocorrelation test in linear models. *Biometrika* **81**, 776–780.

[20] 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.

[21] 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`

[22] Pollock, D. (1999) *A Handbook of Time-Series Analysis, Signal Processing and Dynamics.* New York: Academic Press.

[23] 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.

[24] Taniguchi, M. and Kakizawa, Y. (2000) *Asymptotic Theory of Statistical Inference for Time Series.* Springer-Verlag, New York.

[25] Tziritas, G. (1987) On the distribution of positive-definite Gaussian quadratic forms. *IEEE Transactions on Information Theory* **33**, 895–906.

Table 2. 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

Table 2: 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 3. 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	.0044	.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

Table 3: 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 $n = 120, 180, 240$.

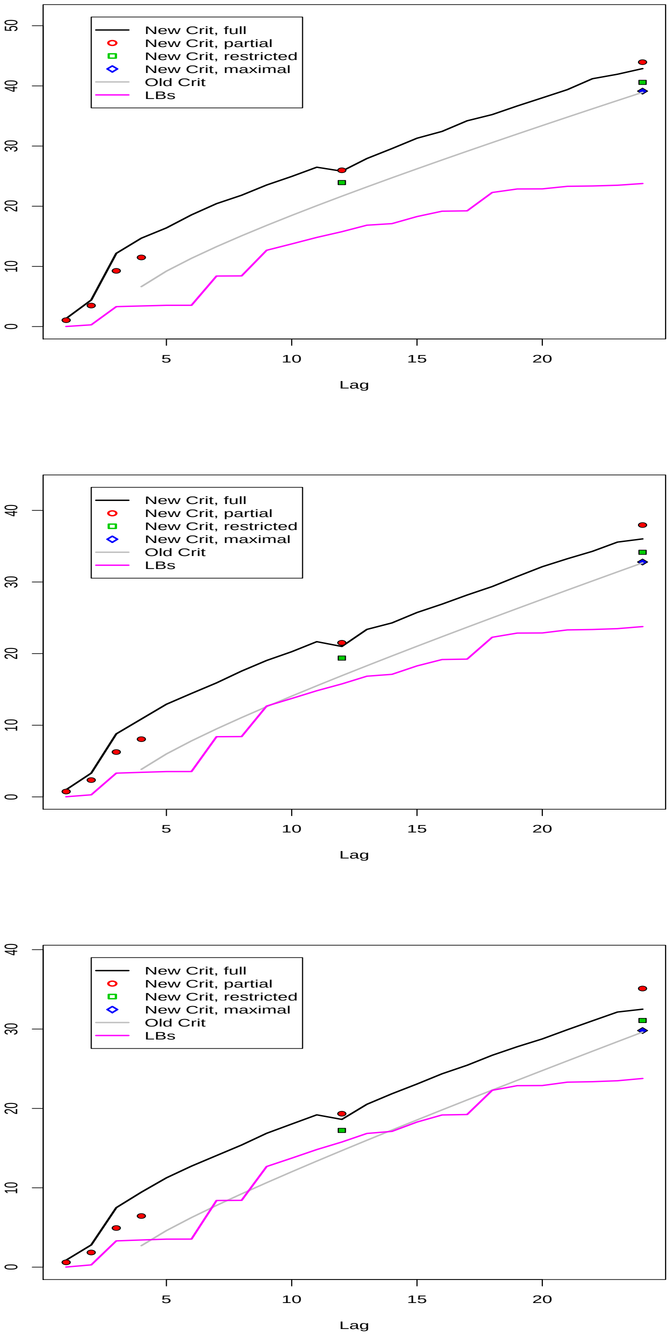

Figure 1: Critical values and LB statistics for Motor series. From top to bottom, the procedure's Type I error rates are .01, .05, and .10.

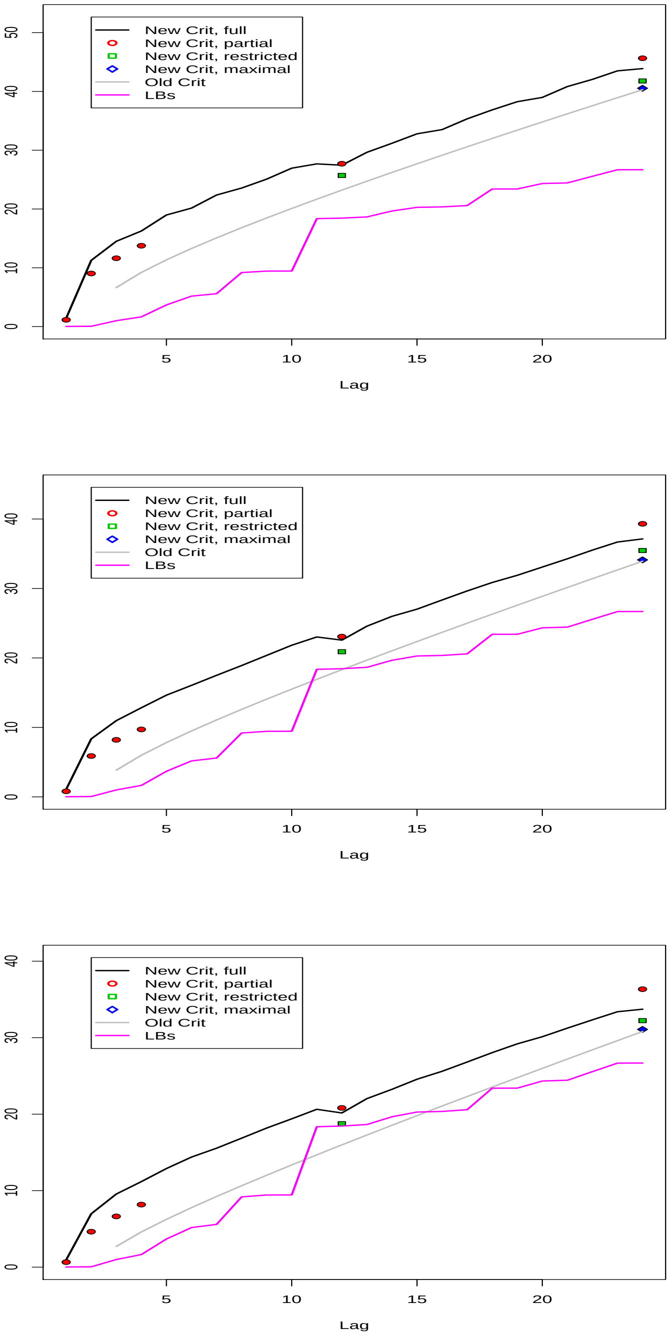

Figure 2: Critical values and LB statistics for Food series. From top to bottom, the procedure's Type I error rates are .01, .05, and .10.

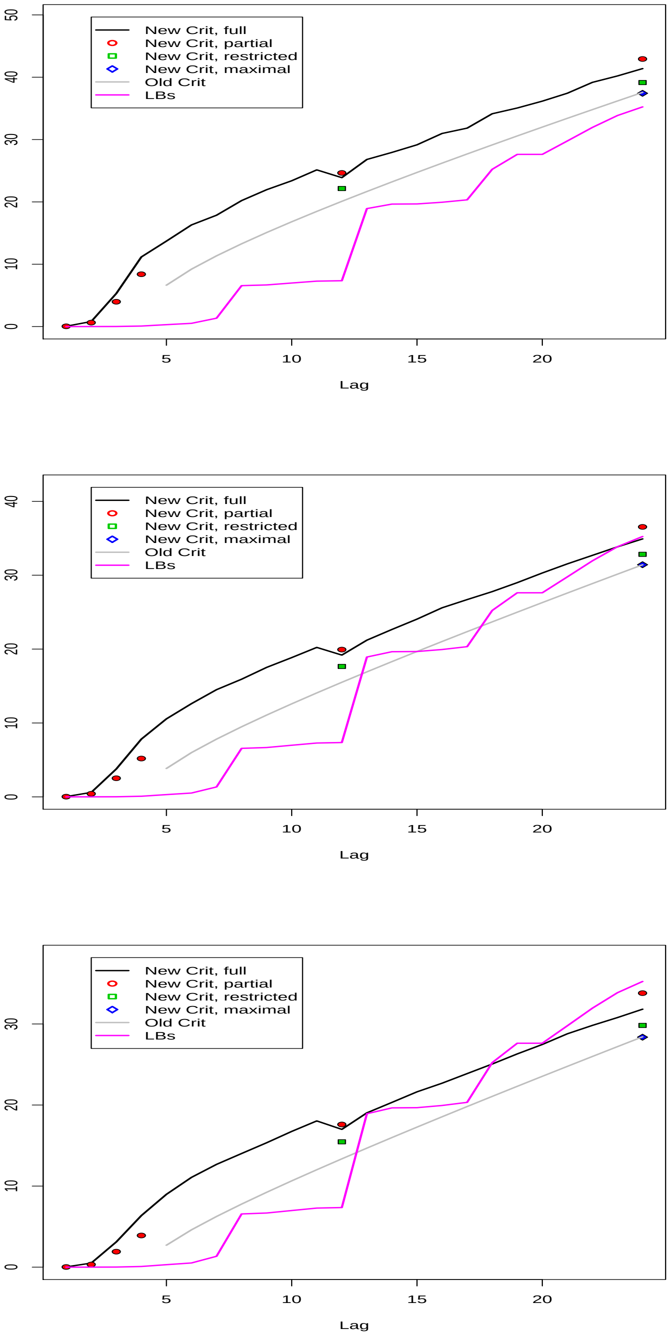

Figure 3: Critical values and LB statistics for Elec series. From top to bottom, the procedure's Type I error rates are .01, .05, and .10.

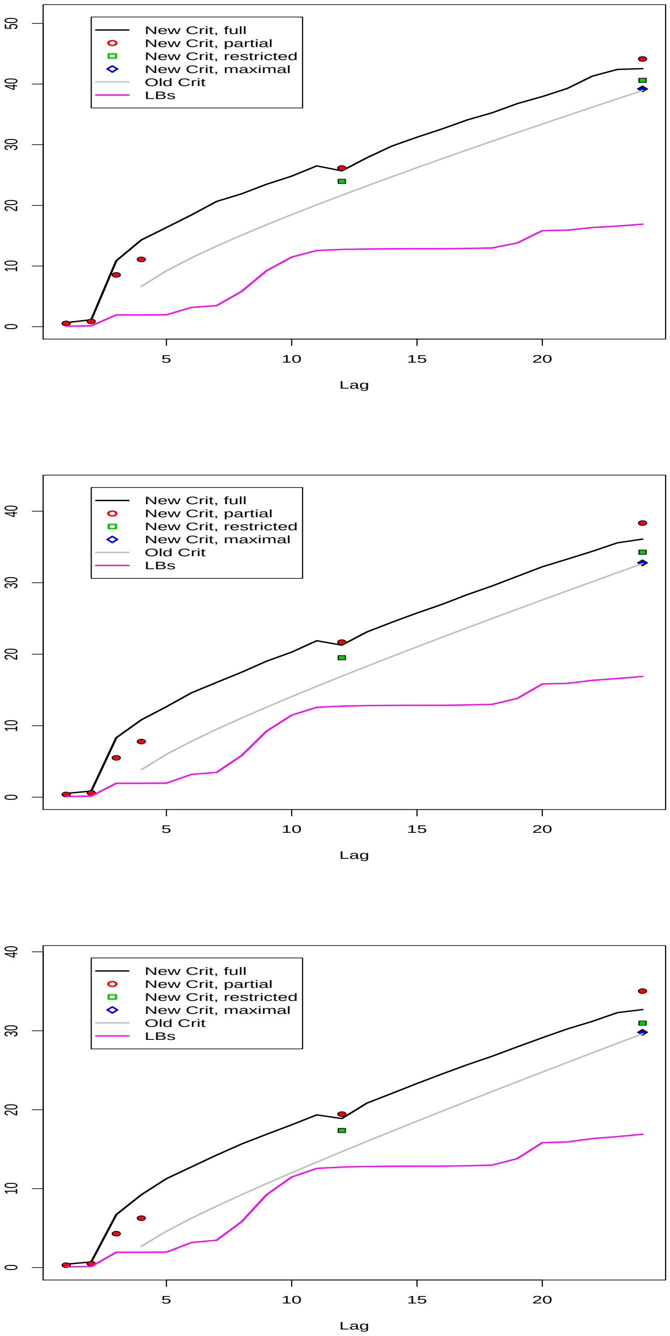

Figure 4: Critical values and LB statistics for Furn series. From top to bottom, the procedure's Type I error rates are .01, .05, and .10.

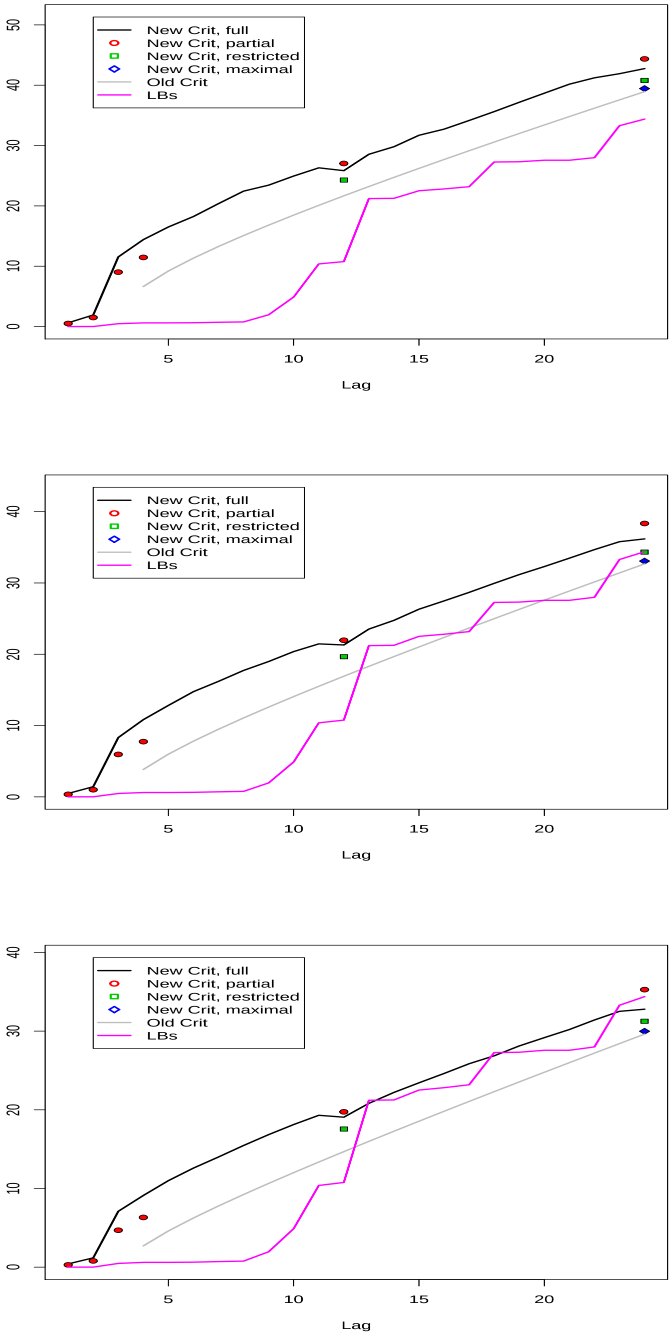

Figure 5: Critical values and LB statistics for Gas series. From top to bottom, the procedure's Type I error rates are .01, .05, and .10.

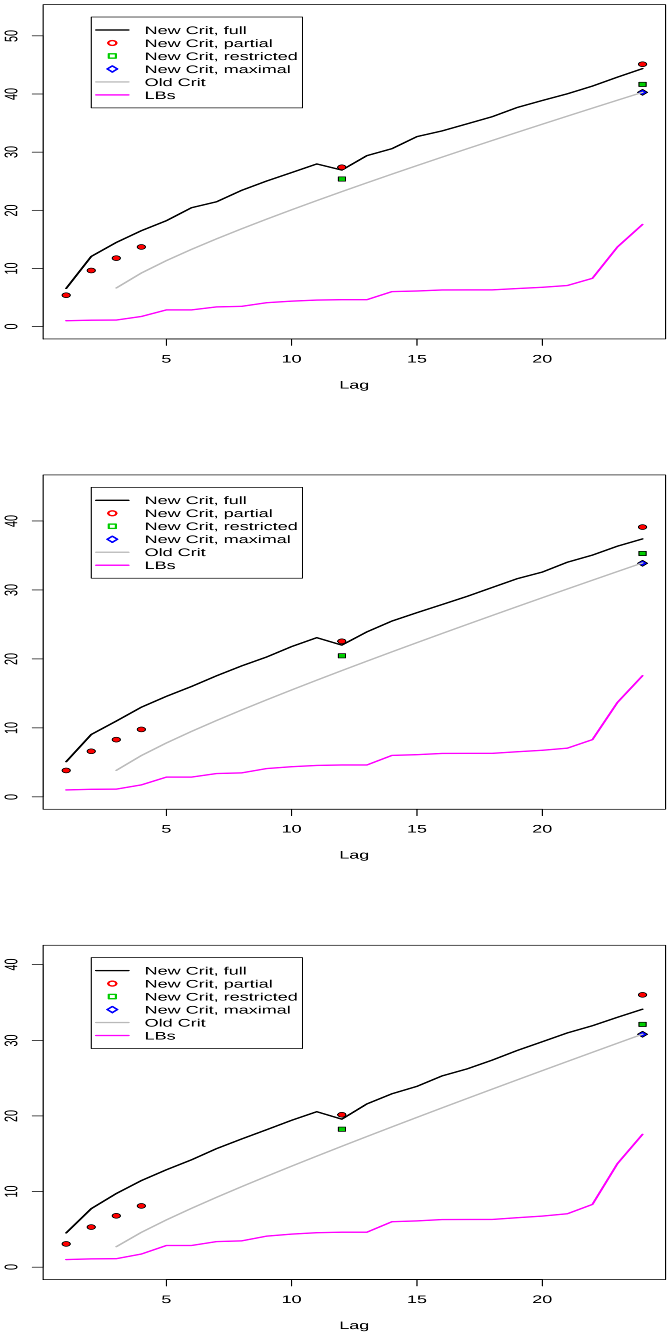

Figure 6: Critical values and LB statistics for GenMerch series. From top to bottom, the procedure's Type I error rates are .01, .05, and .10.

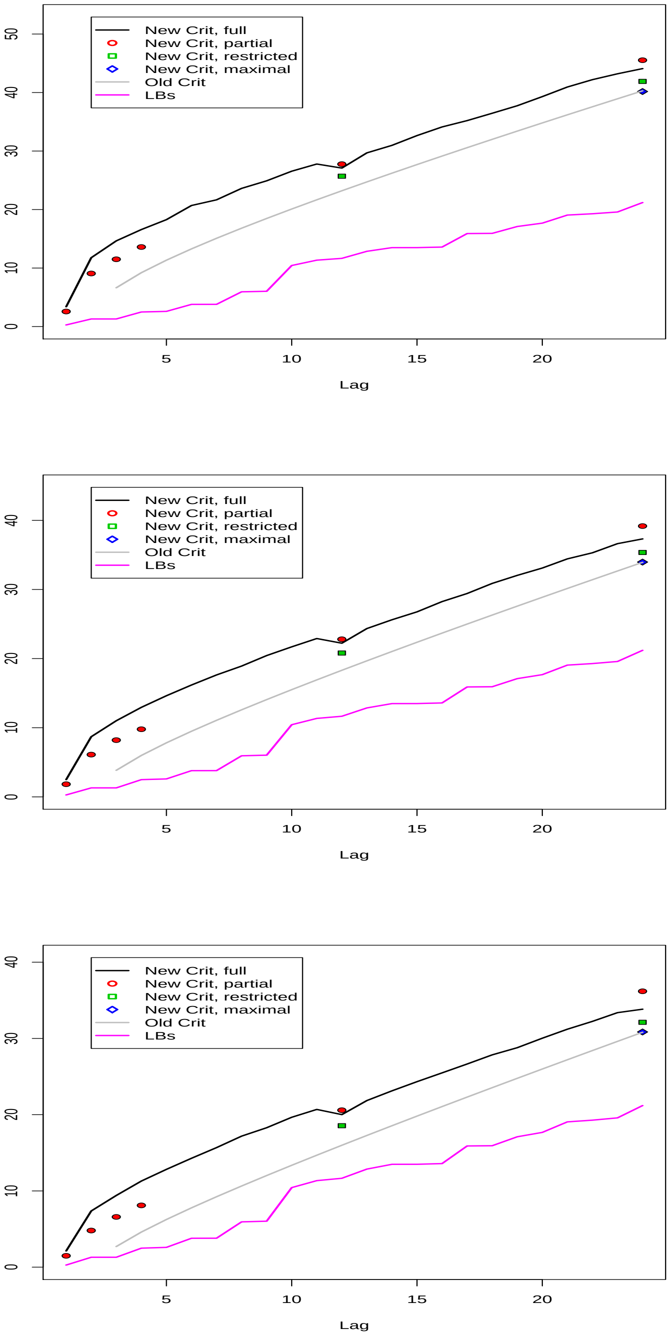

Figure 7: Critical values and LB statistics for Groc series. From top to bottom, the procedure's Type I error rates are .01, .05, and .10.

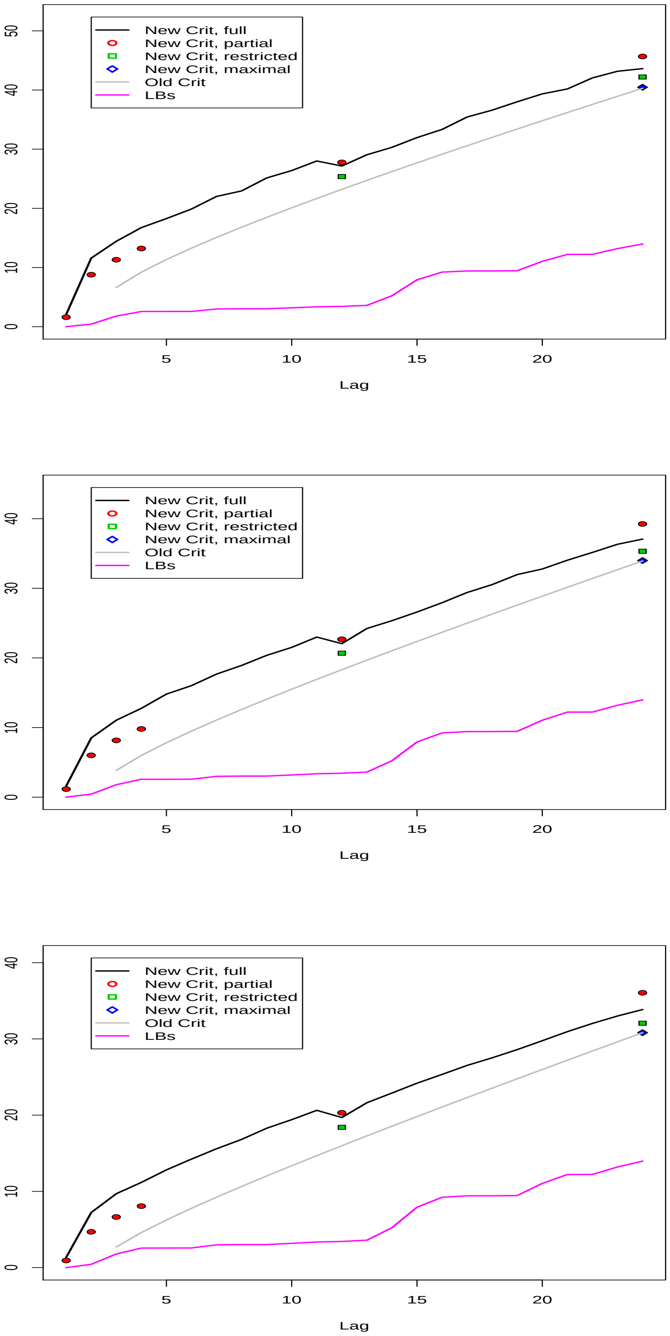

Figure 8: Critical values and LB statistics for MenCloth series. From top to bottom, the procedure's Type I error rates are .01, .05, and .10.

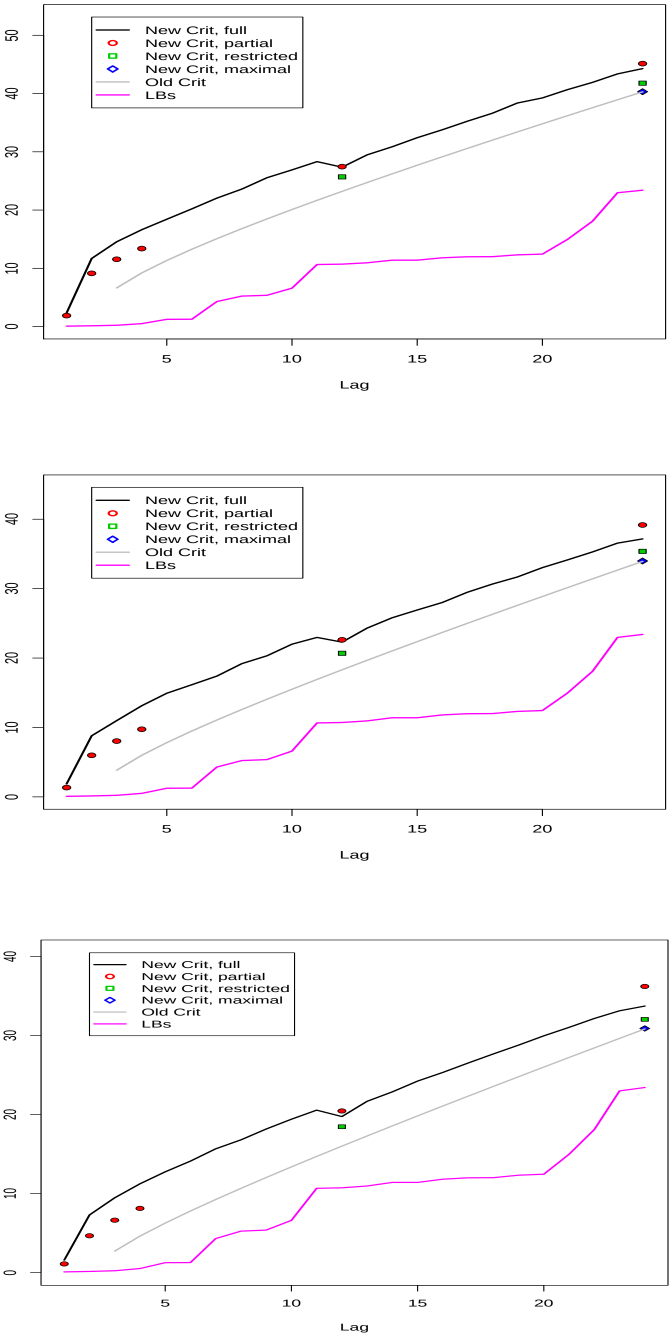

Figure 9: Critical values and LB statistics for WomCloth series. From top to bottom, the procedure's Type I error rates are .01, .05, and .10.