Critical Values for Time Series Diagnostics

Jeremy Penzer

Department of Statistics, London School of Economics
Houghton Street, London, WC2A 2AE, UK.

March 12, 2001

Abstract

Problems involving calibration of sequences of correlated statistics arises in time series diagnostics. Critical values are required to locate unusual points from sequences of diagnostic statistics. We put forward simple criteria, which reliably identify significant outliers, level shifts and changes in seasonal pattern. The circumstances under which the correlation of sequences of statistics becomes important in determining critical values are discussed. These tests are applied to real data sets to detect behaviour previously overlooked.

KEYWORDS: Bonferonni inequalities; Change point detection; Order statistics; Outliers; Level shifts; Structural breaks; Structural time series models.

1 Introduction

Sequences of diagnostic statistics are used to detect unusual behaviour in time series. The statistics may be associated with particular interventions (Tsay 1986; Tsay 1988; Chang, Tiao, and Chen 1988; De Jong and Penzer 1998), measures of influence (Peña 1991; Atkinson, Koopman, and Shephard 1997) or smoothed estimates of disturbance terms (Harvey and Koopman 1992). Some (Peña 1991; De Jong and Penzer 1998) put forward distributions for individual statistics. Others suggest arbitrary benchmarks against which the significance of the diagnostic statistics can be measured, for example;

- "c is a prespecified constant with typical values for c of 3.0, 3.5 or 4.0" (Box, Jenkins, and Reinsel 1994, p. 472),
- "indications of outliers and/or structural changes arise for values greater than 2 in absolute value" (Harvey and Koopman 1992).

Box, Jenkins, and Reinsel (1994) provide no justification for their choice of c. The value chosen by Harvey and Koopman (1992) would give a significance level of approximately 5% for a single test statistic with a standard normal distribution. When, as with time series diagnostics, many statistics are considered simultaneously, this criterion defines a test with a much higher significance level. For n independent standard normal statistics, using the critical value 2, the probability of a type I error is approximately $1-(0.95)^n$. Although the approaches described provide useful diagnostic information, better benchmarks against which significance can be measured are clearly desirable.

De Jong and Penzer (1998) put forward smoother-based statistics for detecting shocks in time series. They propose two types of diagnostic statistic; τ_i, which is the analogue of the usual regression t-statistic, and the quadratic form, τ_i^2. For a series of length n, the index i takes values between 1 and n. The statistics τ_i and τ_i^2 include as special cases most of the other diagnostics described in the literature and will be referred to throughout this paper as the diagnostic statistics.

2 Two approaches to simultaneous testing problems

Two broad approaches to simultaneous testing problems, based on the last order statistic, are identified below. Using the last order statistic has the advantage of preserving the location of any unusual behaviour.

1. **Distribution of largest order statistic**: If $X_1, \ldots, X_n$ is a sequence of independent identically distributed continuous random variables and $X_{(n)}$ is the largest order statistic from this sequence, it is well known (see for example, Stuart and Ord (1987, Volume 1, p. 446)) that
$$P(X_{(n)} < x) = P(X_1 < x)^n.$$

Thus, if c is the $100(1-\alpha)$ percentile for $X_{(n)}$, that is $P(X_{(n)} > c) = \alpha$, and F is the distribution function of X_1,

$$F(c) = (1-\alpha)^{1/n}.$$

This has an application in simultaneous testing. If we observed n independent identically distributed random variables, c is used as a critical value for the maximum, yielding a test with significance level α. This can readily be adapted to two-tailed tests. Tables 1 and 2 give respectively values of c for standard normal and chi-squared on one degree of freedom. Note that c is an approximately linear function of $\log n$.

Table 1: Critical values, c, where $X_j \sim iN(0,1)$ for $j = 1, \ldots, n$

	n		
$100 \times \alpha$	100	1000	10000
0.1	4.265	4.753	5.199
0.5	3.890	4.417	4.891
1.0	3.718	4.264	4.752
2.5	3.477	4.053	4.562
5.0	3.283	3.884	4.412
10.0	3.075	3.706	4.253

Table 2: Critical values, c, where $X_j \sim i\chi^2_1$ for $j = 1, \ldots, n$

	n		
$100 \times \alpha$	100	1000	10000
0.1	19.511	23.933	28.436
0.5	16.444	20.834	25.274
1.0	15.127	19.502	23.924
2.5	13.389	17.741	22.145
5.0	12.069	16.400	20.790
10.0	10.732	15.038	19.412

2. **Bonferroni inequality**: David (1981, p. 91) describes a method for approximating the upper percentage points of the maxima of a sequence of dependent random variables using the Bonferroni inequalities. For any collection of events $A_1, \ldots, A_n$, the first Bonferroni inequality is

$$\sum_i P(A_i) - \sum_{i<j} \sum_j P(A_i A_j) \leq P(\bigcup_{i=1}^{n} A_i) \leq \sum_i P(A_i). \tag{1}$$

If $X_1, \ldots, X_n$ are random variables with identical marginal distributions, and A_i is the event $\{X_i > k\}$, then $P(\bigcup_{i=1}^{n} A_i) = P(X_{(n)} > k)$, and (1) becomes;

$$nP(X_1 > k) - \sum_{i<j} \sum_{j} P(X_i > k, X_j > k) \leq P(X_{(n)} > k) \leq nP(X_1 > k). \qquad (2)$$

If k is a point such that $P(X_1 > k) = \alpha/n$ then $P(X_{(n)} > k) \leq \alpha$. Hence, using k as a critical value for the maximum gives a test with significance level less than or equal to α. Tables 3 and 4 give values of k such that $P(X_1 < k) = 1 - \alpha/n$.

Table 3: Bonferroni approximate critical values, k, where $X_j \sim N(0,1)$ for $j = 1, \ldots, n$

	n		
$100 \times \alpha$	100	1000	10000
0.1	4.265	4.753	5.199
0.5	3.891	4.417	4.892
1.0	3.719	4.265	4.753
2.5	3.481	4.056	4.565
5.0	3.291	3.891	4.417
10.0	3.090	3.719	4.265

Table 4: Bonferroni approximate critical values, k, where $X_j \sim \chi^2_1$ for $j = 1, \ldots, n$

	n		
$100 \times \alpha$	100	1000	10000
0.1	19.512	23.934	28.437
0.5	16.448	20.839	25.279
1.0	15.137	19.512	23.934
2.5	13.412	17.765	22.170
5.0	12.116	16.448	20.839
10.0	10.828	15.137	19.512

The two approaches yield similar results across a range of sample sizes and significance levels. In principal, neither provide suitable critical values for time series diagnostics. The diagnostic statistics are correlated so the key assumption of the first approach does not hold. The Bonferroni approximate critical value k could be used. However, the lower bound of equation (2) is hard to evaluate and the approach provides no other means of gauging how much smaller than α the significance level will be. To get some indication of the practical effect of correlation on the critical values, a simulation experiment is conducted.

3 Simulated critical values for structural time series models

Simulation is used to investigate the way in which correlation in the diagnostics influences the position of critical values. Two models are considered, the local level and the basic structural model.

Local level model

The local level model is defined by

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, & \varepsilon_t &\sim iN(0, \sigma_\varepsilon^2), \\ \mu_{t+1} &= \mu_t + \eta_t, & \eta_t &\sim iN(0, \sigma_\eta^2). \end{aligned} \tag{3}$$

where $\{\varepsilon_t\}$ and $\{\eta_t\}$ are mutually uncorrelated and $q = \sigma_\eta/\sigma_\varepsilon$ is the signal-noise ratio. Simulations are conducted using Ssfpack (Koopman, Shephard, and Doornik 1999) and Ox (Doornik 1998). For ease of exposition we concentrate on τ^2-statistics. There are two natural diagnostic statistics for the local level model; the measurement τ^2 which is associated with outliers and the level τ^2 which is used to detect level shifts. For each given value of q, 10000 series are simulated and the maximal τ^2 value for each replication stored. Percentiles are calculated from the ordered vector of stored maximal values. Averages from repeating this experiment 30 times, for series of length 100, are reported in tables 5 and 6.

Table 5: Percentiles for measurement τ^2 for local level model

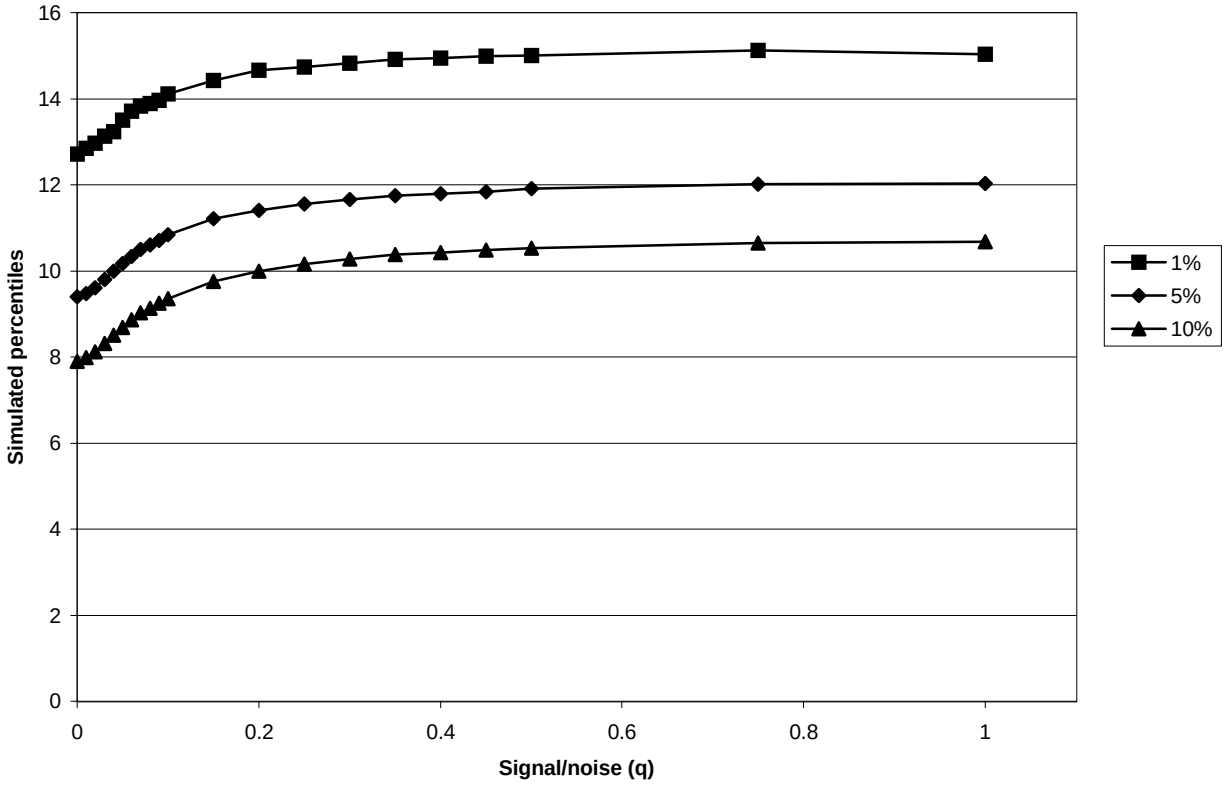
	q						
$100 \times \alpha$	0.0	0.05	0.1	0.5	1	5	∞
1.0	15.16	15.11	15.12	15.15	15.07	15.07	15.07
5.0	12.07	12.03	12.07	12.08	12.03	11.99	12.00
10.0	10.73	10.72	10.72	10.72	10.71	10.65	10.64

Table 6: Percentiles for level τ^2 for local level model

	q						
$100 \times \alpha$	0.0	0.05	0.1	0.5	1	5	∞
1.0	12.71	13.51	14.11	15.01	15.04	15.12	15.11
5.0	9.39	10.18	10.83	11.91	12.03	12.08	12.06
10.0	7.90	8.68	9.36	10.53	10.68	10.73	10.73

The percentiles for the measurement τ^2-statistics are approximately constant with q and, for practical purposes, identical to those given by assuming that the statistics are an independent

Figure 1: Simulated percentiles for level τ^2-statistic



sequence. In contrast, the percentiles for level τ^2-statistics increase as the signal-noise ratio increases, as illustrated in figure 1. For $q > 0.5$, the percentiles are, once again, practically identical to those given by assuming the statistics are independent.

An intuitive explanation of the behaviour of the percentiles is provided by looking at the correlation structure of the diagnostic statistics. If $\rho_m(.)$ and $\rho_\ell(.)$ are, respectively, the auto-correlation functions for measurement and level τ-statistics, then

$$\rho_m(h) \approx -\frac{1}{2}(1-\lambda)\lambda^{h-1}, \quad \rho_\ell(h) \approx \lambda^h,$$

where $\lambda = -(\sqrt{q^2+4q}-2-q)/2$. Measurement τ^2-statistics are very weakly positively correlated and it is not surprising that the percentiles are close to those for an uncorrelated sequence. The level τ^2-statistic has strong positive correlation when λ is large, that is, when q is small. The diagnostics have constant mean and variance over time. Thus, as q decreases and the degree of correlation goes up, the level τ^2-statistic series becomes smoother with smaller extremes. This causes the observed reduction in the percentiles.

Basic structural model

An example of a basic structural model (BSM) is

$$\begin{array}{rcrcrcll} y_t & = & \mu_t & + & \gamma_t & + & \varepsilon_t, & \varepsilon_t \sim iN(0,\sigma_\varepsilon^2), \\ \mu_{t+1} & = & \mu_t & + & \beta_t & + & \eta_t, & \eta_t \sim iN(0,\sigma_\eta^2), \\ \beta_{t+1} & = & & & \beta_t & + & \zeta_t, & \zeta_t \sim iN(0,\sigma_\zeta^2), \\ \gamma_{t+1} & = & -\gamma_t & \ldots & -\gamma_{t-s+2} & + & \omega_t, & \omega_t \sim iN(0,\sigma_\omega^2), \end{array} \tag{4}$$

where the disturbances ε_t, η_t, ζ_t and ω_t are mutually uncorrelated. Model (4) has a dummy seasonal component although a trigonometric seasonal component could also be used (see Harvey (1989)). The parameters of importance in determining the behaviour of diagnostics are the signal-noise ratios $q_\eta = \sigma_\eta/\sigma_\varepsilon$, $q_\zeta = \sigma_\zeta/\sigma_\varepsilon$ and $q_\omega = \sigma_\omega/\sigma_\varepsilon$. In addition to measurement and level diagnostics, we consider seasonal τ^2-statistics which are used to detect sudden changes in seasonal pattern. Initial investigations indicate that measurement τ^2-statistics are not affected by changes in any of the parameters and seasonal statistics are only influenced by changes in q_ω. Statistics for detecting level changes are affected by both q_η and q_ζ. These observations hold for trigonometric seasonality, and for different seasonal patterns and series lengths.

Simulations, using 10 000 replications and averaging over 30 experiments, provide further insight into the behaviour of the diagnostics. Results at 5% significance, for a series of length 100, using a quarterly dummy BSM, are given in tables 7 and 8. The values of the percentiles for seasonal τ^2 are only substantially lower than those for an independent sequence when q_ω is small. For level τ^2-statistics a marked reduction only occurs when both q_η and q_ζ are small.

Table 7: Percentiles for level τ^2 in BSM

		q_η			
		0.01	0.1	1.0	10.0
q_ζ	0.01	10.80	10.89	11.83	11.97
	0.1	11.56	11.60	11.87	11.99
	1.0	11.93	11.90	11.95	11.98
	10.0	11.92	11.93	11.92	11.95

Table 8: Percentiles for seasonal τ^2 in BSM

		q_ω			
		0.01	0.1	1.0	10.0
seasonal component	1	10.68	11.12	11.87	11.95
	2	10.03	10.38	11.64	11.92
	3	10.71	11.13	11.89	11.97

4 Illustrative examples

Flow of the Nile

A series which has been used extensively in the study of outliers and level shifts (Cobb 1978; Balke 1993; Harvey, Koopman, and Penzer 1998) is the flow of the Nile at Aswan (10^8 m^3) from 1871 to 1970, see figure 2. The data are well fitted by a local level model with $\hat{q} = 0.31$. Figures 3 and 4 show respectively the measurement and level τ^2-statistics with critical values generated by simulation marked. The only apparent shock is a level shift in 1899 which is significant at 10%. A discussion of choice of significance levels is in the conclusion. Including the appropriate intervention, the model has $\hat{q} = 0.00$ and no significant outliers or structural breaks.

Gas consumption by other final users

Figure 5 shows quarterly gas consumption (logs) by other final users, which includes domestic consumption, between 1960 quarter 1 and 1986 quarter 4. Durbin and Koopman (2000) use a non-Gaussian model for this series. The hyperparameter estimates from fitting a BSM with quarterly dummy seasonal component are given in table 9. In this instance, the critical values for measurement and seasonal τ^2-statistics differ little from those for an independent sequence. The approximate 5% and 10% critical values are respectively 12.3 and 10.9. Durbin and Koopman (2000) identify a positive outlier in 1970 quarter 3 followed by a negative outlier in 1970 quarter 4. The measurement τ^2-statistics and critical values in figure 6 clearly pick out this feature. Using the seasonal diagnostics put forward in Penzer (2000) we can identify features not mentioned by other authors. Figure 7 shows the τ^2-statistics associated with the second seasonal component once the outliers in 1970 quarters 3 and 4 have been taken into account. The peak in 1970 quarter 2 corresponds to a seasonal shift with a relative increase in consumption in quarter 4 and a decrease in quarter 2 from 1970 quarter 4 onwards. This may be a result of the introduction of North Sea gas and the accompanying increased use of gas for domestic heating. A further decrease in quarter 2 and increase in quarter 4 from 1980 quarter 2 onwards is indicated by the significant τ^2-statistic associated with the second seasonal component in 1979 quarter 4. Hyperparameter estimates and t-statistics for the final fitted model are given in table 9.

	Initial model	With interventions	Intervention	t-statistic
$\hat{\sigma}_\varepsilon$	0.0427	0.0219	outlier 1970Q3	8.754
$\hat{\sigma}_\eta$	0.0000	0.0174	outlier 1970Q4	-9.383
$\hat{\sigma}_\zeta$	0.0028	0.0022	seasonal shift 1970Q2	6.416
$\hat{\sigma}_\omega$	0.0575	0.0300	seasonal shift 1979Q4	-4.2066

Table 9: Gas example – estimated hyperparameters and t-statistics for interventions

Figure 2: Nile example – flow of the Nile at Aswan (10^8 m^3)

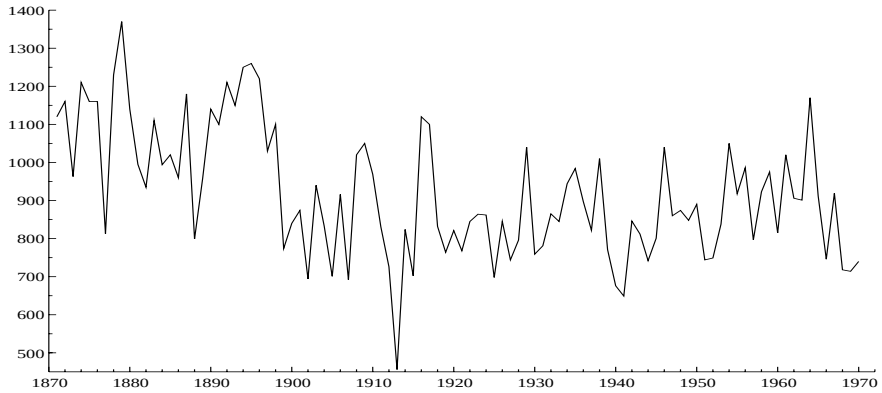


Figure 3: Nile example – measurement τ^2-statistic

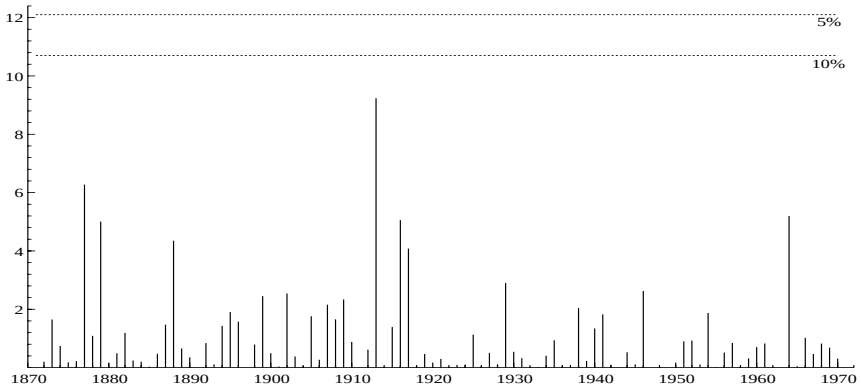


Figure 4: Nile example – level shift τ^2-statistic

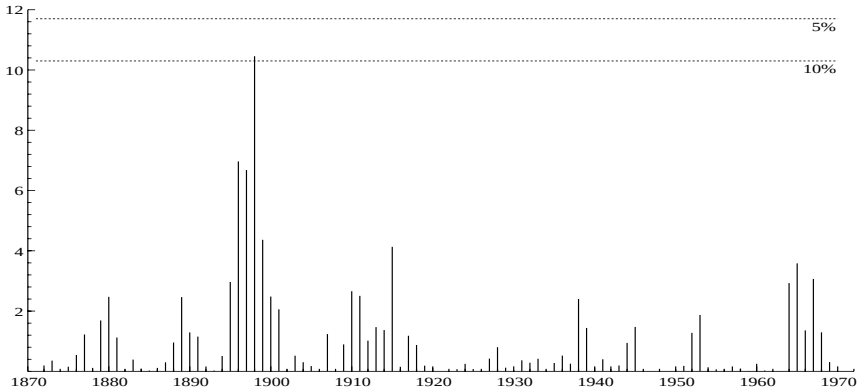


Figure 5: Gas example – gas consumption by other final users (logs)

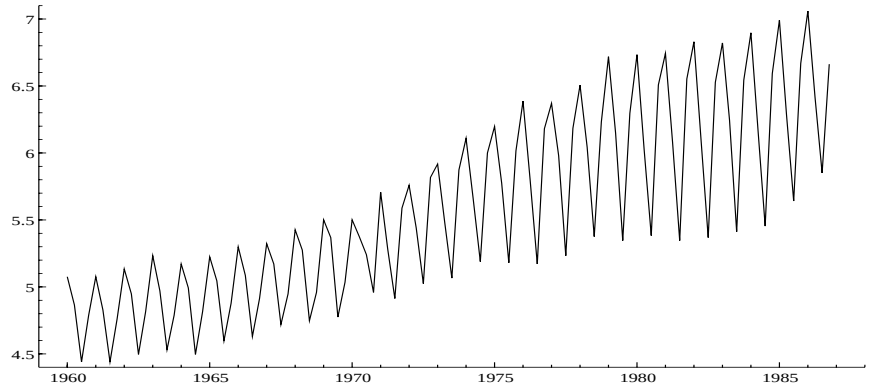


Figure 6: Gas example – measurement τ^2-statistic for initial model

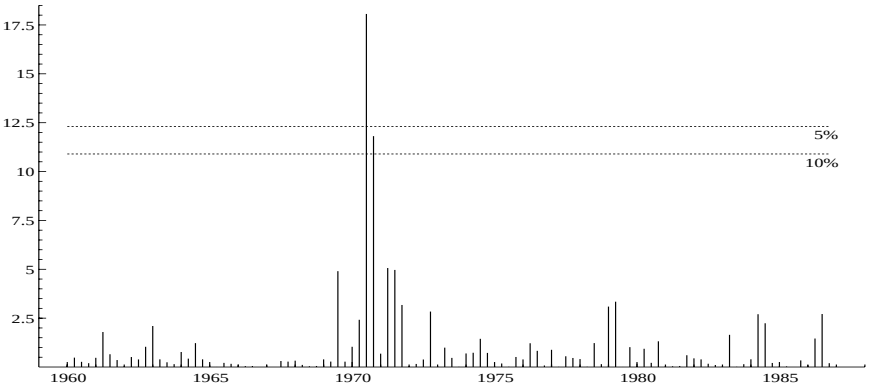
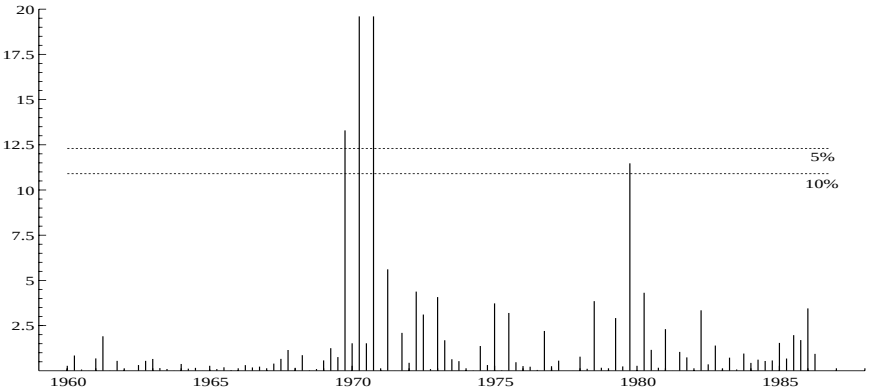


Figure 7: Gas example – second seasonal τ^2-statistic for model with measurement interventions



5 Conclusion

Plots of diagnostic statistics are an effective means of detecting outliers and structural changes. The addition of critical values provides benchmarks against which the significance of the statistics can be judged. In many instances, the statistics, although correlated, can be treated as independent when calculating percentiles. Under these circumstances the critical values are entirely determined by the level of significance, the null distribution of the diagnostic statistic and the sample size. For some parameter values, the degree of dependence in the diagnostics is sufficient to require smaller critical values. For example, in a BSM, when both q_η and q_ζ are less than 0.5, the level τ^2-statistics can no longer be treated as independent.

An appropriate choice of significance level is crucial in determining the outcome of any diagnostic procedure. It could be argued that there are few disadvantages in false positive identification of outliers or structural breaks. In addition, estimating hyperparameters under the null hypothesis may mask the impact of shocks. Both of these observations suggest larger values are appropriate significance levels. However, over-fitting of interventions reduces the model's value as a predictive tool. A significance level of 5% or 10% usually provides a reasonable compromise. In many instances we would like to check for several types of shock simultaneously. This will increase the critical values although the exact effect is dependent on the scheme used for detection. The critical values provided by this work open up the possibility of developing an automatic stepwise procedure for detecting and fitting outliers and structural breaks.

References

Atkinson, A. C., S. J. Koopman, and N. Shephard (1997). Detecting shocks: outliers and breaks in time series. *Journal of Econometrics 80*, 387–422.

Balke, N. S. (1993). Detecting level shifts in time series. *Journal of Business and Economic Statistics 11*, 81–92.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis: Forecasting and Control* (3rd ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Chang, I., G. C. Tiao, and C. Chen (1988). Estimation of time series parameters in presence of outliers. *Technometrics 30*, 193–204.

Cobb, G. W. (1978). The problem of the Nile: conditional solution to a change point problem. *Biometrika 65*, 243–251.

David, H. A. (1981). *Order Statistics* (2nd ed.). New York: John Wiley.

De Jong, P. and J. R. Penzer (1998). Diagnosing shocks in time series. *Journal of the American Statistical Association 93*, 796–806.

Doornik, J. A. (1998). *Ox: an Object-Oriented Matrix Programming Language*. West Wickham, Kent: Timberlake Consultants Press.

Durbin, J. and S. J. Koopman (2000). Time series analysis of non-gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society, Series B 62*, 3–56.

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

Harvey, A. C. and S. J. Koopman (1992). Diagnostic checking of unobserved component time series models. *Journal of Business and Economic Statistics 10*, 377–389.

Harvey, A. C., S. J. Koopman, and J. Penzer (1998). Messy time series: a unified approach. *Advances in Econometrics 13*, 103–144.

Koopman, S. J., N. Shephard, and J. A. Doornik (1999). Statistical algorithms for models in state space form using Ssfpack 2.2. *Econometrics Journal 2*, 113–166.

Peña, D. (1991). Measuring influence in dynamic regression models. *Technometrics 33*, 93–101.

Penzer, J. R. (2000). Fast diagnostics for seasonal shifts in time series. LSE Department of Statistics, Research Report.

Stuart, A. and J. K. Ord (1987). *Kendall's Advanced Theory of Statistics* (5th ed.). London: Griffin.

Tsay, R. S. (1986). Time series model specification in the presence of outliers. *Journal of the American Statistical Association 81*, 132–141.

Tsay, R. S. (1988). Outliers, level shifts and variances changes in time series. *Journal of Forecasting 7*, 1–20.