RESEARCH REPORT SERIES
(Statistics #2018-13)

A Time Series Model Information Criterion Based Upon Residual Entropy

James Livsey
Tucker McElroy

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

Report Issued: December 17, 2018

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.

A Time Series Model Information Criterion Based Upon Residual Entropy

James Livsey and Tucker McElroy

Abstract

The time series model selection problem is strongly rooted in residual analysis. In a typical application many competing models are fitted, wherein the model features are so divergent that direct comparison statistics, such as the likelihood ratio or Akaike Information Criterion, are meaningless. This is because a time series model involves identification of transformation, unit roots, fixed regression effects, and serial correlation patterns. Given this difficulty, we instead advocate an indirect comparison through the entropy of the residual series, assessed through serial correlation and marginal Gaussianity. The ideal residual – namely, a residual series that behaves like Gaussian white noise – is our benchmark against which all model residuals are compared, and provides the basis by which to judge diverse models' superiority. We propose an entropy information criterion that is minimized at the ideal residual, and increases with deviations from either a Gaussian marginal or a non-white serial correlation pattern. Distribution theory is provided, with supporting empirical studies as well as illustrations on retail data. It is known that underfitting (the failure to specify enough model structure) is flagged through the presence of residual serial correlation; a key finding of our research is that overfitting (specifying superfluous model structure) is flagged through the presence of light-tailed 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 modeling of time series can be envisioned as the task of identifying a transformation that reduces the data vector to a residual with maximum entropy. Whereas Jaynes (1957a, b) discusses the maximum entropy principle in broad terms, recent literature on model-free prediction (Politis 2015) has stimulated interest in entropy-maximizing data transformations (henceforth, entropic transformations). Such transformations typically involve marginal transforms (c.f., Box and Cox (1964), Breiman and Friedman (1985)), dummy regressors or stochastic covariates, unit-root filters, and whitening transformations arising from fitted models (Box and Jenkins (1970)). The implicit goal of such efforts is to produce a residual series with Gaussian marginal structure without serial correlation, i.e., a Gaussian white noise sequence. This objective is discussed under the rubric of the model-free prediction principle, in Remark 2.3 of Politis (2013). Because the Gaussian distribution has maximal entropy among contenders with a given first and second moment structure, and because entropy is increased by decreasing serial correlation, it follows that the ideal residual has maximum entropy (Park and Bera, 2009).

Furthermore, as a practical matter the classical methods for fitting time series models yield the best results – in terms of forecast performance and interpretability – when the residual series has maximum entropy. This is because residual serial correlation demands a refined stochastic model (this follows from the Wold decomposition, and basic approximation results for stationary time series, as discussed in Brockwell and Davis (2009)), whereas heavy tails in the residual marginal indicate outliers or extremes corresponding to unexplained phenomena (see discussion in McElroy (2016a)). Remaining serial correlation or excess kurtosis will each contribute in different ways to an increased forecast error variance, reducing the effectiveness of the model. Therefore, it is commonplace in frequentist time series methodology and software to encounter model diagnostics formulated as tests of residual serial correlation (Box and Pierce, 1970) and Gaussian marginal structure (Shapiro and Wilk, 1965). So we may view the residual diagnostic procedures of TRAMO-SEATS (Gómez, V. and A. Maravall, 1996) and X-12-ARIMA (U.S. Census Bureau, 2015) as tests of residual maximum entropy.

Although an entropy-maximizing transformation exists mathematically – known as the Rosenblatt transformation (Rosenblatt, 1952) – its actual identification can be quite challenging; the budget for this task can be audited in terms of all the time series modeling effort being exerted worldwide (Politis, 2013). Complicating the picture is the empirical reality that many contending entropic transformations exist for any given time series, and although each residual series may be quite similar in its entropy characterization, the actual transformations utilized can be radically different. For instance, one entropic transformation may involve a logarithmic transformation together with a SARIMA model identification and fitting, whereas another such transform for the same series might involve no functional transform, instead using a large number of additive outlier dummy regressors together with a non-seasonal ARIMA model specification.

Given the wide variety of entropic transformations available to a practitioner, it is vital to have a disciplined method of selecting the best – although another possibility is to employ them all, through a device of model averaging (Hoeting et al., 1999). In particular, given two models deemed adequate from the standpoint of having a maximum entropy residual, which is to be preferred upon purely empirical grounds? (That is, apart from aesthetic criteria, which might lead a scientist to prefer marginal transformations to using many dummies.) A classical approach to this question might involve a likelihood ratio test statistic, or more broadly a comparison of information criteria. There are limitations to the likelihood approach, as discussed in McElroy (2016b): the models being compared only describe the stationary aspects of the data, parsed through the serial correlation pattern. The Akaike Information Criterion (AIC) is broader, allowing multiple model comparisons, but is still limited to time series models with a common unit root specification (Findley et al., 1998). If comparing entropic transformations that involve different choices of unit root specification, covariates, and/or marginal transformation, the available methods are either outright impossible to utilize or have unknown statistical properties.

This paper adopts a new approach to the broad model comparison problem, recognizing that likelihood-based comparison statistics are too narrow for real applications. In fact, our study was motivated by a modeling challenge posited by the authors of a competing time series software: they had devised a model for the time series of Motor Vehicles and Parts Dealers involving 26 additive outliers and 2 level shifts, whereas X-12-ARIMA identified a logarithmic transformation with a single additive outlier and level shift. The subsequent SARIMA models identified were (2,1,4)(0,0,2) for the competitor, and (1,1,1)(2,0,0) for X-12-ARIMA. Both residual sequences seemed to be fairly adequate from the standpoint of entropy – so which entropic transformation is preferred? Spurred by this challenge, we have developed a framework for assessing and comparing the entropy of residual series arising from quite different modeling paradigms. (We return to the Motor story in Section 5.)

In particular, our work provides a new information criterion based upon logged p-values of two statistics that each measure serial correlation and departures from Gaussianity. Because the distribution for these test statistics is derived under a null hypothesis of a maximum entropy residual – and not a null hypothesis of correct model specification – parameter uncertainty and corrections for degrees of freedom are dispensed with (c.f., Ljung and Box, 1978). As an aside, the treatment of Box-Pierce statistics under a correct model null – where the asymptotic distribution then depends upon the number of estimated parameters – cannot be rigorously extended to the case of dummy regressors or marginal transformations (McElroy and Monsell, 2014). Rather than attempting to extend distributional results to these more complicated entropic transformations, we instead utilize the device of a maximum entropy residual null hypothesis. However, it is important to protect against overfitting, which can arise from using too many dummy regressors or too many parameters in the stationary time series model specification (e.g., an AR(p) model with p too high). A key empirical finding of our work is that overfitting results in light-tailed residuals, and therefore can be detected through abnormally low marginal kurtosis.

The two sufficient facets of maximum entropy for a time series are serial correlation and Gaussianity, so we propose a Box-Pierce statistic for the former and a sum of squared Hermite coefficients for the latter. Other measures have been proposed in Granger and Lin (1994), and Hong and White (2005). For the marginal structure, recall that any continuous random variable can be expressed as a function of a Gaussian variable through the probability integral transform, and the Hermite coefficients of this transform correspond to zero in the case of an identity transformation, i.e., when the variable is Gaussian. The framework and estimation methodology of Janicki and McElroy (2016) is adopted to testing the residual marginal for Gaussianity. Each test, for serial correlation and Gaussianity, is formulated as a Wald statistic taking into account the maximum entropy null hypothesis, and the asymptotic distributions derived in Section 3 are used to obtain p-values and ultimately our proposed entropy information criterion.

The end result is a technique of model comparison that assesses any residual series against the ideal maximum entropy residual, and ranks competitors through the proposed entropy information criterion according to statistically significant deviance from optimality. Section 2 provides the underlying framework for maximal entropy comparisons, including a discussion of Hermite coefficients and serial correlation, while Section 3 provides definitions for both Wald test statistics, along with asymptotic theory and computational facets. Section 4 validates the methodology through simulations of test size and power, while Section 5 provides empirical analyses, returning to the Motor example. The appendix contains details for calculation of our autocovariance matrices in a stable and efficient way.

2 Framework for Maximum Entropy Comparisons

A time series $\{\epsilon_t\}$ has maximal entropy if there is no serial correlation and the marginal is Gaussian. If we entertain a slightly larger space of time series with lower entropy, we can construct measures upon this class to assess discrepancies from the ideal of Gaussian white noise. As we are interested in applying our procedures to residual series from some postulated entropic transformation, it is reasonable to expect that our class need be no larger than the space of stationary time series with absolutely continuous marginal distribution. (A residual sequence exhibiting non-stationarity, such as heteroscedasticity or unit roots, or exhibiting point masses in its empirical distribution, would indicate immediate prior rejection of the entropic transformation.) Conceptually, we imagine a parameter S (i.e., a function of the cumulant functions of $\{\epsilon_t\}$) taking values in the positive quadrant of the plane, where the first component assesses marginal structure and the second component assesses serial correlation. If two residual time series $\{\epsilon_t^{(1)}\}$ and $\{\epsilon_t^{(2)}\}$ are in play, we obtain two parameter vectors S_1 and S_2; see Figure 1.

The origin corresponds to Gaussian white noise, the case of maximal entropy in this class of processes. Increases to the first component of parameter S correspond to departures from a Gaussian marginal structure, while maintaining the same level of serial correlation. Conversely, increasing the second component corresponds to additional serial correlation, while maintaining the same marginal distribution. We might compare S_1 and S_2 through their magnitudes alone, if we are indifferent to the angular portion in their polar decompositions.

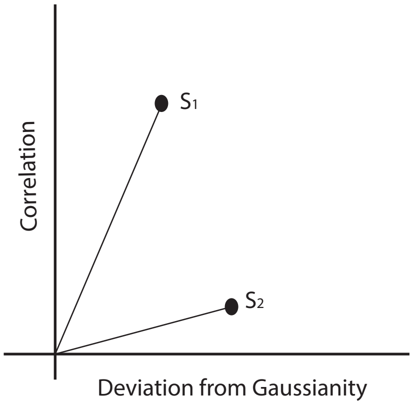


Figure 1: Example plot showing two hypothetical scores from two separate sets of residual values.

We can describe the entire class of residual processes $\{\epsilon_t\}$ as satisfying

$$\epsilon_t = g(Z_t), \tag{1}$$

where $\{Z_t\}$ is a mean zero stationary Gaussian process with unit variance, and g is a monotonic function satisfying $\int g^2(x)\phi(x)dx < \infty$. Here, ϕ and Φ represents the density function and cumulative distribution function of the standard normal distribution respectively. Any continuous marginal distribution F for ϵ_t can be entertained, taking $g = F^{-1} \circ \Phi$. A wide variety of serial correlation structures for $\{\epsilon_t\}$ can be entertained, as the autocovariance function for the residual can be described in terms of the autocovariance function of $\{Z_t\}$, together with the Hermite coefficients (as described in (6) below).

The normalized Hermite polynomials (Samarodnitsky and Taqqu, 1994) are defined via

$$H_k(x) = (-1)^k e^{x^2/2} \frac{d^k}{dx^k} e^{-x^2/2} \tag{2}$$

for $k \geq 0$. The set of functions given by (2) are a complete orthonormal basis for $L^2(d\Phi)$ with respect to the inner product given by

$$\langle f, h \rangle = \int_{-\infty}^{\infty} f(x)h(x)\phi(x)dx. \tag{3}$$

Since $g \in L^2$ we can write

$$g(x) = \sum_{k=0}^{\infty} \eta_k \, H_k(x). \tag{4}$$

The Hermite coefficients in (4) are given by

$$\eta_k = \langle g, H_k \rangle. \tag{5}$$

As $H_0(x) = 1$ and $H_1(x) = x$, the marginal distribution of ϵ_t is Gaussian if and only if g is an affine mapping, which occurs if and only if $\eta_k = 0$ for $k \geq 2$. Therefore, we might define the first component of S to be given by $\sum_{k\geq 2} \eta_k^2$.

On the other hand, the serial correlation structure is governed by the autocovariance function $\gamma_h(Z)$ of the process $\{Z_t\}$. It follows from a result of Hannan (1970, p.83) that

$$\gamma_h(\epsilon) = \sum_{k\geq 1} k!\, \eta_k^2\, \gamma_h^k(Z), \tag{6}$$

which shows how serial correlation is transmitted through g via the Hermite coefficients. In particular, if $\{Z_t\}$ is i.i.d. then $\{\epsilon_t\}$ is serially uncorrelated, no matter whether g is the identity or another transformation. However, as we cannot directly measure the serial correlation of the posited $\{Z_t\}$ series, it is more straight-forward to measure the serial correlation of $\{\epsilon_t\}$ through its own autocovariance function. Therefore, we might define the second component of S to be given by $\sum_{k\geq 1} \gamma_k^2(\epsilon)$. Henceforth we just write γ_k for $\gamma_k(\epsilon)$, as we shall not be concerned with the autocovariances of $\{Z_t\}$.

In terms of a hypothesis testing framework, we wish to adopt a null hypothesis that $\{\epsilon_t\}$ is Gaussian white noise, which is the intersection of two hypotheses:

$$H_G : \{\epsilon_t\} \text{ marginally Gaussian} \tag{7}$$

$$H_C : \{\epsilon_t\} \text{ uncorrelated} \tag{8}$$

Thus, our null hypothesis is $H_C \cap H_G$ and the alternative hypothesis is the space of strictly stationary short memory time series with continuous marginal distributed, denoted H_Ω. We can parametrize large portions of this space through the measures $\sum_{k\geq 2} \eta_k^2$ and $\sum_{k\geq 1} \gamma_k^2$. This concept has been explored in independent component analysis where mutual information (the Kullback-Leibler divergence from a random vector's joint distribution to the product of its marginal distributions) is decomposed, under linear transforms, as the sum of two terms; one term expressing the decorrelation of components of a random vector and one term expressing the deviation from Gaussianity (Cardoso 2003). In practice, we only consider a finite set of the Hermite coefficients and autocovariances, i.e., we choose J and K such that $\underline{\eta} = [\eta_2, \eta_3, \ldots, \eta_J]'$ and $\underline{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_K]'$ are the focus. The true Hermite and acf vectors will be denoted $\widetilde{\underline{\eta}}$ and $\widetilde{\underline{\gamma}}$. Under H_G and H_C respectively, these are zero vectors. This suggests a Gaussianity test statistic and a correlation test statistic each based upon estimators $\widehat{\underline{\eta}}$ and $\widehat{\underline{\gamma}}$ and standardized according to the appropriate null hypotheses. Then a significant value of the Gaussian test statistic indicates rejection of H_G only, whereas a significant value of the correlation test statistic indicates rejection of H_C only.

Of course, alternative measures of correlation and Gaussianity could be utilized; there are frequency domain tests of whiteness (Drouiche (2007), McElroy and Holan (2009)) as well as QQ-norm plots, Shapiro-Wilk tests, or Kolmogorov-Smirnov tests for Gaussianity (Shapiro and Wilk (1965), Kolmogorov (1933), Smirnov (1948), Massey Jr (1951)). Our methods have a fundamentally different approach from these. We advocate testing Gaussianity and whiteness with a model agnostic approach where the hypothesis reflects understanding that model fitting was done. This means the input data (residuals) may not be perfectly Gaussian or perfectly white noise. Clarifying, in our tests for Gaussianity, the null hypothesis understands the input residuals may have a slight amount of correlation left. This is seen in that variance-covariance matrix V (in equation (15) below) is not diagonal. Our methods do not rely on tuning parameters or any knowledge of the models used to produce the residual inputs. This allows models from vastly differing classes to be compared with an equivalent magnitude relation. The results of our tests also have tractable asymptotic theory and intuitive implications.

3 Testing

We propose estimating $\underline{\gamma}$ with the sample autocovariances (SACV) of the residual sample $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$, and estimating $\underline{\eta}$ with the nonparametric quantile estimators of Janicki and McElroy (2016). The SACV for observed residuals at lags $k = 0, 1, \ldots, (T-1)$ is

$$\hat{\gamma_k} = \frac{1}{T} \sum_{t=1}^{T-k} (\epsilon_{t+k} - \bar{\epsilon})(\epsilon_t - \bar{\epsilon}) \tag{9}$$

where $\bar{\epsilon} = \sum_{t=1}^{T} \epsilon_t / T$ is the sample mean. Our estimator of $\underline{\eta}$ is based on the inner product (5). Namely, $\eta_k = \int_0^1 Q(u) H_k(\Xi(u))\, du$ and we utilize a nonparametric estimate of Q, the order statistics of $\{\epsilon_t\}$, denoted $\{\epsilon_{(t)}\}$. Let $\Xi = \Phi^{-1}$, the Gaussian quantile function. Also, F and Q respectively denote the cumulative distribution function and quantile function of ϵ_t.

$$\hat{\eta_k} = \sum_{t=1}^{T} \epsilon_{(t)} \int_{(t-1)/T}^{t/T} H_k(\Xi(u))\ du. \tag{10}$$

The estimator (10) can be viewed as a weighted sum of the order statistics $\{\epsilon_{(t)}\}$. Moreover, these weights only depend on T and hence alleviate any computational complexity of doing numerical integration during every estimation. These weights only need to be explicitly calculated once for a single data analysis problem. We next review the distribution theory for the corresponding test statistics obtained for the set of processes belonging to H_Ω.

Proposition 3.1 *If $\{\epsilon_t\}$ is a stationary two-sided linear process, i.e. there exists absolutely summable coefficients $\{\psi_j\}_{j\in\mathcal{Z}}$ and a white noise sequence $\{w_t\}_{t\in\mathcal{Z}}$ with mean zero, variance σ_w^2, and $E[w_t^4] = \tau\sigma_w^4 < \infty$ such that $\epsilon_t = \sum_{j=-\infty}^{\infty}\psi_j w_{t-j}$, then for non-negative integer K*

$$\sqrt{T}\,\left(\underline{\widehat{\gamma}} - \underline{\widetilde{\gamma}}\right) \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0, W)$$

as $n \to \infty$. The (j,k)th entry of the asymptotic covariance matrix W for $j,k = 1,\ldots,K$ is

$$[W]_{jk} = (\tau - 3)\gamma(j)\gamma(k) + \sum_{\ell=-\infty}^{\infty}\left[\gamma(\ell)\gamma(\ell - j + k) + \gamma(\ell + k)\gamma(\ell - j)\right]$$

Because $\underline{\widetilde{\gamma}}$ is the zero vector under H_C, the test statistic

$$T\,\underline{\widehat{\gamma}}'\,W^{-1}\,\underline{\widehat{\gamma}} \tag{11}$$

is proposed, where W is computed under the null H_C given in (8). It follows from Proposition 3.1 that this test statistic (11) has an asymptotic χ^2 distribution on K degrees of freedom under H_C. However, if H_C is false – so that the true $\underline{\widetilde{\gamma}}$ is nonzero – then instead $(\underline{\widehat{\gamma}} - \underline{\widetilde{\gamma}})'\,W^{-1}\,(\underline{\widehat{\gamma}} - \underline{\widetilde{\gamma}}) = O_P(T^{-1})$, and the asymptotic power of the test statistic is driven by T times the term

$$\underline{\widetilde{\gamma}}'\,W^{-1}\,\underline{\widetilde{\gamma}}, \tag{12}$$

known as the efficacy. Next, we consider the Hermite estimation. When H_G is true, $F = \Phi$ and $Q = \Xi$. We will use the following assumptions at times in what follows.

Assumption 1 (A1) *F is twice differentiable on (a,b) where $-\infty \le a \le b \le \infty$, $a = sup\{x; F(x) = 0\}$ and $b = inf\{x; F(x) = 1\}$.*

Assumption 2 (A2) *$F' = f(x) > 0$ on (a,b).*

Assumption 3 (A3) *For some $\gamma > 0$,*

$$\underset{a<x<b}{sup}\, F(x)\,(1 - F(x))\,\frac{|f'(x)|}{f^2(x)} = \underset{0<t<1}{sup}\, t(1-t)\frac{|f'(Q(t))|}{f^2(Q(t))} \le \gamma$$

Assumption 4 (A4) *If $A = \lim_{x\searrow a} f(x) < \infty$ and $B = \lim_{x\nearrow b} f(x) < \infty$ then either $min(A,B) > 0$ or if $A = 0$ $(B = 0)$ then f is nondecreasing (nonincreasing) on an interval to the right of a (to the left of b).*

Assumption 5 (A5) *$\{F(\epsilon_t), t \ge 1\}$ is a stationary ρ-mixing sequence of uniform $[0,1]$ random variables with $\sum_{t=1}^{\infty}\rho(2^t) < \infty$.*

Proposition 3.2 *Let $\{\epsilon_t\}$ be a stationary sequence of random variables with common distribution function F that is invertible, with $F^{-1} = Q$. Assume A1 - A5. Then we have*

$$\sqrt{T}\,\left(\underline{\widehat{\eta}} - \underline{\widetilde{\eta}}\right) \overset{\mathcal{L}}{\Longrightarrow} \mathcal{N}(0, V) \tag{13}$$

as $n \to \infty$, where the (j,k)th element of V is

$$V_{jk} = \int_0^1 \int_0^1 q(u)q(v)\, H_j(\Xi(u))\, H_k(\Xi(v))\, R(u,v)\, du\, dv. \tag{14}$$

Here $R(u,v) = \sum_{k=-\infty}^{\infty} Cov(\mathbf{1}_{[F(\epsilon_0)\leq u]}, \mathbf{1}_{[F(\epsilon_k)\leq v]})$ and $q = \dot{Q}$ is the derivative of the quantile function of the residual process.

It should be noted that (13) and (14) are expressed for all values $\eta_0, \eta_1, \ldots, \eta_J$. In our testing we will simply omit η_0 and η_1 as discussed at the end of Section 2.

Remark 3.1 The calculation of V requires a computation of the summands defining $R(u,v)$, which is explained in the Appendix.

Again, because $\underline{\widetilde{\eta}}$ is the zero vector under H_G, the test statistic

$$T\,\underline{\widehat{\eta}}'\, V^{-1}\, \underline{\widehat{\eta}} \tag{15}$$

is proposed, where V is computed as described in Remark 3.1 above. It follows from Proposition 3.2 that this test statistic (15) has an asymptotic χ^2 distribution on $J-1$ degrees of freedom under H_G. However, if H_G is false then instead $[\underline{\widehat{\eta}} - \underline{\widetilde{\eta}}]'\, V^{-1}\, [\underline{\widehat{\eta}} - \underline{\widetilde{\eta}}] = O_P(T^{-1})$, and the asymptotic power of the test statistic is driven by T times the term

$$\underline{\widetilde{\eta}}'\, V^{-1}\, \underline{\widetilde{\eta}}, \tag{16}$$

known as the efficacy. Both efficacies (12) and (16) offer a suitable notion of distance, analogous to the Mahalanobis distance (Mitchell and Krzanowski, 1985). That is, we may think of the first and second components of S in Figure 1 as being given by the efficacies (16) and (12).

Performing the tests (11) and (15) require a prespecified number of coefficients to be used. For Gaussian testing, $\eta_2, \eta_3, \ldots, \eta_J$ are used since standardized residuals have approximately $\eta_0 = 0$ and $\eta_1 = 1$. Including these values in (11) provides no added value. Moreover, inclusion of these terms may even lead to a lower power test due to an inflated degrees of freedom. The correlation test uses $\gamma_1, \gamma_2, \ldots, \gamma_K$. When looking at diagnostics for a given set of residuals, multiple values of J and K can be examined; this is similar to looking at many values of the maximal lag in the Ljung-Box test for autocorrelation. It may not be immediately clear why many lags of η and γ are used when expanding g.

It is clear that the test statistics (11) and (15) have power to detect mis-specified models that result from *under-fitting*, in the sense that the entropic transformation is unable to map the data vector to a maximum entropy residual. However, it is vital to examine both correlation and Gaussianity, because it is not uncommon that under-specified models yield serially correlated Gaussian marginals, such that only the correlation test (15) is significant; this can be seen in the data analysis of section 5.2. The inclusion of additional covariates, or more nuanced modeling of the serial correlation, may improve the entropic transformation, thereby resolving the under-fit. On the other hand, *overfitting* can arise from including extraneous components of the entropic transformation, such as spurious covariates or insignificant ARMA coefficients. These model inefficiencies present a more subtle problem, essentially driving down the variability in sample autocovariances and yielding marginal distributions that are overly light-tailed. Therefore, the Gaussianity test (11) is important to detect overfitting.

4 Laboratory Studies

4.1 Size and Power of Test Statistics

To determine the power of tests derived in Section 3, we simulated series of varying lengths from a residual process with known marginal and correlation structure. By allowing these attributes to deviate away from our hypothesized Gaussian white noise assumption we can determine the statistical power of our methods. Clearly this can be done by choosing nonzero vectors $\underline{\widetilde{\eta}}$ and $\underline{\widetilde{\gamma}}$ that generate large efficacies, as the efficacy drives the asymptotic power as discussed in Section 3. But in order to parse our power results in terms of known distributions, we consider a Student's t marginal with ν degrees of freedom and an M-dependent autocorrelation structure.

This power study requires the ability to control the autocorrelation and the marginal distribution of a set of hypothetical time series residuals. This task is accomplished by allowing the marginal distribution of simulated residuals to be t_ν. Thus, as ν increases, the marginal structure moves toward the null of H_G. This task is easily accomplished with our methods by simply setting Q in (14) as the quantile function of the t_ν distribution. This allows calculation of true $\underline{\eta}$ values for a given residual process. For the study that follows, values of ν used were 2, 5, 10, 20, 40 and ∞ where $\nu = \infty$ represents H_G.

Similarly, to allow for a single parameter to control the amount of autocorrelation, each simulated set of residuals was M-dependent with exponentially decaying autocovariance as a function of θ, i.e.

$$\mathrm{Corr}(Z_t, Z_{t+h}) = \theta^h \tag{17}$$

for $h = 0, 1, \ldots, M$ and the correlation is zero otherwise. For the simulation that follows M was chosen to be five. As θ decreases we move toward H_C. Values of θ used in the simulation were 0, 0.05, 0.1, 0.2, 0.25, 0.3 and 0.4 (which all correspond to a positive definite ACF function). Tables 1 through 6 provide results of this power study. Each table is for a fixed level of ν and all values of θ. The values provided were for 1000 simulations of each series for the given values ν and θ. The length of each series, n, was also changed; values of n were doubled from 50 through 800. The values displayed in the table are the proportion of rejections out of the 1000 simulated series. Ideally we can see power increase in all dimensions. That is, the power of our testing framework increases as n increases, θ increases (we move farther from uncorrelated assumption) or ν decreases (we move farther from the Gaussian assumption).

Table 1: Power study results for Gaussianity (Gaus) tests and Correlation (Corr) test for fixed t-distribution level $\nu = 2$ and varying levels of correlation θ

	$\nu = 2$		$\nu = 2$		$\nu = 2$		$\nu = 2$		$\nu = 2$		$\nu = 2$		$\nu = 2$	
	$\theta = 0$		$\theta = .05$		$\theta = .1$		$\theta = .2$		$\theta = .25$		$\theta = .3$		$\theta = .4$	
n	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr
50	0.973	0.027	0.968	0.021	0.967	0.033	0.968	0.046	0.964	0.079	0.972	0.110	0.938	0.229
100	1.000	0.029	0.999	0.045	1.000	0.048	1.000	0.126	0.997	0.179	1.000	0.264	0.998	0.496
200	1.000	0.047	1.000	0.061	1.000	0.079	1.000	0.229	1.000	0.381	1.000	0.525	1.000	0.818
400	1.000	0.059	1.000	0.057	1.000	0.140	1.000	0.429	1.000	0.627	1.000	0.779	1.000	0.934
800	1.000	0.054	1.000	0.087	1.000	0.194	1.000	0.752	1.000	0.873	1.000	0.959	1.000	0.993

Table 2: Power study results for Gaussianity (Gaus) tests and Correlation (Corr) test for fixed t-distribution level $\nu = 5$ and varying levels of correlation θ

	$\nu = 5$		$\nu = 5$		$\nu = 5$		$\nu = 5$		$\nu = 5$		$\nu = 5$		$\nu = 5$	
	$\theta = 0$		$\theta = .05$		$\theta = .1$		$\theta = .2$		$\theta = .25$		$\theta = .3$		$\theta = .4$	
n	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr
50	0.557	0.039	0.566	0.030	0.546	0.049	0.570	0.069	0.561	0.156	0.567	0.215	0.571	0.406
100	0.819	0.047	0.806	0.042	0.841	0.065	0.811	0.212	0.824	0.338	0.816	0.515	0.792	0.817
200	0.970	0.048	0.975	0.055	0.970	0.121	0.971	0.502	0.976	0.717	0.961	0.907	0.963	0.994
400	0.999	0.034	1.000	0.071	1.000	0.256	1.000	0.854	1.000	0.974	0.999	0.999	0.996	1.000
800	1.000	0.051	1.000	0.135	1.000	0.518	1.000	0.991	1.000	1.000	1.000	1.000	1.000	1.000

Table 3: Power study results for Gaussianity (Gaus) tests and Correlation (Corr) test for fixed t-distribution level $\nu = 10$ and varying levels of correlation θ

	$\nu = 10$		$\nu = 10$		$\nu = 10$		$\nu = 10$		$\nu = 10$		$\nu = 10$		$\nu = 10$	
	$\theta = 0$		$\theta = .05$		$\theta = .1$		$\theta = .2$		$\theta = .25$		$\theta = .3$		$\theta = .4$	
n	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr
50	0.201	0.036	0.173	0.046	0.200	0.055	0.219	0.116	0.184	0.154	0.222	0.237	0.248	0.438
100	0.340	0.028	0.331	0.037	0.333	0.093	0.353	0.225	0.377	0.373	0.379	0.560	0.407	0.883
200	0.548	0.054	0.574	0.063	0.584	0.147	0.551	0.486	0.592	0.751	0.585	0.914	0.593	0.998
400	0.812	0.054	0.807	0.100	0.831	0.250	0.810	0.874	0.836	0.982	0.832	0.998	0.839	1.000
800	0.985	0.045	0.985	0.159	0.984	0.536	0.981	0.997	0.984	1.000	0.981	1.000	0.985	1.000

Table 4: Power study results for Gaussianity (Gaus) tests and Correlation (Corr) test for fixed t-distribution level $\nu = 20$ and varying levels of correlation θ

	$\nu = 20$		$\nu = 20$		$\nu = 20$		$\nu = 20$		$\nu = 20$		$\nu = 20$		$\nu = 20$	
	$\theta = 0$		$\theta = .05$		$\theta = .1$		$\theta = .2$		$\theta = .25$		$\theta = .3$		$\theta = .4$	
n	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr
50	0.068	0.039	0.068	0.040	0.074	0.050	0.119	0.114	0.085	0.152	0.116	0.240	0.139	0.451
100	0.110	0.039	0.112	0.042	0.125	0.067	0.143	0.240	0.159	0.388	0.179	0.571	0.206	0.860
200	0.193	0.048	0.185	0.057	0.199	0.135	0.193	0.528	0.238	0.750	0.250	0.910	0.304	0.999
400	0.282	0.041	0.293	0.102	0.354	0.266	0.346	0.868	0.369	0.986	0.369	0.998	0.406	1.000
800	0.496	0.050	0.519	0.120	0.544	0.535	0.546	0.996	0.591	1.000	0.565	1.000	0.600	1.000

Table 5: Power study results for Gaussianity (Gaus) tests and Correlation (Corr) test for fixed t-distribution level $\nu = 40$ and varying levels of correlation θ

	$\nu = 40$		$\nu = 40$		$\nu = 40$		$\nu = 40$		$\nu = 40$		$\nu = 40$		$\nu = 40$	
	$\theta = 0$		$\theta = .05$		$\theta = .1$		$\theta = .2$		$\theta = .25$		$\theta = .3$		$\theta = .4$	
n	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr
50	0.030	0.039	0.037	0.037	0.055	0.040	0.060	0.102	0.088	0.177	0.064	0.252	0.108	0.465
100	0.055	0.040	0.061	0.037	0.066	0.073	0.082	0.232	0.097	0.390	0.105	0.566	0.169	0.864
200	0.089	0.044	0.111	0.064	0.107	0.127	0.128	0.527	0.148	0.782	0.160	0.921	0.197	0.996
400	0.135	0.046	0.160	0.082	0.149	0.288	0.186	0.882	0.197	0.980	0.227	1.000	0.255	1.000
800	0.202	0.057	0.196	0.141	0.222	0.557	0.214	0.997	0.250	1.000	0.247	1.000	0.310	1.000

Table 6: Power study results for Gaussianity (Gaus) tests and Correlation (Corr) test for fixed t-distribution level $\nu = \infty$ and varying levels of correlation θ

	$\nu = \infty$		$\nu = \infty$		$\nu = \infty$		$\nu = \infty$		$\nu = \infty$		$\nu = \infty$		$\nu = \infty$	
	$\theta = 0$		$\theta = .05$		$\theta = .1$		$\theta = .2$		$\theta = .25$		$\theta = .3$		$\theta = .4$	
n	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr	Gaus	Corr
50	0.026	0.042	0.031	0.042	0.032	0.042	0.031	0.108	0.047	0.157	0.077	0.236	0.091	0.471
100	0.052	0.050	0.034	0.057	0.056	0.078	0.062	0.268	0.066	0.395	0.090	0.587	0.133	0.863
200	0.076	0.041	0.070	0.063	0.080	0.141	0.098	0.518	0.105	0.781	0.126	0.907	0.185	0.994
400	0.116	0.054	0.110	0.088	0.112	0.275	0.124	0.869	0.156	0.973	0.146	1.000	0.215	1.000
800	0.139	0.057	0.171	0.149	0.149	0.551	0.201	0.999	0.184	1.000	0.218	1.000	0.256	1.000

4.2 Assessing Overfitting Through Kurtosis

Assessment of overfitting is more nuanced from a model agnostic approach such as ours. This is a direct result of most current overfitting assessments being done through penalizing model parameters. Adding extraneous terms that do not improve the model fit will hurt the likelihood value. The problem is this methodology does not address the fundamental problem: it is difficult to compare model fits from differing prior adjustments. If data is log-transformed, modified for calendar effects (such as trading day or moving holidays) or any other adjustments prior to a model being fit, the resulting typical diagnostics will be non-nested and on differing scales. Our methods can do these comparisons directly. Moreover, our methods detect light tailed or extreme residuals, both of which can result from overfitting. A common statistical measure that compares the tail of a distribution to that of the Gaussian distribution is the kurtosis (Ruppert, 1987). We do not have a direct one-to-one relationship between our Hermite coefficients and kurtosis, $\kappa = E[\epsilon^4]/E[\epsilon^2]^2$. This is seen by considering a slight pertubation from standard normal in the third Hermite coefficient. Let

$$\boldsymbol{\eta} = (0\ 1\ 0\ \Delta\ 0\ 0 \cdots)'.$$

Then

$$\epsilon_t = (1 - 3\Delta) Z_t + \Delta Z_t^3. \tag{18}$$

Hence any fourth moment expression for (18) will involve high order, even moments of the standard normal. For a standard normal random variable Z, the kth even moment is $E[Z^k] = 2^{-k/2} \frac{k!}{(k/2)!}$.

We break down the types of overfitting into two situations. The first is, when a model follows the finite sample data too closely, results in model residuals with very light tails, or $\kappa < 3$. This issue is a common pitfall whereby the model not only fits the underlying dependent variable, but also fits the noise unique to each observed sample (Lever et al., 2016). Hence, the resulting residuals become light tailed. We simulate residuals from this process by truncating the tails of a standard normal distribution at a cutoff value. The results are shown in Figure 2. We see our test for Gaussianity has high power for lightly tailed residuals and the power converges to the type I error rate as the tails of the Gaussian return to normal. One example of the estimates $\underline{\hat{\eta}}$ for this example is given in Table 7. Following the logic above, we see the largest coefficient in magnitude is indeed $\hat{\eta}_3$, but it is not the only non-zero coefficient. This simulated residual series produced a p-value of 0.0092, and failed the residual test at a Type I error rate of 0.05. Moreover, it has $\kappa = 2.614516$ and the coefficient estimates reiterate the discussion above that no one single η_k coefficient dominates detection of kurtosis; together they form a powerful test again the null of normality.

The second overfitting situation is when a model follows the mean too closely and hence produces heavy tailed residuals such that $\kappa > 3$. Of course we are assuming here the raw transformed series

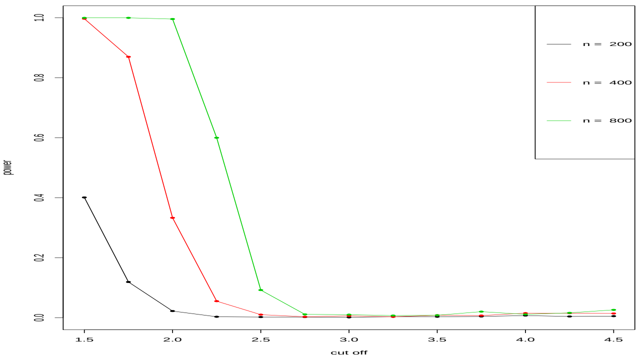


Figure 2: Power of Gaussian testing procedure for light tailed distributions.

Table 7: Parameter estimates for simulation with cutoff value of 2.5 and sample size $n = 800$.

$\hat{\eta}_2 = .00578$	$\hat{\eta}_3 = -.05095$	$\hat{\eta}_4 = -.00060$	$\hat{\eta}_5 = -.01208$
$\hat{\eta}_6 = -.00352$	$\hat{\eta}_7 = .00595$	$\hat{\eta}_8 = .00498$	

does not itself look like Gaussian white noise. Without this assumption the overfitting paradigm discussed here and a suitable model overlap. In this situation some residuals, especially from economic models or any data which reacts to possible latent shock and volatility effects, will be extreme. Any observed data value that is out of the ordinary in magnitude will be ignored by the model and result in heavy tailed residuals. Our testing proceedue performs well in this overfitting scenario. This can be seen in the power study of section 4.1. Each of the Tables 1 through 6 represents a different level of t-distribution degrees of freedom. Smaller values of ν correspond to a residual process with heavier tails than the assumed standard normal distribution.

5 Data analysis

Our methods are oriented toward users who want to automatically compare residuals coming from competing, well fit, models. In this line of thinking, all of our residual analysis is done from an automated fitting point-of-view. In Section 5.1 we compare data transformations prior to model fitting and in Section 5.2 we show our methods will also alert users to overfitted models.

5.1 Motor Vehicles and Parts Dealers

This series is of interest due to two suggested prior adjustments from automatic model fitting algorithms. Models with dramatically different prior adjustments can be difficult to compare e.g) pre-transformed model residuals no longer come from nested models. Autobox and X-13ARIMA-SEATS are two software programs that can do automatic model selection (Automatic Forecasting Systems, Inc. 1986 and US Census Bureau 2015). Hence, both softwares make automatic choices about data transformations. For example, log or no log transformation in X-13ARIMA-SEATS is decided by fitting an airline model to each and selecting the series transformation that produced a preferable likelihood.

We will look at monthly vehicle and parts dealers retail sales. The Motor Vehicle and Parts Dealers subsector is part of the retail trade sector and this series is from fixed point-of-sale locations. Establishments in this subsector typically operate from a showroom or an open lot where vehicles are on display. A plot of the raw retail series is shown in Figure 3. It is clear that the raw series has some values that change levels of magnitude and will need to be adjusted prior to model fitting. When the series is passed through both automated time series modeling routines vastly different prior transformations are suggested [1]. One program suggests taking a logarithmic transformation, including a single additive outlier and including one level-shift for the recession in the mid 2000's. The other program suggests no logarithmic transformation; instead the level of magnitude change is handled via 26 additive outliers and two level shifts. It should be emphasized that this is not a comparison of the two software products (X-13ARIMA-SEATS and Autobox) but a demonstration of the testing procedures in a real world situation. Plots of the prior adjustments are shown in Figure 4.

In order to avoid bias of fitting a time series model to data that was prior adusted with the same software, the two competing prior adjusted series will be fit with the automatic arima model identification routines in **R**'s `forecast` package. Each series will be passed to the `auto.arima()` function with the specified prior transformation already preformed and outliers and level shift supplied at those predetermined locations (Hyndman and Khandakar 2008). In what follows we will use a superscript *log* to indicate components of the log-adjusted series and a superscript *ao* to indicate any component from the preadjustment of adding 26 AO's. For example, x_t^{log} and ϵ_t^{log} will represent the original series after being log adjusted and the resulting residuals after log adjustment. The automated **R** routine found the best fitting model for x_t^{log} to be SARIMA$(1,1,1)(2,0,0)_{12}$ while the best fitting model for x_t^{ao} was a SARIMA$(2,1,4)(0,0,2)_{12}$. The coefficients in the fitted models are given in Table 8.

Table 8: Coefficients for models chosen by auto.model for the Motor series but two different prior adjustments. "Log model" refer to the series which has been log transformed. "AO model" refers to the series which has had 26 additive outliers included but no log transform

	ϕ_1	ϕ_2	θ_1	θ_2	θ_3	θ_4	Φ_1	Φ_2	Θ_1	Θ_2
Log model	0.1157		-0.5181				0.6662	0.1773		
AO model	0.1111	0.5749	-0.3897	-0.5624	0.3198	-0.2777			0.786	0.4796

The fitted residuals scaled by their sample standard deviation for both models are plotted in

[1] Since the interest here lies in analysis of the resulting residuals from model fitting we will not distinguish what transformation each program (AutoBox or X-13ARIMA-SEATS) suggested.

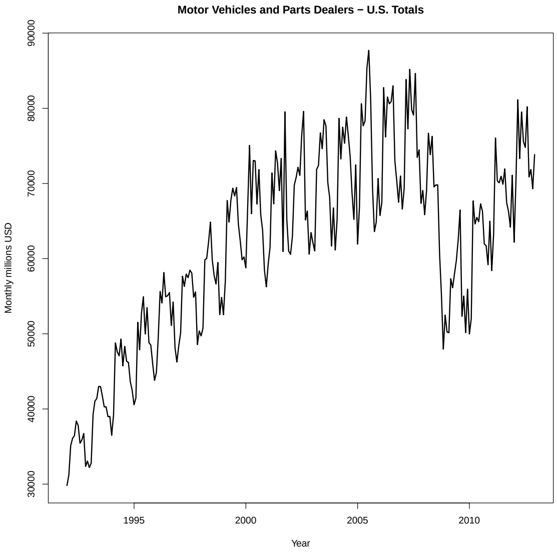


Figure 3: raw retail series

Figure 5. These residual values are then passed through our hypothesis testing framework.

The test for residual autocorrelation given by (15) gives a test statistic of 13.587 (p-value 0.093) for ϵ_t^{log} and a test statistic value of 10.281 (p-value 0.246) for ϵ_t^{ao}. Here, K was chosen to be eight. The test for residual normality given by (11) give a test statistic of 4.29 (p-value 0.7456) and 5.22 (p-value 0.6336) for ϵ_t^{log} and ϵ_t^{ao} respectively. Here $M = 8$. For further intuition the values of $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\gamma}}$ for each series are given in Table 9. Figure 6 provides a visual representation of the results of the hypothesis testing. Our methodology suggests the transform using a multitude of additive outlier regressors produces residuals closer to Gaussian white noise, albeit slightly. Expanding, Figure 6 shows ϵ^{log} deviates less from the Gaussian distribution but has more autocorrelation than ϵ^{ao}.

Table 9: Parameter estimates

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
γ_j^{log}	0.0055	-0.0629	0.1507	-0.1076	0.0508	-0.071	-0.0776	0.0448
γ_j^{ao}	-0.0065	0.0156	-0.0209	0.0197	0.0419	-0.1660	-0.0883	-0.0509
η_j^{log}		0.0167	-0.0007	0.0130	0.0004	0.0013	-0.0281	-0.0115
η_j^{ao}		-0.0011	-0.0055	-0.0335	-0.0313	0.0292	0.0083	-0.0187

5.2 Over and under fitting

In section 5.1 we saw a model with many additive outlier regressors produce residuals that more closely mimicked Gaussian white noise than its log transformed counterpart. A natural question is when will having too many regressors in your model be detrimental to the residuals? Our testing framework also has the added advantage of being able to detect over-fit models. Over-fitting is currently being used (and possible abused) in many modeling situations.

Consider the following two retail sales series, automobile dealers and clothing stores. Each series consists of monthly observations from January 1992 through June 2015. These series are plotted in Figure 7. In order to demonstrate the need for testing Gaussianity and autocorrelation concurrently we fit two separate models to each of the series. First the series are differenced as well as seasonally differenced. Moreover, an automated model fitting procedure was then used to make sure only moderate seasonal lags were required. We mimic underfitting with a low order autoregressive process. An AR(1) is fit and the resulting residuals are testing using the methods suggested by (11) and (15). We imitate overfitting with an excessively high order autoregressive process. An AR(50) is fit and the same testing procedure applied the the resultant residuals.

The underfit model is identified by our correlation test for both series. The test statistic (15) is 22.74 (8.9e-4) and 79.04 (5.6e-15) for the automobile dealers and clothing stores respectively with p-values given in parenthesis. Both tests indicate a strong presence of residual autocovariance presence. The test (11) gave values of 55.08 (0) and 13.14 (0.04). For the clothing store retail series even a simple AR(1) model produces residuals with poor evidence against a normally distributed null.

A more interesting result comes for the overfitting case. After overfitting each series the correlation tests given by (15) gave 0.1746 (0.99) for the automobile dealers series and .1333 (0.99) for the clothing stores. Failure to reject here indicates these models produce uncorrelated residuals. However, the normality test (11) picks up on the overfitting. The test statistics (p-values) for these series are 26.91 (0.0002) and 18.59 (0.0049).

This dichotomy of model fitting issues require a two faceted approach to detection of model misspecification. Many model residuals, depending on the type of misspecification, will still pass univariate diagnostic tools. Our methods are able to detect both under and overfit model residuals.

6 Concluding Remarks

It is more important than ever to have fundamentally sound tools to compare and diagnose models. Statistical models are being fit using ever more sophisticated techniques and a plethora of software and software libraries/packages. Having tools to compare, contrast and diagnose these model fits, which traverse a wide variety of model paradigms, is paramount. The ideas and tests advocated in this paper can be applied to any model comparison where that produces residuals.

We have provided a statistical test that has highly applicable interpretation and is model agnostic. We do not rely on penalty terms, a model specification or knowledge of the optimization routine used in fitting. All that is needed is the residuals from a model or set of residuals from competing models. The guiding principle that a better fitting model will produce residuals closer to maximum entropy will traverse all residual producing modeling paradigms. These models can now be compared; regardless of the pre-transformations, unit roots, fixed effects or serial correlation patterns.

Our exposition has focused on the case of ideal residuals following a Gaussian white noise assumption. This is by far the most widely used assumption on the residuals of competing time series models. However, there is room for future research to extend our methods to other marginal distributions. This can be accomplished by changing the function g in (4). For example, if uniformly distributed residuals are optimal, simply pick $g = \Phi$. For χ^2_2 marginals, choosing $g(Z) = -2\log(\Phi(Z))$ would suffice. Further discussion of choices of g and structures created can be seen in Janicki and McElroy (2016). Our work could also be extended to different correlation structures in a similar fashion.

Another extension of our work to be explored is allowing an expression of efficacy to depend on the angular portion of their polar decomposition as discussion in Section 2. If we accept the perfect entropic reduction transformation premise of Rosenblatt (1952), and further that finding the exact transformation is unlikely, the model fitting exercise become one of minimizing a distance between perfection and practicality. In this mind set, such a representation could allow tests which are biased toward an analysts preference for deviating away from maximum entropy in the direction of Gaussianity or correlation. This approach could increase forecasting performance and/or statistical inference depending on the ultimate goal of the analysis.

A Gaussian Quadrant Probability

We only need to compute V when H_G is true, so that $F = \Phi$. Then for $k \geq 0$

$$\mathrm{Cov}(\mathbf{1}_{[F(\epsilon_0)\leq u]}, \mathbf{1}_{[F(\epsilon_k)\leq v]}) = \mathbb{P}[\epsilon_0 \leq \Xi(u),\, \epsilon_k \leq \Xi(v)] - \mathbb{P}[\epsilon_0 \leq \Xi(u)]\,\mathbb{P}[\epsilon_k \leq \Xi(v)]$$

Observe the first term on the right hand side of the above equation, $\mathbb{P}[\epsilon_0 \leq \Xi(u),\, \epsilon_k \leq \Xi(v)]$, is a quadrant probability for a bivariate Gaussian random variable with mean $[-\Xi(u), -\Xi(v)]'$, variance γ_0 and lag k correlation $\rho_k = \gamma_k/\gamma_0$. This bivariate Gaussian quadrant probability can be expressed in terms of confluent hypergeometric functions. The confluent hypergeometric function of the first kind (Abramowitz and Stegun (1964, Chapter 13)) with parameters a and b is

$$F_1(x; a, b) = \sum_{n=0}^{\infty} \frac{(a)_n x^n}{(b)_n n!}. \tag{19}$$

Here $(a)_n$ notates the rising factorial function, i.e., for a positive integer value n, $(a)_0 = 1$ and $(a)_n = a(a+1)(a+2)\cdots(a+n-1)$. If $a = 0$ or $x = 0$ then $F_1(x; a, b) = 1$. The confluent hypergeometric function is a special case of the generalized hypergeometric functions (Muller, 2001). In Chandramouli and Ranganathan (1999) the bivariate quadrant probability is derived for the case of equal means, and Livsey, McElroy, and Lund (2017) generalizes this formula to the case of orthant probabilities for Gaussian variables with unequal means. Let $[Z_1, Z_2]'$ be a bivariate Gaussian random vector with mean $[\mu_1, \mu_2]'$, equal variances γ_0 and correlation ρ. Then

$$\mathbb{P}[Z_1 \leq 0,\, Z_2 \leq 0] = \frac{1}{4}\sum_{\ell=0}^{\infty}\left\{\frac{(2\rho)^\ell}{\ell!}\left[\frac{F_1(\frac{\ell}{2};\frac{1}{2};\frac{\mu_1^2}{2\gamma_0})}{\Gamma(1-\ell/2)} - \frac{\sqrt{2}\mu_1}{\sqrt{\gamma_0}}\frac{F_1(\frac{\ell+1}{2};\frac{3}{2};\frac{-\mu_1^2}{2\gamma_0})}{\Gamma((1-\ell)/2)}\right]\right. \\ \left. \cdot\left[\frac{F_1(\frac{\ell}{2};\frac{1}{2};\frac{\mu_2^2}{2\gamma_0})}{\Gamma(1-\ell/2)} - \frac{\sqrt{2}\mu_2}{\sqrt{\gamma_0}}\frac{F_1(\frac{\ell+1}{2};\frac{3}{2};\frac{-\mu_2^2}{2\gamma_0})}{\Gamma((1-\ell)/2)}\right]\right\}. \tag{20}$$

We denote $\mathbb{P}[Z_1 \leq 0,\, Z_2 \leq 0] = \phi_Q(\mu_1, \mu_2, \gamma_0, \rho)$. In (20), $\Gamma(\cdot)$ denotes the gamma function, i.e., $\Gamma(\lambda) = \int_0^\infty x^{\lambda-1}e^{-x}dx$. The gamma function has poles at negative integers, so care should be taken when evaluating this expression. If either μ_1 or μ_2 are zero then only odd values of ℓ will yield a nonzero term in the probability expression. It follows that

$$R(u,v) = \min(u,v) - uv + 2\sum_{k\geq 1}\left\{\phi_Q(-\Xi^{-1}(u), -\Xi^{-1}(v), \gamma_0, \rho_k) - uv\right\}.$$

Computation of (20) can be done in a computationally stable and efficient manner. The ultimate utility will come from the accuracy of the finite approximation to the infinite sum. Rapid convergence makes the approximation very accurate. Expanding, the confluent hypergeometric (19) is stable for moderate size $|\frac{xa}{b}|$. This requirement does not hinder our methods since the convergence of (20) depends on the decay rate of $(2\rho)^\ell$. For our problem, in the summation (20),

$$\left|\frac{xa}{b}\right| \propto \frac{\gamma_0}{2\mu}\ell.$$

If for large γ_0 a user runs into numerical instability, writing the power series expansion of the confluent hypergeometric function in its continued fraction counter part can extend the range of numerical accuracy. Details of this continued fraction representation are provided in Jones and Thron (1980). Hence, for moderate values of ρ we expect the number of terms needed for accurate evaluation of (20) to be small. Since in our application ρ will be the ACVF between residual lags, it will be small (a null of zero).

References

[1] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical function: with formulas, graphs and mathematical tables*, Vol. 55, Courier Corporation.

[2] Automatic Forecasting Systems, Inc. (1986). AUTOBOX: The User's Guide. Automatic Forecasting Systems, Inc. Hatboro, PA.

[3] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *J. R. Statist. Soc., Ser. B* **26**, 211–252.

[4] Box, G. E. and Jenkins, G. (1970). *Time series analysis: Forecasting and control.* San Francisco: Holden-Day.

[5] Box, G. E. and Pierce, D. A. (1970). "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models". *Journal of the American statistical Association*, 65(332), 1509-1526.

[6] Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80**, 580–597.

[7] Brockwell, P. J. and Davis, R. A. (2009). *Time series: theory and methods.* New York: Springer.

[8] Cardoso, J. F. (2003) "Dependence, correlation and gaussianity in independent component analysis." *The Journal of Machine Learning Research*, 4, 1177-1203.

[9] Chandramouli, R. and Ranganathan, N. (1999) "Computing the Bivariate Gaussian Probability Integral." *Signal Processing Letters, IEEE 6.6.* 129-131.

[10] Drouiche, K. (2007). A test for spectrum flatness. *Journal of Time Series Analysis* **28**(6), 793–806.

[11] 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 (with Discussion). *J. Bus. Econ. Statist.* **16**, 127–177.

[12] Gòmez, V. and A. Maravall (1996). Programs TRAMO and SEATS : Instructions for the user. Banco de España, Servicio de Estudios, DT 9628. (Updates and additional documentation can be found at http://www.bde.es/webbde/es/secciones/servicio/software/econom.html.).

[13] Gómez, V. and A. Maravall (2001). *Automatic modeling methods for univariate series.* In D. Pena, G. C. Tiao, and R. S. Tsay (Eds.), A Course in Time Series Analysis. New York, NY: J. Wiley and Sons.

[14] Granger, C. and Lin, J.L. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *Journal of Time Series Analysis* **15**(4), 371–384.

[15] Hannan, E.J. (1970) *Multiple Time Series.* New York: Wiley.

[16] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382-401.

[17] Hong, Y. and White, H. (2005). Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica* **73**, 837–901.

[18] Hyndman R., and Khandakar Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, **26**(3), 1-22. http://www.jstatsoft.org/article/view/v027i03.

[19] Janicki, R. and McElroy, T. (2016). Hermite expansion and estimation of monotonic transformations of Gaussian data. *Journal of Nonparametric Statistics* **28**(1), 207–234.

[20] Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review, Series II* **106**(4), 620–630.

[21] Jaynes, E. T. (1957b). Information theory and statistical mechanics II. *Physical Review, Series II* **108**(2), 171–190.

[22] Jones, W. B., and Thron, W. J. (1980). Continued fractions. *Encyclopedia of Mathematics and its Applications.* **11**, 37.

[23] Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione *G. Ist. Ital. Attuari* **4**, 83–91.

[24] Lever, J., Krzywinski, M., and Altman, N. (2016). "Points of significance: model selection and overfitting." 703.

[25] Livsey, J., McElroy, T., and Lund, R. (2017). Gaussian orthant probabilities. *Mimeo.*

[26] Ljung, G. M. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika* **65**(2), 297–303.

[27] Massey Jr, Frank J. (1951) The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* **46**, 68–78.

[28] McElroy, T. (2016a). On the measurement and treatment of extremes in time series. *Extremes* **19**(3), 467–490.

[29] McElroy, T. (2016b). Non-nested model comparisons for time series. *Biometrika* **103**, 905–914.

[30] McElroy, T. and Holan, S. (2009). Spectral domain diagnostics for testing model proximity and disparity in time series data. *Statistical Methodology* **6**, 1–20.

[31] McElroy, T. and Monsell, B. (2014) The multiple testing problem for Box-Pierce statistics. *Electronic Journal of Statistics* **8**, 497–522.

[32] Mitchell, A. F., and Krzanowski, W. J. (1985). The Mahalanobis distance and elliptic distributions. *Biometrika*, 464-467.

[33] Muller, K. E. (2001). Computing the confluent hypergeometric function, $M(a, b, x)$. *Numerische Mathematik* **90**, 179-196.

[34] Park, S.Y. and Bera, A.K. (2009). Maximmum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics* **150**, 219–230.

[35] Politis, D.N. (2013). Model-free model-fitting and predictive distributions. *Test* **22**(2), 183–221.

[36] Politis, D.N. (2015). *Model-free Prediction and Regression: a Transformation-based Approach to Inference.* New York: Springer.

[37] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.* **23**, 470–472.

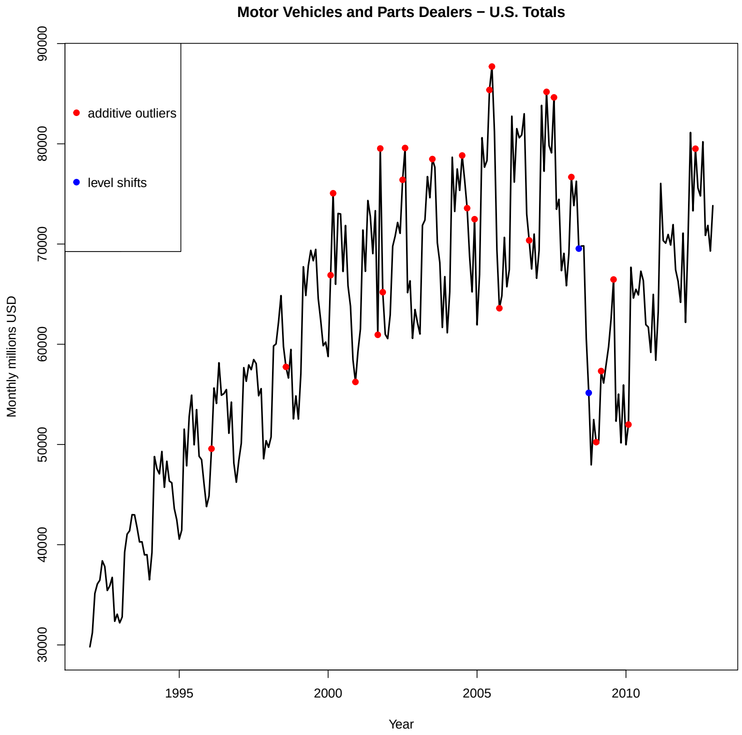
[38] Ruppert, D. (1987). What is kurtosis? An influence function approach. *The American Statistician.* **41**, 1-5.

[39] Samorodnitsky, G. and Taqqu, M.S. (1994) *Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variances*, New York: Chapman & Hall.

[40] Shapiro, S. S. and Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". *Biometrika*, 591-611.

[41] Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* **19**, 279–281.

[42] U.S. Census Bureau (2015), X-13ARIMA-SEATS Reference Manual, U.S. Census Bureau, U.S. Department of Commerce, Washington, DC. `http://www.census.gov/ts/x13as/docX13AS.pdf`



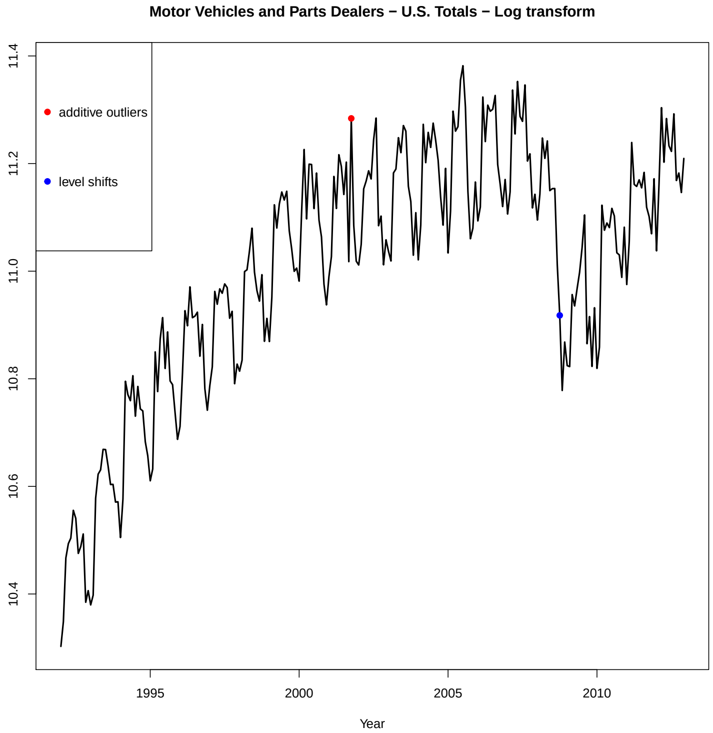


Figure 4: Comparison of pre-adjustments to motor vehicles and parts dealers series. Top plot shows location of 26 additive outlier regressors and two level shift regressors. Bottom plot shows a log transform along with two additive outlier regressors and a level shift.

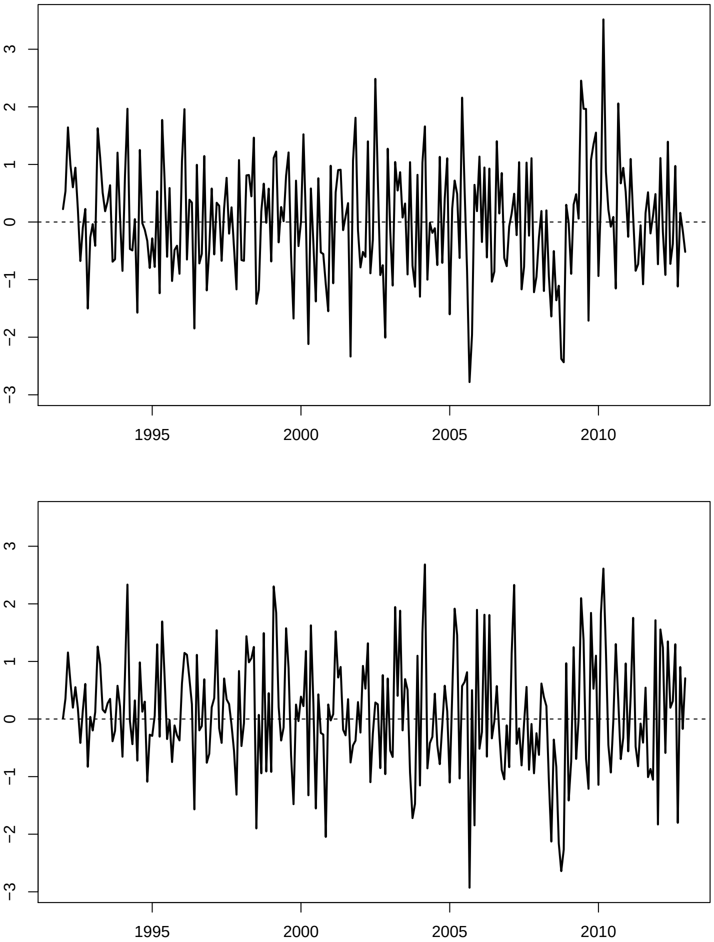


Figure 5: raw residuals from automated model selection routines. Top plot is ϵ_t^{log}. Bottom plot is ϵ_t^{ao}.

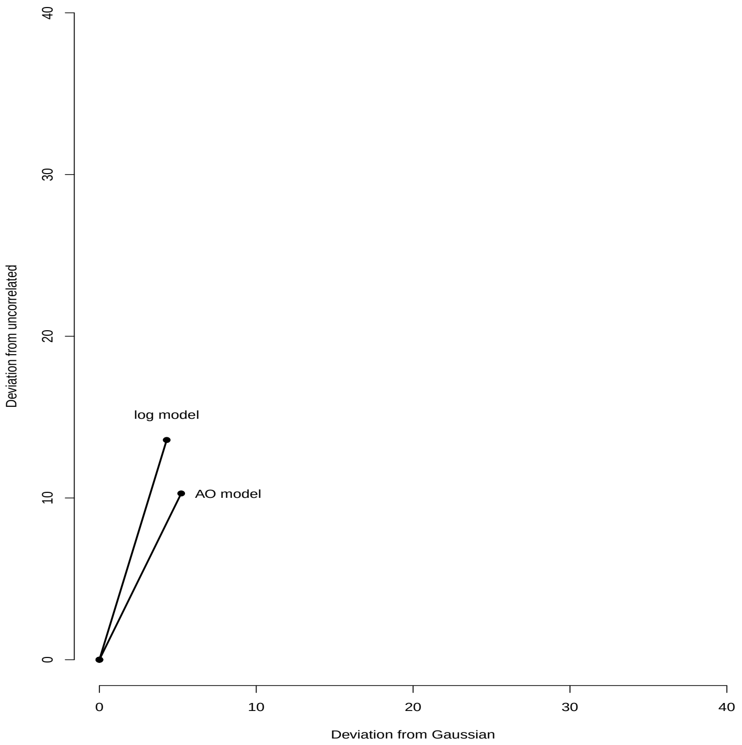


Figure 6: Results from Automobile Dealers series residuals. Longer line is from ϵ_t^{log}. Shorter line is from ϵ_t^{ao}.

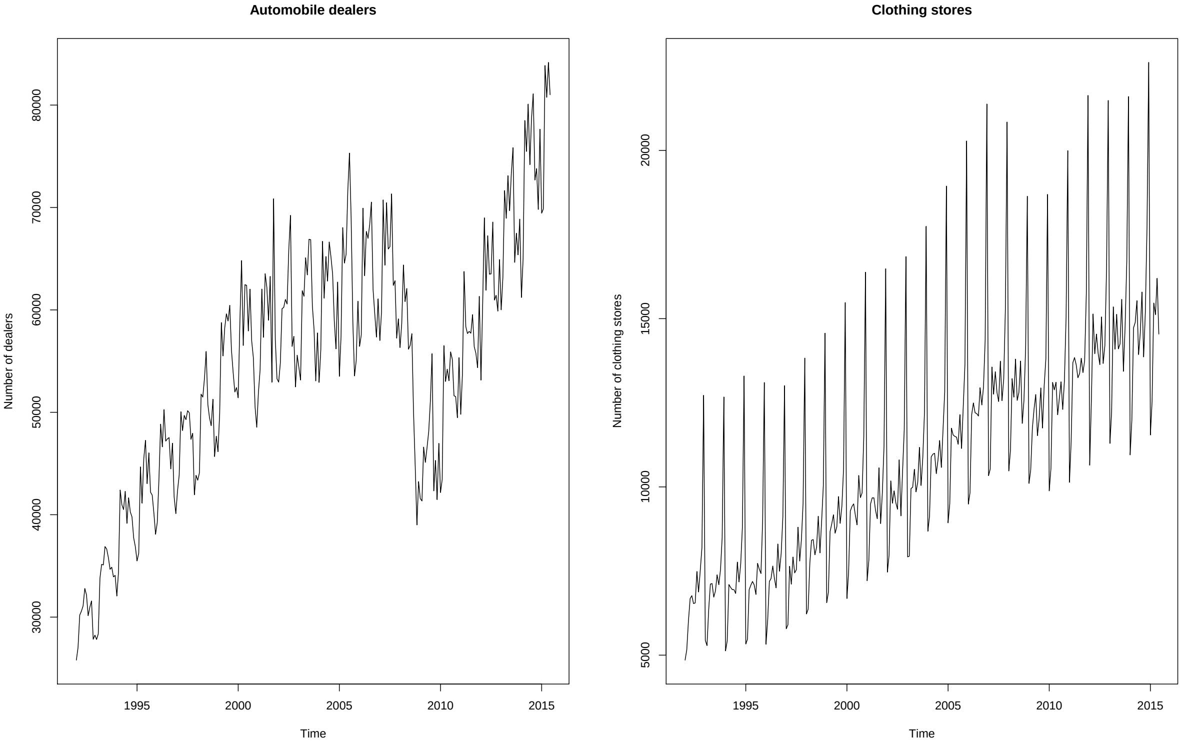


Figure 7: Retail series