Edgeworth's Time Series Model: Not AR(1) But Same Covariance Structure

Stephen Portnoy[1]

It is a profound pleasure to be able to contribute a paper to honor Roger Koenker, whose friendly collegiality over innumerable coffees has been the spark for so much of my research. The results here are especially appropriate: Roger introduced me to Edgeworth's paper.

November 20, 2017

Abstract

In an 1886 paper, Edgeworth developed a method for simulating time series processes with substantial dependence. A version of this process with normal errors has the same means and covariance structure as an AR(1) process, but is actually a mixture of a very large number of processes, some of which are not stationary. That is, joint distributions of lag 3 or greater are not normal but are mixtures of normals (even though all successive pairs are bivariate normal). Thus, it serves as a cautionary example for time series analysis: though the AR(1) process can not be distinguished from the Edgeworth Process by second order properties, inferences based on an AR(1) assumption can fail under the Edgeworth model. This model has many additional surprising features, among which is that it has Markov structure, but is not generated by a one-step transition operator.

[1] Professor, Department of Statistics, University of Illinois at Urbana-Champaign
corresponding email: sportnoy@illinois.edu

1 Introduction.

Most econometricians and statisticians develop a jungle (or at least a zoo) of wild beasts they can use as counterexamples to overly optimistic application of theoretical statistical results. For example, one often wants an example of a bivariate distribution with normal marginals that is not bivariate normal. The example here is a substantial generalization of this to time series models. A version originally appeared in Edgeworth (1886) as an early attempt to simulate economic processes with dependence structures. Edgeworth's paper substantially predates the early work of Yule (1927) and others on time series methods (e.g., see Tsay (2000)). Even before Edgeworth's paper, actuaries were also interested in insurance processes involving increments and decrements occurring "randomly" in time, but the earliest stochastic models for actuarial processes were not developed until a 1903 Ph.D. thesis of Filip Lundberg ("Approximations of the probability function: reinsurance of collective risks", see Bühlmann (1997)). The formal mathematical development of such models did not occur until the second quarter of the twentieth century, leading to a variety of classes of models (Compound Poisson, branching processes, and generalizations) that would provide much better and more natural models for the kinds of economic processes Edgeworth considered.

To generate random innovations, Edgeworth took digits at random ("from pages in a mathematical table"). To generate dependence, he look each observation X_t to be a sum of $m = 20$ such innovations; but rather than as an MA(1), at each time t he generated an independent random choice of an integer in $1{:}m$ and replaced the corresponding innovation by a new independent one. That is, each successive pair of observations had $m-1$ overlapping innovations and only 1 new one; thus generating a highly dependent series.

Here we will replace the innovation distribution by the more modern standard normal and will develop some of the rather wild properties of this beast. With some abuse of nomenclature, the process with normal innovations will still be called the "Edgeworth" process. The purpose here is to provide a cautionary example emphasizing the critical nature of assumptions and hypotheses under which statistical procedures are developed. As noted above, the Edgeworth example is unlikely to be a natural or useful alternative to more modern classes of stochastic process models (including time-domain models), but it does provide an example under which standard time-domain analysis would be invalid even though the process has exactly the same mean and covariance structure as an AR(1) model. Thus, it could not be distinguished from an AR(1) by any methods based on second order properties. This is a lesson of crucial importance to both the theory and the practice of statistical data analysis. In all but the simplest data analysis cases, there will generally be numerous alternative models that would be hard (or impossible) to detect without prior knowledge of where to look, but whose use would lead to substantially different conclusions. Thus, the use of models chosen because they are convenient or appear in standard textbooks should be avoided unless clear scientific justification is available.

To define the Edgeworth Process formally, let

$$e_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\,1) \quad t = -m,\, -(m-1),\, \cdots \tag{1}$$

$$V_t \overset{i.i.d.}{\sim} \mathcal{U}(1:m) \quad ; \quad \delta_{t,j} \equiv I(V_t = j) \quad t = 0,\, 1,\, \cdots \tag{2}$$

where $\mathcal{U}(1:m)$ denotes the uniform distribution on the integers 1 through

m, and I denotes the indicator function. Now define $e_{t,j}$ and the process Y_t recursively for $t = 0, 1, \cdots$ and for $j = 1, \cdots, m$ as follows:

$$\begin{aligned} e_{0,j} &\equiv e_{-(m-j+1)} \quad ; \quad e_{t,j} \equiv e_{t-1,j}\,(1-\delta_{t,j}) + e_t\,\delta_{t,j} & (3) \\ Y_t &\equiv \sum_{j=1}^{m} e_{t,j} \ . & (4) \end{aligned}$$

Some elementary properties are as follows:

Basic Properties:

1. $Y_t \sim \mathcal{N}(0,\, m)$.

   Since $\{e_t\}$ and $\{V_t\}$ are independent, each Y_t is a sum of m independent unit normals.

2. $\{Y_t\}$ is stationary.

   Consider the vector-valued process,

   $$\{U_t \equiv (e_{t-m}, \cdots, e_t\, ;\, V_t\,) \;:\; t = 0,\, 1,\, \cdots\} \,.$$

   Since $\{e_t\}$ and $\{V_t\}$ are i.i.d. and are also independent of each other, $\{U_t \,:\, t = 0, 1, \cdots\}$ is stationary. Thus, since $Y_t = f(U_t)$ for a fixed function, f (independent of t), $\{Y_t \,:\, t = 0, 1, \cdots\}$ is stationary.

3. The joint distribution of (Y_1, Y_2) (equivalently, (Y_t, Y_{t+1})) is bivariate normal:

   $$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}_2(0, \begin{pmatrix} m & m-1 \\ m-1 & m \end{pmatrix}) \,. \qquad (5)$$

   This holds since Y_1 is a sum of m independent normals; and, no matter what V_2 is, Y_2 is the sum of $(m-1)$ of the normals summands in Y_1 plus one independent normal. That is, if w_1 is the sum of the $(m-1)$

e_t's in Y_1 that are not in Y_2, w_2 is the e_t in Y_1 that was replaced, and w_3 is the e_t that replaces w_2 in Y_2, then

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \begin{pmatrix} w_1 + w_2 \\ w_1 + w_3 \end{pmatrix}$$

where the w's are independent with $w_1 \sim \mathcal{N}(0,\, m-1)$ and w_2 and w_3 are $\mathcal{N}(0,\, 1)$. Thus (5) follows.

4. The Edgeworth Process does not satisfy a time-domain model. Basically, the joint distribution of k successive observations is **not** multivariate normal for $k \geq 3$, but is a (non-trivial) mixture of multivariate normals. Specifically, consider $k = 3$ and $m = 2$, and note that $(Y_t,\, Y_{t+1},\, Y_{t+2}) \sim (Y_1,\, Y_2,\, Y_3)$. The components of the distribution of $(Y_1,\, Y_2,\, Y_3)$ depend only on whether or not $V_3 = V_2$. If $V_3 = V_2$, Y_2 and Y_3 share exactly one normal component with Y_1; while if $V_3 \neq V_2$, Y_2 shares a component with Y_1, but Y_3 shares no components with Y_1 and so is independent of Y_1. Thus,

$$(Y_1,\, Y_2,\, Y_3) \sim \frac{1}{2} \otimes \mathcal{N}(0, \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}) + \frac{1}{2} \otimes \mathcal{N}(0, \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}), \quad (6)$$

where the notation "$\frac{1}{2} \otimes$" means that the mixing probability is $\frac{1}{2}$.

The generalization to arbitrary k and m depends on occupancy theory and is described in Property 5 below. Details for the case $m = 2$ appear in Section 2. To help anticipate the general result (which is somewhat complicated to write), the distribution for $k = 4$ and $m = 2$ is a mixture of 4 multivariate normals with equal mixing probabilities (1/4), and with means 0 and covariance matrices:

$$\begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} \tag{7}$$

Deeper and Wilder Properties:

5. (a) The distribution of $(Y_1, Y_2, \cdots, Y_k)$ (or $(Y_t, Y_{t+1}, \cdots, Y_{t+k-1})$) is a mixture of k-variate normal distributions. Considering $(e_{t,j})$ as occupying m cells, the distribution depends only on the occupancy distribution as defined by the $\{V_t\}$. As described in more detail for the $m = 2$ case in the next section, given specific values for $\{V_t\}$, the Y_t's are all sums of m specific e_t's; and so only the covariance matrix (conditional on the V_t-values) is needed. The covariance of Y_1 and Y_k depends only on how many e_t's are common in their sums. Specifically, there is a common e_t in the sums for Y_1 and Y_k (for $k > 1$) if and only if the corresponding cell (defined by $(V_1, \cdots, V_k)$) is empty. If there are R such cells, then $\mathrm{Cov}(Y_1, Y_k|R) = R$; and so $\mathrm{Cov}(Y_1, Y_k) = ER$. From "occupancy theory" (*e.g.*, see Feller (1968)), $ER = m(1 - 1/m)^{k-1}$. This is also clear since the probability that cell j is empty at time k is just the probability that $\{V_2, \cdots, V_k\}$ all differ from j, which is $((m-1)/m)^{k-1}$; and since R is just the sum of the indicator functions that cell j is empty at time k.

(b) As a consequence:

$$\begin{aligned} \mathrm{Cov}(Y_t,\, Y_{t+k}) &= m(1-1/m)^{k-1} && (8) \\ \mathrm{Corr}(Y_t,\, Y_{t+k}) &= (1-1/m)^{k-1}\ . && (9) \end{aligned}$$

(c) Thus the Edgeworth Process has exactly the same covariance structure as an AR(1) process with autocorrelation $1-1/m$ and marginal variance m; but it is clearly not an AR(1) since the k-variate distributions are mixtures of normals if k is greater than 2.

6. Perhaps the deepest properties of the Edgeworth Process concern its Markovian properties. It is clear from the definition (see (1) and (3)) that the distribution of $(Y_t,\, Y_{t+1},\, \cdots)$ is determined entirely by $\{e_{t,j} \,:\, j = 1, ..., m\}$ and the future e's and V's: $\{(e_k,\, V_k) \,:\, k = t,\, t+1,\, \cdots\}$. That is, letting $\mathcal{S}_t$ be the (joint) sigma field generated by both $\{e_{t,j} : j = 1, ..., m\}$ and V_t, the Edgeworth Process satisfies

$$\{Y_t,\, t > k\} \,|\, \{S_t,\, S_{t-1},\, \cdots\} \sim \{Y_t,\, t > k\} \,|\, S_t\,\}\ , \qquad (10)$$

and thus the Edgeworth Process is a Markov process adapted to the sigma fields $\{\mathcal{S}_t\}$.

However, the Edgeworth Process is not itself a Markov chain. Informally, since the joint distribution of $(Y_t,\, Y_{t+1})$ is the same as that of an AR(1), it can not be generated by a one-step transition operator (since the operator would generate the AR(1) process). More explicitly, from the joint distribution described in the following Section, the conditional distribution of Y_t given $\{Y_{t-1},\, Y_{t-2},\, \cdots\}$ does not depend only on Y_{t-1}. Specifically, from (6) , the conditional distribution of

Y_3 given $(Y_1,\, Y_2)$ is a mixture of two normal distributions with (conditional) means $\frac{1}{3}Y_1 + \frac{1}{3}Y_2$ and $-\frac{1}{3}Y_1 + \frac{2}{3}Y_2$ respectively (as calculated from the covariance matrices in (6)). So the conditional distribution does not depend only on Y_2.

7. From Properties 4 and 5 above (and Section 2), it is clear that the prediction distribution is rather more complicated for the Edgeworth Process than for an AR(1). Let $m = 2$ and consider prediction of $Y_{N+\ell}$ from $Z \equiv (Y_1,\, \cdots,\, Y_N)$. From Section 3, the distribution of $(Z,\, Y_{N+\ell})$ is a mixture of $K = 2^{n+\ell-2}$ multivariate normal components with mixing probabilities p_k and covariance matrices Σ_k for $k = 1,\, \cdots,\, K$ (and zero means); and so the conditional (prediction) distribution becomes:

$$f(Y_{N+k}|Z) = \frac{\sum_{k=1}^{K} p_k \varphi_{Y_{N+\ell}\,|\,Z}(y,\, z)\varphi_k(z)}{\sum_{k=1}^{K} p_k \varphi_k(z)} \tag{11}$$

where $\varphi_{Y_{N+\ell}\,|\,Z}$ is the conditional density of the k-th component based on Σ_k and $\varphi_k(z)$ is the marginal density of Z for the k-th component. More precisely, let Σ_k be partitioned into 4 submatrices: $\Sigma_k^{1,1}$ equal to the upper $N \times N$ submatrix corresponding to Z, $\Sigma_k^{2,1}$ equal to the (row) vector of elements $(Cov(Y_i,\, Y_{N+\ell})\, :\, i = 1,\, \cdots,\, N)$, the transpose $\Sigma_k^{1,2}$, and the lower right entry, $\Sigma_k^{2,2} = 2$. Then $\varphi_{Y_{N+\ell}\,|\,Z}$ is the density for $\mathcal{N}(\Sigma_k^{2,1}(\Sigma_k^{1,1})^{-1}y,\, 2 - \Sigma_k^{2,1}(\Sigma_k^{1,1})^{-1}\Sigma_k^{1,2})$, and φ_k is the density for $\mathcal{N}(0,\, \Sigma_k^{1,1})$.

Generally, this will depend strongly on all of Z and will have no simple closed-form expression. However, for $N = 2$ and $\ell = 1$ the distribution of $(Y_1,\, Y_2,\, Y_3)$ has two components (see (6)). Since the marginal distribution of $(Y_1,\, Y_2)$ is the same for both components, this marginal density will factor (and cancel); and so the conditional density of $Y_3 \,|\, (Y_1,\, Y_2)$ is just the mixture of the conditional densities for each of the two components. Thus, computing the conditional distributions for each of the two components yields:

$$Y_3 \,|\, (Y_1 = y_1,\, Y_2 = y_2) \sim \frac{1}{2} \otimes \mathcal{N}\left(\frac{1}{3}(y_1 + y_2),\, \frac{5}{3}\right) + \frac{1}{2} \otimes \mathcal{N}\left(\frac{2}{3} y_2,\, \frac{5}{3}\right) \ . \quad (12)$$

Since conditional normal means are linear and (hence) are defined solely by the second order properties, the point prediction of Y_3 (given $(Y_1,\, Y_2)$ will be exactly the same as for an AR(1) process (specifically, $Y_2/2$).

For $N \geq 3$, the conditional distribution of $Y_{N+\ell}$ given $Z = (Y_1,\, \cdots,\, Y_N)$ is no longer even a mixture of conditional normals; and in fact, for each component, the conditional mean of $Y_{N+\ell}$ depends on all the conditioning variables $(Y_1,\, \cdots,\, Y_N)$. The coefficients are constants defined from Σ_k, but the conditional means are weighted by the density $\varphi_k(z)$. Thus, it may seem rather surprising that, in fact, the coefficients of $(Y_1,\, \cdots,\, Y_{N-1})$ all cancel and that $E[Y_{N+\ell} \,|\, Z\,] = (\frac{1}{2})^{\ell}\, Y_N$. That is, the point predictor under the Edgeworth model is always the same as that under the AR(1) model.

To show this, let $\mathcal{S}_N$ be the sigma-field generated by all the e's and V's at or before time N. Then since $\mathcal{S}_N$ determines $Z = (Y_1,\, \cdots,\, Y_N)$, it is a larger sigma-field. Thus

$$E[Y_{N+\ell} \,|\, Z\,] = E[E[Y_{N+\ell} \,|\, \mathcal{S}_N\,] \,|\, Z\,] \ . \qquad (13)$$

To compute the inner conditional expectation, note that $Y_N = w_1 + w_2$, where w_1 and w_2 are earlier innovations. Then, if $\{V_{N+1},\, \cdots,\, V_{N+\ell}\}$ all equal 1, $Y_{N+\ell} = w_2 + e_{N+\ell}$; if $\{V_{N+1}, \cdots, V_{N+\ell}\}$ all equal 2, $Y_{N+\ell} = w_1 + e_{N+\ell}$; and otherwise $Y_{N+\ell} = e_1^* + e_2^*$, where these two summands are future innovations and are independent of $\mathcal{S}_N$. Therefore, since $V_{N+1}, \cdots, V_{N+\ell}$ equal a fixed index with probability $(\frac{1}{2})^\ell$

$$\begin{aligned} E[Y_{N+\ell} \,|\, \mathcal{S}_N\,] &= 2^{-\ell}\, w_1 \,+\, 2^{-\ell}\, w_2 \,+\, (1 - 2^{-(\ell-1)}) \times 0 \\ &= 2^{-\ell}\,(w_1 + w_2) \,=\, \left(\frac{1}{2}\right)^\ell \, Y_N \;. \end{aligned} \tag{14}$$

Thus, since Y_N is part of Z, (13) implies that the point predictor under the Edgeworth model is also $(\frac{1}{2})^\ell \, Y_N$, the same as for an AR(1).

To see the source of the cancelation underlying this result, consider predicting Y_4 from $(Y_1,\, Y_2,\, Y_3)$. The four components of the distribution of $(Y_1,\, Y_2,\, Y_3),\, Y_4)$ can be obtained (recursively) from the components of the distribution of $(Y_1,\, Y_2,\, Y_3)$ as follows: for each 3×3 matrix (say, Σ_0) in (6) , there are two unique 3-vectors a and b (depending on Σ_0) such that two component matrices in (7) are

$$\begin{pmatrix} \Sigma_0 & a \\ a' & 2 \end{pmatrix} \text{ and } \begin{pmatrix} \Sigma_0 & b \\ b' & 2 \end{pmatrix} \;. \tag{15}$$

Since the marginal distribution of $(Y_1,\, Y_2,\, Y_3)$ is the same for these two components, the conditional densities add, and the conditional means also add when computing the point predictor. While neither of the conditional means is $Y_3/2$, the coefficients of Y_1 and Y_2 cancel in the sum, of the conditional means for each pair of components . Thus, the sum is $Y_3/2$, which can be factored out of (11) to provide the desired result. A bit more complicated computation shows the same pairwise cancellation for $N = 4$, and I conjecture that such pairwise cancellation always holds.

Note that the conditional distribution of $Y_{N+\ell}$ given $(Y_1, \cdots, Y_N)$ is still a very complicated mixture and is quite different from the simple normal distribution under the AR(1) process. Some discussion of the practical implication of the difference is provided in Section 3.

2 On the joint distribution when $m = 2$

When $m = 2$ it is possible to describe the joint distribution of the Edgeworth series in some detail, though it is not clear that there is a feasible way to calculate the density numerically. As noted above, the joint distribution of a sample of length n is a mixture of n-variate normals, where the covariance matrices of the components are determined by the occupancy distribution. For $m = 2$, this is given simply by the runs distribution.

First, from (8) , the main tri-daigonal of the covariance matrix for any component is as follows: the main diagonal is always 2, and the sub- and super- diagonals are all 1. Now, consider the first row of a covariance matrix (as defined by the V_t's) and let t_1 be the length of the fist run (of V's) starting with V_2. Generalizing from Property 4 (equation (6)), if $t_1 = 1$, V_2 and V_3 differ and Y_1 and Y_3 have independent summands. Thus, $\text{Cov}(Y_1, Y_3) = 0$, and also $\text{Cov}(Y_1, Y_t) = 0$ for $t \geq 3$. Otherwise, V_2 and V_t will be equal up to $t = t_1+1$, and the corresponding pairs of Y's will share exactly one summand. Thus, $\text{Cov}(Y_1, Y_t) = 1$, for $t = 2, \cdots, t_1+1$. At time t_1+1 a different value of V (specifically, V_{t_1+2}) appears, and all subsequent covariance's will be zero. That is, the upper $(t_1 + 1) \times (t_1 + 1)$ submatrix will be simply the interclass covariance matrix with diagonals equal to 2 and all other entries equal to 1.

The remaining upper $(t_1+1)\times(n-t_1-1)$ submatrix will be all 0 **except** for the $(t_1+1,\, t_1+2)$ element, which is on the super-diagonal and is 1. The lower $(n-t_1-1)\times(t_1+1)$ submatrix will be defined by symmetry (based on the length of the first run).

Continuing recursively, let t_2 be the length of the next run. Then the next $t_2\times t_2$ diagonal submatrix of the covariance matrix will be determined (in the same manner as above), as will the subsequent off-diagonal entries. This continues until reaching the lower 2×2 submatrix, which is known to be $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Therefore, each runs distribution will determine a unique $(n\times n)$ covariance matrix, and the associated probability will be associated with the corresponding n-variate normal component. From the runs distribution, there will be 2^{n-2} multivariate normal components each with probability $2^{-(n-2)}$. Thus, if n is not small attempts to compute the density would lead to serious round-off complications.

3 Distinguishing AR(1) from Edgeworth

When faced with real time-series data, the good statistician will work with the scientist who took the data to develop an appropriate model. There are a variety of models that may apply when the data arises from a process with random increments and decrements occurring in time, but as noted previously it is not likely that the Edgeworth model would be reasonable. In cases where no physical model suggests itself, the statistician will quite generally try standard ARIMA fitting. After addressing possible non-stationarity (perhaps through a unit-root test or detrending), the statistician will estimate the autocorrelation function, and try to fit an ARIMA model. Suppose the data were generated by an Edgeworth Process (say, with $m = 2$). The autocovariance function would be exactly the same as for an AR(1), and so it is very likely that the series would be identified as an AR(1) (at least if the series is long enough). Standard diagnostics for adequacy of fit would fail to cast doubt on the AR(1) model. The spectrum is exactly the same; and attempts to check for non-normal innovations would also fail since the residuals are also approximately normal for the Edgeworth Process. A scatter plot of $(y_{t+1},\, y_t)$ would also be (exactly) the same as for an AR(1) process. Only by analyzing sets of three or more successive observations would it be possible to distinguish the Edgeworth Process from an AR(1).

Nonetheless, the processes are different, and use of an AR(1) assumption when the Edgeworth Process holds can be significant. Consider prediction intervals in the $m = 2$ case: under an AR(1), for predicting Y_{N+2} the (conditional) prediction distribution is $\mathcal{N}(Y_N/4,\, 15/8)$. As developed in property 7, the Edgeworth distribution of $Y_{N+2}\,|\,Y_N]$, can be found to be $\frac{1}{2}\otimes\mathcal{N}(Y_N/2,\, 3/2)+\frac{1}{2}\otimes\mathcal{N}(0,\, 2)$. As noted, the conditional mean is the same: $Y_N/4$; but the conditional distribution and coverage can be quite different if Y_N is moderately large. If one wants to report the conditional variance, the difference is more remarkable: for $Y_N = 4$, the (true) Edgeworth prediction variance is over 50% larger than the nominal AR(1) value.

As a specific numeric example, consider a Edgeworth Process of length 1000, and suppose one wants to find a prediction interval two steps ahead following an observation greater than 5. Recall that the variance is 2, so an observation this large with n = 1000 is not unexpected. Specifically, a random Edgeworth Process of length 1000 was simulated, and there was one observation greater than 5: $y[755] = 5.6642$. If a 95% prediction interval were calculated using an AR(1) model, the coverage under the Edgeworth model would be .83 if the exact (AR(1)) variance of $\sqrt{15/8}$ were used and only .76 if an AR(1) estimate from the simulated data were used. The difference from .95 seems sufficiently large to be disconcerting.

More generally, how well can a formal test of AR(1) vs. Edgeworth distinguish the processes? By the Neyman-Pearson Lemma, the best test would be based on the likelihood ratios; but as noted above, the Edgeworth (mixed) density is essentially impossible to compute. Thus, as an alternative test, consider testing consecutive sets of 3 observations $(Y_t,\, Y_{t+1},\, Y_{t+2})$ by the Neyman-Pearson test and then taking the maximum over the $n-2$ such test statistics. To assess distinguishability, consider the test with equal Type I and Type II error probabilities; that is, choose the critical value to give equal Type I and Type II errors (as estimated by simulations). The error probability is then compared to that for some common well-known tests. Specifically, the error probability (for the test above based on the maximum of 3-observation tests) was estimated empirically using a simulation of 1000 Edgeworth and AR(1) samples of length 800. Table 1 gives the (empirical) equal-tailed error probability (for the test above) for sample sizes $n = 100$, 200, 400, and 800; and lists the mean difference Δ for the equal-tailed test of $\mathcal{N}(0,\, 1)$ against $\mathcal{N}(\Delta,\, 1)$ with the same (equal) error probabilities (based on an i.i.d. sample of length n). Since Δ is greater than .1 when $n = 100$, the Edgeworth and AR(1) processes can be distinguished better than a mean difference of 1 standard error in i.i.d. normal samples of this size. The best

(Neyman-Pearson) test would perform even better.

Perhaps of more interest, we can compare "distinguishable differences" for parameters under dependent samples. For example, for testing a mean with AR(1) errors with $\rho = .5$, the standard error of the mean is $\sqrt{2}$ times greater, and so the equivalent mean difference would also be $\sqrt{2}$ larger. Alternatively, for testing the autocorrelation, ρ, in an AR(1) the equivalent difference (in ρ) would be nearly the same as in Table 1, since the standard error of $\hat{\rho}$ is approximately $1/\sqrt{n}$.

Table 1: Mean Difference Δ equivalent to test of AR(1) vs. Edgeworth

n	100	200	400	800
error prob	0.292	0.244	0.198	0.174
Δ (SD units)	.1095	.0981	.0872	.0849

Note that derivations of Kullback-Leibler distances could provide a less empirical measure of deviation. The Kullback-Leibler distance between two distributions P and Q with densities p and q is

$$\mathrm{KL}(P,\, Q) \equiv \int \log\left(\frac{p(x)}{q(x)}\right)\, p(x)\, dx \ .$$

While the Edgeworth mixture density is too complicated to permit analytic integration, the Kullback-Leibler distances between the Edgeworth Process and the corresponding AR(1) can be approximated by simulation. For $m = 2$, the distances were KL(AR(1), Edge) = .0046 and KL(Edge, AR(1)) = .0053 (based on simulations with 40,000 replications, for which the standard error is approximately .0005). The Kullback-Leibler distance between $\mathcal{N}(0,\, 1)$ and $\mathcal{N}(0,\, \Delta)$ is $1/2\, \Delta^2$. Thus, the discrepancies equivalent to the two Edgeworth and AR(1) KL-distances are $\Delta = .095$ and $\Delta = .103$, respectively; which seem quite similar to the values in the table. Kullback-Leibler distances were also calculated (by the same simulations) for $m =$ 3, 4, 5, 10 and 20. They were clearly decreasing in m, thus corroborating the comment at the beginning of this Section that the $m = 2$ case is the easiest to distinguish.

4 Extensions

Even more dangerous beasts can be constructed by allowing non-normal distributions, by defining more complex versions of the Edgeworth Process, or by generalizing the innovation process. While any one such process may be relatively easily distinguished from an assumed model (*e.g.*, a time domain model), it seems highly unlikely that there is an omnibus test that would be informative for all such models. Two simple extensions are described below, but investigating such beasts in greater generality is likely to prove intriguing and perhaps frightening.

For one extension, consider replacing the normal innovations, $\{e_t\}$, with negative exponential ones; and consider the case $m = 2$. The process will no longer have normal joint distributions, and the means will each be 2 (as a sum of 2 negative exponentials); but the process will still be stationary and have the same covariance structure as an AR(1) with $\rho = .5$).

Successive observations will again share exactly one innovation; and so using equation (3) with $e_{t,1} = e_{t-1}$ and $e_{t+1,1} = e_t$ (for example), we have

$$Y_t \sim e_{t-1} + e_t \quad ; \quad Y_{t+1} \sim e_t + e_{t+1} = WY_t + e_{t+1} \ , \tag{16}$$

where $W \equiv e_t/(e_{t-1}+e_t)$. Note that W is uniform on (0, 1) and is independent of $e_{t-1}+e_t = Y_t$, but it is not independent of Y_{t+1} . Allowing $\{W_t\}$ to be i.i.d. will define the following random-coefficient AR(1):

$$Y_t = W_t\, Y_{t-1} \,+\, e_t \qquad W_1,\, W_2,\, \cdots \quad i.i.d.\ \mathcal{U}(0,\, 1)\,; \tag{17}$$

which should be a better approximation to the Edgeworth Process than the constant-coefficient AR(1). Note that (17) is a Quantile Autoregression Model; see Koenker and Xiao (2006). It is straightforward (though a bit tedious) to compute the conditional densities of Y_3 given Y_1 for the Edgeworth and constant–coefficient processes. The conditional density for the random-coefficient AR(1) is somewhat more difficult to obtain, and so a density estimate based on 10,000 simulated triples (Y_1, Y_2, Y_3) was used. These densities are plotted in Figure 1. Clearly, the constant-coefficient AR(1) will be easily distinguished from the Edgeworth version, though again the two processes would share the same covariance structure. The random-coefficient AR(1) would be much more difficult to distinguish unless one knows exactly what to look for.

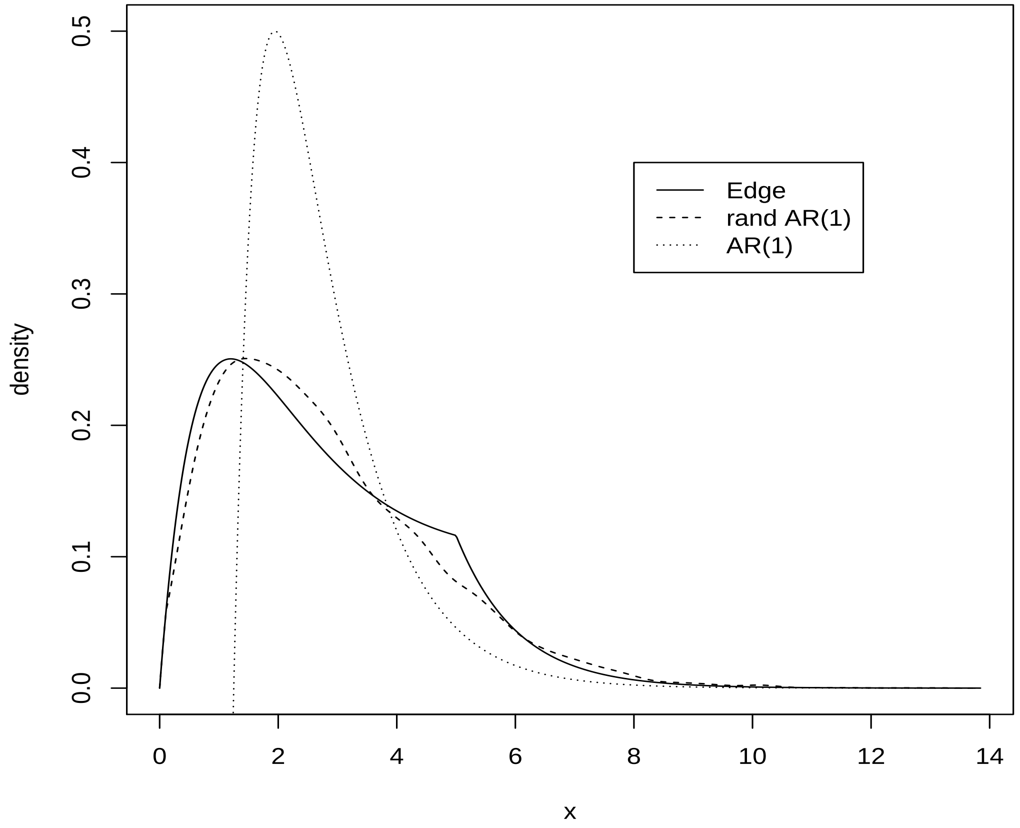

Figure 1: Conditional Densities: $Y_3 \,|\, Y_1 = 5$

A second example addresses the question of whether there are Edgeworth-like (Gaussian) processes that are even closer to the AR(1). Specifically, are there processes such that triples (or larger sets) of successive observations share the same joint distribution. The following example answers the question in the affirmative: generate the e_t's and V_t's as before, but Y_t will now be the sum of the innovations in each of 4 locations L1, L2, L3, and L4. At each t, there will be four independent (negative exponential) innovations in these locations. To define Y_{t+1}, redefine the innovations in the locations as follows. If $V_t = 1$, replace the innovation in L1 by a new one, then interchange the innovations in L3 and L4, and replace the innovation in L3 by a new (independent) one. If $V_t = 2$, do the analogous replacements in L2 and L4 after an interchange (between L3 and L4). Then entries in L1 and L2 look exactly as for the original Edgeworth Process, but entries in L3 and L4 remain the same over two trials only if the V_t's alternate. It is not hard to show that Y_t and Y_{t+1} always have exactly 2 of the 4 e_t's in common, while Y_t and Y_{t+2} will always have exactly 1 of the 4 e_t's in common. Thus, $(Y_t,\, Y_{t+1},\, Y_{t+2})$ will have a trivariate normal distribution with mean zero and covariance matrix:

$$\begin{pmatrix} 4 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 4 \end{pmatrix},$$

which is exactly the same as the distribution for an AR(1) with variance 4 and correlation .5.

Clearly, there is an enormous number of such processes, as well as various Markovian and non-Markovian processes that will also have the same or very similar covariance structures. It seems clear that no finite data set could ever distinguish an AR(1) from all of them.

5 Conclusions

Despite its historical antiquity, the Edgeworth Process provides a serious cautionary warning about the overly enthusiastic application of standard time series theory and methodology. Specifically:

1. Even for (marginally) normal observations, second-order properties do not identify the process. Only when the entire series is jointly normal do second order properties suffice. As the Edgeworth Process shows, there are processes that would pass any standard model-adequacy tests for an AR(1) (based on second-order properties), but that differ from AR(1).

2. Even if the time series looks very much like an AR process, inferences based on this assumption may be quite inappropriate. For the Edgeworth Process, the two-step-ahead prediction interval following an unusually large observation is a specific example, but other inferences involving higher-order properties would also be affected.

3. A stationary Markov process adapted to a sequence of sigma-fields (sets of auxiliary random vectors) need not be generated by a one-step transition operator. This is actually a general property of hidden Markov processes, of which the Edgeworth Process is an example; and it emphasizes the importance of considering the states of the hidden variables (here, the V_t's) when using the model.

The Edgeworth Process is also useful in a classroom setting. It is accessible and well-behaved (stationary and Markovian), and yet provides a counterexample to a variety of theoretical results when the hypothesized time-domain model does not hold. The process is also sufficiently simple that it can provide a source for intriguing and enlightening homework problems, even in an introductory course. Though careful pedagogic developments go beyond the scope of this article, it may be noted that this is where mathematics is most valuable in scientific research: the conclusions of theorems are rarely surprising, but the mathematical development establishes the connections with essential hypotheses. If these hypotheses fail, then inference may be highly compromised; and careful consideration of the hypotheses can often lead to appropriate measures for detecting and adjusting for their failure.

To conclude, examples like the Edgeworth Process demonstrate clearly the dangers in assuming models simply because they are convenient or appear in standard textbooks. Generally (especially for more complex data), there will be an extremely wide range of alternative models that will be very hard (or even impossible) to distinguish from the nominal model unless external information suggests what to look for. Nonetheless, statistical analyses and conclusions may be substantially different under these alternatives. The Edgeworth Process is a specific example cautioning the data analyst to employ models only when the underlying model assumptions have clear and convincing scientific justification.

References

[1] Bühlmann, H. (1997). The actuary: the role and limitations of the profession since the mid-19th century. *Astin Bulletin, 27(02)*, 165-171.

[2] Edgeworth, F.Y. (1886). Problems in Probabilities, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Vol 22 – Fifth Series*, Taylor and Francis, Red Lion Court, Fleet Street, London; 371 - 384.

[3] Feller, W. (1968). An introduction to probability theory and its applications: volume I, Wiley, 1959.

[4] Koenker, R. and Xiao, Z. (2006). Quantile Autoregression, *Journal of the American Statistical Association, 101*, 980-990.

[5] Tsay, R.S. (2000). Time Series and Forecasting: Brief History and Future Research, *Journal of the American Statistical Association, 95*, 638-643.