Modeling Time Series with both Permanent and Transient Components using the Partially Autoregressive Model

Matthew Clegg*

January 28, 2015

Abstract

A time series model is discussed that incorporates both permanent and transient effects. Estimation techniques are given, and the power of the likelihood ratio test is assessed. When applied to the monthly price/earnings series of the S&P 500 over the period 1871–2013, both permanent and transient components are found, although the transient component is by far the more dominant effect. When applied to the daily returns of the individual components of the S&P 500 over the period 2002–2013, evidence is found that a statistically significant proportion of these series exhibit both permanent and transient effects.

Introduction

In this paper, I revisit a model that was previously considered in Summers [9] and also in Poterba and Summers [6]. The model was proposed as a means for describing financial time series that have both permanent and transient components, and is given by the following specification[1]:

*Author's email: matthewcleggphd@gmail.com

[1]The notation has been changed from the original papers.

$$
\begin{aligned}
X_t &= M_t + R_t \\
M_t &= \rho M_{t-1} + \epsilon_{M,t} \\
R_t &= R_{t-1} + \epsilon_{R,t} \\
\epsilon_{M,t} &\sim NID(0, \sigma_M^2) \\
\epsilon_{R,t} &\sim NID(0, \sigma_R^2) \\
-1 &< \rho < 1
\end{aligned}
\tag{1}
$$

Thus, the series X_t is a sum of a random walk R_t, representing the permanent component, and a mean-reverting series M_t, representing the transient component. The coefficient of mean reversion is ρ.

Poterba and Summers were concerned with the problem of determining whether or not the monthly price series of the S&P 500 index exhibited evidence of mean reversion over scales of 3–7 years, and so they fixed ρ at 0.98. For this value of ρ, the power of the tests that they developed were found to be unacceptably small. Perhaps for this reason, they discarded the model, concluding that, "Even the best possible tests have very low power."

However, with the advent of high frequency trading, available data sets have become much larger. Consequently, it is appropriate to reconsider models that were previously rejected because of their low power at smaller sample sizes.

Moreover, there seems to be a widespread belief that many financial and economic time series contain both mean-reverting and random walk components. Thus, there is a need to be able to model such series adequately, and the model presented in equation (1) is perhaps the simplest possible model having both a mean-reverting and a random walk component. This makes it the natural starting point for a study of such time series. This paper represents an attempt to undertake such a study.

There does not seem to be a great deal of literature on models of time series with both permanent and transient components. As an alternative, one might consider an ARIMA model. Although the ARIMA model does allow for unit roots, one of the advantages of the formulation considered here is that it allows the dynamics of the mean-reverting process M_t to be independent of the random walk process R_t.

The contributions of this paper are as follows. Two different methods are given for identifying the parameters of the model. The first is a fast

Electronic copy available at: http://ssrn.com/abstract=2556957

but imprecise calculation based upon lagged variances, while the second is a more accurate calculation based upon maximum likelihood estimation of an associated Kalman filter. The power of the likelihood ratio test is then examined. Then, a study of the S&P 500 is undertaken. When the model is applied to the monthly returns of the S&P 500 index, no evidence for a mean reverting component is found. However, when it is applied to the daily returns of the individual constituents of the S&P 500, a statistically significant proportion of these series are found to contain both random walk and mean-reverting components.

1 Parameter Estimation via Lagged Variances

To simplify terminology, the term *partially autoregressive*, or just PAR, will be used henceforth to refer to series of the form (1). This nomenclature suggests the perhaps optimistic point of view that were it not for the (hopefully small) random walk component, the series would be manageable. The mean-reverting portion M_t could in principle be taken to be autoregressive of order k, or a more general ARMA(k, l) series, or perhaps even fractionally integrated. Thus, one could speak of a series that is partially autoregressive of order k, or that is partially ARMA(k, l), and so on. For this paper, though, attention will be limited to the simplest case, namely when M_t is autoregressive of order one.

Because X_t contains a random walk, it is not stationary, and consequently its variance is undefined. For this reason, the lagged differences of the series $(1-B)X_t, (1-B^2)X_t, (1-B^3)X_t, ...$ will be investigated. Because $\epsilon_{M,t}$ and $\epsilon_{R,t}$ are independent, there is the relationship

$$\mathrm{Var}[(1-B)X_t] = \mathrm{Var}[(1-B)M_t] + \mathrm{Var}[(1-B)R_t]$$

Since $(1-B)R_t = \epsilon_{R,t}$, it follows that $\mathrm{Var}[(1-B)R_t] = \sigma_R^2$. Calculation of $\mathrm{Var}[(1-B)M_t]$ is slightly more involved.

Since M_t is assumed to be an autoregressive sequence of order 1 with $\rho < 1$, it is stationary, and its variance is easily calculated from the steady state equation as

$$\mathrm{Var}M_t = \frac{\sigma_M^2}{1-\rho^2}$$

See Brockwell and Davis [3] for details. From this, the variance of $(1-B)M_t$ is then found to be

$$\begin{aligned}
\text{Var}[(1-B)M_t] &= \text{Var}[\rho M_{t-1} + \epsilon_{M,t} - M_{t-1}] \\
&= \text{Var}[(\rho-1)M_{t-1} + \epsilon_{M,t}] \\
&= (\rho-1)^2 \frac{\sigma_M^2}{1-\rho^2} + \sigma_M^2 \\
&= \frac{[(\rho-1)^2 + (1-\rho^2)]\sigma_M^2}{1-\rho^2} \\
&= \frac{2}{\rho+1}\sigma_M^2
\end{aligned}$$

Consequently, it follows that

$$\text{Var}[(1-B)X_t] = \frac{2}{\rho+1}\sigma_M^2 + \sigma_R^2$$

At this point it seems relevant to introduce the notion of the *proportion of variance attributable to mean reversion*, R_{MR}^2. This is defined as

$$R_{MR}^2 = \frac{\text{Var}[(1-B)M_t]}{\text{Var}[(1-B)X_t]}$$

This quantity will have a value between zero and one. When it is zero, X_t is a pure random walk with no mean reverting component, and when it is one, X_t is a pure AR(1) process with no random walk component. Thus, R_{MR}^2 can be viewed as a measure of how close X_t is to being a pure random walk. Based upon the foregoing, the proportion of variance attributable to mean reversion is seen to be[2]

$$R_{MR}^2 = \frac{2\sigma_M^2}{2\sigma_M^2 + (1+\rho)\sigma_R^2}$$

Attention is now focused on the higher order lagged variances. It is readily seen that

[2]This is essentially the same as the quantity δ given in Poterba and Summers.

$$\begin{aligned}\mathrm{Var}[(1-B^k)R_t] &= \mathrm{Var}[\epsilon_{R,t}+\epsilon_{R,t-1}+\cdots+\epsilon_{R,t-k+1}] \\ &= k\sigma_R^2\end{aligned} \tag{2}$$

By induction, it is also seen that

$$\begin{aligned}M_t - M_{t-k} = (\rho^k-1)M_{t-k} + \epsilon_{M,t} + \rho\epsilon_{M,t-1} + \rho^2\epsilon_{M,t-2} \\ + \cdots + \rho^{k-1}\epsilon_{M,t-k+1}\end{aligned}$$

and consequently

$$\begin{aligned}\mathrm{Var}[(1-B^k)M_t] &= \mathrm{Var}[(\rho^k-1)M_{t-k}+ \\ &\qquad \epsilon_{M,t}+\rho\epsilon_{M,t-1}+\rho^2\epsilon_{M,t-2}+\cdots+\rho^{k-1}\epsilon_{M,t-k+1}] \\ &= (\rho^k-1)^2\mathrm{Var}[M_{t-k}]+ \\ &\qquad (1+\rho^2+\rho^4+\cdots+\rho^{2(k-1)})\sigma_M^2 \\ &= (\rho^k-1)^2\left(\frac{\sigma_M^2}{1-\rho^2}\right)+\frac{1-\rho^{2k}}{1-\rho^2}\sigma_M^2 \\ &= \frac{(\rho^k-1)^2+(1-\rho^{2k})}{1-\rho^2}\sigma_M^2\end{aligned}$$

Putting this together with equation (2), the following expression for $\mathrm{Var}[(1-B^k)X_t]$ is obtained:

$$\mathrm{Var}[(1-B^k)X_t] = \frac{(\rho^k-1)^2+(1-\rho^{2k})}{1-\rho^2}\sigma_M^2 + k\sigma_R^2 \tag{3}$$

Let the right hand side of this equation be denoted v_k. Making use of the equations for v_1, v_2 and v_3, it is possible to solve for ρ, σ_M^2 and σ_R^2. At this point, a computer algebra system becomes helpful. In any event, further algebra shows that

$$\begin{aligned}\rho &= -\frac{v_1-2v_2+v3}{2v_1-v_2} \\ \sigma_M^2 &= \frac{1}{2}\left(\frac{\rho+1}{\rho-1}\right)(v_2-2v_1) \\ \sigma_R^2 &= \frac{1}{2}\left(v_2-2\sigma_M^2\right)\end{aligned} \tag{4}$$

The condition $2v_1 - v_2 = 0$ is satisfied exactly when $\rho = 1$, and since it is assumed that $\rho < 1$, these equations are well-defined. Consequently, this establishes that each PAR series has a unique parameterization.

When $\rho = 1$, this uniqueness property breaks down. In this case, the parameterization $(\rho, \sigma_M, \sigma_R)$ is equivalent to the parameterization (ρ, τ_M, τ_R) when $\sigma_M^2 + \sigma_R^2 = \tau_M^2 + \tau_R^2$.

The above procedure can be used to directly compute estimates of the parameters. However, for smaller sample sizes, the estimation errors can be substantial. Experience shows that maximum likelihood estimation of the Kalman filter yields better results in practice. This method will be explained next.

2 State-Space Representation

A state-space representation is a mathematical model of a physical system as a set of input, output and state variables related by first-order differential equations. While the state space formulation originates in control theory, it has proven to have widespread applicability in statistics, finance and machine learning, among other areas. The state space representation consists of two equations, a state equation and an observation equation, given as follows:

$$\begin{aligned} Z_t &= F_{t-1} Z_{t-1} + G_{t-1} U_{t-1} + W_{t-1} \\ X_t &= H_t Z_t + V_t \end{aligned}$$

The state of the system is given by Z_t, which may not be directly observable. The state is assumed to follow a linear dynamic and it may be influenced by a control input U_{t-1}. The term W_{t-1} is a noise term, which has covariance matrix Q_t. The observable portion of the system is represented by X_t. It is assumed to have a linear dependence on the hidden state Z_t, given by H_t, and to be influenced by its own noise term V_t, whose covariance matrix is R_t.

Kalman Filter

Despite the generality of this formulation, it is a remarkable fact that there is a procedure for determining the optimal estimate of the hidden state Z_t based upon previous observations of the system and assumed values of the

parameters. This is the Kalman filter algorithm. For a starting point into the extensive Kalman filter literature, see Brockwell and Davis [3] or Simon [8].

The Kalman filter algorithm is given by a simple set of equations for updating the estimate $\hat{Z}_t$ of Z_t, based upon the most recent observation X_t. It can be thought of as proceeding in two steps. In the first step, an *a priori* estimate $\hat{Z}_t^-$ is constructed based upon the information that was available at time $t-1$. Then, when the new observation X_t is recorded, an *a posteriori* estimate $\hat{Z}_t^+$ is constructed based upon the newly available information. The updating procedure also maintains an estimate P_t of the covariance of the hidden state Z_t. The Kalman filter equations are given as follows:

$$
\begin{aligned}
P_t^- &= F_{t-1} P_{t-1}^+ F_{t-1}^T + Q_{t-1} \\
K_t &= P_t^- H_t^T (H_t P_t^- H_t^T + R_t)^{-1} \\
\hat{Z}_t^- &= F_{t-1} \hat{Z}_{t-1}^+ + G_{t-1} U_{t-1} \\
\hat{Z}_t^+ &= \hat{Z}_t^- + K_t (X_t - H_t \hat{Z}_t^-) \\
P_t^+ &= (I - K_t H_t) P_t^- (I - K_t H_t)^T + K_t R_t K_t^T
\end{aligned}
$$

A crucial component of the Kalman filter equations is the Kalman gain matrix K_t. This matrix determines the influence that a new observation has upon the estimate of the hidden state Z_t.

Estimation of the States of a PAR Sequence

The partially autoregressive model has a particularly simple state space representation, given as follows:

$$
\begin{aligned}
Z_t &= \begin{bmatrix} M_t \\ R_t \end{bmatrix} = \begin{pmatrix} \rho & 0 \\ 0 & 1 \end{pmatrix} \begin{bmatrix} M_{t-1} \\ R_{t-1} \end{bmatrix} + \begin{pmatrix} \epsilon_{M,t} \\ \epsilon_{R,t} \end{pmatrix} \\
X_t &= \begin{bmatrix} 1 & 1 \end{bmatrix} Z_t \\
Q &= \begin{pmatrix} \sigma_M^2 & 0 \\ 0 & \sigma_R^2 \end{pmatrix}
\end{aligned}
$$

In particular, the partially autoregressive model assumes a steady state system, and consequently the matrices F_t, H_t, Q_t, P_t and K_t are constant.

Thus, the subscripts will be dropped when subsequently referring to these matrices. Also the measurement error V_t is taken to be zero. After making these simplifications, the Kalman filter equations reduce to

$$P^- = FP^+F^T + Q \tag{5}$$
$$K = P^-H^T(HP^-H^T)^{-1} \tag{6}$$
$$\hat{Z}_t^- = F\hat{Z}_{t-1}^+ \tag{7}$$
$$\hat{Z}_t^+ = \hat{Z}_t^- + K(X_t - H\hat{Z}_t^-) \tag{8}$$
$$P^+ = (I - KH)P^-(I - KH)^T \tag{9}$$

The covariance matrix P^- can be calculated as follows. Begin by substituting equations (9) and (6) into equation (5). After simplification, the following is obtained:

$$P^- = FP^-F^T - FP^-H^T(HP^-H^T)^{-1}HP^-F^T + Q \tag{10}$$

Given specific values for ρ, σ_M and σ_R, these can then be substituted into the above equation (10) and a steady state solution can be found. This can then be used to obtain a steady state solution for the Kalman gain matrix K, by substituting the result into equation (6). After performing these steps and simplifying, the steady state Kalman gain is found to be

$$K = \begin{pmatrix} \frac{2\sigma_M^2}{\sigma_R\left(\sqrt{(\rho+1)^2\sigma_R^2+4\sigma_M^2}+\rho\sigma_R+\sigma_R\right)+2\sigma_M^2} \\ \frac{2\sigma_R}{\sqrt{(\rho+1)^2\sigma_R^2+4\sigma_M^2}-\rho\sigma_R+\sigma_R} \end{pmatrix} \tag{11}$$

A procedure for calculating estimates of the mean-reverting series M_t and random walk series R_t can then be given as follows. As input, the algorithm is given the observed price series X_t and the parameters ρ, σ_M and σ_R. The algorithm first calculates the Kalman gain K using ρ, σ_M and σ_R. The initial state of the system is taken to be $M_1 = 0$ and $R_1 = X_1$. The algorithm then iterates through each of the remaining observations in the input sequence X_t. For each observation, the Kalman update equations (7) and (8) are used to produce an estimate of the hidden state Z_t at that time. Code for the algorithm in the R language is given in listing 1.

Listing 1: Procedure for Estimating States of a PAR Sequence

```
kalman_estimate <- function(X, rho, sigmaM, sigmaR) {
    # Calculates estimates of the mean-reverting
    # series M[t] and random walk series R[t].
    #
    # Input values:
    #    X:       A series of observations of a
    #             partially autoregressive sequence
    #    rho:     The coefficient of mean reversion
    #    sigmaM: Standard deviation of the innovations
    #             of the mean-reverting component.
    #    sigmaR: Standard deviation of the innovations
    #             of the random walk component.
    #
    # Output values:
    #    M:    Estimate of the mean-reverting series
    #    R:    Estimate of the random walk weries

    # See equation (11) for calculation of K
    K <- kalman_gain(rho, sigmaM, sigmaR)

    M <- 0
    R <- X[1]
    for (i in 2:length(X)) {
        # Predicted value of X[i]
        xhat <- rho * M[i-1] + R[i-1]
        # Prediction error
        e <- X[i] - xhat
        M[i] <- rho * M[i-1] + e * K[1]
        R[i] <- R[i-1] + e * K[2]
    }

    return (list(M, R))
}
```

Maximum Likelihood Estimation

The preceding section shows that given a sequence X_t and parameters ρ, σ_M and σ_R, it is possible to form estimates of the hidden states M_t and R_t. This is fine if ρ, σ_M and σ_R are known, but what should be done if they are not? One approach is to use the estimates of these values produced by the lagged variance method given earlier. In this section, another method for estimating ρ, σ_M and σ_R is described.

Let $X_1, X_2, \ldots, X_t$ be the observed sequence of a partially AR(1) process, and $\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_t$ be the inferred states of the generating process. The probability density function is given as

$$\begin{aligned} p(X_1, X_2, \ldots, X_t | \rho, \sigma_M, \sigma_R) &= p(X_1 | \rho, \sigma_M, \sigma_R) \prod_{k=2}^{t} p(X_k | X_{k-1}, \rho, \sigma_M, \sigma_R) \\ &= p(X_1 | \rho, \sigma_M, \sigma_R) \prod_{k=2}^{t} p(e_k | X_{k-1}, \rho, \sigma_M, \sigma_R) \end{aligned}$$

where

$$\begin{aligned} e_k &= X_k - E(X_k | X_{k-1}, \rho, \sigma_M, \sigma_R) \\ &= X_k - HF\hat{Z}_{k-1}. \end{aligned}$$

Moreover, $e_k \sim N(0, \sigma_M^2 + \sigma_R^2)$. Consequently, the log likelihood function can be written as

$$\ln[L(\rho, \sigma_M, \sigma_R)] = -\frac{t-1}{2} \ln[2\pi(\sigma_M^2 + \sigma_R^2)] - \frac{1}{2(\sigma_M^2 + \sigma_R^2)} \sum_{k=2}^{t} e_k^2 \qquad (12)$$

Thus, the algorithm for maximum likelihood estimation is as follows. Initial estimates $\rho_0, \sigma_{M,0}$ and $\sigma_{R,0}$ are first obtained by using the equations (4). Starting from these initial estimates, a quasi-Newton method is then used to search for values of ρ, σ_M and σ_R that maximize the likelihood function (12). The errors e_k are obtained from line 27 of the Kalman estimation procedure.

Robust Parameter Estimation

It is well known that the daily returns of equities have a heavy-tailed distribution and that the normal distribution provides a poor fit to them. There have been extensive efforts to identify alternative distributions that would be more suitable for modeling these returns. One choice that seems to be effective and for which there is some theoretical justification is Student's t distribution [4].

A possible approach is to assume that the error terms e_k follow a scaled Student's t distribution. That is to say, it is assumed that $e_k = (\sigma_R + \sigma_M)\tau_k$, where $\tau_k \sim t(\nu)$, and $t(\nu)$ is a Student's t distribution with ν degrees of freedom. If this assumption is made, then the log likelihood function becomes

$$\ln[L(\rho, \sigma_M, \sigma_R)] = (t-1)\log\left[\Gamma\left(\frac{\nu+1}{2}\right)\right] - (t-1)\log\left[\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)\right] - (t-1)(\sigma_M + \sigma_R) - \frac{\nu+1}{2}\left[\sum_{k=2}^{t}\log\left(1 + \frac{e_k^2}{\nu(\sigma_M+\sigma_R)^2}\right)\right]$$

As will be seen, when this model is applied to the daily price movements of individual equities with ν fixed at $\nu = 5$, the fits obtained are superior to those obtained under the assumption of normality.

Nonetheless, this approach does have an imperfection. The assumption that the error terms follow a Student's t distribution says little about the distribution of the innovations of the component sequences M_t and R_t. In fact, in general the innovations of M_t and R_t *cannot* be t-distributed, as the sum of two t distributions is not usually a t distribution.

This motivates a search for other candidate heavy-tailed distributions. A popular choice in the finance community in recent years has been the normal inverse gaussian distribution [1]. However, the available implementation was much slower than that of the Student's t distribution, so the choice was made to live with this imperfection.

3 Hypothesis Testing

When assessing appropriateness of fit, there seem to be two natural hypotheses against which the assumption of a partially autoregressive sequence should be compared. One hypothesis is that $\{X_t\}$ is simply a pure random

	RW Null			AR(1) Null		
RW Samples	p=0.01	0.05	0.10	p=0.01	0.05	0.10
n = 50	-4.65	-2.87	-2.16	-2.95	-1.61	-1.03
100	-4.65	-2.96	-2.20	-3.30	-1.81	-1.22
250	-4.58	-2.99	-2.24	-3.59	-2.05	-1.39
500	-4.72	-3.15	-2.39	-3.85	-2.26	-1.56
1000	-4.78	-3.09	-2.35	-3.96	-2.30	-1.55
2500	-4.80	-3.09	-2.36	-3.95	-2.36	-1.63
AR(1) Samples						
$\rho = 0.90$	p=0.01	0.05	0.10	p=0.01	0.05	0.10
n = 50	-6.53	-4.68	-3.86	-2.58	-1.23	-0.67
100	-8.19	-6.18	-5.32	-2.44	-0.99	-0.43
250	-13.04	-10.88	-9.78	-1.91	-0.55	-0.07
500	-20.88	-18.45	-17.11	-1.63	-0.30	-0.00
1000	-36.16	-32.68	-30.98	-1.42	-0.13	-0.00
2500	-78.99	-74.66	-72.30	-1.29	-0.00	-0.00
$Q[\chi^2]$	-4.61	-3.00	-2.30	-3.32	-1.92	-1.35

Table 1: Quantiles of Likelihood Ratio Function

walk. The other hypothesis is that $\{X_t\}$ is a pure autoregressive sequence of order one. Thus, the null hypothesis is a union of the conditions:

$\mathcal{H}_0^R$: $\{X_t\}$ is a pure random walk

$\mathcal{H}_0^M$: $\{X_t\}$ is a pure autoregressive sequence of order one

To accept the alternative hypothesis that $\{X_t\}$ is PAR, both parts of this null hypothesis must be rejected.

The likelihood ratio test can be used to test either of these nulls. The random walk hypothesis can be tested by fitting $\{X_t\}$ to the model (1) where σ_M has been fixed at $\sigma_M = 0$. Similarly, the pure autoregressive hypothesis can be tested by fitting $\{X_t\}$ to this model when σ_R has been fixed at $\sigma_R = 0$.

Simulations were performed to obtain critical values of the likelihood ratio function for each of these cases. For the random walk null, 10,000 random walks were generated of length 50, 100, 250, etc. For each such

sequence, a maximum likelihood fit to the PAR model was found using a steady state Kalman filter. Also, maximum likelihood fits of the steady state Kalman filters were obtained under the assumptions that $\sigma_M = 0$ (the random walk null hypothesis) and $\sigma_R = 0$ (the AR(1) null hypothesis). The log likelihood ratios were then computed, and the quantiles were found. An identical procedure was also performed for the AR(1) null.

The results are given in Table 1. This table is divided into six panels, which are presented in three rows and two columns. The lefthand column contains the critical values for the test of the null hypothesis that the sequence is a pure random walk. The righthand column contains the critical values for the test of the null hypothesis that the sequence is pure autoregressive of order one.

The top two panels tabulate the results obtained when performing maximum likelihood fits to random walk sequences. (The length of each sequence is given by n.) For each random walk, maximum likelihood fits were performed and the log likelihood ratio score was then computed. Thus, the upper left panel records the critical values for the likelihood ratio test when applied to the null hypothesis that the series is a pure random walk.

The middle two panels tabulate the results obtained when performing maximum likelihood fits to random AR(1) sequences, where the coefficient of mean reversion was fixed at $\rho = 0.90$. Thus, the middle right panel records the critical values for the likelihood ratio test when applied to the null hypothesis that the series is pure AR(1) with $\rho = 0.90$.

Various different values of ρ were investigated. As ρ approaches 1, the values in the middle panel converge to those in the upper panel. The value $\rho = 0.90$ seemed to represent a reasonable compromise.

The bottom two panels record the theoretical values of the likelihood ratio test. Let $\Lambda_{\mathcal{H}}$ be the log likelihood ratio function for testing the hypothesis $\mathcal{H}$. If the assumptions of Wilks' theorem [10] are satisfied, then $-\Lambda_{\mathcal{H}}/2$ converges to a χ^2 distribution with degrees of freedom equal to the difference in dimensionality between the null hypothesis and the alternative hypothesis. The quantiles predicted by Wilks' theorem are a reasonable match for the simulated values in the upper left panel for testing the random walk null hypothesis. However, the quantiles predicted for testing the pure AR(1) null hypothesis are much higher than the values that were actually observed.

The acceptance procedure for determining whether or not a sequence is PAR is given as algorithm 1. Let $\Lambda_R(X_t)$ by the value of the log likelihood ratio function for sequence X_t when the null hypothesis is a random walk,

and let $\Lambda_M(X_t)$ be the value when the null hypothesis is an AR(1) sequence. The probability that $\{X_t\}$ will be accepted by this algorithm can be written as

$$\begin{aligned} p(\text{PAR}) &= p(\neg\text{RW})p(\neg\text{AR1}|\neg\text{RW}) \\ &= p(\Lambda_R(X_t) < C_R(\alpha)) \\ &\quad p(\Lambda_M(X_t) < C_M(\alpha)|\Lambda_R(X_t) < C_R(\alpha)) \end{aligned}$$

Data: A time series $\{X_t\}$ and a confidence level α
Result: RW if the pure random walk hypothesis is accepted;
AR1 if the pure autoregressive hypothesis is accepted;
PAR otherwise

```
Compute the likelihood ratio score Λ_R(X_t)
if Λ_R(X_t) > C_R(α) then
    Output "RW"
end
Compute the likelihood ratio score Λ_M(X_t)
if Λ_M(X_t) > C_M(α) then
    Output "AR1"
end
Output "PAR"
```

Algorithm 1: Procedure for Testing Whether a Sequence is PAR

This procedure depends crucially on the selection of proper values for $C_R(\alpha)$ and $C_M(\alpha)$. If $\{X_t\}$ is either RW or AR(1), then the desired outcome is that $p(\text{PAR}) = \alpha$. A search procedure was implemented and was found to produce satisfactory values of $C_M(\alpha)$ and $C_R(\alpha)$ for $\alpha \leq 0.20$.

The power of a statistical test is the probability that it will reject the null hypothesis, given that the null hypothesis is false. Power can be viewed as a function of the distributional parameters. To calculate the power of the PAR test, random PAR sequences were generated for various values of ρ and σ_M, holding σ_R constant at one. For each randomly generated sequence, the PAR test was applied with a critical value of $p = 0.05$.

A graph of the power of the likelihood ratio test is given in Figure 1. The upper panel displays the power curves when $\sigma_M = 0.5$ and $\sigma_R = 1.0$.

When $\rho = 0.5$, the proportion of variance attributable to mean reversion is $R^2_{MR} = 0.25$, and when $\rho = 0.9$, $R^2_{MR} \approx 0.21$. In this range, the power of the PAR test is not very high, even for sample sizes of 1,000. The middle panel displays the power curves when $\sigma_M = \sigma_R = 1.0$. For $\rho = 0.5$, the power is acceptable for $n \geq 250$. For this value of ρ, $R^2_{MR} \approx 0.59$. The lower panel displays the power curves when $\sigma_M = 2$. For $\rho = 0.5$, $R^2_{MR} \approx 0.84$. The power when $\rho = 0.9$ still suffers, although it starts to enter the acceptable range for $n \geq 500$. As these curves demonstrate, the power of the test is highly dependent upon ρ. As ρ approaches 1, it becomes increasingly difficult to distinguish a PAR sequence from a pure random walk.

An alternative presentation of this data is given in Figure 2. In this graph, each curve is obtained by holding R^2_{MR} constant and graphing the power as a function of the coefficient of mean reversion ρ. The power is seen to increase as R^2_{MR} increases, and it decreases as ρ gets larger. If $\rho \geq 0.8$ and $R^2_{MR} \leq 0.5$, the power is less than 0.5.

One way of interpreting these results is to consider them in terms of the half-life of mean reversion. For an AR(1) series having value M_t at time t, the expected value of the series k periods in the future is $E[M_{t+k}|M_t] = \rho^k M_t$. The half-life of mean reversion is the number of periods λ such that the expected value of the series will be no more than half of the value at time t. It is easily worked out that $\lambda = \log(0.5)/\log(\rho)$. When $\rho = 0.5$, the half-life is one period. When $\rho = 0.707$, it is two periods. And when $\rho = 0.84$, it is about four periods in the future. Thus, the PAR test is most powerful when the process being modeled has a relatively short half-life of mean reversion.

One area of concern is that the values of the log likelihood ratio function Λ_M are so close to zero when X_t is AR(1). It was thought that this might have diminished the power of the PAR test. Therefore, an alternative to the likelihood ratio test for the AR(1) null was sought. Kwiatkowski *et. al.* [5] give a test of the null hypothesis that a series is stationary (the KPSS test). The KPSS statistic was implemented as an alternative to Λ_M. However, the power of the resulting test was not found to be as great as the power obtained when using Λ_M.

4 Analysis of S&P 500

Robert Shiller [7] makes available on his web site the monthly price series of the S&P 500 going back to 1871. The PAR model was fitted to both the price

series and the logged price series, and likelihood ratio tests were performed. The results are given in Table 2. In addition, fits were performed to the price/earnings series and to the CAPE ratio, which is a price/earnings ratio based upon inflation-adjusted earnings over the preceding ten years. Because the CAPE ratio requires ten years of historical earnings, the series for the CAPE ratio begins in January 1881.

Series	Dates	n	Λ_R	Λ_M	p	Assessment
Price	1/1871 - 12/2013	1716	0.00	0.00	1.000	RW
log(Price)	1/1871 - 12/2013	1716	0.00	0.00	0.859	RW
P/E	1/1871 - 12/2013	1716	-8.19	-0.39	0.039	PAR
CAPE	1/1881 - 12/2013	1596	-1.87	0.00	0.728	RW

Table 2: Tests of PAR fits to Shiller's Monthly S&P 500 Data

As can be seen from this table, the random walk hypothesis cannot be rejected for the price series nor for the log of the price series. For the price/earnings series, both the random walk hypothesis and the AR(1) hypothesis can be rejected at the 5% confidence level, so this series is classified as PAR. Interestingly, the series for the CAPE ratio could not be distinguished from a random walk.

Further details on the maximum likelihood fit of a PAR model to the price/earnings series are given in Table 3. Standard errors are given in parenthesis below the estimated parameters. The coefficient of mean reversion ρ is found to be 0.9785, which corresponds to a half-life of mean reversion of 32 months. The proportion of variance attributable to mean reversion is 0.9967, indicating that almost all of the variation in the series is due to the AR(1) component. This is further confirmed by the fact that the value of σ_M is within one standard error of zero. Thus, the price/earning series is found to be a mean-reverting process with a half-life of 32 months, combined with a small component of random drift.

An investigation was then conducted of the individual components of the S&P 500 over the period 2002–2013. A list of the S&P 500 components was downloaded from the Standard and Poor's web site on August 13, 2013. On September 18, 2014, the adjusted closing prices for each of these components was downloaded from Yahoo! for the twelve year period January 1, 2002

Series	ρ	σ_M	σ_R	R^2_{MR}	Log Likelihood
Price/Earnings	0.9785	1.73	0.10	0.9967	-3381.58
	(0.0057)	(0.03)	(0.12)		

Table 3: Fit of PAR Model to Monthly Price/Earnings of S&P 500

through December 31, 2013. The data was then divided into six two-year periods. Within each two-year period, a security was only considered for inclusion if a complete price series was available for the security in that period.

For each two year period, a PAR fit was calculated to the logged price series of each eligible security. A summary of the fits that were obtained is given in Table 4. The column labels are as follows. N is the number of securities for which fits were computed. p_{RW} is the proportion of fits for which the random walk hypothesis was accepted. p_{AR1} is the proportion of fits for which the random walk hypothesis was rejected and the AR(1) hypothesis was accepted. p_{PAR} is the proportion of fits for which both the random walk and AR(1) hypotheses were rejected. $\bar{\rho}$ is the average of the fitted values of ρ. $\bar{R}^2_{MR}$ is the average estimate of the proportion of variance attributable to mean reversion, and LL is the average log likelihood score. For $\bar{\rho}, \bar{R}^2_{MR}$ and LL, these averages are computed only for those fits that were identified as PAR.

Years	N	p_{RW}	p_{AR1}	p_{PAR}	$\bar{\rho}$	$\bar{R}^2_{MR}$	LL
2002-2003	427	0.74	0.01	0.25	-0.11	0.32	1243.90
2004-2005	439	0.92	0.01	0.07	0.16	0.47	1459.78
2006-2007	453	0.77	0.01	0.23	0.33	0.42	1428.45
2008-2009	460	0.55	0.01	0.44	0.28	0.44	1041.23
2010-2011	467	0.64	0.02	0.34	-0.35	0.25	1321.22
2012-2013	476	0.96	0.00	0.04	0.37	0.48	1449.09
Total	2722	0.76	0.01	0.23	0.06	0.37	1245.28

Table 4: Fit of PAR Model to S&P 500 Constituents

As can be seen from the table, only 76% of series are found to be random walks, while 23% are found to be PAR. Under the null hypothesis that all

series are random walks, one would expect no more than 5% of series to be identified as PAR. Thus, a much larger proportion of series are found to be PAR than would be expected. Nonetheless, it is a concern that only 4% of sequences were found to be PAR in the most recent period (2012–2013).

An important question is whether or not the PAR property is persistent. In other words, if a price series is found to be PAR in one period, how likely is it that it will be persistent in the subsequent period? If the PAR model is to be useful as predictor of future behavior, there should be evidence of persistence.

Let PAR(k) be the set of those price series that are found to be partially autoregressive in period k. If the PAR property is persistent, then it is expected that

$$Pr(x \in \text{PAR}(k+1) | x \in \text{PAR}(k)) > Pr(x \in PAR(k+1)).$$

Alternatively, if being partially autoregressive in period k is independent of being partially autoregressive in period $k+1$, then it is to be expected that these two quantities will be unrelated. A χ^2-test can be used to test the null hypothesis that being partially autoregressive in period k is independent of being partially autoregressive in period $k+1$.

Table 5 reports the results that were obtained. The column labeled N reports the total number of sequences considered in that period. The columns labeled A_t and A_{t-1} report the number of sequences that were identified as PAR in the current period and the previous period. The column labeled $A_{t|t-1}$ reports the number of sequences that were found to be PAR both periods t and $t-1$. The column labeled p_t reports the proportion of sequences in period t that were found to be PAR, while the colum labeled $p_{t|t-1}$ reports the proportion of sequences from period $t-1$ which were found to be PAR in both periods t and $t-1$. If asterisks are shown, this represents the significance level of the χ^2-test. Three asterisks are given for results that are significant at the 1% level, two asterisks for results that are significant at the 5% level, and one asterisk for results that are significant at the 10% level.

As can be seen from the table, the degree of persistence was significant at the 10% level in four of the five periods, and it was significant at the 1% level in two of the five periods. When the data is aggregated, the average frequency with which sequences are found to be PAR is 22%, whereas sequences that are PAR in the preceding period have a 32% probability of being PAR in the succeeding period. This is significant at the 1% level.

Period	N	A_t	A_{t-1}	$A_t \vert A_{t-1}$	p_t	$p_{t\vert t-1}$
2004-2005	439	32	106	17	0.073	0.160***
2006-2007	453	102	32	12	0.225	0.375*
2008-2009	460	202	102	58	0.439	0.569**
2010-2011	467	160	202	97	0.343	0.480***
2012-2013	476	18	160	9	0.038	0.056
Total	2295	514	602	193	0.224	0.321***

Table 5: Persistence of PAR Fits to S&P 500 Constituents

4.1 Results of Robust Estimation

In this section, results are presented for estimation of PAR models using robust estimation. Robust PAR models were fit to each of the sequences in the study data set. The statistics on the fits that were obtained are given in Table 6. As can be seen, the average log likelihood score for a fit using robust estimation was 1268.19. This is to be compared to an average log likelihood score of 1245.28 when using standard estimation. Thus, there is quite a significant difference in average likelihood scores, suggesting that the robust estimation technique in general gives a much better fit. Nonetheless, the proportion of sequences found to be PAR is nearly the same in both models. Also, the average values found for ρ and R^2_{MR} are also quite similar.

Years	N	p_{RW}	p_{AR1}	p_{PAR}	$\bar{\rho}$	$\bar{R}^2_{MR}$	LL
2002-2003	427	0.68	0.01	0.30	-0.12	0.29	1262.47
2004-2005	439	0.90	0.00	0.10	0.21	0.40	1481.44
2006-2007	453	0.80	0.01	0.19	0.31	0.38	1452.21
2008-2009	460	0.60	0.03	0.37	0.15	0.38	1066.18
2010-2011	467	0.75	0.02	0.23	-0.15	0.29	1307.24
2012-2013	476	0.93	0.01	0.06	-0.13	0.24	1474.28
Total	2722	0.78	0.01	0.21	0.05	0.34	1268.19

Table 6: Fit of Robust PAR Model to S&P 500 Constituents

Measures of persistence were also computed. Results are given in Table 7. When robust estimation was used, 18.9% of sequences were found to be PAR, and the rate of persistence was 26.4%. While this is still highly significant,

the rate of persistence is lower than the rate that was found using standard estimation.

Period	N	A_t	A_{t-1}	$A_t \vert A_{t-1}$	p_t	$p_{t\vert t-1}$
2004-2005	439	42	130	22	0.096	0.169**
2006-2007	453	87	42	9	0.192	0.214
2008-2009	460	170	87	43	0.370	0.494**
2010-2011	467	106	170	62	0.227	0.365***
2012-2013	476	29	106	5	0.061	0.047
Total	2295	434	535	141	0.189	0.264***

Table 7: Persistence of Robust PAR Fits to S&P 500 Constituents

4.2 Analysis of Mean Reverting Components

In this section, further analysis is given of the mean reverting components of the (non-robust) PAR models fitted to the S&P 500. For those series that are found to be PAR, the average proportion of variance attributable to mean reversion was 37%. The average coefficient of mean reversion was found to be only 0.06. For a series with $\rho = 0.06$ and $R^2_{MR} = 0.37$, σ_M would be about $0.56\sigma_R$. The Kalman gain matrix would be about $[0.19\ 0.81]^T$. Thus, about 19% of the price fluctuation in a given day would be attributable to the mean-reverting component.

A graph of the distribution of ρ is given in the lefthand panel of figure 3. The distribution appears to be bimodal, with a relatively sharp peak at $\rho = -0.75$, and a second relatively flat peak at $\rho = 0.4$ with a long right tail. Of the 620 sequences that were identified as PAR, there were 382 (62%) for which $\rho > 0$ and 238 (38%) for which $\rho < 0$. It was surprising to find so many sequences for which $\rho < 0$, and no explanation is known for this finding. When $\rho < 0$, the mean reverting component will oscillate rapidly around zero.

Summary statistics for the cases $\rho > 0$ and $\rho < 0$ are presented in the following table (Table 8). The column K_{MR} is the average Kalman gain of the mean reverting component, and $p_{y|y-1}$ is the probability of persistence. As can be seen from this table, the proportion of variance attributable to mean reversion is higher when $\rho > 0$, and the probability of persistence is also higher.

	N	$\bar{\rho}$	R^2_{MR}	K_{MR}	$p_{y\|y-1}$
$\rho > 0$	382	0.458	0.489	0.275	0.393
$\rho < 0$	238	-0.590	0.173	0.080	0.181
All	620	0.056	0.368	0.192	0.311

Table 8: Summary Statistics for S&P 500 Sequences Identified as PAR

To gain a better understanding of the potential economic value of the mean-reverting component, an attempt was made to obtain a rough estimate of the value of the mean-reverting component expressed in dollars. For a given security, let $\bar{P}$ be the mean value of its price, and let ρ and σ_M be the parameters of the PAR fit of the mean-reverting component of the logged price series. Then, the following value was computed as an estimate of the standard deviation of the mean-reverting component of the price:

$$\tau = \bar{P}\sqrt{\frac{\sigma_M^2}{1-\rho^2}}$$

The quantity involving the square root is the standard deviation of the mean-reverting component of the logged price series. This could possibly be viewed as a measure of the daily volatility of the mean-reverting component. This is multiplied by $\bar{P}$ to obtain an estimate in units of dollars.

The set of PAR sequences were divided into two groups, those for which $\rho < 0$ and those for which $\rho > 0$. For each sequence, the associated value of τ was computed. The density of the distribution of values of τ was then plotted. The density curves are shown in figure 4. It is immediately apparent that the two density curves are quite distinct. For $\rho < 0$, the density curve has a high peak at about $\tau = \$0.10$. For $\rho > 0$, the density curve has a peak at about $\tau = \$0.30$ and a long and heavy right tail.

4.3 Effects of Bid-Ask Bounce

In the U.S. equity markets, as well as in many other markets, an order book of pending orders is maintained. A potential buyer of stock may submit a buy order along with the highest price that the buyer is willing to pay (the bid price), or the buyer may state that he is willing to buy at the currently prevailing price (a market order). Similarly, a potential seller of stock may

submit an order to the market specifying the minimum price she is willing to receive (the ask price), or she may state that she is willing to sell at the currently prevailing price (a market order). If the current highest bid price is at least as high as the current lowest ask price, then a transaction occurs. Otherwise, the order is placed in the order book and may be filled when a subsequent order arrives. In quiescence, there is always a gap between the highest bid price and the lowest ask price. This is called the bid-ask spread.

A large bid-ask spread can have a noticeable effect on the observed distribution of prices. If the final order of the day (or measurement period) is a market order to sell, then the closing price will reflect the highest bid price at the time. On the other hand, if the final order of the day is a market order to buy, then it will reflect the lowest ask price at the time. When observing a sequence of trades spaced closely in time, an oscillation in price will be seen, reflecting the bid-ask spread. This phenomenon is known as bid-ask bounce.

If the closing price of a security is X, there is no guarantee that a buyer would have been able to purchase the security at price X. This is because the final order of the day may have been a market order to sell, which would have reflected the bid price at that time. A potential buyer would have had to pay X plus the bid-ask spread. For this reason, trading models that are based upon closing prices can potentially lead to the prediction of illusory profits. Therefore, it is important to understand how the partially autoregressive model behaves in the face of bid-ask bounce.

One possible model for this phenomenon is to model the closing price X_t as the sum of the true price R_t and a binomially distributed error reflecting the bid-ask bounce. Thus, the following model is considered:

$$\begin{aligned} X_t &= R_t + (\gamma_t - \frac{1}{2})s \\ R_t &= R_{t-1}e^{\epsilon_t} \end{aligned} \tag{13}$$

In this model, γ_t is a binomially distributed random variable having value 0 or 1, s is the bid-ask spread, and the log returns ϵ_t of the true price process R_t are assumed to be normally distributed.

To estimate the effects of bid-ask bounce, a number of random sequences were generated according to the above dynamic, and fits were performed of the PAR model to the logged sequences. Statistics were then collected on the frequency with which the sequence was identified as PAR and on the estimated values of the parameters.

Figure 5 displays the frequency with which sequences were identified as PAR for one such set of simulations. In these simulations, the starting price X_0 was chosen as 100, and the volatility was chosen as 1%, e.g., $\epsilon_t \sim N(0, 0.01^2)$. The bid-ask spread s was allowed to vary. For each value of s, a number of sequences of length 500 were generated and fit to the PAR model. The proportion of sequences identified as PAR is graphed.

As can be seen, when the bid-ask spread is small, there is a low probability that a sequence of the form (13) will be identified as PAR. However, as the bid-ask spread grows, the probability of misidentification increases. When the bid-ask spread is equal to the daily volatility, nearly 100% of sequences are identified as PAR. It is clear from this that the presence of bid-ask bounce can be a significant issue when applying the PAR model.

In a second study, the distribution of the fitted value of ρ was examined. In this study, 1,000 random sequences of the form (13) were generated from the parameters $X_0 = 10.00, s = 0.10$ and $\epsilon_t \sim N(0, 0.01^2)$. Over 90% of such sequences were identified as PAR. For those that were identified as PAR, the distribution of ρ was plotted. This is given in Figure 6.

As can be seen from the plot, the fitted values of ρ have a peak at zero. Of the 1,000 sequences that were generated, there were 943 cases in which the fitted value of ρ was within two standard errors of zero. By contrast, for the S&P 500 data, 276 of the 632 sequences (43.7%) identified as PAR had a fitted value of ρ that was within two standard errors of zero. Moreover, the density plot of ρ for the S&P 500 data does not contain a peak at zero. Thus, while bid-ask bounce may explain a substantial number of the PAR fits, it does not seem to fully explain the number of fits obtained to the PAR model.

4.4 Effects of Volatility Clustering

It has been known for some time that the prices of U.S. equities, as well as the prices of many other securities, experience transient periods of high volatility. There is a large literature concerned with modeling and explaining this behavior. A popular model of this phenomenon is the generalized autoregressive conditional heteroscedasticity (GARCH) model of Bollerslev [2]. A price series following the GARCH(1,1) model can be given by the following specification:

$$
\begin{aligned}
\log X_t &= \log X_{t-1} + \epsilon_t \\
\epsilon_t &= \sigma_t z_t \\
\sigma_t^2 &= \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\
z_t &\sim N(0,1) \\
\alpha_0 &> 0, \alpha_1 \geq 0, \beta \geq 0
\end{aligned} \tag{14}
$$

To test whether or not a time series potentially contains GARCH effects, the Ljung-Box portmanteau test can be used on the squared return series. When applied to the data set of S&P 500 stocks, 1257 of 2702 (46.5%) series were found to exhibit GARCH effects. When applied to just those series that were identified as PAR, 452 out of 632 (71.5%) were found to exhibit GARCH effects.

GARCH models were then fitted to all of the PAR series that were identified as potentially having GARCH effects. The median values of the fitted parameters were $\alpha_0 = 9.42 \times 10^{-6}, \alpha_1 = 0.0956$, and $\beta = 0.8828$. Ten thousand random GARCH sequences were then generated having these parameters, and the PAR model was fitted to each such sequence. The number of sequences that were identified as PAR was 848 (8.48%). The expected number of sequences identified as PAR would have been 500 (5%). Thus, the presence of GARCH effects increases the probability that a sequence will be falsely identified as PAR.

For those randomly generated GARCH sequences that were identified as PAR, the distribution of fitted values of ρ was plotted. This plot is given in Figure 7. As can be seen, the density has two peaks, one at about $\rho = -0.9$ and the other at about $\rho = 0.9$. The left peak somewhat resembles the left peak of the density of ρ for PAR fits to the S&P 500, but the right peak cannot be found on the S&P 500 density plot.

To obtain a better understanding of the frequency with which GARCH effects may result in misidentification as PAR, a number of random GARCH sequences were generated for various values of α_1, and the frequency of misidentification as PAR was plotted. For each value of α_1, one thousand GARCH sequences were generated using the parameters $\alpha_0 = 1 \times 10^{-5}$ and $\beta = 0.95 - \alpha_1$. The graph that was obtained is given if Figure 8. As can be seen, when α_1 is close to zero, there is no noticeable effect. The frequency of misidentification increases as α_1 increases, reaching a peak of about 0.34.

It is an interesting question to ask whether or not bid-ask bounce and GARCH effects fully explain the number of PAR fits found in the S&P 500 data. To give a definitive answer to this question would probably require a detailed analysis of actual bid-ask spreads and a more precise analysis of the impact of GARCH effcts. Having said that, the following preliminary assessment can be given. The S&P 500 data set consists of 2,862 series in total, of which 632 were identified as PAR. Of those identified as PAR, it was found that 276 had a fitted value of ρ within two standard errors of zero. The sequences for which bid-ask bounce led to misidentification as PAR are presumably a subset of this group. This leaves 2,862 - 276 = 2,586 sequences for which bid-ask bounce is not presumed to have a large effect. It was found that for the median values of the fitted GARCH parameters, the rate of false positives was 8.48%. Thus, of the remaining 2,586 sequences, it may be inferred that approximately 219 additional false positives may have occurred. Thus, the total number of sequences accounted for through bid-ask bounce and GARCH effects is 495. This leaves 137 sequences unaccounted for. Thus, the data tentatively suggests that there is another effect at work that has not yet been identified.

5 Conclusion

This paper has provided a detailed look at a simple time series model that includes both permanent and transient components. Two different techniques for estimation of model parameters have been given. The likelihood ratio test has been shown to provide reasonable power over a range of parameter values.

When applied to the monthly price/earnings series of the S&P 500, strong evidence for mean reversion is found. The half life of mean reversion is found to be 32 months, and the price/earnings series is also found to contain a small drift component.

When applied to the daily log prices of individual components of the S&P 500, a significant fraction of the securities are found to contain both permanent and transient components. Moreover, a significant fraction of partially autoregressive series are persistent. It is unknown whether or not this phenomena is economically important.

When fitting the PAR model, it is important to check whether or not the bid-ask spread is large in relation to volatility. This may lead to the series being improperly identified as PAR. In addition, volatility clustering can lead

to misidentification.

Several questions are left open for further research.

First, one might try to further elaborate the partially autoregressive model. It may be useful to develop a model that directly incorporates bid-ask bounce and volatility clustering. Mean reversion may occur on multiple different time scales. Thus, it is conceivable that the price series of a security may be a superposition of various processes, one of which reverts to the mean over a time scale of a few days, while another reverts to the mean over a time scale of months or even years. In order to investigate this question, it would be necessary to develop a more elaborate model.

Another direction for future work would be to look at incorporating additional data sources. For example, perhaps news events can be classified as to whether they primarily affect the mean-reverting component of the price or the random walk component.

Finally, it would be useful to examine whether or not hedging can be used to further isolate the mean reverting component of a price series. If the mean-reverting component belongs primarily to the idiosyncratic portion of the price series, then there is a possibility that much of the permanent component can be hedged away, leaving a residual series that is predominantly mean-reverting. This could have significant economic consequences.

References

[1] Barndorff-Nielsen, O., and Blæsild, P. "Hyperbolic distributions." In *Encyclopedia of Statistical Sciences*, eds., Johnson, N. L., Kotz, S., and Read, C. B., Vol. 3, pp. 700–707. New York: Wiley, 1983.

[2] Bollerslev, Tim (1986). "Generalized Autoregressive Conditional Heteroskedasticity". Journal of Econometrics 31:3 (1986): 307–327.

[3] Brockwell, Peter J., and Richard A. Davis, eds. *Introduction to time series and forecasting*. Vol. 1. Taylor & Francis, 2002.

[4] Clark, Peter K., "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices," *Econometrica,* 41:1 (Jan. 1973), 135-155.

[5] Kwiatkowski, Denis, Phillips, Peter C.B., Schmidt, Peter, and Yongcheol Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *Journal of Econometrics* 54 (1992): 159-178.

[6] Poterba, James M., and Lawrence H. Summers. "Mean reversion in stock prices: Evidence and implications." *Journal of financial economics* 22:1 (1988): 27-59.

[7] Shiller, Robert. "U.S. Stock Markets 1871–Present and CAPE Ratio," http://www.econ.yale.edu/ shiller/data/ie_data.xls

[8] Simon, Dan. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons, 2006.

[9] Summers, Lawrence H. "Does the stock market rationally reflect fundamental values?" *The Journal of Finance* 41:3 (1986): 591-601.

[10] Wilks, S. S. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *The Annals of Mathematical Statistics* 9:1 (1938): 60–62.

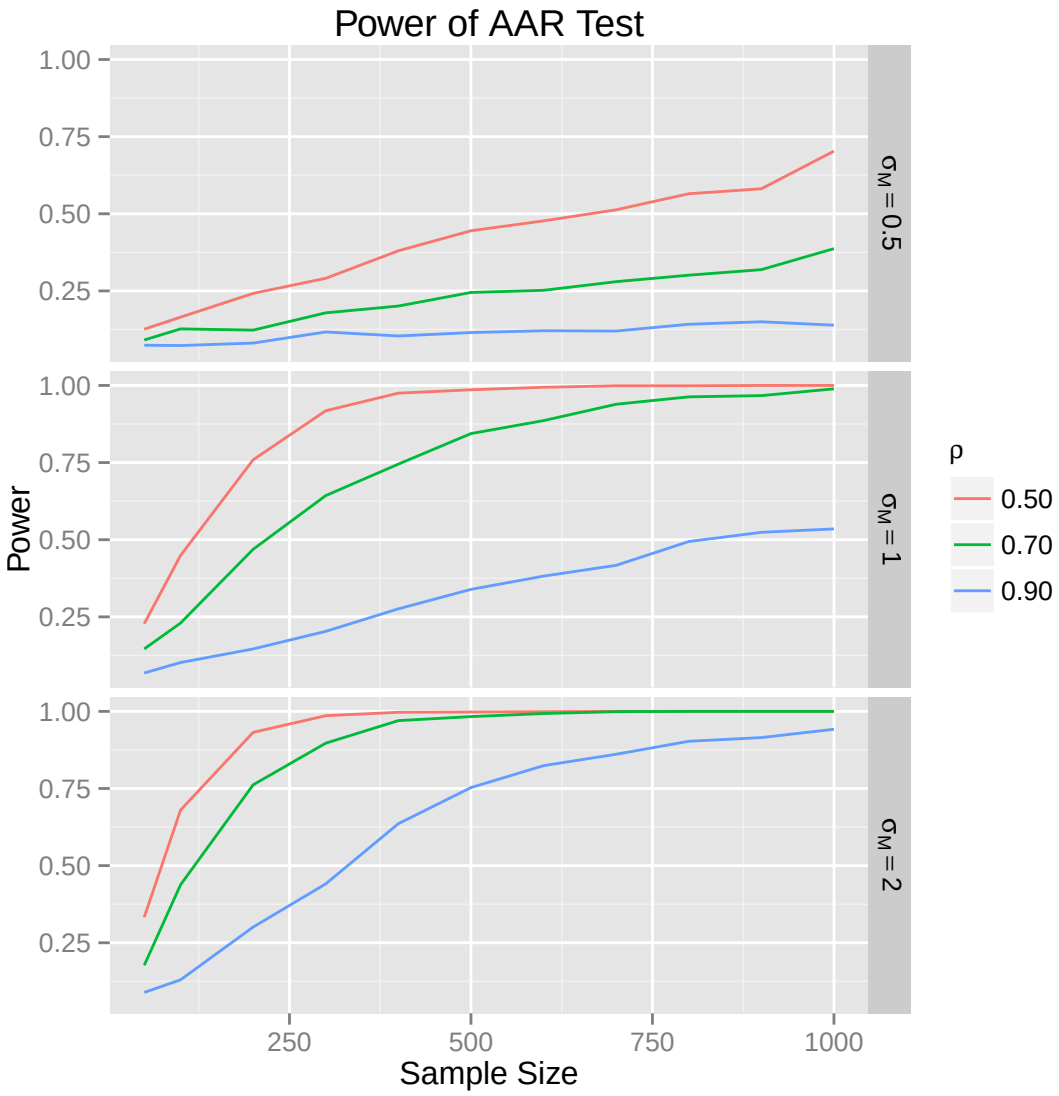

Figure 1: For various combinations of ρ and σ_M and for various sample sizes, a number of random PAR sequences were generated, where σ_R was held constant at 1. For each random PAR sequence, a likelihood ratio test was performed. The critical value used for acceptance was $p = 0.05$.

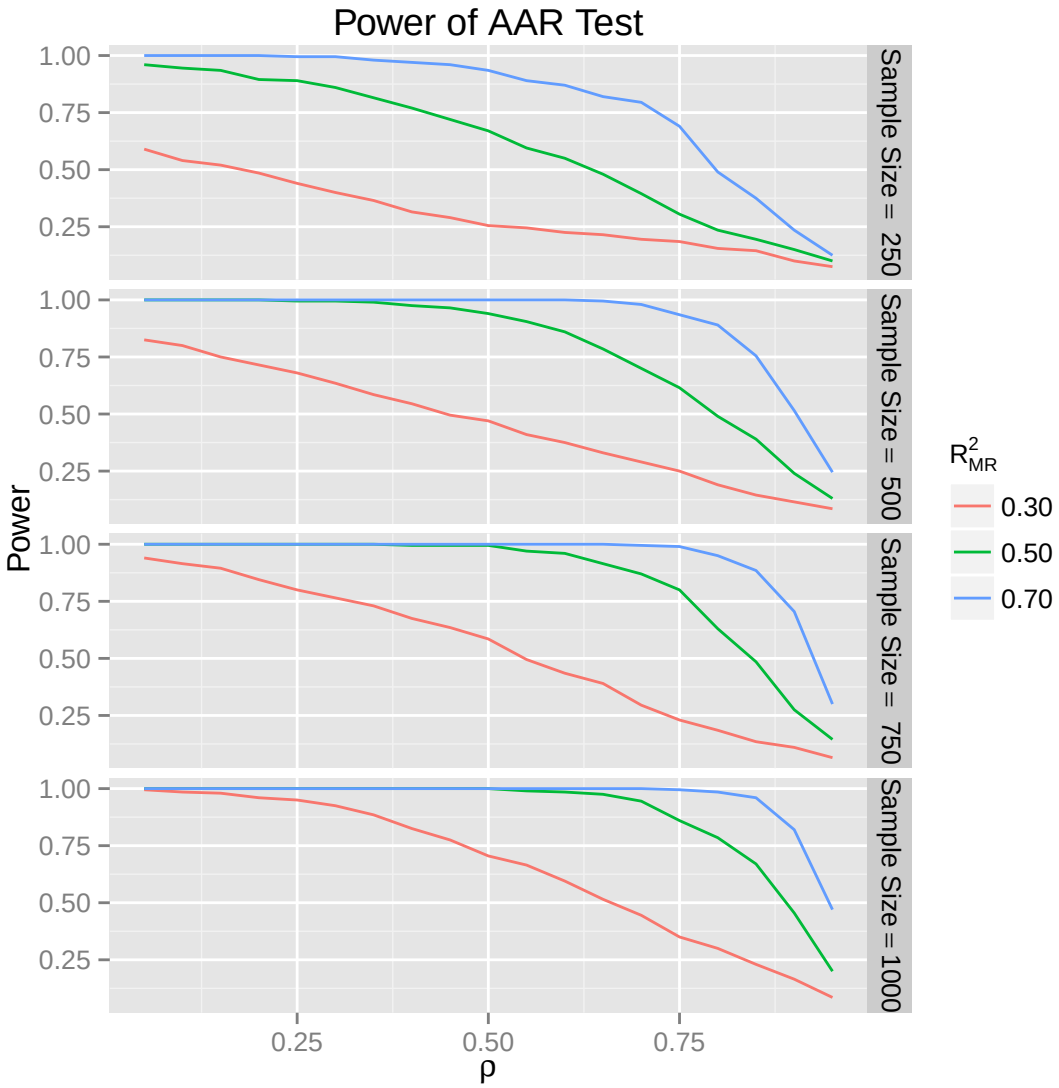

Figure 2: This graph displays the power of the PAR test as a function of the proportion of variance attributable to mean reversion (R^2_{MR}). For various combinations of ρ and R^2_{MR} and for various sample sizes, a number of random PAR sequences were generated. For each random PAR sequence, a likelihood ratio test was performed. The critical value used for acceptance was $p = 0.05$.

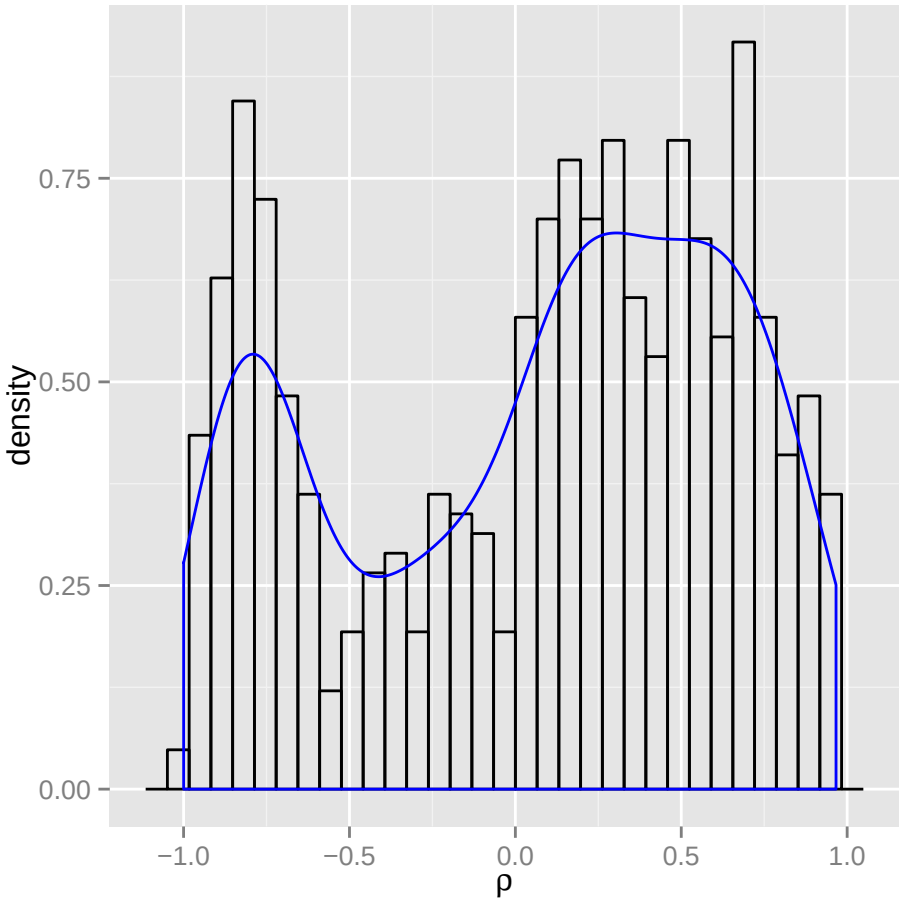

Figure 3: Density of ρ for those sequences in the S&P 500 which were identified as PAR

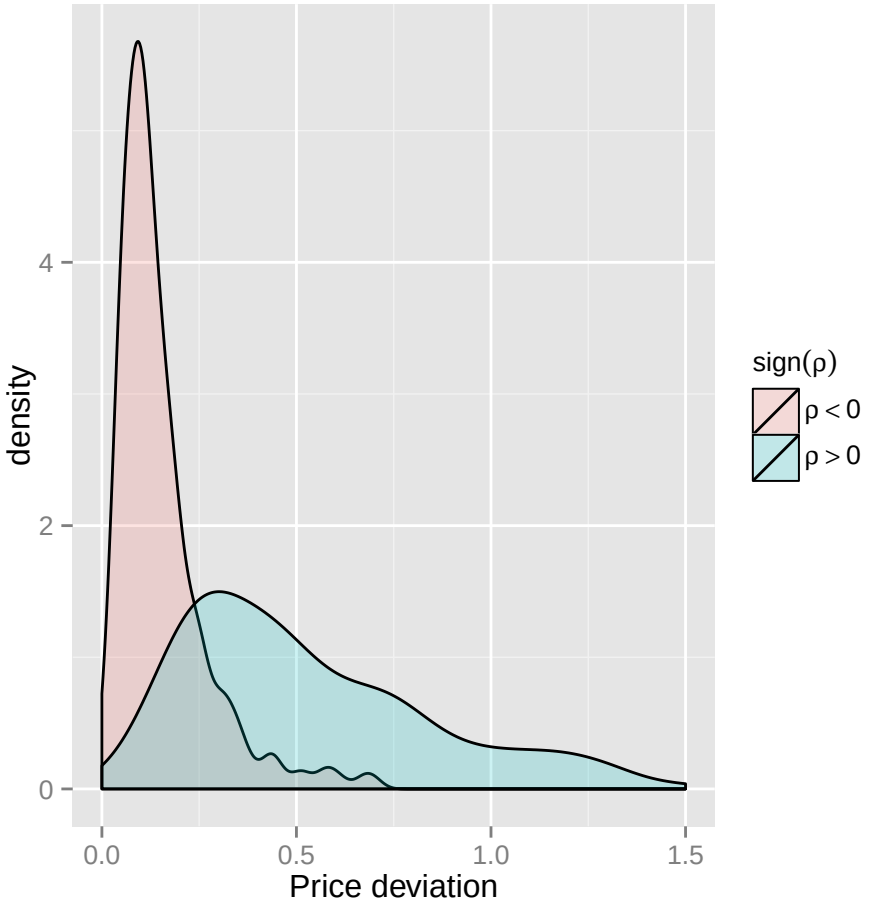

Figure 4: Distribution of Price Deviations. Two density curves are given, one corresponding to the case where $\rho > 0$, and the other to the case where $\rho < 0$. In each curve, the distribution of the price deviation of the mean reverting component is plotted.

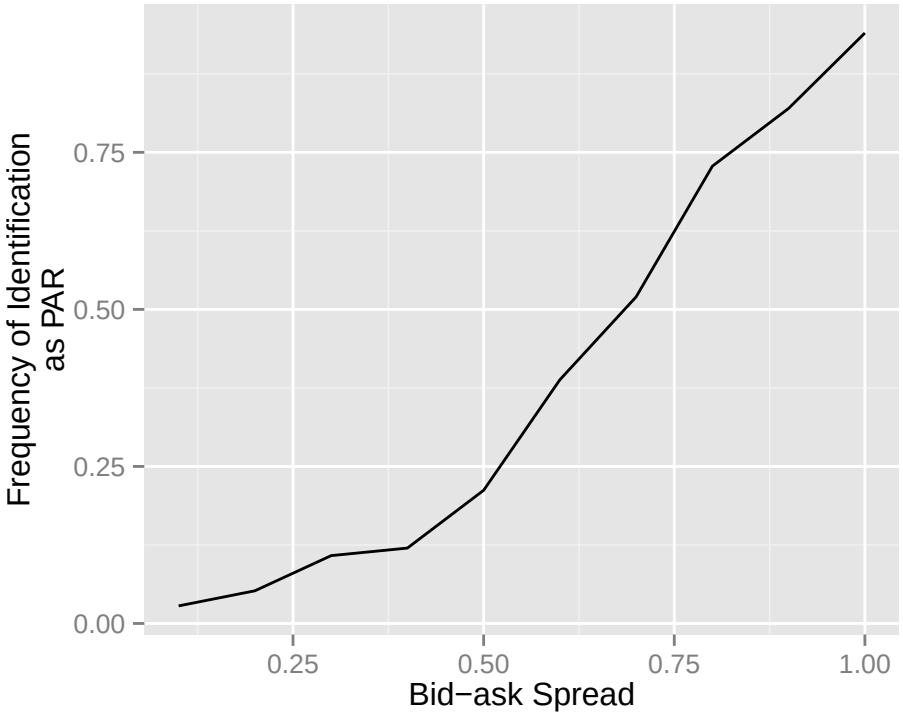

Figure 5: Effects of Bid-Ask Spread on Idenfication as PAR. A number of random sequences were generated incorporating bid ask bounce, using the model given in Equation (13). In these simulations, the starting price X_0 was chosen as 100, and the volatility was chosen as 1%, e.g., $\epsilon_t \sim N(0, 0.01^2)$. The bid-ask spread s was allowed to vary. For each value of s, a number of sequences of length 500 were generated and fit to the PAR model. The proportion of sequences identified as PAR is graphed.

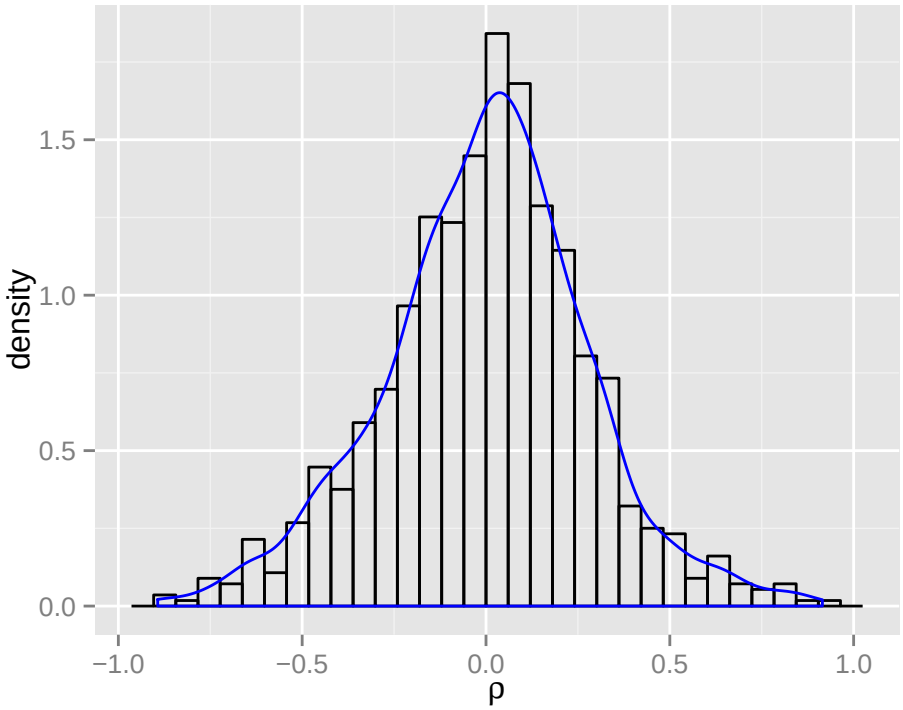

Figure 6: Distribution of ρ When PAR Model to Random Walks with Bid-Ask Bounce. In this study, 1,000 random sequences of the form (13) were generated from the parameters $X_0 = 10.00, s = 0.10$ and $\epsilon_t \sim N(0, 0.01^2)$. Over 90% of such sequences were identified as PAR. For those that were identified as PAR, the distribution of ρ was plotted.

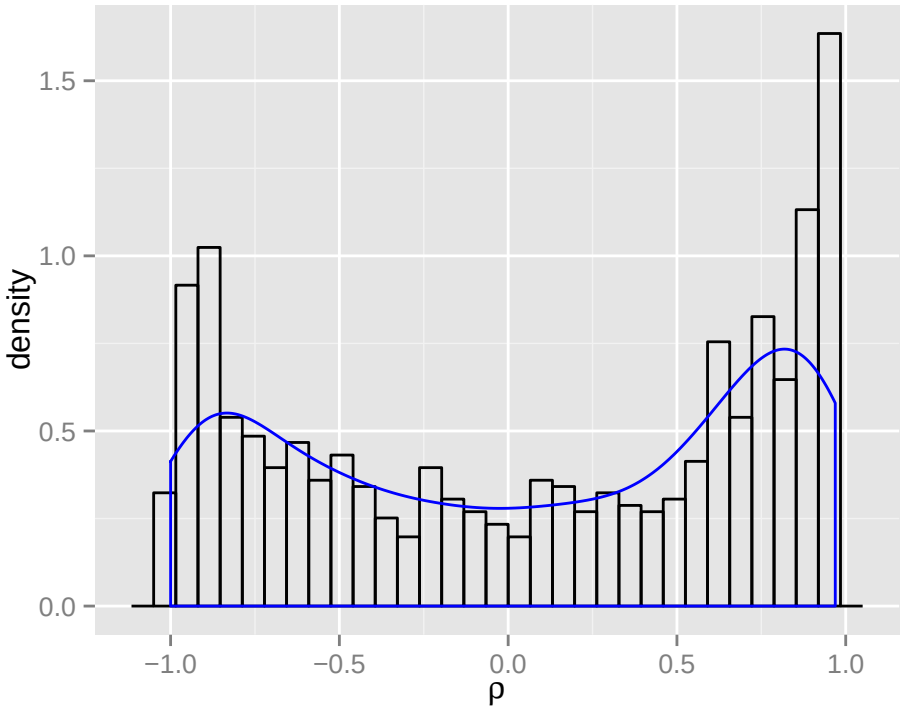

Figure 7: Density of ρ for GARCH sequences identified as PAR. Ten thousand random GARCH sequences of length 500 were generated using the parameters $\alpha_0 = 9.42 \times 10^{-6}, \alpha_1 = 0.0956$, and $\beta = 0.8828$. Each sequence was then fitted to the PAR model. For those sequences that were identifed as PAR, the distribution of ρ is plotted.

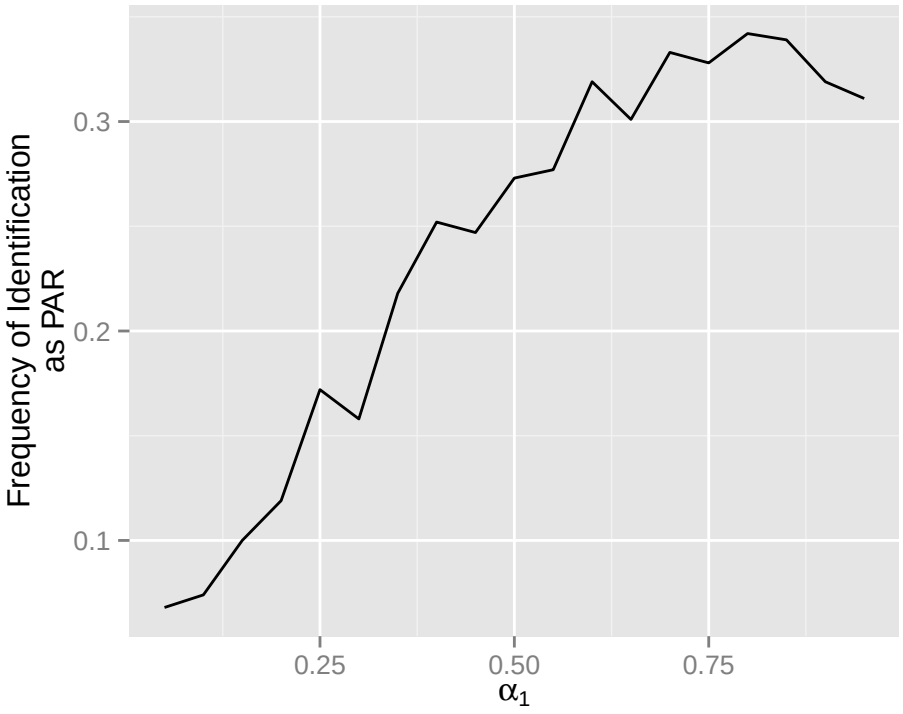

Figure 8: Frequency of misidentification of GARCH series as PAR. For each value of α_1, one thousand GARCH sequences were generated using the parameters $\alpha_0 = 1 \times 10^{-5}$ and $\beta = 0.95 - \alpha_1$. For each such sequence, a check was performed to see if the sequence is identified as PAR. The frequency of (mis)identification as PAR is then plotted.