The Auto-Correlation Function For a Two-Valued Time Series

Richard G. Clegg and Maurice Dodson,
Department of Mathematics, University of York, York,
United Kingdom, YO10 5DD, *email: richard@manor.york.ac.uk*

March 29, 2004

Abstract

Two simple theorems are proved about the auto-correlation function (ACF) of weakly stationary time series which can only take two values. The first theorem shows how the ACF can be expressed in terms of the conditional probability of occurrence of a single value of the time-series. The second theorem shows how the ACF can be expressed in terms of the variance of the number of occurences of one of the values.

Keywords: Autocorrelation, binary series

1 Introduction

For a weakly-stationary time series, the ACF as a function of lag k is given by

$$\rho(k) = \frac{\gamma(k)}{\sigma^2} = \frac{\mathrm{E}\left[(X_i - \mu)(X_{i+k} - \mu)\right]}{\sigma^2},$$

where $\gamma(k)$ is the autocovariance at lag k and σ^2 is the variance ($\sigma^2 = \gamma(0)$).

Consider a two-valued time series, $\{X_t : t \in \mathbb{N}\}$ where $X_t \in \{a, b\}$ and $a \neq b$.

This type of time series is of interest in a number of different applications. Many of the results in Feller (1949) can be seen in terms of a time series with two values. Such time-series are used in a variety of fields including statistical mechanics (Wang 1989) modelling Long-Range Dependence in telecommunications traffic (Arrowsmith et al. 2001) and estimates of rainfall and stock-transactions (Hyndman 1999). Two simple theorems are proved here which do not seem to appear in the literature in the area.

The following notation will be used throughout this paper

$$\begin{aligned} P_a(k) &= \mathbb{P}\left[X_{t+k} = a | X_t = a\right] \\ P_b(k) &= \mathbb{P}\left[X_{t+k} = b | X_t = b\right] \\ p &= \mathbb{P}\left[X_t = a\right]. \end{aligned}$$

Note that if the time series is to have two values, these three quantities must be in the range $(0, 1)$.

The mean μ is given by

$$\mu = \mathrm{E}\left[X_t\right] = pa + (1-p)b. \tag{1}$$

and the variance σ^2 by

$$\sigma^2 = \mathrm{E}\left[(X_t - \mu)^2\right] = \mathrm{E}\left[X_t^2\right] - \mu^2 = p(1-p)(a-b)^2.$$

2 On the ACF of a Weakly Stationary Time Series

Theorem 1. For a weakly-stationary time series $\{X_t : t \in \mathbb{N}\}$, which can only take two values a and b, the autocorrelation function $\rho(k)$ is given by

$$\rho(k) = P_a(k) + P_b(k) - 1 = \frac{P_a(k) - p}{(1-p)} = \frac{P_b(k) - (1-p)}{p}.$$

Proof. Equation (1) can be rearranged to give

$$a - \mu = (1-p)(a-b), \tag{2}$$

and

$$b - \mu = p(b-a). \tag{3}$$

Since the series has only two values,

$$P_a(k) = 1 - \mathbb{P}\left[X_{t+k} = b | X_t = a\right],$$

and

$$P_b(k) = 1 - \mathbb{P}\left[X_{t+k} = a | X_t = b\right].$$

The auto-covariance $\gamma(k)$ is given by

$$\begin{aligned}
\gamma(k) &= \mathrm{E}\left[(X_{t+k} - \mu)(X_t - \mu)\right] \\
&= \mathbb{P}\left[X_{t+k} = a, X_t = a\right](a-\mu)^2 + \mathbb{P}\left[X_{t+k} = b, X_t = a\right](a-\mu)(b-\mu) \\
&\quad + \mathbb{P}\left[X_{t+k} = b, X_t = b\right](b-\mu)^2 + \mathbb{P}\left[X_{t+k} = a, X_t = a\right](a-\mu)(b-\mu) \\
&= \mathbb{P}\left[X_{t+k} = a | X_t = a\right]p(a-\mu)^2 + \mathbb{P}\left[X_{t+k} = b | X_t = a\right]p(a-\mu)(b-\mu) \\
&\quad + \mathbb{P}\left[X_{t+k} = b | X_t = b\right](1-p)(b-\mu)^2 + \mathbb{P}\left[X_{t+k} = a | X_t = b\right](1-p)(a-\mu)(b-\mu) \\
&= P_a(k)p(a-\mu)^2 + (1 - P_a(k))p(a-\mu)(b-\mu) \\
&\quad + (1 - P_b(k))(1-p)(a-\mu)(b-\mu) + P_b(k)(1-p)(b-\mu)^2.
\end{aligned}$$

Thus using (2) and (3),

$$\begin{aligned}
\gamma(k) &= P_a(k)[p(1-p)^2(a-b)^2 + p^2(1-p)(a-b)^2] - p^2(1-p)(a-b)^2 \\
&\quad + P_b(k)[(1-p)p^2(a-b)^2 + p(1-p)^2(a-b)^2] - p(1-p)^2(a-b)^2 \\
&= P_a(k)p(1-p)(a-b)^2 + P_b(k)p(1-p)(a-b)^2 - p(1-p)(a-b)^2 \\
&= \sigma^2[P_a(k) + P_b(k) - 1].
\end{aligned}$$

Therefore,

$$\rho(k) = P_a(k) + P_b(k) - 1, \tag{4}$$

which is the first part of the theorem.

Again, taking the autocovariance gives

$$\begin{aligned}
\gamma(k) &= \mathrm{E}\left[(X_{t+k} - \mu)(X_t - \mu)\right] = \mathrm{E}\left[X_{t+k}X_t\right] - \mu^2 \\
&= P_a(k)pa^2 + (1 - P_a(k))pab + (1 - P_b(k))(1-p)ab + P_b(k)(1-p)b^2 - \mu^2 \\
&= a(a-b)p\left[P_a(k) - p\right] - b(a-b)(1-p)\left[P_b(k) - (1-p)\right] \\
&= \frac{\sigma^2}{a-b}\left[\frac{a}{(1-p)}\left[P_a(k) - p\right] - \frac{b}{p}\left[P_b(k) - (1-p)\right]\right].
\end{aligned}$$

This gives the ACF

$$\rho(k) = \frac{a\left[P_a(k) - p\right]}{(1-p)(a-b)} - \frac{b\left[P_b(k) - (1-p)\right]}{p(a-b)}. \tag{5}$$

Setting this equal to (4) and rearranging gives

$$P_a(k) + P_b(k) - 1 = \frac{aP_a(k) - ap}{(1-p)(a-b)} - \frac{bP_b(k) - b(1-p)}{p(a-b)}$$

and

$$\begin{aligned}
p(1-p)(a-b)P_b(k) + bP_b(k)(1-p) &= apP_a(k) - p(1-p)(a-b)P_a(k) \\
&\quad + p(1-p)(a-b) - ap^2 + b(1-p)^2 \\
(1-p)[P_b(k) - 1] &= p[P_a(k) - 1] \\
\frac{P_a(k) - p}{(1-p)} &= \frac{P_b(k) - (1-p)}{p}.
\end{aligned}$$

Substituting this into (5) gives

$$\rho(k) = \frac{a\left[P_b(k) - (1-p)\right]}{p(a-b)} - \frac{b\left[P_b(k) - (1-p)\right]}{p(a-b)} = \frac{P_b(k) - (1-p)}{p},$$

which, in view of (4) completes the proof. □

Definition 1. Let $I(X_t)$ be an indicator function which has the value 1 if $X_t = a$ and 0 otherwise.

Definition 2. Let A_n where $n \in \mathbb{N}$ be the number of occurences of the value a in a weakly stationary, two-valued time-series. Let $A_n(t)$ be the number of occurences of a between X_{t+1} and X_{t+n}, i.e.,

$$A_n(t) = \sum_{i=1}^{n} I(X_{t+i}).$$

Theorem 2. Given the conditions of the previous theorem then, for $k > 2$,

$$\rho(k) = \frac{\text{var}\,(A_{k+1}) - 2\text{var}\,(A_k) + \text{var}\,(A_{k-1})}{2p(1-p)}.$$

Proof. Expanding the variance of $A_n(t)$ in terms of expectation values gives

$$\begin{aligned}
\text{var}\,(A_n(t)) &= \text{E}\left[A_n(t)^2\right] - \text{E}\left[A_n(t)\right]^2 \\
&= \text{E}\left[\left(\sum_{i=1}^{n} I(X_{t+i})\right)^2\right] - \text{E}\left[\sum_{i=1}^{n} I(X_{t+i})\right]^2 \\
&= \text{E}\left[\left(\sum_{i=1}^{n} I(X_{t+i})\right)^2\right] - \left(\sum_{i=1}^{n} \text{E}\left[I(X_t + i)\right]\right)^2.
\end{aligned}$$

Since the series is stationary, $\text{E}\left[X_t\right] = \text{E}\left[X_0\right]$ and also $\text{E}\left[X_t X_{t+k}\right] = \text{E}\left[X_0 X_k\right]$. Therefore $\text{E}\left[I(X_t)\right] = \text{E}\left[I(X_0)\right]$. Similarly $\text{var}\,(A_n) = \text{var}\,(A_n(t))$. Substituting and rearranging the first sum gives

$$\text{var}\,(A_n) = \text{E}\left[2\left(\sum_{i=0}^{n-1}(n-i)I(X_0)I(X_i)\right) - \left(\sum_{i=1}^{n} I(X_0)I(X_0)\right)\right] - \left(\sum_{i=1}^{n}\text{E}\left[I(X_0)\right]\right)^2.$$

Clearly, $\text{E}\left[I(X_0)\right] = p$ and $\text{E}\left[I(X_0)^2\right] = p$. Also

$$\begin{aligned}
\text{E}\left[I(X_t)I(X_{t+k})\right] &= \text{E}\left[I(X_0)I(X_k)\right] \\
&= \mathbb{P}\left[X_k = a, X_0 = a\right] \\
&= \mathbb{P}\left[X_k = a | X_0 = a\right]\mathbb{P}\left[X_0 = a\right] \\
&= P_a(k)p.
\end{aligned}$$

Making these substitutions and rearranging the sums gives

$$\begin{aligned}
\text{var}\,(A_n) &= \left(2\sum_{i=0}^{n-1}(n-i)\text{E}\left[I(X_0)I(X_i)\right]\right) - \sum_{i=1}^{n}\text{E}\left[I(X_0)I(X_0)\right] - \left(\sum_{i=0}^{n-1} p\right)^2 \\
&= 2\left(\sum_{i=1}^{n}(n-i)P_a(i)p\right) - np - n^2p^2.
\end{aligned}$$

By the same process,

$$\text{var}\,(A_{n+1}) = 2\left(\sum_{i=0}^{n}(n+1-i)P_a(i)p\right) - (n+1)p - (n+1)^2p^2.$$

Taking the first difference gives

$$\text{var}\,(A_{n+1}) - \text{var}\,(A_n) = 2\left(\sum_{i=0}^{n} P_a(i)p\right) - p - 2np^2 - p^2.$$

Similarly

$$\text{var}\,(A_n) - \text{var}\,(A_{n-1}) = 2\left(\sum_{i=0}^{n-1} P_a(i)p\right) - p - 2(n-1)p^2 - p^2,$$

where $n \geq 2$. Taking the second difference gives

$$\text{var}\,(A_{n+1}) - 2\text{var}\,(A_n) + \text{var}\,(A_{n-1}) = 2p(P_n - p).$$

Therefore, subsituting this into Theorem 1 gives, for $n \geq 2$,

$$\rho(n) = \frac{\text{var}\,(A_{n+1}) - 2\text{var}\,(A_n) + \text{var}\,(A_{n-1})}{2p(1-p)}$$

□

3 Discussion

Models like those in Arrowsmith et al. (2001) and Wang (1989) consider time series with two values in terms of ON periods and OFF periods where the distribution of the lengths of the ON period is different to that of the OFF period. It might be thought that as a result of this, the correlations in the ON period (designated here as $P_k(a)$) might fall off at a different rate to the correlations in the OFF period ($P_k(b)$). The result proved in Theorem 1 shows that this cannot be the case. Indeed Theorem 1 may find some use in the mathematics of Markov chains since, the two states of the time series could be interpreted as an indication of whether the chain was or was not in some subset of states. The quantities $P_a(k)$ and $P_b(k)$ then become statements about the k-step transition probabilities.

The second theorem may be of use when the behaviour of the autocorrelation function of a system is needed but cannot be easily determined analytically. Feller (1949) proves results about the asymptotic variance of a quantity which is asymptotically equal to $\text{var}\,(A_n)$ in a wide variety of circumstances.

References

1 ARROWSMITH, D. K., MONDRAGÓN, R. J. AND PITTS, J. M. (2001). Chaotic Maps for traffic modelling and queueing performance analysis. *Performance Analysis*, 42:223–240.

2 FELLER, W. (1949). Fluctuation Theory of Recurrent Events. *Trans. of the Amer. Math. Soc.*, 67(1):94–119.

3 HYNDMAN, R. J. (1999). Nonparametric additive regression models for binary time series. In *Proceedings, Australasian Meeting of the Econometric Society University of Technology, Sydney.*

4 WANG, X. J. (1989). Statistical Physics Of Temporal Intermittency. *Phys. Rev. A*, 40(11):6647–6661.