Bernoulli **8**(1), 2002, 117–122

A note on the equivalence of two approaches for specifying a Markov process

K.-S. CHAN[1] and H. TONG[2]

[1]*Department of Statistics and Actuarial Science, University of Iowa, Iowa City IA 52242-1409, USA. E-mail: kchan@stat.uiowa.edu*
[2]*Department of Statistics, London School of Economics, Houghton Street, London, WC2A 2AE, UK and University of Hong Kong, Pokfulam, Hong Kong. E-mail: h.tong@lse.ac.uk*

The probabilistic structure of a discrete-time (high-order) vector Markov process may be studied using two approaches. In the first approach, the Markov process is specified by the transition probability and the initial distribution. An alternative approach is via a stochastic difference equation. We have proved that these two approaches are equivalent under very mild conditions.

Keywords: ARMA model; equivalence of distribution functions; stationarity; stochastic difference equation; transition probability

1. Introduction

Consider a discrete-time Markov process. We can study its structure with reference to its conditional probabilities, i.e. its transition probabilities, in addition to its initial distribution. This approach will be referred to as the direct approach. An alternative approach is via a stochastic difference equation. In the linear Gaussian case, the two approaches are clearly equivalent. The state-space approach to nonlinear time series typically adopts the transition probability approach, while the nonlinear autoregressive approach is clearly within the nonlinear stochastic difference framework. Rather curiously, the two approaches seem to be developing without much interaction. It is therefore relevant to investigate whether they are theoretically equivalent. Now, for a real-valued Markov chain and in the first-order case, the two approaches can be shown to be equivalent under very general conditions; see Rosenblatt (1971, Lemma 2, p. 169) and Tong (1990, Lemma 3.1, p. 97). Whether this equivalence can be generalized to the case of higher-order vector-valued Markov processes is, as far as we are aware, an open problem. This open problem deserves careful attention as the equivalence of the two approaches, if true, has the following useful consequences. First, it implies that whether we should use one approach in preference to the other becomes a matter of either taste or convenience. Second, it is then guaranteed that one can convert freely between the two approaches. While the direct approach facilitates likelihood calculations, the stochastic difference equation approach yields a direct description of the dynamics of the process, which is of relevance for forecasting purposes. Moreover, it provides a recipe for recovering the noise terms under certain conditions, which terms are useful for (not necessarily likelihood-

1350–7265

based) model estimation and diagnostics. Hence, the equivalence of the two approaches would be of some practical value.

The purpose of this paper is to address this open problem. We show in Section 2 that under very general conditions, the two approaches are indeed equivalent.

2. Main result

We recall the well-known device of writing a pth-order Markov chain, say $\{Y_n\}$, as a first-order vector-valued Markov chain via the stacking operation $X_n = (Y_n^{\mathrm{T}}, Y_{n-1}^{\mathrm{T}}, \ldots, Y_{n-p+1}^{\mathrm{T}})^{\mathrm{T}}$, where X_n is k-dimensional and k is equal to p times the dimension of Y_n. Henceforth, it suffices to consider first-order k-dimensional Markov chains.

Let $\{X_n = (X_{n,1}, X_{n,2}, \ldots, X_{n,k})^{\mathrm{T}}\}$ be a k-dimensional homogeneous Markov chain. We say that $\{X_n\}$ satisfies a stochastic difference equation of the *normal form* if

$$X_{n+1,1} = G_1(X_n, \epsilon_{n+1,1}), \tag{1}$$

$$X_{n+1,2} = G_2(X_n, X_{n+1,1}, \epsilon_{n+1,2}), \tag{2}$$

$$X_{n+1,3} = G_3(X_n, X_{n+1,1}, X_{n+1,2}, \epsilon_{n+1,3}), \tag{3}$$

$$\vdots$$

$$X_{n+1,j} = G_j(X_n, X_{n+1,i}, i < j, \epsilon_{n+1,j}), \tag{4}$$

$$\vdots$$

$$X_{n+1,k} = G_k(X_n, X_{n+1,i}, i < k, \epsilon_{n+1,k}), \tag{5}$$

where $\epsilon_n = (\epsilon_{n,1}, \epsilon_{n,2}, \ldots, \epsilon_{n,k})^{\mathrm{T}}$, $\{\epsilon_n\}$ is a sequence of independent and identically distributed (i.i.d.) random vectors, each of which consists of independent components, and the Gs are measurable functions. By substituting, in order, (1) into (2), (2) into (3), and so on, we see that $X_{n+1} = G(X_n, \epsilon_{n+1})$, for some function G. The normal-form stochastic difference equation, if it exists, is not unique for the underlying process, as it depends on the order of the components in X_n. Even when this order is fixed, the normal-form stochastic difference equation is generally not unique, because we can apply a non-singular transformation to ϵ_n componentwise and modify the functional form of the Gs accordingly to preserve the equalities (1)–(5). Equations (1)–(5) are said to be *invertible* if and only if we can invert the equations to express $\epsilon_{n+1,j}$ in terms of X_n, $X_{n+1,i}$, $i \leqslant j$, i.e. there exist measurable functions H_j, $1 \leqslant j \leqslant k$, such that $\epsilon_{n+1,j} = H_j(X_n, X_{n+1,i}, i \leqslant j)$ for all n, j. Notice that a non-degenerate vector Gaussian AR(1) process always admits an invertible normal-form stochastic difference equation that can be derived as follows. Assume that $\{X_n\}$ is a non-degenerate vector Gaussian AR(1) process. Then, given past Xs, $X_{n+1} \sim N(AX_n, \Sigma)$, where A is a $k \times k$ constant matrix and Σ a positive definite matrix. Consider the Choleski decomposition $\Sigma = R^{\mathrm{T}} D R$, where R is a $k \times k$ upper-triangular matrix with unit diagonal elements and D is

a diagonal matrix. Then $(R^{\mathrm{T}})^{-1}X_{n+1} \sim N((R^{\mathrm{T}})^{-1}AX_n, D)$. Notice that $(R^{\mathrm{T}})^{-1}$ is a lower-triangular matrix with unit diagonal elements. Hence, $(R^{\mathrm{T}})^{-1}X_{n+1} = (R^{\mathrm{T}})^{-1}AX_n + \epsilon_{n+1}$, where $\epsilon_{n+1} \sim N(0, D)$, yielding an invertible normal-form stochastic difference equation representation for $\{X_n\}$.

The above definition of invertibility implicitly assumes that the state vectors X_n are observable. In some cases, the state vectors X_n are partially observable, e.g. only the first components of the state vectors are observable, in which case it may be more desirable to rephrase the definition of invertibility by requiring that the noises be expressed as functions of the current and past observable components of the state vectors. However, we shall not pursue this issue here.

Let us now introduce some notation for the general case of k-dimensional Markov chains. Denote by $B_{n+1,j}$ the σ-algebra generated by all Xs up to and including epoch n, and $X_{n+1,i}$, $0 < i \leqslant j$, i.e. the first j components of X_{n+1}. If $\{X_n\}$ satisfies an invertible normal-form stochastic difference equation, then $B_{n+1,j} = B_n \times C_{n+1,1} \times C_{n+1,2,} \times \ldots \times C_{n+1,j}$, where $C_{n+1,j}$ denotes the σ-algebra generated by $\epsilon_{n+1,j}$. This is an explicit requirement on the manner in which the process is generated, namely, the jth component of X_{n+1} is built up by the past Xs and the first j components of the invertible noise term ϵ_{n+1}. If this requirement is satisfied then, for fixed j, the one-dimensional conditional distributions of $X_{n+1,j+1}$ given $X_n = x$, $X_{n+1,i} = x_{n+1,i}$, $i \leqslant j$, are equivalent functions for all x, $x_{n+1,i}$, $i \leqslant j$. (Following Rosenblatt (1971), two unidimensional distribution functions, say F nd G, are said to be equivalent if and only if the discontinuity points of F can be mapped to those of G in a one-to-one manner with the jump size preserved.) It is well known that a distribution function admits at most countably many points of discontinuity. For any distribution function, say H, the set of all discontinuity points of H will be denoted by $D(H)$. Hence, F and G are equivalent if and only if there exists a one-to-one map $\phi : x \in D(F) \to \phi(x) \in D(G)$ preserving the jump sizes, i.e. for all $x \in D(F)$, $F(x) - F(x-) = G(\phi(x)) - G(\phi(x)-)$ where, for example, $F(x-) = \lim_{y \uparrow x} F(y)$. In particular, continuous distribution functions are, by definition, equivalent. Conversely, following the same kind of construction as in Rosenblatt (1971, p. 169), it will be shown below that if the equivalence requirement on the conditional distribution functions holds, there exist $\epsilon_n = (\epsilon_{n,1}, \epsilon_{n,2}, \ldots, \epsilon_{n,k})^{\mathrm{T}}$ and G_j, $j = 1, \ldots, k$, such that $X_{n+1,j+1} = G_j(X_n, X_{n+1,i}, i \leqslant j, \epsilon_{n+1,j+1})$, which is, indeed, an invertible normal-form stochastic difference equation representation, as can be seen from the proof of the lemma below. Consequently, we have $X_{n+1} = G(X_n, \epsilon_{n+1})$ almost everywhere, a stochastic difference equation representation.

We can now summarize the preceding heuristic discussion in the main result below by first posing the following question.

Question Q. *Do there exist independent components $\epsilon_n = (\epsilon_{n,1}, \ldots, \epsilon_{n,k})^{\mathrm{T}}$ such that, for all n and all j, $\epsilon_{n+1,j+1}$ is independent of $B_{n+1,j}$ while $B_{n+1,j+1}$ equals the σ-algebra generated by $B_{n+1,j}$ and $\epsilon_{n+1,j+1}$?*

Lemma. *Let $\{X_n\}$ be a k-dimensional stationary Markov chain with conditional distributions $F_j(x_{j+1}|x, x_{n+1,i}, i \leqslant j) = P(X_{n+1,j+1} \leqslant x_{j+1}|X_n = x, X_{n+1,i} = x_{n+1,i}, i \leqslant j)$. The answer to Question Q is affirmative if and only if, for each j, $F_j(x_{j+1}|x, x_{n+1,i}, i \leqslant j)$, as functions of*

x_{j+1}, are equivalent for almost all x, $x_{n+1,i}$, $i \leqslant j$, with respect to the stationary probability measure of the process $\{X_n\}$.

Remark. A simple sufficient condition for the equivalence of the Fs in the preceding lemma is that the stationary distribution of (X_n, X_{n+1}) admits a positive pdf with respect to the Lebesgue measure on $\mathbb{R}^{2k}$, in which case all the Fs are continuous functions with the real line as their support, and hence they are equivalent. While the positivity condition generally holds for stationary Markov chains with 'non-degenerate' transition probability kernel, it is inapplicable for Markov chains with 'singular' transition probability kernel, e.g. when X_n is the state vector in the Markovian representation of a scalar-valued kth-order Markov chain, say $\{Y_n\}$, so that $X_n = (Y_n, Y_{n-1}, \ldots, Y_{n-p+1})^{\mathrm{T}}$. In the latter case for $j > 1$, the F_j are the distribution functions of some Dirac delta probability measures, and, for $j = 1$, coincide with the conditional distributions of Y_{n+1} given $Y_n, Y_{n-1}, \ldots, Y_{n-k+1}$. We can then adapt the preceding positivity condition to require that the stationary distribution of $(Y_{n+1}, Y_n, \ldots, Y_{n-k+1})$ admits a pdf with respect to the Lebesgue measure on $\mathbb{R}^{k+1}$.

Before we present the proof of this lemma, three additional remarks are in order. First, in the vector case, the stochastic difference equation representation is ordinarily not unique. In fact, the non-uniqueness of the difference equation representation is related to the identifiability of the model. This problem is very challenging and well documented in the case of linear autoregressive moving-average (ARMA) models for vector time series; see Tuan (1978) and Hannan (1979). Second, we have assumed a particular order of the components of X_n. Already in the linear ARMA case, ordering the components differently may yield different stochastic difference equation representations, some of which may be more tractable than others in terms of statistical inference. Third, jumps may occur naturally in the conditional distribution functions. For example, consider an AR(2) model $Y_n = \alpha_1 Y_{n-1} + \alpha_2 Y_{n-2} + e_n$ where the e_n are i.i.d. Let $X_n = (X_{n,1} = Y_n, X_{n,2} = Y_{n-1})^{\mathrm{T}}$. Clearly, the conditional probability distribution of $X_{n+1,2}$ given X_n jumps at the point $X_{n,1}$ from 0 to 1.

Proof. First, we consider the proof of the necessity of the equivalence of the F_j for each j. Let $C_{n,j}$ denote the σ-algebra generated by $\epsilon_{n,j}$. For fixed j, the equality $B_{n+1,j} = B_{n+1,j-1} \times C_{n+1,j}$ implies that, conditional on X_n, $X_{n+1,i}$, $i < j$, the σ-algebra generated by $X_{n+1,j}$ equals $C_{n+1,j}$ and hence the F_j are equivalent functions.

Conversely, suppose that, for all j, the F_j are equivalent functions. For simplicity, first assume that all the F_j are continuous functions with no jumps. Then, let $U_{n+1,j} = F_j(X_{n+1,j}|X_n, X_{n+1,i}, i < j)$, which are uniformly distributed over the unit interval [0, 1]. Hence, $\{U_n = (U_{n,1}, \ldots, U_{n,k})^{\mathrm{T}}\}$ is a sequence of i.i.d. random variables, each component of U_n being uniformly distributed over [0, 1]. Therefore, $X_{n+1,j} = F_j^{-1}(U_{n,j}|X_n, X_{n+1,i}, i < j)$ almost surely. Here, the inverse of a distribution function, say F, is defined as $F^{-1}(x) = \inf\{y: F(y) > x\}$. This demonstrates that there exists a G such that $X_{n+1} = G(X_n, U_{n+1})$, a stochastic difference equation representation.

If the functions F_j have jumps, we adapt below the arguments of Rosenblatt (1971, p. 169) to show that the above difference equation representation still holds, but the

components of U_n need not be uniformly distributed over [0, 1]. For fixed j, let the set of discontinuity points of $F_j(\cdot|x,\ x_{n+1,i},\ 0 < i < j)$ be $D = \{d_1,\ d_2,\ d_3,\ \ldots\}$ where the ds are assumed to be labelled so that their corresponding jump sizes are non-increasing: $p_1 \geqslant p_2 \geqslant p_3 \geqslant \ldots$. (Note that the only accumulation point of the jump sizes must be zero, as the total probability mass equals 1; hence the jump sizes attain their maximum value.) In other words, for all k, $F_j(d_k|x,\ x_{n+1,i},\ 0 < i < j) - F_j(d_k - |x,\ x_{n+1,i},\ 0 < i < j) = p_k$ with the ds labelled sequentially so that the corresponding ps are non-increasing. Moreover, the set of ps depends on j but is independent of x, $x_{n+1,i}$, $0 < i < j$, by the equivalence of the F_j. On the other hand, the set of discontinuity points does depend on both j and the conditioning values, x, $x_{n+1,i}$, $0 < i < j$. We have suppressed the dependence of the ps on j for the sake of conciseness; similarly, we adopt a simpler notation for the ds. Now, define $\epsilon_{n+1,j} = i$ when $X_{n+1,j} = d_i$ given $X_n = x,\ X_{n+1,i} = x_{n+1,i},\ 0 < i < j$; otherwise define $\epsilon_{n+1,j} = F^{-1}(X_{n+1,j}|X_n = x,\ X_{n+1,i} = x_{n+1,i},\ 0 < i < j)$. It is readily verified that $\epsilon_{n,j}$ is a probabilistic mixture such that it equals i with probability p_i and is uniformly distributed on [0, 1] with probability $1 - \sum_{i=1}^{\infty} p_i$. Clearly, there exist measurable functions G_j such that $X_{n+1,j} = G_j(\epsilon_{n+1,j},\ X_n,\ X_{n+1,i},\ 0 < i < j)$. This completes the proof that there exists G such that $X_{n+1} = G(X_n,\ \epsilon_{n+1})$, a stochastic difference equation representation. □

3. Conclusion

In the case of vector ARMA modelling, the non-uniqueness of the stochastic difference equation representation has given rise to much research on convenient parametrization of vector ARMA models; see, for example, Akaike (1976), Tuan (1978), Hannan (1979), Hannan and Deistler (1988) and Tiao and Tsay (1989). This non-uniqueness problem is much more challenging in the case of nonlinear time series, partly because the functional form of the nonlinear model is often unknown. Innovative research on this challenging problem is clearly needed to advance nonlinear multiple time series modelling.

Acknowledgement

H. Tong was partially supported by grants from the Biotechnology and Biological Sciences Research Council and the Engineering and Physical Sciences Research Council of the United Kingdom, and the Research Grants Council of Hong Kong. We thank a referee for very helpful comments leading to an improved version of the paper.

References

Akaike, H. (1976) Canonical correlation analysis of time series and the use of an information criterion. In R.K. Mehra and D.G. Lainiotis (eds), *System Identification: Advances in Case Studies*, pp. 27–96. New York: Academic Press.

Hannan, E.J. (1979) A note on autoregressive-moving average identification. *Biometrika*, **66**, 672–674.

Hannan, E.J. and Deistler, M. (1988) *The Statistical Theory of Linear Systems*. New York: Wiley.
Rosenblatt, R. (1971) *Markov Processes: Structure and Asymptotic Behavior*. Berlin: Springer-Verlag.
Tiao, G.C. and Tsay, R.S. (1989) Model specificaiton in multivariate time series (with discussion). *J. Roy. Statist. Soc. Ser. B*, **51**, 157–213.
Tong, H. (1990) *Non-linear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press.
Tuan, P.D. (1978) On the fitting of multivariate processes of the autoregressive-moving average type. *Biometrika*, **65**, 99–107.

Received February 2001 and revised May 2001.