IN-65-CR
185084
18P

COMPUTATION OF CANONICAL CORRELATION AND BEST PREDICTABLE ASPECT OF FUTURE FOR TIME SERIES

Mohsen Pourahmadi[1]
University of California, Davis[2]

A.G. Miamee[3]
Hampton University
Hampton, Virginia

ABSTRACT: The canonical correlation between the (infinite) past and future of a stationary time series is shown to be the limit of the canonical correlation between the (infinite) past and (finite) future, and computation of the latter is reduced to a (generalized) eigenvalue problem involving (finite) matrices. This provides a convenient and, essentially, finite- dimensional algorithm for computing canonical correlations and components of a time series. An upper bound is conjectured for the largest canonical correlation.

(NASA-CR-184655) COMPUTATION OF CANONICAL CORRELATION AND BEST PREDICTABLE ASPECT OF FUTURE FOR TIME SERIES (California Univ.) 18 p CSCL 12A
N89-14810
Unclas
G3/65 0185084

[1]Research supported by the NSF grant DMS-8601858.

[2]On leave from Northern Illinois University.

[3]Research supported in part by the NASA Grant NAG-1-768.

1. Introduction

For many practical and theoretical problems in time series analysis, cf. Akaike (1975), Tsay and Tiao (1985), Pourahmadi (1985), it is of interest to know or compute ρ the canonical or maximal correlation between the past

$$P = [\ldots, X_{t-1}, X_t]',$$

and the future

$$F = [X_{t+1}, X_{t+2}, \ldots]'$$

of a stationary time series $\{X_t\}$, and the corresponding canonical component or the best predictable aspect of future. Using the familiar ideas from multivariate analysis, this task, requires computation of eigenvalues, eigenvectors and inversion of infinite matrices or operators.

For ARMA processes canonical correlations and components can be computed (exactly) by solving linear systems of algebraic equations, cf. Helson and Szegö (1960) and Yaglom (1983). For general stationary processes, Jewell et. al. (1983) have given an algorithm for computing (approximating) the canonical correlations as the eigenvalues of an infinite-dimensional (Hankel) operator, in the spectral domain.

In this paper, we provide a time domain algorithm for computing (approximating) canonical correlations of a (nondeterministic) stationary process, which requires only solving linear system(s) of algebraic equations, cf. Yaglom (1965). For instance, in our approach, computation of ρ_m, the canonical correlation between the (infinite) past P and (finite) future

$$F_m = [X_{t+1}, \ldots, X_{t+m}]', \quad 1 \le m < \infty,$$

requires solving one (generalized) eigenvalue problem for two mxm matrices G_m and Γ_m, cf. Theorem 2.3.

Our approach relies primarily on the Wold decomposition of a (nondeterministic) stationary process; this makes it possible to reduce the genuinely infinite-dimensional problem of computation of ρ_m or ρ to an, essentially, finite-dimensional problem; in addition to its computational simplicity, this approach also provides a procedure for computing ρ even when it does not exist as an eigenvalue of an operator, cf. Jewell and Bloomfield (1983) and Jewell et. al. (1983).

In section 2, we develop a procedure for computing the best predictable aspect of (finite) future; the main result is Theorem 2.3. This result along with a simple fact about geometry of Hilbert spaces are used in Section 3, to give an algorithm for computing ρ and the best predictable aspect of the entire future. This procedure is applied to the well-known models fitted to the sunspot numbers series; it turns out that even for m=4, ρ_m provides a good approximation for ρ.

An interesting and yet open problem in this area is that of finding a sharp upper bound for ρ; we have conjectured that $\sqrt{1-\frac{\sigma'^2}{\gamma_0}}$ is an upper bound for ρ, where σ'^2 is the interpolation error of a missing value based on the other values of the process. Throughout this paper, we have emphasized computation of the largest canonical correlation; other canonical correlations and components can be computed by following a standard procedure in multivariate analysis, cf. Theorem 2.3.

2. Best Predictable Aspect of (Finite) Future

For many practical and theoretical problems it is of interest to find the best predictable aspect of the future of a system; when the system is modelled by a stochastic process $\{X_t\}$, the problem of interest can be restated as that of finding the best predictable linear functional of the future values of the form

$$X = \sum_{r=1}^{m} c_r X_{t+r}, \tag{2.1}$$

where $m \leq \infty$ and $c_1,\ldots,c_m$ are (necessarily) unknown; when $m = \infty$, (2.1) should be viewed as the limit in the mean of finite linear combinations. In general, this is a hard problem to solve.

For the time being, we deal with the simpler problem of finding the best linear predictor and prediction error of X in (2.1) when m and $c_1,\ldots,c_m$ are known, and then in the next section we show how the solution of this apparently simpler problem can be employed to resolve the more difficult problem of finding the best predictable aspect of the future. The need for prediciton of linear functionals of the form (2.1), with known m and $c_1,\ldots,c_m$, arises when the forecaster is interested not only in a forecast of individual future values but also in forecast of a linear combination of m future values and a confidence interval for it. For example, if sales are recorded monthly, the forecaster might be interested in the forecast of next year's total sales ($m = 12$, $c_1 = \ldots = c_{12} = 1$), or one might be interested in forecasting the average of some future values ($c_1 = \ldots = c_m = 1/m$), etc.

Note that when $c_1 = \ldots = c_{m-1} = 0$ and $c_m = 1$, then $X = X_{t+m}$, and the prediction problem of X_{t+m} can be solved in the time domain by using the Wold decomposition of $\{X_t\}$; in fact, with

$$X_t = \sum_{k=0}^{\infty} b_k \varepsilon_{t-k} + V_t,\ b_o = 1,\ \sum_{k=0}^{\infty} b_k^2 < \infty,\ \sigma^2 = \mathrm{Var}(\varepsilon_t) \tag{2.2}$$

representing the Wold decomposition of $\{X_t\}$, where $\{\varepsilon_t\}$ is the innovation process of $\{X_t\}$ and $\{V_t\}$ a deterministic process uncorrelated with $\{\varepsilon_t\}$, the best linear predictor of X_{t+r} is given by

$$\hat{X}_{t+r} = \sum_{k=r}^{\infty} b_k \varepsilon_{t+r-k} + V_{t+r}, \tag{2.3}$$

and its (mean square) prediction error is

$$\mathrm{Var}\left(X_{t+r} - \hat{X}_{t+r}\right) = \sigma^2 \sum_{k=0}^{r-1} b_k^2. \tag{2.4}$$

Note that (2.2) also gives rise to the following representation of $\Gamma = \left(\gamma_{i-j}\right)_{i,j=1,\infty}$, the covariance matrix of $\{X_t\}$:

$$\Gamma = \sigma^2 TT' + \Gamma_V \tag{2.5}$$

where $T = \left(b_{j-i}\right)_{i,j=1,\infty}$ with $b_j = 0$ for $j < 0$, and Γ_V is the covariance matrix of the deterministic process $\{V_t\}$. As it is expected, the prediction problem of the more general linear functional (2.1) also hinges on the Wold decomposition of $\{X_t\}$. Indeed, from (2.1) and (2.2) we have

$$X = \sum_{k=0}^{\infty} \Big(\sum_{r=1}^{m} c_r b_{r+k-m}\Big) \varepsilon_{t+m-k} + \sum_{r=1}^{m} c_r V_{t+r},$$

$$= \Big(\sum_{k=0}^{m-1} + \sum_{k=m}^{\infty}\Big) \Big(\sum_{r=1}^{m} c_r b_{r+k-m}\Big) \varepsilon_{t+m-k} + \sum_{r=1}^{m} c_r V_{t+r},$$

from this, $\hat{X}$ the best linear predictor of X based on P is

Consider the harmonizable process $\{X_t,\ t \in R\} \subset L_0^2(P)$ given by $X_t = \int_R e^{it\theta} Z(d\theta)$ and its spectral bimeasure which is induced by Z, i.e.,

$$F(A,B) = E\, Z(A)\ \overline{Z(B)}.$$

We claim that the corresponding spectral domain $L^2(F)$ in this case is not complete.

Verification. By our Lemma there exists a nonzero vector in $H_y(-\infty)$ which does not have a series representation as in (4). Take one such vector V. Since V is clearly in $H_y(0)$ there exists a sequence $\Sigma\, a_k^n\, Y_{-k} = V_n$ of finite linear combination of Y_k's; $k \leq 0$ which converges to V in $L_0^2(P)$. We can write

$$v_n = \int_R f_n(\theta) Z(d\theta)$$

where the nonzero functions f_n are defined on positive integers with $f_n(k) = a_k^n$. By our Theorem in section 2 we have

$$||f_n - f_m||_F = ||\, v_n - v_m|| \ .$$

Now since v_n converges to v and hence is Cauchy so is f_n. However this particular sequence f_n of functions in $L^2(F)$ does not converge to any element f in $L^2(F)$. Because otherwise another application of the Theorem in section 2 shows that f is in $L^1(Z)$ and

$$||f_n - f||_F = ||\int_R (f_n - f)\, dZ\ || = ||V_n - \int_R f dZ|| \ .$$

Thus we see that V_n also converges to $\int_R f dz$. So

$$V = \int_R f\ dZ = \sum_{i=0}^{\infty} f(i)\ Z\ (\{i\}) = \sum_{i=0}^{\infty} f(i)\lambda_i\ Y_{-i},$$

which contradicts our choice of V.

REMARK 1. Our example shows that the main result of [7] claiming the completeness of the spectral domain of any multivariate weakly harmonizable process X_t is false even for a univariate strongly harmonizable process.

REMARK 2. We feel that the error in [7] occurs in lines 8 and 9 of the second column of page 4612, where the existence of a "certain projection onto a subspace" is asserted and a reference to page 33 of [9] is made to support it. In view of the results established in this note the results in which is a quadratic form whose matrix is the matrix of prediction errors. For computational purposes, it is important to note that the matrix G_m is, indeed, the upper left mxm submatrix of the matrix

$$G = \sigma^2 T'T, \tag{2.8}$$

where the (infinite) matrix T is as in (2.5), this provides a simple method of computing G_m when the moving average parameters $b_1, b_2, \ldots$ are known or the task of Cholesky factorization of the covariance matrix Γ is accomplished. In the following Γ_m also stands for the upper left mxm submatrix of Γ.

The measure of (linear) predictability of any function X is usually defined as

$$\lambda(X) = 1 - \frac{\mathrm{Var}(X-\hat{X})}{\mathrm{Var}(X)},$$

cf. Jewell and Bloomfield (1983).

Next, we summarize some of the previous results.

Lemma 2.1. Let $\{X_t\}$ be a nondeterministic stationary process with covariance function $\{\gamma_k\}$ and moving average parameters $b_o = 1$, $b_1, b_2, \ldots, X = \sum_{r=1}^{m} c_r X_{t+r}$ where $m < \infty$ and $c_1, \ldots, c_m$ are given real constants. Then, with $\hat{X}$ denoting the best linear predictor of X based on the infinite past $X_t, X_{t-1}, \ldots$, we have

(a) $\mathrm{Var}(X-\hat{X}) = c'G_m c$.

(b) the measure of (linear) predictability of X is given by

$$\lambda(X) = 1 - \frac{c'G_m\, c}{c'\Gamma_m\, c}.$$

Remark 2.2. For m=1, the measure of predictability of $X = X_{t+1}$ has the simple form $\lambda(X_{t+1}) = 1 - \sigma^2/\gamma_o$, cf. Lemma 2.1(b), since $\sigma^2 = \exp\{\int \log f(\lambda)\, d\lambda/2\pi\}$, where $f(\lambda)$ is the spectral density of the process, it follows that $\lambda(X_{t+1})$ can be expressed in terms of the density of the process. However, for m>1, it seems difficult to find expressions for

$\lambda(X)$ in terms of the density.

Next, we find the best predictable aspect of the future for a given m. In view of Lemma 2.1(b) this amounts to finding $c_1, \ldots, c_m$ such that for $X = \sum_{r=1}^{m} c_r X_{t+r}$, $\lambda(X)$ is maximized. The next theorem shows how this can be reduced to a standard (generalized) eigenvalue problem.

<u>Theorem 2.3</u>. Let $\{X_t\}$ be a nondeterministic stationary process and $X = \sum_{r=1}^{m} c_r X_{t+r}$, for $m \geq 1$ fixed. Then X is the best predictable aspect of the m future values $X_{t+1}, \ldots, X_{t+m}$, if $c = [c_1, \ldots, c_m]'$ satisfies

$$(G_m - \lambda \Gamma_m)c = 0, \text{ for some } \lambda \in R. \tag{2.9}$$

More precisely, let $\lambda_1 < \ldots < \lambda_k$, $(k \leq m)$ be the distinct roots of the determinantal equation

$$\det(G_m - \lambda \Gamma_m) = 0 \tag{2.10}$$

and $c_{(1)}, \ldots, c_{(m)}$ be the corresponding orthonormalized eigenvectors, i.e.

$$c'_{(i)} \Gamma_m c_{(j)} = \delta_{i,j}, \quad i,j = 1,2,\ldots,m.$$

Then, $X_{(1)} = c'_{(1)} F_m$, with $F_m = [X_{t+1}, \ldots, X_{t+m}]'$ is the best predictable aspect of future with the measure of predictability

$$\lambda(X_{(1)}) = 1 - \lambda_1,$$

and in general $X_{(i)} = c'_{(i)} F_m$ is the i^{th} best predictable aspect of future with measure of predictability given by

$$\lambda(X_{(i)}) = 1 - \lambda_i.$$

Proof. Note that the problem of maximizing $\lambda(X)$ over the variation of c is equivalent to minimizing $c'G_m c$ subject to the side condition $c'\Gamma_m c = 1$. Now, the results follow either from using the standard Lagrangian multiplier method, cf. Rao (1973, p. 583), or a method based on the Hilbert-Courant maximization Lemma, cf. Johnson and Wichern (1988, p. 441). ∎

For the purpose of computation it is important to note that roots of (2.10) are the same as the eigenvalues of the matrix $S_m G_m S_m'$, where S_m can be chosen to be either the inverse of the symmetric square root of Γ_m or the inverse of the Cholesky factor of Γ_m. In the computation that follows we have used the latter. For a given time series data set $X_1,\ldots,X_T$, the moving average parameters $b_1,b_2,\ldots$ can be estimated either by fitting ARMA models to data or factorizing the estimated spectral density, cf. Jewell et al. (1983).

2.4 Example.

(a) For $X_t = \varepsilon_t + \varepsilon_{t-1}$, $Var(\varepsilon_t) = 1$, $m = 2$, we have

$$\Gamma_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}; \quad G_2 = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \det(G_2 - \lambda\Gamma_2) = (1 - \lambda)(1 - 3\lambda)$$

with roots and corresponding vectors

$$\lambda_1 = 1/3, \qquad c_{(1)} = [-\,2/\sqrt{6},\ 1/\sqrt{6}]',$$

$$\lambda_2 = 1, \qquad c_{(2)} = [\ 0,\ 1/\sqrt{2}]',$$

and best predictable aspects

$$X_{(1)} = -2/\sqrt{6}\ X_{t+1} + 1/\sqrt{6}\ X_{t+2}, \quad \lambda(X_{(1)}) = 2/3$$

$$X_{(2)} = 1/\sqrt{2}\ X_{t+2}, \qquad \lambda(X_{(2)}) = 0.$$

Note that $X_{(2)}$ is actually uncorrelated with $X_t, X_{t-1}, \ldots$.

(b) For $X_t = \varepsilon_t + .216\, \varepsilon_{t-1} - .36\varepsilon_{t-2}$, $Var(\varepsilon_t) = 1$, $m=4$ we have

i	λ_i	$c_{(i)}$
1	.7291	[.817, -.703, .363, -.258]'
2	.9419	[.552, .71 , .061, .21]'
3	1.000	[0 , 0 , 0 , .922]'
4	1.000	[0 , 0 , .929, -.109]'

(c) For $X_t - .216X_{t-1} + .36X_{t-2} = \varepsilon_t$, $Var(\varepsilon_t) = 1$, $m=4$, we have

i	λ_i	$c_{(i)}$
1	.810	[.818, -.573, 0 , 0]'
2	.912	[.499, .736, 0 , 0]'
3	1.000	[-.110, .409, -.511, .952]'
4	1.000	[.343, -.096, .886, .306]'

(d) For $X_t - .216X_{t-1} = \varepsilon_t - .36\varepsilon_{t-1}$, $Var(\varepsilon_t) = 1$, $m=4$, we have

i	λ_i	$c_{(i)}$
1	.9756	[.976, .351, .126, .43]'
2	1.000	[-.552, -.113, .975, .148]'
3	1.000	[-.009, -.001, -.019, .986]'
4	1.000	[-.214, .94 , .230, -.058]'

It is interesting and important to note the pattern of zeros in the $c_{(i)}$'s in examples (b) and (c), and explore their relationship with those in Tsay and Tiao (1985).

3. Canonical Correlation

It is well-known that there is a close connection between prediction (regression) problems and the concept of correlation; the root of this phenomenon can be traced to the following simple property of the geometry of Hilbert spaces: For N a subspace of a Hilbert space H, and X an element of H with P_N^X denoting its orthogonal projection onto N, we have

$$(X,\ P_N^X) = (X,\ X-X+P_N^X) = \|X\|^2 - (X,\ X-P_N^X)$$
$$= \|X\|^2 - \|X-P_N^X\|^2. \qquad (3.1)$$

In this identity, the relationship between the correlation (angle) of X and P_N^X, and their distance $\|X-P_N^X\|^2$ (prediction error) is rather self-evident; in the following we shall make deeper use of (an extension of) this identity in developing an algorithm for computing ρ_m and ρ, the canonical correlation between P and F_m, P and F respectively; which allows us to reduce a genuinely infinite-dimensional problem to a finite-dimensional problem; for this we need the following two useful lemmas which are not necessarily new and their proofs might be around in the literature. Due to the importance of these lemmas in our work, in Section 4 we provide proofs for these lemmas.

<u>Lemma 3.1</u>. Let M and N be any two subspaces of $L^2(\Omega)$, the space of square integrable random variables. Then,

$$\sup_{\substack{X\in M\\ Y\in N}} \left| \mathrm{Corr}(X,Y) \right| = 1 - \inf_{\substack{X\in M\\ \|X\|=1}} \|X-P_N^X\|^2 \quad ,$$

furthermore with $\rho(M,N)$ denoting the above quantity, we have $\rho(M_1,N_1) \le \rho(M,N_1) \le \rho(M,N)$, for any $M_1 \subseteq M$ and $N_1 \subseteq N$, that is $\rho(\cdot,\cdot)$ is an increasing (set) function.

<u>Lemma 3.2</u>. Let $\{X_t\}$ be a stationary process, with P, F, F_m, ρ and ρ_m as before, and $P_m = [X_{t-m+1},\ldots,X_t]$, $m \ge 1$. Then,

(a) $\rho = \rho(P,F) = \lim_{m\to\infty} \rho_m$.

(b) $\rho = \lim_{m\to\infty} \rho(P_m,F_m) = \lim_{m\to\infty} \lim_{n\to\infty} \rho(P_m,\ F_n)$.

Next, we state and prove the main result of this section. It is instructive to compare the result in part (b) with Theorem 1 in Jewell and Bloomfield (1983).

Theorem 3.3. Let $\{X_t\}$ be a nondeterministic stationary process with covariance function $\{\gamma_k\}$ and moving average parameters $b_o = 1, b_1, b_2, \ldots$. Then,

(a) ρ_m, the first canonical correlation between the (infinite) past P and (finite) future F_m, is given by

$$\rho_m = \sqrt{1-\lambda_{1,m}} \quad ,$$

where $\lambda_{1,m}$ is the smallest root of the determinantal equation (2.10).

(b) As $m\to\infty$, $\rho_m \uparrow \rho$, in fact,

$$\rho = \sup_m \rho_m = \sqrt{1-\inf_m \lambda_{1,m}} \quad ,$$

and the best predictable aspect of the (entire) future F is equal to $\lim_{m\to\infty} X_{(1)}$, where $X_{(1)}$ is as in Theorem 2.3.

Proof. Part (a) follows from Lemma 3.1 and Theorem 2.3, by taking N and M as the closed linear span of entries of P and F_m, respectively. (b) follows from Lemma 3.2(a) and Theorem 2.3. ∎

Due to the importance of ρ in many situations, it is desirable to find accurate bounds for it, whenever it is not possible to compute its exact value. This problem has been studied by Jewell et. al. (1983) and some elementary upper bounds for ρ are given in terms of certain components of the spectral density of the process. A sharp lower bound for ρ can be obtained from Lemma 3.2(a) and 3.1 by taking $X = \frac{X_{t+1}}{\sqrt{\gamma_o}} \in F$;

$$\rho \geq \sqrt{1 - \sigma^2/\gamma_o} \text{ , cf. Remark 2.2.} \qquad (3.2)$$

To show that this bound is sharp, note that for an AR(1) process

$$X_t = aX_{t-1} + \epsilon_t, \quad \sigma^2 = 1, \quad |a|<1,$$

we have $\gamma_0 = \frac{1}{1-a^2}$ and the bound $\sqrt{1 - \sigma^2/\gamma_0} = |a|$ is attained by ρ. It is much harder to find a sharp upper bound for ρ; however, motivated by (3.2) we conjecture that when $\rho<1$, then

$$\rho \leq \sqrt{1 - \sigma'^2/\gamma_0}\ , \tag{3.3}$$

where σ'^2 is the interpolation error of X_{t+1} based on $\{X_s;\ s \neq t+1\}$. We note that the bound (3.3) is attained for the aforementioned AR(1) process, since in this case, by using a result of Kolmogorov (1941), we have

$$\sigma'^2 = \left(\int f^{-1}(\theta)d\theta/2\pi\right)^{-1} = 1,$$

where $f(\theta) = |1 - ae^{i\theta}|^{-2}$ is the spectral density of the AR(1) process, and

$$\sqrt{1 - \sigma'^2/\gamma_0} = |a|.$$

A more solid motivation for the bound in (3.3) is the fact that $\rho(\cdot,\cdot)$, cf. Lemmas 3.1 and 3.2, is an increasing (set) function of its arguments; therefore, replacing N(P) by $N_1 = \overline{sp}\{X_s;\ s \neq t+1\}$, one arrives at a bound of the form

$$\sqrt{1 - K\,\sigma'^2/\gamma_0}\ ,$$

for ρ, where K is a constant. Thus, the conjecture amounts to showing that K = 1.

<u>Remark 3.4</u>. The canonical correlation between P and $F_{(k)} = (X_{t+k}, X_{t+k+1}, \ldots)'$, $k\geq1$ fixed, denoted by $\rho(k)$, can be also computed by the procedure developed in this paper. In fact, for any $m > k$, and taking c =

$[0,\ldots,0,c_{k+1},\ldots,c_m]$ one can prove results similar to those in Sections 2 and 3 for $\rho_{m-k}(k)$, which is the largest correlation between P and $[X_{k+1},\ldots,X_m]$. It is evident that $\rho_{m-k}(k) \to \rho(k)$ as $m \to \infty$, cf. Theorem 3.3; in this case $\rho_{m-k}(k) = \sqrt{1 - \lambda_{1,m-k}(k)}$, where $\lambda_{1,m-k}(k)$ is actually the smallest root of

$$\det\left(G'_{m-k} - \lambda\Gamma_{m-k}\right) = 0,$$

where G'_{m-k} is the $(m-k)x(m-k)$ matrix obtained from G_m by deleting its first k rows and k columns.

Example 3.5. The well-known sunspot numbers series has been studied by many people and various models fitted to the data are given in Table 1, cf. Jewell et al. (1983). We have calculated ρ_4, that is the canonical correlation between P and F_4, and the corresponding canonical component, using the method of Theorem 3.3, see Table 2. These results are very close to the results in Table 2 of Jewell et al. (1983) which contains the value ρ for these models; this suggests that the rate of convergence of ρ_m to ρ must be rather fast. For model 2, ρ_4 is far from ρ reported in Jewell et al. (1983), this difference persists even when m is large; it should be noted that model 2 represents a nonstationary process, and it might be that for such processes our approximation may not work well as far as compution of ρ is concerned. Despite this, the canonical component for model 2 is almost the same as that for model 1.

Table 1

	Model	Source
1	$X_t - 1.34X_{t-1} + .65X_{t-2} = \varepsilon_t$	Yule, Box-Jenkins
2	$X_t - 1.62X_{t-1} + X_{t-2} = \varepsilon_t$	Yule
3	$X_t - 1.3X_{t-1} + .54X_{t-2} + .15X_{t-3} - .19X_{t-4} + .24X_{t-5} - .4X_{t-6} = \varepsilon_t$	Bailey
4	$X_t - 1.57X_{t-1} + 1.02X_{t-2} - .21X_{t-3} = \varepsilon_t$	Box-Jenkins
5	$X_t - 1.42X_{t-1} + .72X_{t-2} = \varepsilon_t - .15\varepsilon_{t-1}$	Phadke and Wu
6	$X_t - 1.25X_{t-1} + .54X_{t-2} - .19X_{t-3} = \varepsilon_t$	Morris, Schaerf

Table 2

Model	ρ_4^2	Canonical Component
1	.8566	$X_t - .36X_{t+1}$
2	.99	$X_t - .38X_{t+1}$
3	.8602	$X_t - .268X_{t+1} - .107X_{t+2}$
4	.9149	$X_t - .474X_{t+1} - .082X_{t+2}$
5	.8476	$X_t - .296X_{t+1} - .044X_{t+2}$
6	.8676	$X_t - .409X_{t+1} + .126X_{t+2}$

4. Proofs of the Lemmas

In this section we provide proofs of Lemmas 3.1 and 3.2.

Proof of Lemma 3.1. It is obvious that

$$\{\mathrm{Corr}(X, P_N^X);\ X \in M\} \leq \{\mathrm{Corr}(X,Y);\ X \in M,\ Y \in N\},$$

and therefore,

$$\sup_{X \in M} |\mathrm{Corr}(X,\ P_N^X)| \le \sup_{\substack{X \in M \\ Y \in N}} |\mathrm{Corr}(X,Y)|. \tag{1}$$

Also, for any $X \in M$ we have from (3.1) that

$$|\mathrm{Corr}(X,Y)| \le |\mathrm{Corr}(X,\ P_N^X)|, \text{ for all } Y \in N,$$

and thus,

$$\sup_{\substack{X \in M \\ Y \in N}} |\mathrm{Corr}(X,Y)| \le \sup_{X \in M} |\mathrm{Corr}((X,\ P_N^X)|. \tag{2}$$

Now, from (1) and (2), we get that

$$\sup_{\substack{X \in M \\ Y \in N}} |\mathrm{Corr}(X,Y)| = \sup_{X \in M} |\mathrm{Corr}(X,\ P_N^X)|. \tag{3}$$

For any $X \in M$ we have

$$\mathrm{Corr}(X,\ P_N^X) = \frac{(X, P_N^X)}{\|X\|\,\|P_N^X\|} = \frac{\|P_N^X\|^2}{\|X\|\,\|P_N^X\|} = \frac{\|P_N^X\|}{\|X\|}$$

$$= \frac{\|X\|^2 - \|X-P_N^X\|^2}{\|X\|} = \sqrt{1-\|Y-P_N^Y\|^2} \quad ,$$

where $Y = \frac{X}{\|X\|} \in M$ with $\|Y\| = 1$; furtehrmore, from (3.1) we have

$$\mathrm{Corr}(Y,\ P_N^Y) = \sqrt{1-\|Y-P_N^Y\|^2} \quad ,$$

thus

$$\left\{ \left|\mathrm{Corr}(X,\ P_N^X)\right|;\ X \in M \right\} = \left\{ \sqrt{1-\|Y-P_N^Y\|^2};\ Y \in M,\ \|Y\| = 1 \right\} ,$$

or equivalently,

$$\sup_{X \in M} \left|\mathrm{Corr}(X,\ P_N^X)\right| = \sqrt{1 - \inf_{\substack{X \in M \\ \|X\|=1}} \|X-P_N^X\|^2} \quad . \tag{4}$$

The desired result, now, follows from (3) and (4). ∎

<u>Proof of Lemma 3.2:</u> The sequence $\{\rho_m\}$ is bounded and nondecreasing, thus it is convergent and, in fact,

$$\lim_{m\to\infty} \rho_m = \sup_m \rho_m. \tag{1}$$

Also, since the linear span of F_m is a subset of that of F, we have

$$\rho_m \le \rho, \text{ for all } m \ge 1,$$

and therefore,

$$\lim_m \rho_m \le \rho. \tag{2}$$

To establish equality in (2), note that for any two finite linear combinations $X = \sum_{k=0}^{n} a_k X_{t-k}$, $Y = \sum_{k=1}^{m} b_k X_{t+k}$, we have

$$|\mathrm{Corr}(X,Y)| \le \rho_m \le \sup_m \rho_m . \tag{3}$$

By taking supremum of both sides of (3), over all X and Y as above, we arrive at

$$\rho \le \sup_m \rho_m . \tag{4}$$

The desired result, now, follows from (1), (2) and (4). Proof of (b) is similar to (a).

References

Akaike, H. (1975). Markovian representation of stochastic processes by canonical variables. SIAM J. Control 13, 162-173.

Helson, H., Szegö, G. (1960). A problem in prediction theory. Ann. Mat. Pura. Appl. 51, 107-138.

Jewell, N.P., Bloomfield, P., Bartmann, F.C. (1983). Canonical correlations of past and future for time series: Bounds and computation. Ann. Statist. 11, 848-855.

Jewell, N.P., Bloomfield, P. (1983). Canonical correlations of past and future for time series: Definitions and theory. Ann. Statist. 11, 837-847.

Johnson, R.A., Wichern, D.W. (1988). Applied Multivariate Statistical Analysis. Prentice Hall, New Jersey.

Piccolo, D., Tunnicliffe Wilson, G. (1984). A unified approach to ARMA model identification and preliminary estimation. J. of Time Series Analysis, 5, 183-204.

Pourahmadi, M. (1985). A matricial extension of the Helson-Szegö theorem and its applications in multivariate prediction. J. of Multivariate Analysis, 16, 265-275.

Rao, C.R. (1973). Linear Statistical Inference and Its Applications. John Wiley & Sons, New York.

Tsay, R.S., Tiao, G.C. (1985). Use of canonical analysis in time series model identification. Biometrika 72, 299-315.

Yaglom, A.M. (1965). Stationary Gaussian processes satisfying the strong mixing condition and best predictable functionals. Proc. Int. Research Seminar of the Statistical Laboratory, University of California, Berkeley, 1963, 241-252. Springer-Verlag, New York.