Generalized Durbin-Levinson and Burg Algorithms

P.J. Brockwell
Department of Statistics
Colorado State University

and

R. Dahlhaus
Institut für Angewandte Mathematik
Universität Heidelberg

Abstract

We develop recursive algorithms for subset modelling and prediction which generalize the well-known Durbin-Levinson and Burg algorithms and include the univariate version of the subset Whittle algorithm of Penm and Terrell (1982). The results are derived using a basic property of orthogonal projections which leads to very simple derivations of the standard versions of the algorithms. As an application of the results, we obtain new and easily applied algorithms for the recursive calculation of the best linear h-step predictors (for any fixed $h > 0$) of an arbitrary stationary process with known mean and covariance function.

KEYWORDS: *recursive autoregression, Durbin-Levinson algorithm, Burg algorithm, linear prediction, subset modelling, multistep prediction.*

1 Introduction

A fundamental problem in time series analysis is the determination, for a series $\{X_t, t = 0, \pm 1, \ldots\}$ with $E(X_t^2) < \infty$ for all t, of the best linear predictor,

$$\hat{X}_{n+1}(m) = \sum_{j=1}^{m} \phi_m(j) X_{n+1-j} = \boldsymbol{\phi}_m' \mathbf{X}_n(m), \tag{1.1}$$

where $\boldsymbol{\phi}_m = (\phi_m(1), \ldots, \phi_m(m))'$, $\mathbf{X}_n(m) = (X_n, \ldots, X_{n+1-m})'$ and "best" means that $\boldsymbol{\phi}_m$ is chosen to minimize the mean squared error $E[(X_{n+1} - \hat{X}_{n+1}(m))^2]$. It is well-known (see e.g. Brockwell and Davis, 1991) that the mean squared error is minimized if $\boldsymbol{\phi}_m$ is any solution of the equation

$$\Gamma_m \boldsymbol{\phi}_m = \boldsymbol{\gamma}_m, \tag{1.2}$$

where $\boldsymbol{\gamma}_m = E[\mathbf{X}_n(m) X_{n+1}]$ and $\Gamma_m = E[\mathbf{X}_n(m)\mathbf{X}_n(m)']$. The corresponding mean squared error is

$$v_m = E(X_{n+1}^2) - \boldsymbol{\gamma}_m' \Gamma_m^{-1} \boldsymbol{\gamma}_m, \tag{1.3}$$

where Γ_m^{-1} is any generalized inverse of Γ_m. All solutions $\boldsymbol{\phi}_m = \Gamma_m^{-1}\boldsymbol{\gamma}_m$ of (1.2) give the same linear predictor $\hat{X}_{n+1}(m) = \boldsymbol{\phi}_m' \mathbf{X}_n(m)$.

In the case when $\{X_t\}$ is a stationary process, the Durbin-Levinson algorithm allows us to compute the predictors $\hat{X}_{n+1}(m)$ recursively in m, eliminating the need for matrix inversions (see Algorithm 1 in Section 3). An overview of the Durbin-Levinson algorithm and related algorithms may be found in Morettin (1984).

The Yule-Walker equations and the Durbin-Levinson algorithm also play an important role in the *modelling* of time series data. In order to fit a zero-mean stationary autoregression of order m to a series of observations $x_1, \ldots, x_T$ of a time series at times $1, \ldots, T$ respectively, we replace the expected products $E(X_{t+u}X_t)$ by sample analogues and then determine the corresponding $\boldsymbol{\phi}_m$ and v_m. Thus the Yule-Walker AR(m) model for the data is obtained by replacing each expected product $E(X_{t+u}X_t)$ in equations (1.2) and (1.3) by the corresponding sample value $\hat{\gamma}(u) = (\sum_{t=1}^{T-u} x_{t+u}x_t)/T$, $0 \leq u < T$, or its tapered analogue (see Remark 2 in Section 3). If we denote the resulting sample analogues of $\boldsymbol{\gamma}_m$ and Γ_m by $\hat{\boldsymbol{\gamma}}_m$ and $\hat{\Gamma}_m$ respectively, then the Yule-Walker AR(m) model for the data is

$$X_t = \hat{\boldsymbol{\phi}}_m' \mathbf{X}_{t-1}(m) + Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \hat{v}_m), \tag{1.4}$$

where

$$\hat{\boldsymbol{\phi}}_m = \hat{\Gamma}_m^{-1}\hat{\boldsymbol{\gamma}}_m \quad \text{and} \quad \hat{v}_m = \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}_m'\hat{\Gamma}_m^{-1}\hat{\boldsymbol{\gamma}}_m. \tag{1.5}$$

In this case we note (see Brockwell and Davis, 1991, Problem 7.11) that except in the trivial and uninteresting case when $\hat{\gamma}(0) = 0$, the matrix $\hat{\Gamma}_m$ is non-singular for each $m = 1, 2, \ldots$ and consequently the matrix $\hat{\Gamma}_m^{-1}$ in (1.5) is the unique inverse of $\hat{\Gamma}_m$. To fit a non-zero mean stationary series to the observations $y_1, \ldots, y_T$, we can use the same procedure to fit a zero-mean model to the mean-corrected data $x_1 = y_1 - \overline{y}, \ldots, x_T = y_T - \overline{y}$. The model for the original data is then obtained by replacing each X_t in the model for the mean corrected data by $Y_t - \overline{y}$.

Just as in the calculation of the best linear predictor $\hat{X}_{n+1}(m)$, the Durbin-Levinson algorithm can also be used to determine the coefficient vectors $\hat{\boldsymbol{\phi}}_m$ recursively in m, thus eliminating the need for any matrix inversions.

It is also possible to determine $\hat{\boldsymbol{\phi}}_m$ and $\hat{v}_m$ (and hence the Yule-Walker AR(m) model for the data) using recursions based on forward and backward empirical residuals as described below in Section 5. The idea of using forward and backward empirical residuals is due to Burg (1968). However Burg used the empirical residuals in a different way, leading to an AR(m) model for the data which differs from the Yule-Walker model. Further details and generalizations of these algorithms are given in Section 5.

The purpose of this paper is to establish a basic result for orthogonal projections which leads to very simple derivations of the standard Durbin-Levinson and Burg algorithms. It also plays a key role in the development of the generalized versions considered later in the paper. These include the univariate version of the subset Whittle algorithm due to Penm and Terrell (1982) and a corresponding subset version of the Burg algorithm. As a byproduct of the analysis we also obtain two easily applied recursive algorithms for determining the best h-step linear predictors of an arbitrary stationary sequence. Multivariate versions of the algorithms can be derived analogously, but we shall consider these in a subsequent

paper. The recursive algorithms are even more advantageous in the multivariate than in the univariate case.

In Section 2 we prove the basic algorithm on orthogonal projections. In Section 3 we show how it leads at once to the classical Durbin-Levinson algorithm. We also derive and discuss a sample version of the algorithm and the role of tapering in improving the parameter estimates derived from it. In Section 4 we derive the subset Durbin-Levinson algorithm and its empirical counterpart and derive two algorithms for h-step prediction. Section 5 deals with corresponding subset generalizations of Burg's algorithm.

2 The Basic Algorithm

Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be random column vectors, all of whose components have finite second moments. For any two such random vectors, say $\mathbf{X}$ and $\mathbf{Y}$, we define the matrix of inner products,

$$< \mathbf{X}, \mathbf{Y} >:= E(\mathbf{X}\mathbf{Y}')$$

and we say that $\mathbf{X}$ and $\mathbf{Y}$ are orthogonal if $< \mathbf{X}, \mathbf{Y} >= 0$ or equivalently if $< \mathbf{Y}, \mathbf{X} >= 0$.

The best linear predictor of $\mathbf{X}$ in terms of $\mathbf{Y}$ and $\mathbf{Z}$ is defined to be the linear combination (with matrix coefficients) of $\mathbf{Y}$ and $\mathbf{Z}$, whose components each have minimum mean-square distance from the corresponding component of $\mathbf{X}$. If we denote this best predictor by

$$\hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z}) = A_1\mathbf{Y} + A_2\mathbf{Z},$$

then we know (see e.g. Brockwell and Davis, 1991, Chapter 10) that $\hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z})$ is uniquely determined by the requirement that the error vector $\mathbf{X} - A_1\mathbf{Y} - A_2\mathbf{Z}$ is orthogonal to $\mathbf{Y}$ and to $\mathbf{Z}$, i.e. by the equations

$$\begin{bmatrix} < \mathbf{Y}, \mathbf{Y} > & < \mathbf{Y}, \mathbf{Z} > \\ < \mathbf{Z}, \mathbf{Y} > & < \mathbf{Z}, \mathbf{Z} > \end{bmatrix} \begin{bmatrix} A_1' \\ A_2' \end{bmatrix} = \begin{bmatrix} < \mathbf{Y}, \mathbf{X} > \\ < \mathbf{Z}, \mathbf{X} > \end{bmatrix}. \tag{2.1}$$

Note that although $\hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z})$ is determined uniquely by (2.1), there may be more than one pair of coefficients A_1 and A_2 satisfying the equations. If this is the case then all of these solutions will still give one and the same predictor. The second moment matrix of the prediction error vector is

$$\begin{aligned} v_{\mathbf{X}|\mathbf{Y},\mathbf{Z}} &=< \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z}), \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z}) > \\ &=< \mathbf{X}, \mathbf{X} > -A_1 < \mathbf{Y}, \mathbf{X} > -A_2 < \mathbf{Z}, \mathbf{X} > . \end{aligned} \tag{2.2}$$

Analogously to $\hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z})$, we define $\hat{\mathbf{X}}(\mathbf{Y})$ and $\hat{\mathbf{Z}}(\mathbf{Y})$ to be the best linear predictors in terms of $\mathbf{Y}$ of $\mathbf{X}$ and $\mathbf{Z}$ respectively. Writing

$$\hat{\mathbf{X}}(\mathbf{Y}) = B\mathbf{Y}$$

and

$$\hat{\mathbf{Z}}(\mathbf{Y}) = C\mathbf{Y},$$

we find that the equations analogous to (2.1) and (2.2) for $\hat{\mathbf{X}}(\mathbf{Y})$ are

$$< \mathbf{Y}, \mathbf{Y} > B' =< \mathbf{Y}, \mathbf{X} >, \tag{2.3}$$

$$v_{\mathbf{X}|\mathbf{Y}} =< \mathbf{X}, \mathbf{X} > -B < \mathbf{Y}, \mathbf{X} >, \tag{2.4}$$

and for $\hat{\mathbf{Z}}(\mathbf{Y})$ are

$$< \mathbf{Y}, \mathbf{Y} > C' =< \mathbf{Y}, \mathbf{Z} >, \tag{2.5}$$

$$v_{\mathbf{Z}|\mathbf{Y}} =< \mathbf{Z}, \mathbf{Z} > -C < \mathbf{Y}, \mathbf{Z} > . \tag{2.6}$$

We note also that

$$B < \mathbf{Y}, \mathbf{Z} >=< \mathbf{X}, \mathbf{Y} > C', \tag{2.7}$$

since both sides are equal to $B < \mathbf{Y}, \mathbf{Y} > C'$.

The following theorem shows how the predictor $\hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z})$ can be determined in a simple way from the predictors $\hat{\mathbf{X}}(\mathbf{Y})$ and $\hat{\mathbf{Z}}(\mathbf{Y})$.

Theorem 1

If $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *are random vectors whose components all have finite second moments and if the best linear predictors of* $\mathbf{X}$ *and* $\mathbf{Z}$ *in terms of* $\mathbf{Y}$ *are* $\hat{\mathbf{X}}(\mathbf{Y}) = B\mathbf{Y}$ *and* $\hat{\mathbf{Z}}(\mathbf{Y}) = C\mathbf{Y}$, *with prediction-error second moment matrices* $v_{\mathbf{X}|\mathbf{Y}}$ *and* $v_{\mathbf{Z}|\mathbf{Y}}$ *respectively, then the best linear predictor of* $\mathbf{X}$ *in terms of* $\mathbf{Y}$ *and* $\mathbf{Z}$ *is*

$$\hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z}) = A_1\mathbf{Y} + A_2\mathbf{Z}, \tag{2.8}$$

where

$$A_2 =< \mathbf{X} - B\mathbf{Y}, \mathbf{Z} > v_{\mathbf{Z}|\mathbf{Y}}^{-1}, \tag{2.9}$$

$$A_1 = B - A_2C \tag{2.10}$$

and $v_{\mathbf{Z}|\mathbf{Y}}^{-1}$ *is any generalized inverse of* $v_{\mathbf{Z}|\mathbf{Y}}$. *The corresponding second moment matrix of the prediction errors is*

$$v_{\mathbf{X}|\mathbf{Y},\mathbf{Z}} = v_{\mathbf{X}|\mathbf{Y}} - A_2 v_{\mathbf{Z}|\mathbf{Y}} A_2'. \tag{2.11}$$

Proof

The best linear predictor of $\mathbf{X}$ in terms of $\mathbf{Y}$ and $\mathbf{Z}$ can be expressed as

$$\hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z}) = \hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y})) = \hat{\mathbf{X}}(\mathbf{Y}) + \hat{\mathbf{X}}(\mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y})) \tag{2.12}$$

since $\mathbf{Y}$ and $\mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y})$ are orthogonal. But

$$\hat{\mathbf{X}}(\mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y})) = A_2(\mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}))$$

where A_2 satisfies

$$A_2 < \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}),\ \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}) >=< \mathbf{X},\ \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}) >,$$

i.e.

$$A_2 =< \mathbf{X},\ \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}) > v_{\mathbf{Z}|\mathbf{Y}}^{-1}$$

$$=< \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}),\ \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}) > v_{\mathbf{Z}|\mathbf{Y}}^{-1}$$

$$=< \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}),\ \mathbf{Z} > v_{\mathbf{Z}|\mathbf{Y}}^{-1},$$

where $v_{\mathbf{Z}|\mathbf{Y}}^{-1}$ is any generalized inverse of $v_{\mathbf{Z}|\mathbf{Y}}$. Substituting for $\hat{\mathbf{X}}(\mathbf{Y})$ and $\hat{\mathbf{X}}(\mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}))$ in (2.12) then gives

$$\hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z}) = B\mathbf{Y} + A_2(\mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y})), \tag{2.13}$$

where A_2 is defined by (2.9). Replacing $\hat{\mathbf{Z}}(\mathbf{Y})$ by $C\mathbf{Y}$ gives the required result (2.8). To determine $v_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$ we rewrite (2.13) in the form

$$\mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}) = (\mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z})) + A_2(\mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y})).$$

Using the orthogonality of the two terms on the right-hand side, we obtain

$$v_{\mathbf{X}|\mathbf{Y}} =< \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}),\ \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}) >= v_{\mathbf{X}|\mathbf{Y},\mathbf{Z}} + A_2 v_{\mathbf{Z}|\mathbf{Y}} A_2',$$

which is the result (2.11). □

Corollary 1

Under the conditions of Theorem 1,

$$\hat{\mathbf{Z}}(\mathbf{Y}, \mathbf{X}) = D_1\mathbf{Y} + D_2\mathbf{X}, \tag{2.14}$$

where

$$D_2 =< \mathbf{Z} - C\mathbf{Y},\ \mathbf{X} > v_{\mathbf{X}|\mathbf{Y}}^{-1}, \tag{2.15}$$

$$D_1 = C - D_2 B, \tag{2.16}$$

$$v_{\mathbf{Z}|\mathbf{X},\mathbf{Y}} = v_{\mathbf{Z}|\mathbf{Y}} - D_2' v_{\mathbf{X}|\mathbf{Y}} D_2, \tag{2.17}$$

and $v_{\mathbf{X}|\mathbf{Y}}^{-1}$ is any generalized inverse of $v_{\mathbf{X}|\mathbf{Y}}$. In addition we can write

$$v_{\mathbf{X}|\mathbf{Y},\mathbf{Z}} = v_{\mathbf{X}|\mathbf{Y}} - A_2 D_2 v_{\mathbf{X}|\mathbf{Y}} \tag{2.18}$$

and

$$v_{\mathbf{Z}|\mathbf{X},\mathbf{Y}} = v_{\mathbf{Z}|\mathbf{Y}} - D_2 A_2 v_{\mathbf{Z}|\mathbf{Y}}. \tag{2.19}$$

Proof

Equations (2.14–2.17) are obtained immediately from the theorem by interchanging the roles of $\mathbf{X}$ and $\mathbf{Z}$. Equation (2.18) follows from (2.11) and the relation

$$D_2 v_{\mathbf{X}|\mathbf{Y}} =< \mathbf{Z} - C\mathbf{Y},\ \mathbf{X} >=< \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}),\ \hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z}) >= v_{\mathbf{Z}|\mathbf{Y}} A_2',$$

which follows from (2.15) and (2.13). Equation (2.19) is obtained in exactly the same way by interchanging $\mathbf{X}$ with $\mathbf{Z}$ and D_2 with A_2. □

Remark 1. If $\mathbf{X}$ and $\mathbf{Y}$ have zero mean vectors then $< \mathbf{X},\ \mathbf{Y} >$ is the matrix whose (i,j) element is the covariance of the i^{th}component of $\mathbf{X}$ and the j^{th} component of $\mathbf{Y}$. In the non-zero mean case, we usually wish to compute $\hat{\mathbf{X}}(\mathbf{1},\mathbf{Y},\mathbf{Z})$ where $\mathbf{1}$ is a vector of ones. This is easily reduced to a zero-mean problem by observing that

$$\hat{\mathbf{X}}(\mathbf{1},\mathbf{Y},\mathbf{Z}) = E\mathbf{X} + \hat{\mathbf{X}}^*(\mathbf{Y} - E\mathbf{Y}, \mathbf{Z} - E\mathbf{Z}),$$

where $\mathbf{X}^* = \mathbf{X} - E\mathbf{X}$.

Remark 2. The predictors $\hat{\mathbf{X}}(\mathbf{Y},\mathbf{Z})$, $\hat{\mathbf{Z}}(\mathbf{Y},\mathbf{X})$, $\hat{\mathbf{X}}(\mathbf{Y})$ and $\hat{\mathbf{Z}}(\mathbf{Y})$ are all uniquely defined. However the coefficients A_1, A_2, B, C, D_1 and D_2 may not be uniquely defined since the matrices $v_{\mathbf{X}|\mathbf{Y}}$ and $v_{\mathbf{Z}|\mathbf{Y}}$ may be singular. The coefficients will be unique if there is no linear relation between any of the components of $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, i.e. if

$$\text{Var}[(\mathbf{X}', \mathbf{Y}', \mathbf{Z}')\mathbf{d}] \neq 0 \quad \text{for all } \mathbf{d} \neq \mathbf{0}.$$

Remark 3. In the course of the proof of Theorem 1 we showed that

$$A_2 = < \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}),\ \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}) > v^{-1}_{\mathbf{Z}|\mathbf{Y}} \tag{2.20}$$

and interchanging the roles of $\mathbf{X}$ and $\mathbf{Z}$, we see also that

$$D_2 = < \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}),\ \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}) > v^{-1}_{\mathbf{X}|\mathbf{Y}}. \tag{2.21}$$

Equation (2.20) states that A_2 is the partial regression coefficient of $\mathbf{X}$ onto $\mathbf{Z}$ given $\mathbf{Y}$. This interpretation is also evident from (2.13). Similarly D_2 is the partial regression coefficient of $\mathbf{Z}$ onto $\mathbf{X}$ given $\mathbf{Y}$.

Remark 4. The conclusions of Theorem 1 can also be derived by direct algebraic manipulation of equations (2.1) to (2.6), showing that if equations (2.1) to (2.6) hold then so do equations (2.9) to (2.11) for some generalized inverse $v^{-1}_{\mathbf{Z}|\mathbf{Y}}$ of $v_{\mathbf{Z}|\mathbf{Y}}$. (*Any* generalized inverse of $v_{\mathbf{Z}|\mathbf{Y}}$ in (2.9) will however give the same linear predictor $A_1\mathbf{Y} + A_2\mathbf{Z}$.) This means that equations (2.9) to (2.11) can be used to solve the equations obtained from (2.1) when sample estimates are substituted for the matrices $E(\mathbf{YY}')$, $E(\mathbf{YZ}')$, $E(\mathbf{ZY}')$, $E(\mathbf{ZZ}')$, $E(\mathbf{YX}')$ and $E(\mathbf{ZX}')$. In Section 3 we shall discuss the solution of these equations from a different point of view using the following results, the proofs of which are identical to those of Theorem 1 and Corollary 1.

Theorem 2

Let $\mathbf{X}$, $\mathbf{Y}$ *and* $\mathbf{Z}$ *be finite-dimensional column vectors, all of whose components are elements of the same inner-product space* S. *For any two such vectors, e.g.* $\mathbf{X}$ *and* $\mathbf{Y}$, *we define the matrix of inner products,*

$$< \mathbf{X}, \mathbf{Y} > = [< X_i, Y_j >]_{i,j},$$

where $< X_i, Y_j >$ *is the inner product of the components* X_i *of* $\mathbf{X}$ *and* Y_j *of* $\mathbf{Y}$. *The projection* $\hat{\mathbf{X}}(\mathbf{Y},\mathbf{Z})$ *of* $\mathbf{X}$ *onto the span of* $\mathbf{Y}$ *and* $\mathbf{Z}$ *is defined to be the linear combination*

(with matrix coefficients) of $\mathbf{Y}$ *and* $\mathbf{Z}$*, whose components each have minimum mean-square distance from the corresponding component of* $\mathbf{X}$*. The corresponding squared-error matrix is defined to be*

$$v_{\mathbf{X}|\mathbf{Y},\mathbf{Z}} = < \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y},\mathbf{Z}), \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y},\mathbf{Z}) > .$$

If $\hat{\mathbf{X}}(\mathbf{Y}) = B\mathbf{Y}$ *and* $\hat{\mathbf{Z}}(\mathbf{Y}) = C\mathbf{Y}$ *are the projections of* $\mathbf{X}$ *and* $\mathbf{Z}$ *onto the span of* $\mathbf{Y}$*, with corresponding squared-error matrices* $v_{\mathbf{X}|\mathbf{Y}}$ *and* $v_{\mathbf{Z}|\mathbf{Y}}$ *respectively, then* $\hat{\mathbf{X}}(\mathbf{Y},\mathbf{Z})$ *and* $v_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$ *satisfy the same equations (2.8)–(2.11) as in Theorem 1.*

Corollary 2

Equations (2.14)–(2.21) remain valid in the context of Theorem 2.

Remark 6. Although we have not found references to Theorems 1 and 2 in the forms stated above, the underlying idea is not new. It has been used for example in the context of hierarchical regression. Thus if $\mathbf{X}$ is regressed linearly on $\mathbf{Y}$ and then a new regressor $\mathbf{Z}$ is introduced, Theorem 1 gives the updating equations in Theorem 3.7(i) and (ii) of Seber (1977). (We are grateful to Lutz Dümbgen for this comment.) However it seems not to have been widely exploited in time series analysis in spite of the fact that its power and simplicity lead to very simple derivations of well-known results as illustrated below.

The idea in all of the applications is to compute predictors recursively by making use of the solutions of simpler prediction problems.

3 The Durbin-Levinson algorithm

In this section we show how the Durbin-Levinson algorithm follows at once from Theorem 1 and derive and discuss the corresponding empirical version used in the Yule-Walker estimation of autoregressive coefficients.

Suppose that $\{X_t, t = 0, \pm 1, \ldots\}$ is a zero-mean stationary process with $E(X_{t+u}X_t) = \gamma(u)$, $t, u \in \{0, \pm 1, \ldots\}$. We restrict attention to the zero-mean case since corresponding results for the non-zero mean case are easily derived using Remark 1 of Section 2.

Algorithm 1 (Durbin-Levinson)

Let $\{X_t\}$ *be a zero-mean stationary process and let*

$$\hat{X}_{n+1}(m) = \sum_{j=1}^{m} \phi_m(j) X_{n+1-j} \tag{3.1}$$

be the best linear 1-step predictor of X_{n+1} *given* $X_n, \ldots, X_{n+1-m}$*, with corresponding mean squared error* v_m*. Then the vectors* $\boldsymbol{\phi}_m = (\phi_m(1), \ldots, \phi_m(m))'$ *and mean squared errors*

v_m, $m = 1, 2, \ldots$, *satisfy the following recursions.*

$$\begin{aligned}
\phi_{m+1}(m+1) &= [\gamma(m+1) - \textstyle\sum_{j=1}^{m} \phi_m(j)\gamma(m+1-j)]v_m^{-1}, \\
\phi_{m+1}(j) &= \phi_m(j) - \phi_{m+1}(m+1)\phi_m(m+1-j), \quad j = 1, \ldots, m, \\
v_{m+1} &= (1 - \phi_{m+1}^2(m+1))v_m,
\end{aligned}$$

where the generalized inverse $v_m^{-1} = 0$ *if* $v_m = 0$. *The initial conditions are* $v_0 = \gamma(0)$ *and* $\phi_1(1) = \gamma(1)/\gamma(0)$.

Proof

Apply Theorem 1 with

$$\mathbf{X} = X_{n+1}, \quad \mathbf{Y} = (X_n, X_{n-1}, \ldots, X_{n+1-m})' \text{ and } \mathbf{Z} = X_{n-m},$$

noting that

$$B = (\phi_m(1), \ldots, \phi_m(m)), \quad C = (\phi_m(m), \ldots, \phi_m(1)),$$

and

$$v_{\mathbf{X}|\mathbf{Y}} = v_{\mathbf{Z}|\mathbf{Y}} = v_m. \qquad \square$$

Remark 1. The Whittle multivariate version of the Durbin-Levinson algorithm can be derived by much the same argument from Theorem 1 and Corollary 1. We shall study algorithms for multivariate processes in a forthcoming paper.

Remark 2. (Yule-Walker estimation/data tapers) Suppose we are given observations $x_1, \ldots, x_T$ of a zero-mean stationary time series $\{X_t\}$. For any $m < T$, the Yule-Walker AR(m) model for the series is obtained by substituting the sample covariances $\hat{\gamma}(u) = \frac{1}{T}\sum_{t=1}^{T-u} x_t x_{t+u}$ for the covariances $\gamma(u)$, $u = 0, \ldots, m$, in equations (1.2) and (1.3) and solving for $\phi_m(1), \ldots, \phi_m(m)$ and v_m. If we denote the solution by $\hat{\boldsymbol{\phi}}_m = (\hat{\phi}_m(1), \ldots, \hat{\phi}_m(m))'$ and $\hat{v}_m$, then the fitted model is

$$X_t = \hat{\phi}_m(1)X_{t-1} + \cdots + \hat{\phi}_m(m)X_{t-m} + Z_t, \quad \{Z_t\} \sim \text{WN}(0, \hat{v}_m),$$

where (cf. (1.2) and (1.3))

$$\hat{\Gamma}_m \hat{\boldsymbol{\phi}}_m = \hat{\boldsymbol{\gamma}}_m \tag{3.2}$$

and

$$\hat{v}_m = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}_m' \hat{\boldsymbol{\gamma}}_m'. \tag{3.3}$$

By Remark 4 of Section 2 we can determine the fitted model using Algorithm 1 with covariances replaced by the corresponding sample covariances. A more illuminating interpretation of Yule-Walker estimation however is obtained as follows. We extend x_t by defining $x_t = 0$ for $t < 0$ or $t > T$. With a slight abuse of notation (since it does not depend on t) we then write $\mathbf{x}_t$ for the *sequence* $\{x_t, t = 0, \pm 1, \ldots\}$, that is

$$\mathbf{x}_t = (\ldots, 0, x_1, \ldots, x_T, 0, \ldots). \tag{3.4}$$

The shifted sequence $\{x_{t-j}, t = 0, \pm 1, \ldots\}$ will be denoted $\mathbf{x}_{t-j}$. The space $\mathcal{S}$ of real sequences with finitely many non-zero components becomes an inner-product space if we define, for any $\mathbf{u}, \mathbf{v} \in \mathcal{S}$,

$$< \mathbf{u}, \mathbf{v} >= T^{-1} \sum_t u_t v_t. \tag{3.5}$$

With this inner product we see that

$$< \mathbf{x}_{t+u}, \mathbf{x}_t >= \hat{\gamma}(u) = \frac{1}{T} \sum_{t=1}^{T-h} x_t x_{t+u} \tag{3.6}$$

and hence that the sample Yule-Walker equations (3.2) and (3.3) are precisely the equations determining the projection

$$\hat{\mathbf{x}}_{n+1}(m) = \sum_{j=1}^{m} \hat{\phi}_m(j) \mathbf{x}_{n+1-j} \tag{3.7}$$

of the sequence $\mathbf{x}_{n+1}$ onto the set of linear combinations (with real-valued scalar coefficients) of the sequences $\mathbf{x}_{n+1-j}$, $j = 1, \ldots, m$. (Analogously equations (1.2) and (1.3) determine the projection of the random variable X_{n+1} onto the set of linear combinations of X_{n+1-j}, $j = 1, \ldots, m$, under the inner product $< U, V >= E(UV')$.) If we define the sequence $\mathbf{x}_t$ by

$$\mathbf{x}_t = (\ldots, 0, h_1 x_1, \ldots, h_T x_T, 0, \ldots) \tag{3.8}$$

where h_t is a data taper with $\sum_{t=1}^{T} h_t^2 = T$ (usually one takes $h_t = h\left(\frac{t-1/2}{T}\right) \cdot c_T$ where $c_T = T / \sum_{t=1}^{T} h\left(\frac{t-1/2}{T}\right)^2$ and $h : [0,1] \to [0, \infty]$ is a smooth function with $h(0) = h(1) = 0$), then we obtain as the solution of the same projection problem again the equation (3.7) where now the coefficients $\hat{\phi}_m(j)$ are the solutions of the tapered Yule-Walker equations, that is of (3.2) and (3.3) with the tapered covariances

$$\hat{\gamma}(u) =< \mathbf{x}_{t+u}, \mathbf{x}_t >= \frac{1}{\sum_{t=1}^{T} h_t^2} \sum_{t=1}^{T-u} h_t x_t h_{t+u} x_{t+u}. \tag{3.9}$$

The tapered Yule-Walker estimates have much better small-sample properties (Dahlhaus, 1988). The above arguments offer an explanation for this improvement: In (3.7) all components of $\mathbf{x}_{n+1}$ are "predicted" with the same $\hat{\phi}_m(j)$ from the corresponding components of $\mathbf{x}_{n+1-j}$ leading to problems at the edges (for example when 0 is "predicted" from $x_T, \ldots, x_{T+1-m}$). This problem is relaxed by using an adequate taper. By using Theorem 2 we now get the empirical version of the Durbin-Levinson algorithm.

Algorithm 2 (Empirical Durbin-Levinson)

With $\{\mathbf{x}_t\}$ and $\hat{\gamma}(u)$ defined as in (3.4) and (3.6) or as in (3.8) and (3.9) respectively and inner products as in (3.5), let

$$\hat{\mathbf{x}}_{n+1}(m) = \sum_{j=1}^{m} \hat{\phi}_m(j) \mathbf{x}_{n+1-j} \tag{3.10}$$

be the projection of $\mathbf{x}_{n+1}$ *on the span of* $\mathbf{x}_n, \ldots, \mathbf{x}_{n+1-m}$*, with squared error* $\hat{v}_m =< \mathbf{x} - \hat{\mathbf{x}}_{n+1}(m), \mathbf{x} - \hat{\mathbf{x}}_{n+1}(m) >$*. Then the vectors* $\hat{\boldsymbol{\phi}}_m = (\hat{\phi}_m(1), \ldots, \hat{\phi}_m(m))'$ *and squared errors* $\hat{v}_m$*,* $m = 1, 2, \ldots$*, satisfy the recursions,*

$$\begin{aligned} \hat{\phi}_{m+1}(m+1) &= [\hat{\gamma}(m+1) - \textstyle\sum_{j=1}^{m} \hat{\phi}_m(j)\hat{\gamma}(m+1-j)]\hat{v}_m^{-1}, \\ \hat{\phi}_{m+1}(j) &= \hat{\phi}_m(j) - \hat{\phi}_{m+1}(m+1)\hat{\phi}_m(m+1-j), \quad j = 1, \ldots, m, \end{aligned}$$

and

$$\hat{v}_{m+1} = (1 - \hat{\phi}_{m+1}^2(m+1))\hat{v}_m.$$

The initial conditions are $\hat{v}_0 = \hat{\gamma}(0)$ *and* $\hat{\phi}_1(1) = \hat{\gamma}(1)/\hat{\gamma}(0)$*.*

Proof

Apply Theorem 2 with

$$\mathbf{X} = \mathbf{x}_{n+1}, \quad \mathbf{Y} = (\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n+1-m})' \text{ and } \mathbf{Z} = \mathbf{x}_{n-m},$$

noting that

$$B = (\hat{\phi}_m(1), \ldots, \hat{\phi}_m(m)), \quad C = (\hat{\phi}_m(m), \ldots, \hat{\phi}_m(1)),$$

and

$$v_{\mathbf{X}|\mathbf{Y}} = v_{\mathbf{Z}|\mathbf{Y}} = \hat{v}_m. \qquad \square$$

4 Subset prediction and autoregression

We next consider a generalization of the Durbin-Levinson algorithm for calculating the best linear predictor of X_{n+1} in terms of a *subset* of the observations, $X_1, \ldots, X_n$. The same algorithm can also be applied to fit a subset autoregression to observations of a stationary time series and has been developed in this context by Penm and Terrell (1982).

A simple but important special case is h-step prediction, which requires the determination of the best predictor of X_{n+1} in terms of $\{X_{n+1-h}, X_{n+1-h-1}, \ldots\}$. More generally we are concerned with prediction in terms of an arbitrarily specified subset of $\{X_n, X_{n-1}, \ldots\}$. It is straightforward enough to write down the equations analogous to (1.2), which determine the best linear predictor of X_{n+1} in terms of $\{X_{n+1-k_1}, \ldots, X_{n+1-k_m}\}$. Our aim here however is to use Theorem 1 to derive a recursive algorithm for their solution. A more complicated derivation of the analogue of this algorithm in the multivariate case was given by Penm and Terrell (1982), who used it for the fitting of subset autoregressions. Note that Penm and Terrell have formulated their algorithm in terms of the missing lags while we formulate it in terms of the present lags. Before stating the algorithm we introduce some simplifying notation.

Suppose that $K = \{k_1, \ldots, k_m\} \subset \{1, \ldots, n\}$, where $1 \leq k_1 < k_2 < \cdots < k_m \leq n$. We can then write the best linear predictor of X_{n+1} in terms of $\{X_{n+1-j}, j \in K\}$ as

$$\hat{X}_{n+1}(K) = \sum_{j \in K} \phi_K(j) X_{n+1-j} \tag{4.1}$$

for some coefficients $\phi_K(j)$, $j \in K$. Letting

$$\mathbf{X}_n(K) = (X_{n+1-k_1}, \ldots, X_{n+1-k_m})'$$

and

$$\boldsymbol{\phi}_K = (\phi_K(k_1), \ldots, \phi_K(k_m))',$$

we can rewrite (4.1) in the vector form,

$$\hat{X}_{n+1}(K) = \boldsymbol{\phi}_K' \mathbf{X}_n(K). \tag{4.2}$$

If L is any non-empty subset $\{l_1, \ldots, l_r\}$ of K with $l_1 < l_2 < \cdots < l_r$, we define

$$\boldsymbol{\phi}_K(L) = (\phi_K(l_1), \ldots, \phi_K(l_r))'$$

and use the notation $\overline{\boldsymbol{\phi}}_K$ and $\overline{\boldsymbol{\phi}}_K(L)$ for the vectors $\boldsymbol{\phi}_K$ and $\boldsymbol{\phi}_K(L)$ with the order of their components reversed. If we define

$$\boldsymbol{\gamma}_K = (\gamma(k_1), \ldots, \gamma(k_m))'$$

and

$$\Gamma_K = E(\mathbf{X}_n(K)\mathbf{X}_n(K)')$$

then the coefficient vector $\boldsymbol{\phi}_K$ in (4.2) satisfies

$$\Gamma_K \boldsymbol{\phi}_K = \boldsymbol{\gamma}_K \tag{4.3}$$

and the mean squared error of the best predictor $\boldsymbol{\phi}_K' \mathbf{X}_n(K)$ is

$$v_K = \gamma(0) - \boldsymbol{\phi}_K' \boldsymbol{\gamma}_K. \tag{4.4}$$

We now define the sets of indices

$$J = \{k_1, \ldots, k_{m-1}\} \text{ and } J^* = \{k_m - k_{m-1}, \ldots, k_m - k_1\}$$

and apply Theorem 1 with

$$\mathbf{X} = X_{n+1}, \quad \mathbf{Y} = \mathbf{X}_n(J)', \quad \mathbf{Z} = X_{n+1-k_m},$$

$$B = \boldsymbol{\phi}_J', \; C = \overline{\boldsymbol{\phi}}_{J^*}', \; v_{\mathbf{X}|\mathbf{Y}} = v_J \text{ and } v_{\mathbf{Z}|\mathbf{Y}} = v_{J^*}.$$

This immediately gives the following algorithm expressing the best predictor $\hat{X}_{n+1}(K)$ and its mean squared error v_K in terms of $\boldsymbol{\phi}_J$, $\boldsymbol{\phi}_{J^*}$, v_J, v_{J^*}, $\boldsymbol{\gamma}_{J^*}$ and $\gamma(k_m)$. It is the univariate version of the subset Whittle algorithm derived by Penm and Terrell (1982).

Algorithm 3

If $\{X_t\}$ is a zero-mean stationary process with covariance function γ and if, in the notation introduced above, $\hat{X}_{n+1}(J) = \boldsymbol{\phi}_J' \mathbf{X}_n(J)$ and $\hat{X}_{n+1}(J^*) = \boldsymbol{\phi}_{J^*}' \mathbf{X}_n(J^*)$, with mean squared errors v_J and v_{J^*} respectively, then $\hat{X}_{n+1}(K) = \boldsymbol{\phi}_K' \mathbf{X}_n(K)$ where

$$\phi_K(k_m) = (\gamma(k_m) - \boldsymbol{\phi}_J' \overline{\boldsymbol{\gamma}}_{J^*}) v_{J^*}^{-1}, \tag{4.5}$$

$$\phi_K(J) = \phi_J - \phi_K(k_m)\overline{\phi}_{J^*} \tag{4.6}$$

and the mean squared error of $\hat{X}_{n+1}(K)$ *is*

$$v_K = v_J - \phi_K(k_m)^2 v_{J^*}, \tag{4.7}$$

where the generalized inverse $v_{J^*}^{-1}$ *is zero if* $v_{J^*} = 0$. *The initial conditions for the recursions are specified in Remark 3 below.*

Remark 1. Algorithm 1 is the special case of Algorithm 3 with $K = \{1, \ldots, m\}$.

Remark 2. The lags in the set J^* enter the recursions because they are the time lags between the observations $X_{n+1-k_1}, \ldots, X_{n+1-k_{m-1}}$ and the *earliest* observation X_{n+1-k_m} just as J is the set of time lags between the *predicted* observation X_{n+1} and the observations $X_{n+1-k_1}, \ldots, X_{n+1-k_{m-1}}$. In the Durbin-Levinson algorithm the sets J and J^* are identical.

Remark 3. In order to apply the algorithm to compute all subset predictors with maximum lag less than or equal to r, we proceed as follows.

1. Compute all predictors based on a single observation, i.e. with $J = \{j\}$, for $1 \leq j \leq r$. For the predictor with $J = \{j\}$ we have

   $$\phi_J = \gamma(j)/\gamma(0) = \rho(j),$$

   where ρ is the correlation function of $\{X_t\}$, and

   $$v_J = \gamma(0)(1 - \rho(j)^2).$$

2. Use Algorithm 3 to compute all predictors based on m observations with maximum lag r from those based on $m-1$ observations with maximum lag r.

3. Continue until reaching the best predictor based on $X_n, \ldots, X_{n+1-r}$. This corresponds to the subset $K = \{1, \ldots, r\}$.

This procedure requires the storage of $\binom{r}{m}$ coefficient vectors and mean squared errors at the end of step 2 in order to take the next step in the iteration. If we are using the empirical version to fit a subset Yule-Walker model as described below in Remark 4, then a "best" model can be selected by minimizing an order-selection criterion such as AIC. As we proceed through the iterations we must then keep track of the current best model and its corresponding AIC value. We should also use tapered covariances (Dahlhaus, 1988) to obtain estimates with better small-sample properties.

Remark 4. (Yule-Walker subset estimation) As in Remark 2 of Section 3, the Yule-Walker *subset* autoregressive model for the data $x_1, \ldots, x_T$ with lags $k_1, \ldots, k_m$ is

$$X_t = \hat{\phi}_K(k_1)X_{t-k_1} + \cdots + \hat{\phi}_K(k_m)X_{t-k_m} + Z_t, \quad \{Z_t\} \sim \mathrm{WN}(0, \hat{v}_K),$$

where the vector $\hat{\boldsymbol{\phi}}_K = (\hat{\phi}_{k_1}, \ldots, \hat{\phi}_{k_m})'$ and white noise variance $\hat{v}_K$ satisfy the equations

$$\hat{\Gamma}_K \hat{\boldsymbol{\phi}}_K = \hat{\boldsymbol{\gamma}}_K \tag{4.8}$$

and

$$\hat{v}_K = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}_K' \hat{\boldsymbol{\gamma}}_K' \tag{4.9}$$

(cf. (4.3) and (4.4)), where $\hat{\Gamma}_K$ and $\hat{\gamma}_K$ are obtained from Γ_K and γ_K respectively by replacing each covariance $\gamma(u)$ by the corresponding sample covariance $\hat{\gamma}(u)$ as given in (3.6) or (3.9). We can therefore determine the fitted model using Algorithm 3 with covariances replaced by the corresponding sample covariances.

By the same argument used in Section 3, we obtain the following subset analogue of Algorithm 2. Like Algorithm 2, it provides us with a direct empirical interpretation of the fitted Yule-Walker model.

Algorithm 4 (Empirical subset algorithm)

With $\mathbf{x}_t$ and $\hat{\gamma}(u)$ defined as in (3.4) and (3.6) or as in (3.8) and (3.9) respectively and with inner products defined as in (3.5), let

$$\hat{\mathbf{x}}_{n+1}(K) = \sum_{j=1}^{m} \hat{\phi}_K(k_j)\mathbf{x}_{n+1-k_j} \tag{4.10}$$

be the projection of $\mathbf{x}_{n+1}$ on the span of $\mathbf{x}_{n+1-k_1}, \ldots, \mathbf{x}_{n+1-k_m}$. Then the vectors $\hat{\boldsymbol{\phi}}_K = (\hat{\phi}_K(k_1), \ldots, \hat{\phi}_K(k_m))'$ and squared errors $\hat{v}_K = < \mathbf{x} - \hat{\mathbf{x}}_{n+1}(K), \mathbf{x} - \hat{\mathbf{x}}_{n+1}(K) >$ satisfy the recursions,

$$\hat{\phi}_K(k_m) = (\hat{\gamma}(k_m) - \hat{\boldsymbol{\phi}}_J' \overline{\hat{\boldsymbol{\gamma}}}_{J^*})\hat{v}_{J^*}^{-1}, \tag{4.11}$$

$$\hat{\boldsymbol{\phi}}_K(J) = \hat{\boldsymbol{\phi}}_J - \hat{\phi}_K(k_m)\overline{\hat{\boldsymbol{\phi}}}_{J^*} \tag{4.12}$$

and

$$\hat{v}_K = \hat{v}_J - \hat{\phi}_K(k_m)^2 \hat{v}_{J^*}, \tag{4.13}$$

where the index sets J and J^* are defined as in Algorithm 3 and the generalized inverse $\hat{v}_{J^*}^{-1}$ is zero if $\hat{v}_{J^*} = 0$. The initial conditions are as in Remark 3 with $\hat{\gamma}(j)$ replacing $\gamma(j)$ and $\hat{\rho}(j) = \hat{\gamma}(j)/\hat{\gamma}(0)$ replacing $\rho(j)$.

Remark 5. (Causality) It is well-known (see e.g. Brockwell and Davis, 1991) that the solution of the empirical Yule-Walker equations (3.2) always gives a causal fitted model, i.e. the roots of the equation $1 - \hat{\phi}_{m1}z - \cdots - \hat{\phi}_{mm}z^m = 0$ lie outside the unit circle in the complex plane. This is an important result since the derivation of the Yule-Walker equations (1.2) requires stationarity of $\{X_t\}$ and zero correlation between Z_t and $X_s, s \leq t$. These conditions are not satisfied if the model is not causal. Unfortunately the solution of the empirical Yule-Walker subset equations (4.8) does *not* necessarily give a causal model. For example, suppose we take a sample from the stationary causal ARMA process defined by $X_t - X_{t-1} + 0.5X_{t-2} = Z_t + Z_{t-1}$, where $\{Z_t\}$ is white noise, and suppose that the sample autocorrrelations at lags 1,2 and 3 happen to coincide with those of the generating ARMA process, i.e. $\hat{\rho}(1) = 3/4$, $\hat{\rho}(2) = 1/4$ and $\hat{\rho}(3) = -1/8$. If we then attempt to fit a subset AR(3) model to the sampled data with non-zero coefficients at lags 2 and 3 only, we find

that the solution of (4.8) with $K = \{2, 3\}$ is $\hat{\phi}_K(2) = \frac{11}{14}$ and $\hat{\phi}_K(3) = -\frac{5}{7}$. However the subset model

$$X_t - \frac{11}{14}X_{t-2} + \frac{5}{7}X_{t-3} = Z_t, \ \{Z_t\} \sim WN(0, \sigma^2) \tag{4.14}$$

is not causal. The autocorrelations $\rho(1)$, $\rho(2)$ and $\rho(3)$ of the stationary solution of (4.14) are *not* equal to 3/4, 1/4 and $-1/8$ respectively. This is because (unlike $\hat{\rho}(1)$, $\hat{\rho}(2)$ and $\hat{\rho}(3)$) they do *not* satisfy the Yule-Walker equations,

$$\begin{bmatrix} 1 & \rho(1) \\ \rho(1) & 1 \end{bmatrix} \begin{bmatrix} 11/14 \\ -5/7 \end{bmatrix} = \begin{bmatrix} \rho(2) \\ \rho(3) \end{bmatrix}.$$

This is because of the non-causality of the model (4.14). Consequently (4.14) cannot be interpreted as a model for the the data whose correlations at lags 1, 2 and 3 match the corresponding sample autocorrelations. The same problem of interpretation arises whenever the equations (4.8) give a non-causal model. It simply means that the data cannot be well-fitted by a subset autoregression with the specified set of lags, K. In spite of this shortcoming however, we can still correctly infer from the fitted model (4.14) (assuming that the sample autocorrelations $\hat{\rho}(1)$, $\hat{\rho}(2)$ and $\hat{\rho}(3)$ are accurate estimates of the true correlations) that the best linear predictor of X_t in terms of X_{t-2} and X_{t-3} is

$$\hat{X}_t = \frac{11}{14}X_{t-2} - \frac{5}{7}X_{t-3}$$

Subset modelling is thus an appropriate method of generating best linear predictors in terms of subsets of the data (based on estimated covariances) in spite of the fact that the subset model may not be an appropriate model for describing the entire series. It simply gives the best linear predictor within a prescribed class. A class of predictors which is of particular interest is that of the h-step predictors. The remainder of this section is devoted to this class, specifically to the recursive determination of h-step predictors for a stationary process with known covariance function.

Instead of calculating best subset predictors for *all* subsets with maximum lag r, it is frequently of interest to calculate the best subset predictor for one particular subset or for a family of subsets. In such cases the number of steps required in the application of Algorithm 3 may be reduced considerably. An example is the calculation of the best linear h-step predictor of X_{n+1} in terms of $X_{n+1-h}, \ldots, X_{n+1-h-m}$. This is just the predictor $\hat{X}_{n+1}(K)$ with $K = \{h, h+1, \ldots, h+m\}$. In order to determine $X_{n+1}(K)$ from Algorithm 3, it suffices to determine $\boldsymbol{\phi}_J$, $\boldsymbol{\phi}_{J^*}$ v_J and v_{J^*}, where

$$J = \{h, h+1, \ldots, h+m-1\} \ \text{ and } \ J^* = \{1, 2, \ldots, m\}.$$

To determine these quantities from the algorithm we need only $\boldsymbol{\phi}_J$, $\boldsymbol{\phi}_{J^*}$ v_J and v_{J^*}, where now

$$J = \{h, h+1, \ldots, h+m-2\} \ \text{ and } \ J^* = \{1, 2, \ldots, m-1\}.$$

Continuing to work backwards we see that we can begin the iterations with the predictors based on $\{h\}$ and $\{1\}$ only, and then apply Algorithm 3 one step at a time, computing two

sets of coefficients and two mean squared errors at each step, arriving after m stages at the required predictor $\hat{X}_{n+1}(K)$. We make this process more explicit in the following algorithm.

Algorithm 5 (h-step prediction)

Let $\{X_t\}$ be a zero-mean stationary process and let

$$\hat{X}_{n+h}^{(h)}(m) = \sum_{j=1}^{m} \phi_m^{(h)}(j) X_{n+1-j} \tag{4.15}$$

be the best linear predictor of X_{n+h} given $X_n, \ldots, X_{n+1-m}$, with corresponding mean squared error $v_m^{(h)}$. Then the best linear predictor of X_{n+h} given $X_n, \ldots, X_{n-m}$ can be expressed as

$$\hat{X}_{n+h}^{(h)}(m+1) = \sum_{j=1}^{m+1} \phi_{m+1}^{(h)}(j) X_{n+1-j},$$

where the coefficients and mean squared errors satisfy

$$\phi_{m+1}^{(h)}(m+1) = [\gamma(m+h) - \sum_{j=1}^{m} \phi_m^{(h)}(j)\gamma(m+1-j)]\, v_m^{-1}, \tag{4.16}$$

$$\phi_{m+1}^{(h)}(j) = \phi_m^{(h)}(j) - \phi_{m+1}^{(h)}(m+1)\phi_m(m+1-j), \quad j = 1, \ldots, m, \tag{4.17}$$

$$v_{m+1}^{(h)} = v_m^{(h)} - \phi_{m+1}^{(h)}(m+1)^2 v_m, \tag{4.18}$$

with initial conditions,

$$\phi_1^{(h)}(1) = \rho(h) \quad \text{and} \quad v_1^{(h)} = \gamma(0)(1-\rho(h)^2).$$

The function ρ is the correlation function of $\{X_t\}$ and the one-step prediction coefficients $\phi_m(j)$, $j = 1, \ldots, m$, and mean-squared errors v_m are defined as in Algorithm 1.

Proof With the above definition of K, J and J^* Algorithm 5 immediately follows from Algorithm 3. Note that for simplicity of notation we have shifted all variables, in particular $\phi_m^{(h)}(j) = \phi_j(j+h-1)\,(j = 1, \ldots, m)$ and $\phi_{m+1}^{(h)}(j) = \phi_K(j+h-1)\,(j = 1, \ldots, m+1)$. □

Remark 6. Note that the recursive calculations for determining $\phi_{m+1}^{(h)}$ and ϕ_{m+1} can be performed in parallel. It also follows from Corollary 1 that the mean squared errors $v_m^{(h)}$ satisfy the recursions,

$$v_{m+1}^{(h)} = (1 - \phi_{m+1}^{(h)}(m+1)\phi_{m+1}(m+1))v_m^{(h)}. \qquad \square$$

The above algorithm gives a recursion for getting the coefficients $\phi_{m+1}^{(h)}(j)$ from the $\phi_m^{(h)}(j)$. We now prove another algorithm which allows the calculation of the coefficients $\phi_m^{(h+1)}(j)$ from the $\phi_m^{(h)}(j)$ - an algorithm which for practical purposes is of even higher importance. This algorithm does not follow directly from Algorithm 3 but with repeated application of the basic algorithm from Theorem 1.

Algorithm 6 (h-step prediction)

Let $\{X_t\}$ be a zero-mean stationary process and let $\phi_m^{(h)}(j)$ $(j = 1, \ldots, m)$ be the coefficients of the best linear predictor of X_{n+h} given $X_n, \ldots, X_{n+1-m}$ with mean squared error $v_m^{(h)}$ as given in (4.15). Then the coefficients $\phi_m^{(h+1)}(j)$ $(j = 1, \ldots, m)$ of the best linear predictor of X_{n+h+1} given $X_n, \ldots, X_{n+1-m}$ with mean squared error $v_m^{(h+1)}$ satisfy

$$\phi_m^{(h+1)}(m) = [\gamma(m+h) - \sum_{j=1}^{m-1} \phi_{m-1}(j)\gamma(m+h-j)]v_{m-1}^{-1}, \tag{4.19}$$

$$\phi_m^{(h+1)}(j) = \phi_m^{(h)}(j+1) + \phi_m^{(h)}(1)\phi_m(j) - \phi_m^{(h+1)}(m)\phi_m(m-j),\ j = 1, \ldots, m-1, \tag{4.20}$$

$$v_m^{(h+1)} = v_m^{(h)} + [\phi_m^{(h)}(1)^2 - \phi_m^{(h+1)}(m)^2]v_{m-1}, \tag{4.21}$$

where the starting values $\phi_m^{(1)}(j) = \phi_m(j)$ and $v_m^{(1)} = v_m$ can be calculated with Algorithm 1 or from (1.2) and (1.3) directly (since m is fixed). Furthermore $v_{m-1} = v_m/[1 - \phi_m(m)^2]$ (see Algorithm 1).

Proof Application of Theorem 1 with

$$\mathbf{X} = X_{n+h},\ \mathbf{Y} = (X_{n-1}, \ldots, X_{n+1-m})' \text{ and } \mathbf{Z} = X_n$$

yields

$$\phi_m^{(h)}(j+1) = \phi_{m-1}^{(h+1)}(j) - \phi_m^{(h)}(1)\phi_m(j),\ j = 1, \ldots, m-1 \tag{4.22}$$

and

$$v_m^{(h)} = v_{m-1}^{(h)} - \phi_m^{(h)}(1)^2 v_{m-1}. \tag{4.23}$$

A second application of Theorem 1 and of (2.7) with

$$\mathbf{X} = X_{n+h+1},\ \mathbf{Y} = (X_u, \ldots X_{n+2-m})' \text{ and } \mathbf{Z} = X_{n+1-m}$$

leads to

$$\phi_m^{(h+1)}(m) = < \mathbf{Z} - c\mathbf{Y}, \mathbf{X} > v_{\mathbf{Z}|\mathbf{Y}}^{-1}$$

$$= [\gamma(m+h) - \sum_{j=1}^{m-1} \phi_{m-1}(j)\gamma(m+h-j)]v_{m-1}^{-1} \tag{4.24}$$

$$\phi_m^{(h+1)}(j) = \phi_{m-1}^{(h+1)}(j) - \phi_m^{(h+1)}(m)\phi_m(m-j),\ j = 1, \ldots, m-1 \tag{4.25}$$

$$v_m^{(h+1)} = v_{m-1}^{(h+1)} - \phi_m^{(h+1)}(m)^2 v_{m-1}. \tag{4.26}$$

(The last two equations are (4.17) and (4.18) with h and m replaced by $h+1$ and $m-1$ respectively). (4.22) and (4.25) now give (4.20) while (4.23) and (4.26) give (4.21). □

5 Prediction error representations and Burg algorithms

In this section we derive a different kind of recursion for the best predictors, based on the "forward" and "backward" residuals, which we now define. Let K, J and J^* be the sets of lags defined in Section 4 and suppose that the best linear predictor of X_{n+1} in terms of the components of $\mathbf{X}_J = (X_{n+1-k_1}, \ldots, X_{n+1-k_{m-1}})'$ is $\hat{X}_{n+1}(J) = \sum_{j \in J} \phi_J(j) X_{n+1-j}$.

Then the forward residual at time t for the lag subset J is defined as

$$\varepsilon_J(t) = X_t - \sum_{j \in J} \phi_J(j) X_{t-j}$$

and the backward residual as

$$\eta_{J^*}(t) = X_t - \sum_{j \in J^*} \phi_{J^*}(j) X_{t+j}.$$

The expected squares of these residuals are the mean squared prediction errors v_j and v_j^* respectively from Section 4. Since the process is univariate and stationary, the backward and the forward mean squared prediction errors are the same. The following proposition summarizes some properties of $\varepsilon_J(t)$ and $\eta_{J^*}(t)$ and gives an update relation which leads to algorithms of Burg's type for prediction and modelling.

Proposition 1

Let $K = \{k_1, \ldots, k_m\} \subset \{1, \ldots, n\}$, *where* $k_1 < k_2 < \cdots < k_m$, *and let* $J = \{k_1, \ldots, k_{m-1}\}$, $J^* = \{k_m - k_1, \ldots, k_m - k_{m-1}\}$ *and* $K^* = J^* \cup \{k_m\}$. *Then*

(i) $< \varepsilon_J(t), \varepsilon_J(t) >= v_J$

(ii) $< \eta_{J^*}(t), \eta_{J^*}(t) >= v_{J^*}$

(iii) $< \varepsilon_J(t), \eta_{J^*}(t - k_m) >= v_J \phi_{K^*}(k_m) = \phi_K(k_m) v_{J^*}$

(iv) $\varepsilon_K(t) = \varepsilon_J(t) - \phi_K(k_m) \eta_{J^*}(t - k_m)$

(v) $\eta_{K^*}(t - k_m) = \eta_{J^*}(t - k_m) - \phi_{K^*}(k_m) \varepsilon_J(t)$

Proof

(i) and (ii) are just the definitions of the mean squared prediction errors.

(iii) is a restatement of equations (2.20) and (2.21) with

$$\mathbf{X} = X_t, \quad \mathbf{Y} = (X_{t-k_1}, \ldots, X_{t-k_{m-1}})', \quad \mathbf{Z} = X_{t-k_m},$$

$$B = \boldsymbol{\phi}_J', \quad C = \overline{\boldsymbol{\phi}}_{J^*}', \quad v_{\mathbf{X}|\mathbf{Y}} = v_J \text{ and } v_{\mathbf{Z}|\mathbf{Y}} = v_{J^*},$$

$$A_2 = \phi_K(k_m) \text{ and } D_2 = \phi_{K^*}(k_m).$$

(iv) is obtained by observing that

$$\varepsilon_J(t) - \phi_K(k_m)\eta_{J^*}(t-k_m) = \mathbf{X} - \boldsymbol{\phi}_J'\mathbf{Y} - \phi_K(k_m)(\mathbf{Z} - \overline{\boldsymbol{\phi}}_{J^*}'\mathbf{Y})$$

and noting from (4.6) that $\boldsymbol{\phi}_J' - \phi_K(k_m)\overline{\boldsymbol{\phi}}_{J^*}' = \phi_K(J)'$, whence

$$\varepsilon_J(t) - \phi_K(k_m)\eta_{J^*}(t-k_m) = \mathbf{X} - \boldsymbol{\phi}_K(J)'\mathbf{Y} - \phi_K(k_m)\mathbf{Z} = \varepsilon_K(t).$$

The result (v) is proved in exactly the same way as (iv). □

We now turn to the empirical equivalent of Proposition 1. In the notation of Remark 2 of Section 3, our aim is the recursive determination, based on empirical prediction errors, of $\hat{\boldsymbol{\phi}}_K$ and $\hat{v}_K$ satisfying

$$\hat{\Gamma}_K\hat{\boldsymbol{\phi}}_K = \hat{\boldsymbol{\gamma}}_K \tag{5.1}$$

and

$$\hat{v}_K = \hat{\gamma}(0) - \hat{\boldsymbol{\phi}}_K'\hat{\boldsymbol{\gamma}}_K'. \tag{5.2}$$

By Algorithm 4, the projection $\hat{\mathbf{x}}_{n+1}(K)$ of the sequence $\mathbf{x}_{n+1}$ on the span of $\mathbf{x}_{n+1-j}$, $j \in K$, is then

$$\hat{\mathbf{x}}_{n+1}(K) = \sum_{j\in K}\hat{\phi}_j(K)\mathbf{x}_{n+1-j}, \tag{5.3}$$

with empirical mean squared error

$$< \mathbf{x}_{n+1} - \hat{\mathbf{x}}_{n+1}(K), \mathbf{x}_{n+1} - \hat{\mathbf{x}}_{n+1}(K) >= \frac{1}{T}\sum_t \left(x_{t+1} - \sum_{j\in K}\hat{\phi}_j(K)x_{t+1-j} \right)^2 = \hat{v}_K.$$

The coefficients $\hat{\phi}_j(K)$ and $\hat{v}_K$ are the solutions of the empirical forward prediction problem. It can easily be seen that the empirical backward prediction problem also leads exactly to the same values. The forward and backward prediction error sequences for the lag subset J are now

$$\hat{\epsilon}_J(t) = \mathbf{x}_t - \sum_{j\in J}\hat{\phi}_J(j)\mathbf{x}_{t-j}$$

$$\hat{\eta}_{J^*}(t) = \mathbf{x}_t - \sum_{j\in J^*}\hat{\phi}_{J^*}(j)\mathbf{x}_{t+j}$$

respectively; the corresponding empirical forward and backward mean squared errors are

$$\hat{v}_J =< \hat{\epsilon}_J(t), \hat{\epsilon}_J(t) > \text{ and } \hat{v}_{J^*} =< \hat{\eta}_{J^*}(t), \hat{\eta}_{J^*}(t) > .$$

We can now state the following analogue of Proposition 1.

Proposition 2

Proposition 1 holds if we replace v by $\hat{v}$, ϕ by $\hat{\phi}$, ϵ by $\hat{\epsilon}$, η by $\hat{\eta}$ and define the inner products as in (3.4).

Proof The proof is identical to the proof of Proposition 1, defining

$$\mathbf{X} = \mathbf{x}_t,\ \mathbf{Y} = (\mathbf{x}_{t-k1}, \ldots, \mathbf{x}_{t-k_{m-1}})',\ \mathbf{Z} = \mathbf{x}_{t-k_m},$$

$$B = \hat{\phi}'_J,\ C = \overline{\hat{\phi}}'_{J^*},\ v_{\mathbf{x}|\mathbf{y}} = \hat{v}_{\mathbf{J}} \text{ and } v_{\mathbf{z}|\mathbf{y}} = \hat{v}_{\mathbf{J}^*},$$

and applying Theorem 2 and Corollary 2 with inner products defined as in (3.4) □

With the aid of Proposition 2 we can now state a set of recursions based on empirical prediction errors for solving equations (5.1) and (5.2).

Algorithm 7 (Burg solution of the subset Yule-Walker equations)

A solution of (5.1) is determined by the recursions

$$\hat{\phi}_K(k_m) = \frac{\sum_t \hat{\epsilon}_J(t)\hat{\eta}_{J^*}(t-k_m)}{\sum_t \hat{\eta}_{J^*}(t)^2}, \tag{5.4}$$

$$\hat{\phi}_K(J) = \hat{\phi}_J - \hat{\phi}_K(k_m)\overline{\hat{\phi}}_{J^*}, \tag{5.5}$$

$$\hat{\phi}_{K^*}(k_m) = \frac{\sum_t \hat{\epsilon}_J(t)\hat{\eta}_{J^*}(t-k_m)}{\sum_t \hat{\epsilon}_J(t)^2}, \tag{5.6}$$

$$\hat{\phi}_{K^*}(J^*) = \hat{\phi}_{J^*} - \hat{\phi}_K(k_m)\overline{\hat{\phi}}_J, \tag{5.7}$$

$$\hat{v}_K = [1 - \phi_K(k_m)\phi_{K^*}(k_m)]\hat{v}_J \tag{5.8}$$

$$\hat{v}_{K^*} = [1 - \phi_K(k_m)\phi_{K^*}(k_m)]\hat{v}_{J^*}, \tag{5.9}$$

where the empirical forward and backward residuals satisfy

$$\hat{\epsilon}_K(t) = \hat{\epsilon}_J(t) - \hat{\phi}_K(k_m)\hat{\eta}_{J^*}(t-k_m) \tag{5.10}$$

and

$$\hat{\eta}_{K^*}(t-k_m) = \hat{\eta}_{J^*}(t-k_m)\hat{\phi}_{K^*}(k_m)\hat{\epsilon}_{J^*}(t), \tag{5.11}$$

with initial conditions,

$$\hat{\epsilon}_\phi(t) = \hat{\eta}_\phi(t) = x_t,\ t = 0, \pm 1, \pm 2, \ldots,$$

and

$$\hat{v}_\phi = \frac{1}{T}\sum_{t=1}^{T} x_t^2.$$

The performance of the estimators $\hat{\phi}_K(k_m)$ and $\hat{\phi}_{K^*}(J^*)$ suffers as a result of the poor quality of $\hat{\epsilon}_J(t)$ and $\hat{\eta}_{J^*}(t-k_m)$ as estimators of $\epsilon_J(t)$ and $\eta_{J^*}(t-k_m)$ at the edges of the observation domain (cf. the discussion in Remark 2 of Section 3 where tapering was suggested as a means of improving the estimators). Algorithm 7 suggests the following possibilities for improving the estimates, the third of which is a direct generalization to subset modelling of Burg's algorithm for fitting autoregressions:

(i) Replace the sums over all t in (5.4) and (5.6) by the sums from $t = k_m + 1$ to $t = T$.

(ii) Estimate the correlation between the backward and forward residuals as

$$\hat{\varrho}_{J,J^*} = \frac{\sum_{t=k_m+1}^{T} \hat{\epsilon}_J(t)\hat{\eta}_{J^*}(t-k_m)}{\left(\sum_{t=k_m+1}^{T}, \hat{\epsilon}_J(t)^2\right)^{1/2} \left(\sum_{t=k_m+1}^{T}, \hat{\eta}_{J^*}(t-k_m)^2\right)^{1/2}}$$

and set

$$\begin{aligned} \hat{\phi}_K(k_m) &= \hat{\varrho}_{J,J^*} \left(\frac{\hat{v}_J}{\hat{v}_{J^*}}\right)^{1/2} \\ \hat{\phi}_{K^*}(k_m) &= \hat{\varrho}_{J,J^*} \left(\frac{\hat{v}_{J^*}}{\hat{v}_J}\right)^{1/2} \end{aligned}$$

This is in the spirit of the multivariate Burg-algorithm as given by Morf et al. (1978).

(iii) Minimize the sum of squares of the forward and backward residuals over the range of data values where these are defined. This corresponds to the original idea of Burg (1968) for fitting AR(m)-models. In the context of subset modelling we find from Proposition 1, parts (iii), (iv) and (v) that the Burg estimate of $\phi_K(k_m)$ should minimize

$$S_K = \sum_{t=k_m+1}^{T} \left[(\hat{\epsilon}_J(t) - \hat{\phi}_K(k_m)\hat{\eta}_{J^*}(t-k_m))^2 + (\hat{\eta}_{J^*}(t-k_m) - \hat{\phi}_K(k_m)\frac{\hat{v}_{J^*}}{\hat{v}_J}\hat{\epsilon}_J(t))^2\right].$$

leading to

$$\hat{\phi}_K(k_m) = \frac{\hat{v}_J(\hat{v}_J + \hat{v}_{J^*})\sum_{t=k_m+1}^{T} \hat{\epsilon}_J(t)\hat{\eta}_{J^*}(t-k_m)}{\sum_{t=k_m+1}^{T} (\hat{v}_J^2\hat{\eta}_{J^*}(t-k_m)^2 + \hat{v}_{J^*}^2\hat{\epsilon}_J(t)^2)},$$

and

$$\hat{\phi}_{K^*}(k_m) = \frac{\hat{v}_{J^*}(\hat{v}_J + \hat{v}_{J^*})\sum_{t=k_m+1}^{T} \hat{\epsilon}_J(t)\hat{\eta}_{J^*}(t-k_m)}{\sum_{t=k_m+1}^{T} (\hat{v}_J^2\hat{\eta}_{J^*}(t-k_m)^2 + \hat{v}_{J^*}^2\hat{\epsilon}_J(t)^2)},$$

Of course all three methods generate different sequences of residuals and estimated coefficients.

We note finally that a Burg version of Algorithm 5 for h-step prediction can also be derived from Algorithm 7.

Acknowledgments

This work was supported partially by NSF Grant DMS 9504596 (P.B.), the Deutsche Forschungsgemeinschaft (Da 187/9-1) and European Union Capital and Mobility Programme (ERB CHRX-CT940963) (R.D.). The support of the Institute of Applied Mathematics, University of Heidelberg, is also gratefully acknowledged by P.B.

References

Brockwell, P.J. and Davis, R.A. (1987). Time Series: Theory and Methods, Springer, New York.

Burg, J.P. (1968). A new analysis technique for time series data. NATO Advanced Study Institute of Signal Processing with emphasis on Underwater Acoustics. In: Childers, D.G. (ed.): Modern Spectrum Analysis, IEEE Press, New York, 1978.

Dahlhaus, R. (1988). Small sample effects in time series analysis: a new asymptotic theory and a new estimate. Ann. Statist. 16, 808-841.

Durbin, J. (1960). The fitting of time-series models. Rev. Internat. Statist. Inst. 28, 233-244.

Morettin, P.A. (1984). The Levinson algorithm and its applications in time series analysis. Int. Statist. Review 52, 83-92.

Morf, M., Vieira, A., Lee, D.T.L. and Kailath, T. (1978). Recursive multichannel maximum entropy spectral estimation. IEEE Trans. Geosci. Electron. GE-16, 85-94.

Penm. Jack H.W. and Terrell, R.D. (1982). On the recursive fitting of subset autoregressions. J. Time Series Analysis, 3, 43-59.

Seber, G.A.F. (1977). Linear Regression Analysis. John Wiley, New York.

Whittle, P.W. (1963). On the fitting of multivariate autoregressions and the approximate canonical factorization of a spectral density matrix. Biometrika 50, 129-134.