Modified Burg Algorithms for Multivariate Subset Autoregression

Peter J. Brockwell [*] Rainer Dahlhaus [†]

A. Alexandre Trindade [‡]

May 20, 2002

Abstract

In this paper we derive subset versions of several lattice algorithms for estimating the parameters of a multivariate autoregression in which some of the coefficient matrices are constrained to be zero. We first establish a recursive prediction-error version of the empirical Yule-Walker equations for the parameter estimates. The estimated coefficients obtained from these recursions are the coefficients of the best linear one-step predictors of the process under the assumption that the autocovariance function is the same as the sample autocovariance function. By modifying the recursions to allow for certain inherent shortcomings, we then derive new estimators which generalize the Vieira-Morf, Nutall-Strand and Burg estimators to the multivariate subset case. We show that the new estimators minimize weighted sums of squares of the forward and backward prediction errors in recursive schemes which closely resemble the original scheme of Burg. The performances of the estimators are compared in a simulation study.

Keywords: lattice algorithm, linear prediction, multistep prediction, multivariate autoregression, recursive autoregression, VAR process.

[*] Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, U.S.A. E-mail: pjbrock@stat.colostate.edu

[†] Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany. E-mail: dahlhaus@statlab.uni-heidelberg.de

[‡] To whom correspondence should be addressed at: Department of Statistics, University of Florida, P.O. Box 118545, Gainesville, FL 32611-8545, U.S.A. Phone: (352)392-1941 ext 216, Fax: (352)392-5175, E-mail: trindade@stat.ufl.edu

1 Introduction

A fundamental problem in the analysis of stationary time series, is the forecasting of future observations based on some subset of past observations. If the covariance function of the process is known, one can obtain "best" (smallest mean squared error) linear predictors by solving the well-known Yule-Walker equations. This is the *prediction problem.* Typically however, one simply has a set of data and no knowledge of the covariance function of the underlying stochastic process that generated the observed realization. Modeling the observed data is then a necessary first step before implementing any type of model-based forecasting technique. This is the *modeling problem.*

By its very nature, the linear forecasting problem has a clear-cut solution which is completely determined by the assumed mean and autocovariance functions. The modeling problem on the other hand is much more complicated, involving as it does quesTjΩ-27jΩ970TDΩ80TDΩ0.0002TcΩ(is)TjΩ110.99980TDΩ-0.0001Tc we can determine the best linear predictor of $\mathbf{X}_t$ in terms of the lagged variables $\{\mathbf{X}_j, j \in K\}$, i.e. ,

$$\hat{\mathbf{X}}_t = \Phi(k_1)\mathbf{X}_{t-k_1} + \cdots + \Phi(k_m)\mathbf{X}_{t-k_m},$$

by finding a solution $\{\Phi_{k_i}, i = 1, \ldots, m\}$ of the Yule-Walker equations,

$$\sum_{i=1}^{m} \Phi(k_i)\Gamma(k_j - k_i) \quad = \quad \Gamma(k_j), \quad j = 1, \ldots, m.$$

The Yule-Walker *estimators* of the coefficient matrices in the autoregression (1) satisfy the same equation with the *sample autocovariance function of the data* replacing the *autocovariance function of the model.* In this sense there is a close connection between the prediction and modeling problems.

The Yule-Walker equations consist of md^2 linear equations for the components of the coefficient matrices. In the case when $k_i = i, i = 1, \ldots, m$, their solution is greatly simplified by the use of Whittle's generalization of the Levinson-Durbin recursions (see e.g. Brockwell and Davis, 1991) which require the inversion of $m \times m$ matrices only. It does however require simultaneous consideration of the "backward" Yule-Walker equations,

$$\sum_{i=1}^{m} \Psi(k_i)\Gamma(k_j - k_i)' \quad = \quad \Gamma(k_j)', \quad j = 1, \ldots, m,$$

which determine (when the autocovariance function $\Gamma(\cdot)$ is known) the best linear estimate of $\mathbf{X}_t$ of the form,

$$\hat{\mathbf{X}}_t = \Psi(k_1)\mathbf{X}_{t+k_1} + \cdots + \Psi(k_m)\mathbf{X}_{t+k_m}.$$

Analogous recursions hold also in the more general subset case (see Penm and Terrell, 1982). We review these recursions in Section 2, and in Section 3 we develop new algorithms for estimation of subset VAR models which extend Burg's algorithm and related algorithms based on empirical prediction errors. The results generalize those of Brockwell and Dahlhaus (2002) to the multivariate case. Unlike the univariate case, the multivariate case is complicated by the fact that the coefficient matrices $\Phi(k)$ and $\Psi(k)$ in the Yule-Walker equations are not equal.

VAR models are frequently used in practice in preference to the more general vector ARMA (VARMA) models, because of their relative simplicity with respect to identification, estimation, interpretation and forecasting. Some applications of subset prediction and modeling are as follows.

Forecasting with missing observations. This is a very natural application of subset prediction when the covariance function of the process is known.

h-step prediction. h-step-ahead prediction based on m successive observations is equivalent to subset prediction with the set of lags $K = \{h, h+1, \ldots, h+m-1\}$. The recursions for the coefficients turn out to be especially simple in this case although subsets of K can also be used.

Modeling of seasonal time series. If B denotes the backward shift operator, i.e. $B^k X_t = X_{t-k}$ for any positive integer k, then causal subset AR models of the form,

$$(1 - \psi B^s)(1 - \phi_1 B - \cdots - \phi_p B^p) X_t = Z_t,$$

will exhibit approximate cyclical behavior for appropriate values of the coefficients $\psi,\ \phi_1, \ldots, \phi_p$, and orders $s,\ p$, as evidenced by sharp peaks in the spectral density. This suggests that some seasonal time series can effectively be modeled as subset AR processes.

Fitting "best" subset models. When fitting a VAR model, one typically searches for the most appropriate order up to some maximum lag, p, guided by an information criterion like AICC, BIC, or MDL. As computing power grows, practitioners are increasingly allowing themselves the flexibility of considering all 2^p possible subset candidate models to adequately describe their data. McClave (1975), was one of the first to consider subset AR modeling, introducing an algorithm adapted from linear regression for best subset identification. Recently, there has been a flurry of activity, mostly in the signal processing literature, on devising efficient algorithms to narrow the list of candidates (see for example, Sarkar and Sharma, 1997, and Terrell and Zhang, 1997).

For the estimation of parameters in VAR (and SVAR) models it is of course desirable to have fast and efficient algorithms. Direct numerical maximization of the Gaussian likelihood is efficient but slow, and is complicated by the large number of parameters to be optimized, as well as the complexity of the likelihood surface. One important application of recursive methods is the rapid generation of "preliminary" models which can then be used to initialize a numerical search algorithm for maximizing the likelihood. For this purpose, it is important that the preliminary models should be close to the maximum likelihood model. For this reason we shall compare the algorithms introduced in the paper in terms of the Gaussian likelihoods of the fitted models.

In section 2 we discuss the subset Levinson-Durbin algorithm of Penm and Terrell (1982), and derive some useful properties of the forward and backward prediction errors. In section 3 we derive analogous properties for the *empirical* prediction errors, and use them to obtain model estimation algorithms of Burg's type. Several different estimators are obtained in this way, including the subset Yule-Walker estimators of Penm and Terrell, and multivariate subset generalizations of the algorithms of Burg, Vieira and Morf, and Nutall and Strand. We conclude in section 4 with a comparison of the performance of these estimators using simulated univariate and bivariate data.

2 The Prediction Problem

If $\mathbf{X}$ and $\mathbf{Y}$ are random column vectors, all of whose components have finite second moments, 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 $< \mathbf{Y}, \mathbf{X} >= 0$.

If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}$ all have finite second moments then the best linear predictor of $\mathbf{Y}$ in terms of $\mathbf{X}_1, \ldots, \mathbf{X}_n$,

$$\hat{\mathbf{Y}} = A_1\mathbf{X}_1 + \cdots + A_n\mathbf{X}_n,$$

is defined to be the linear combination (with matrix coefficients) of $\mathbf{X}_1 \ldots, \mathbf{X}_n$ whose components each have minimum mean-square distance from the corresponding component of $\mathbf{Y}$. We know (see e.g. Brockwell and Davis, 1991, Chapter 11) that $\hat{\mathbf{Y}}$ is uniquely determined by the requirement that the error vector $\mathbf{Y} - A_1\mathbf{X}_1 - \cdots - A_n\mathbf{X}_n$ is orthogonal to $\mathbf{X}_i$ for each $i \in \{1, \ldots, n\}$.

If $\{\mathbf{X}_t,\ t = 0, \pm 1, \pm 2, \cdots\}$ is a zero-mean weakly stationary d-variate time series with autocovariance function, $\Gamma(h) \equiv \mathbf{E}[\mathbf{X}_{t+h}\mathbf{X}_t'], h = 0, \pm 1, \ldots,$ then the best linear predictor of $\mathbf{X}_t$ in terms of $\{\mathbf{X}_{t-k}, k \in K\}$, with $K = \{k_1, \cdots, k_m\}$ and $1 \leq k_1 < k_2 < \cdots < k_m$, is defined as

$$\hat{\mathbf{X}}_t(K) = \sum_{i \in K} \Phi_K(i)\mathbf{X}_{t-i}, \tag{2}$$

and the orthogonality conditions determining $\hat{\mathbf{X}}_t(K)$ are easily found to be the (forward) Yule-Walker (YW) equations,

$$\sum_{i\in K} \Phi_K(i)\Gamma(k-i) = \Gamma(k), \quad k \in K. \tag{3}$$

The covariance matrix of the error vector $\mathbf{X}_t - \hat{\mathbf{X}}_t(K)$ is

$$U_K = \Gamma(0) - \sum_{i\in K} \Phi_K(i)\Gamma(i)'. \tag{4}$$

Analogously, we define the best *backward* linear predictor of $\mathbf{X}_t$ based on the lag subset K, to be $\hat{\mathbf{X}}_t^{(b)}(K) = \sum_{i\in K} \Psi_K(i)\mathbf{X}_{t+i}$. For reasons that will become clear on inspection of the algorithms, we focus instead on the lag subset $K^* = \{k_m - k_{m-1}, \cdots, k_m - k_1, k_m\}$, and the corresponding backward predictor,

$$\hat{\mathbf{X}}_t^{(b)}(K^*) = \sum_{j\in K^*} \Psi_{K^*}(j)\mathbf{X}_{t+j}. \tag{5}$$

The orthogonality conditions give the backward Yule-Walker equations for the coefficients, $\Psi_{K^*}(j)$, i.e.

$$\sum_{j\in K^*} \Psi_{K^*}(j)\Gamma(k-j)' = \Gamma(k)', \quad k \in K^*. \tag{6}$$

The covariance matrix of the backward error vector $\mathbf{X}_t - \hat{\mathbf{X}}_t^{(b)}(K^*)$ is

$$V_{K^*} = \Gamma(0) - \sum_{j\in K^*} \Psi_{K^*}(j)\Gamma(j). \tag{7}$$

(Note the manner in which the sets K and K^* are connected: if we imagine $0, k_1, \ldots, k_m$ as points plotted on the time axis, then K^* is the set of time intervals measured from the rightmost point, k_m, to each of the other points k_{m-1}, $k_{m-2}, \ldots, 0$.)

The Levinson-Durbin algorithm is a recursive algorithm for solving equations (3) and (4) when $K = \{1, \ldots, m\}$ (which simultaneously solves the less interesting equations (6) and (7)). Algorithm 1 below is a subset version of the algorithm, due to Penm and Terrell (1982).

Algorithm 1 (Subset Levinson-Durbin)

$$\begin{aligned}
\Phi_K(k_m) &= \left(\Gamma(k_m) - \sum_{i\in J}\Phi_J(i)\Gamma(k_m - i)\right)V_{J^*}^{-1}\\
\Phi_K(i) &= \Phi_J(i) - \Phi_K(k_m)\Psi_{J^*}(k_m - i), \quad i \in J\\
\Psi_{K^*}(k_m) &= \left(\Gamma(k_m)' - \sum_{j\in J^*}\Psi_{J^*}(j)\Gamma(k_m - j)'\right)U_J^{-1}\\
\Psi_{K^*}(j) &= \Psi_{J^*}(j) - \Psi_{K^*}(k_m)\Phi_J(k_m - j), \quad j \in J^*\\
U_K &= U_J - \Phi_K(k_m)V_{J^*}\Phi_K(k_m)'\\
V_{K^*} &= V_{J^*} - \Psi_{K^*}(k_m)U_J\Psi_{K^*}(k_m)',
\end{aligned}$$

where U_J^{-1} and $V_{J^*}^{-1}$ denote generalized inverses. The subsets J and J^*, are derived from K and K^*, respectively, by omitting k_m. The initial conditions are, $U_\emptyset = \Gamma(0) = V_\emptyset$.

Note that the predictor based on the subset of the current iteration, K, is formed from predictors obtained in earlier iterations based on the subsets J and J^*. Algorithms with this kind of recursive structure require careful computer implementation. Trindade (2001) discusses the recursions and provides guidance for implementing them in a high-level programming language.

We now derive some properties of the forward and backward prediction errors $\boldsymbol{\epsilon}_K(t)$ and $\boldsymbol{\eta}_{K^*}(t)$, defined respectively as

$$\boldsymbol{\epsilon}_K(t) = \mathbf{X}_t - \hat{\mathbf{X}}_t(K), \quad \text{and} \quad \boldsymbol{\eta}_{K^*}(t) = \mathbf{X}_t - \hat{\mathbf{X}}_t^{(b)}(K^*).$$

The covariance matrices of these prediction errors are respectively the matrices U_K and V_{K^*} defined by (4) and (7).

Proposition 1
Let K, K^*, J and J^* be defined as in Algorithm 1 and define the matrix of inner products for any two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with finite second moments as $< \mathbf{X}, \mathbf{Y} >= E(\mathbf{X}\mathbf{Y}')$. Then

(i) $< \boldsymbol{\epsilon}_J(t), \boldsymbol{\epsilon}_J(t) >= U_J,$

(ii) $< \boldsymbol{\eta}_{J^*}(t), \boldsymbol{\eta}_{J^*}(t) >= V_{J^*},$

(iii) $< \boldsymbol{\epsilon}_J(t), \boldsymbol{\eta}_{J^*}(t - k_m) >= U_J\Psi_{K^*}(k_m)' = \Phi_K(k_m)V_{J^*},$

(iv) $\boldsymbol{\epsilon}_K(t) = \boldsymbol{\epsilon}_J(t) - \Phi_K(k_m)\boldsymbol{\eta}_{J^*}(t-k_m)$,

(v) $\boldsymbol{\eta}_{K^*}(t) = \boldsymbol{\eta}_{J^*}(t) - \Psi_{K^*}(k_m)\boldsymbol{\epsilon}_J(t+k_m)$.

Proof
(i) and (ii) are just the definitions of the error covariance matrices.
(iii) is a restatement of equations (26) and (30) of Appendix A.1 with

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

$$B = (\Phi_J(k_1), \ldots, \Phi_J(k_{m-1})), \quad C = (\Psi_{J^*}(k_m-k_1), \ldots, \Psi_{J^*}(k_m-k_{m-1})),$$

$$A_2 = \Phi_K(k_m), \quad D_2 = \Psi_{K^*}(k_m), \quad v_{\mathbf{X}|\mathbf{Y}} = U_J, \text{ and } v_{\mathbf{Z}|\mathbf{Y}} = V_{J^*}.$$

(iv) is obtained by observing that

$$\begin{aligned}\boldsymbol{\epsilon}_J(t) - \Phi_K(k_m)\boldsymbol{\eta}_{J^*}(t-k_m) &= \mathbf{X}_t - \sum_{j\in J}\Phi_J(j)\mathbf{X}_{t-j} \\ &\quad - \Phi_K(k_m)\left(\mathbf{X}_{t-k_m} - \sum_{j\in J}\Psi_{J^*}(k_m-j)\mathbf{X}_{t-j}\right), \qquad (8)\end{aligned}$$

and noting from Algorithm 1 that

$$\Phi_J(j) - \Phi_K(k_m)\Psi_{J^*}(k_m-j) = \Phi_K(j), \; j \in J.$$

(v) is proved in exactly the same way. □

In Section 3.1 we give an empirical version of Proposition 1 which leads to an alternative set of recursions for solving the Yule-Walker equations and which motivates the definition of a number of other recursive prediction-error based algorithms.

3 The Modeling Problem

Given observations $\mathbf{x}_1, \cdots, \mathbf{x}_n$ of a zero-mean stationary time series $\{\mathbf{X}_t\}$, we wish to estimate the parameters Σ and the coefficient matrices $\Phi_K(k_i), i = 1, \ldots, m$ in the subset model (1) for the data. The Yule-Walker estimators are the matrices for which the model autocovariances at lags $0, k_1, \ldots, k_m$, coincide with the *sample autocovariances*, defined as

$$\hat{\Gamma}(h) = \begin{cases} \frac{1}{n}\sum_{t=1}^{n-h}\mathbf{x}_{t+h}\mathbf{x}'_t & \text{, if } h \geq 0, \\ \hat{\Gamma}(-h)' & \text{, if } h < 0. \end{cases}$$

In principle the estimates of $\Phi_K(k_i)$ and Σ can be found directly from equations (3) and (4) by substituting $\hat{\Gamma}(\cdot)$ for $\Gamma(\cdot)$, solving (3) to obtain the estimates $\hat{\Phi}_K(k_i), i = 1, \ldots, m$, and using $\hat{U}_K$, found from (4), as the estimate of Σ.

Algorithm 1 however provides a much more convenient recursive method of arriving at the same estimates, since it reduces the problem to one involving manipulations of $d \times d$ rather than $md^2 \times md^2$ matrices. (At the same time it produces estimates of the coefficients $\Psi_{K^*}(i), i \in K^*$.)

Remark. It is possible that the subset model obtained by this procedure will be non-causal, i.e. that the matrix $I_d - \hat{\Phi}_K(k_1)z^{k_1} - \cdots - \hat{\Phi}_K(k_m)z^{k_m}$ (where I_d is the $d \times d$ identity matrix) will have an eigenvalue with absolute value greater than or equal to 1. This is an indication that the subset model with lags in K is inappropriate for the data. Even in this case however, it remains true that the expression (2) with each $\Phi_K(i)$ replaced by $\hat{\Phi}_K(i)$ gives the best linear predictor of X_t in terms of X_{t-k_i}, $i = 1, \ldots, m$, under the assumption that the sample autocovariances are the true autocovariances.

3.1 Estimation based on empirical prediction errors

In order to develop estimation procedures based on *observed* prediction errors, we need an empirical version of Proposition 1, i.e. a version which is expressed entirely in terms of the data and the Yule-Walker estimators $\{\hat{\Phi}_K(i), i \in K\}$, $\{\hat{\Psi}_{K^*}(i), i \in K^*\}$, $\hat{U}_K$, and $\hat{V}_{K^*}$.

Before stating the required proposition, we introduce some additional notation which is required for a closer analysis of the empirical Yule-Walker equations.

For $t \leq 0$ and for $t > n$ define $\mathbf{x}_t = \mathbf{0}$ and let $\underset{\sim}{\mathbf{x}}$ be the $d \times \infty$ array whose j^{th} column is $\mathbf{x}_j$, $j = 0, \pm 1, \ldots$, i.e.

$$\underset{\sim}{\mathbf{x}} = \{\mathbf{x}_j,\ j = 0, \pm 1, \ldots\}.$$

Now let $\underset{\sim}{\mathbf{x}}_t$ be the array obtained by shifting the columns of $\underset{\sim}{\mathbf{x}}$ by t places to the left, i.e.

$$\underset{\sim}{\mathbf{x}}_t = \{\mathbf{x}_{t+j},\ j = 0, \pm 1, \ldots\},$$

and think of $\underset{\sim}{\mathbf{x}}_t$ as a column vector of d elements, each element being an infinite-dimensional row vector with finitely many non-zero elements. The set of all such row vectors constitutes an inner-product space, if we define the inner-product of any two elements $\mathbf{u} = \{u_j\}$ and $\mathbf{v} = \{v_j\}$ as

$$< \mathbf{u}, \mathbf{v} >= \frac{1}{n} \sum_{j=-\infty}^{\infty} u_j v_j . \tag{9}$$

The d-component column vectors $\underset{\sim}{\mathbf{x}}_t$ are then quite analogous to the random vectors $\mathbf{X}_t$ as vectors of elements of an inner-product space, except that the inner products between components are defined as in (9) instead of as expected products. The factor $1/n$ is included in the definition (9) since the matrix of inner-products $< \underset{\sim}{\mathbf{x}}_{t+h}, \underset{\sim}{\mathbf{x}}_t >$, i.e. the matrix whose (i,j)-element is the inner product of the i^{th} row of $\underset{\sim}{\mathbf{x}}_{t+h}$ with the j^{th} row of $\underset{\sim}{\mathbf{x}}_t$, is then the sample covariance matrix $\hat{\Gamma}(h)$ of the data set $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

The empirical counterpart of equation (3) can now immediately be recognized as the equation for the coefficients $\hat{\Phi}_K(j),\ j \in K$, in the expression

$$\hat{\underset{\sim}{\mathbf{x}}}_t(K) = \sum_{j \in K} \hat{\Phi}_K(j) \underset{\sim}{\mathbf{x}}_{t-j},$$

for the projection of $\underset{\sim}{\mathbf{x}}_t$ onto the span of the rows of $\underset{\sim}{\mathbf{x}}_{t-j},\ j \in K$. This is because (3) simply expresses the orthogonality conditions,

$$< \hat{\underset{\sim}{\mathbf{x}}}_t(K) - \underset{\sim}{\mathbf{x}}_t, \underset{\sim}{\mathbf{x}}_{t-j} >= 0, \quad j \in K.$$

Moreover, the empirical counterpart of equation (4) identifies the Yule-Walker white noise covariance estimate $\hat{U}_K$ as the error product matrix,

$$\hat{U}_K =< \underset{\sim}{\mathbf{x}}_t - \hat{\underset{\sim}{\mathbf{x}}}_t(K), \underset{\sim}{\mathbf{x}}_t - \hat{\underset{\sim}{\mathbf{x}}}_t(K) > .$$

Analogously to $\underset{\sim}{\mathbf{x}}$, we define the $d \times \infty$ arrays of forward and backward empirical prediction errors as

$$\hat{\underset{\sim}{\boldsymbol{\epsilon}}}_K(t) = \underset{\sim}{\mathbf{x}}_t - \hat{\underset{\sim}{\mathbf{x}}}_t(K) = \mathbf{x}_t - \sum_{j \in K} \hat{\Phi}_K(j) \underset{\sim}{\mathbf{x}}_{t-j},$$

and

$$\hat{\underset{\sim}{\boldsymbol{\eta}}}_{K^*}(t) = \underset{\sim}{\mathbf{x}}_t - \hat{\underset{\sim}{\mathbf{x}}}_t^{(b)}(K^*) = \mathbf{x}_t - \sum_{j \in K^*} \hat{\Psi}_{K^*}(j) \underset{\sim}{\mathbf{x}}_{t+j},$$

respectively. The corresponding empirical forward and backward product error matrices are

$$\hat{U}_K =< \hat{\underset{\sim}{\boldsymbol{\epsilon}}}_K(t), \hat{\underset{\sim}{\boldsymbol{\epsilon}}}_K(t) >, \quad \text{and} \quad \hat{V}_{K^*} =< \hat{\underset{\sim}{\boldsymbol{\eta}}}_{K^*}(t), \hat{\underset{\sim}{\boldsymbol{\eta}}}_{K^*}(t) > .$$

We can now state the required empirical analogue of Proposition 1.

Proposition 2

In the notation of Proposition 1, and with the inner product and $d \times \infty$ arrays $\hat{\underset{\sim}{\varepsilon}}_K(t)$, $\hat{\underset{\sim}{\eta}}_{K^*}(t)$ defined as above, the assertions of Proposition 1 hold with $\Phi, \Psi, U, V, \boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ replaced by $\hat{\Phi}, \hat{\Psi}, \hat{U}, \hat{V}$, $\hat{\underset{\sim}{\varepsilon}}$ and $\hat{\underset{\sim}{\eta}}$, respectively.

Proof

By virtue of Theorem A.2 of Appendix A.1, the proof is identical to the proof of Proposition 1 with

$$\mathbf{X} = \underset{\sim}{\mathbf{x}}_t,\ \mathbf{Y} = (\underset{\sim}{\mathbf{x}}'_{t-k_1}, \ldots, \underset{\sim}{\mathbf{x}}'_{t-k_{m-1}})',\ \mathbf{Z} = \underset{\sim}{\mathbf{x}}_{t-k_m},$$

and $\boldsymbol{\Phi}, \boldsymbol{\Psi}, U, V, \boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ replaced by $\hat{\boldsymbol{\Phi}}, \hat{\boldsymbol{\Psi}}, \hat{U}, \hat{V}, \hat{\underset{\sim}{\varepsilon}}$ and $\hat{\underset{\sim}{\eta}}$ respectively. □

From Proposition 1 and Algorithm 1, we immediately obtain the following algorithm which generates, from the empirical prediction errors, a solution of the empirical subset Yule-Walker equations.

Algorithm 2 (Prediction error solution of Yule-Walker equations)

$$\begin{aligned}
\hat{\Phi}_K(k_m) &= \left(\frac{1}{n}\sum_{t=1}^{n+k_m} \hat{\boldsymbol{\varepsilon}}_J(t)\hat{\boldsymbol{\eta}}_{J^*}(t-k_m)'\right)\hat{V}_{J^*}^{-1}, && (10)\\
\hat{\Phi}_K(i) &= \hat{\Phi}_J(i) - \hat{\Phi}_K(k_m)\hat{\Psi}_{J^*}(k_m - i), \quad i \in J \\
\hat{\Psi}_{K^*}(k_m) &= \hat{V}_{J^*}\hat{\Phi}_K(k_m)'\hat{U}_J^{-1} && (11)\\
\hat{\Psi}_{K^*}(j) &= \hat{\Psi}_{J^*}(j) - \hat{\Psi}_{K^*}(k_m)\hat{\Phi}_J(k_m - j), \quad j \in J^* && (12)\\
\hat{U}_K &= \hat{U}_J - \hat{\Phi}_K(k_m)\hat{V}_{J^*}\hat{\Phi}_K(k_m)' \\
\hat{V}_{K^*} &= \hat{V}_{J^*} - \hat{\Psi}_{K^*}(k_m)\hat{U}_J\hat{\Psi}_{K^*}(k_m)' && (13)\\
\hat{\boldsymbol{\varepsilon}}_K(t) &= \hat{\boldsymbol{\varepsilon}}_J(t) - \hat{\Phi}_K(k_m)\hat{\boldsymbol{\eta}}_{J^*}(t-k_m) && (14)\\
\hat{\boldsymbol{\eta}}_{K^*}(t) &= \hat{\boldsymbol{\eta}}_{J^*}(t) - \hat{\Psi}_{K^*}(k_m)\hat{\boldsymbol{\varepsilon}}_J(t+k_m) && (15)
\end{aligned}$$

with initial conditions,

$$\hat{\boldsymbol{\varepsilon}}_{\emptyset}(t) = \mathbf{x}_t = \hat{\boldsymbol{\eta}}_{\emptyset}(t), \quad \textit{and}, \quad \hat{U}_{\emptyset} = \hat{\Gamma}(0) = \hat{V}_{\emptyset}.$$

Remark. (Lattice algorithms). Algorithm 2 provides us with yet another recursive algorithm for computing Yule-Walker estimates. In the engineering literature (14) and (15) are referred to as *lattice equations* and the coefficients $\hat{\Phi}_K(k_m)$ and $\hat{\Psi}_{K^*}(k_m)$ as (estimated) *reflection coefficients*. There is a large literature on such algorithms (see for example Itakura and Saito, 1971, Makhoul, 1977, Morf, Vieira, Lee, and Kailath 1978, and Haykin, 1996). A variety of different estimators has been constructed in the case when $K = \{1, \ldots, m\}$ by modifying equations (10) and (11) in Algorithm 2 in order to improve the estimated reflection coefficients. Subset analogs of several of these are discussed below in Section 3.2.

Burg's (1968) original univariate version of this algorithm, gives causal models which generally provide better resolution of spectral peaks, and have higher Gaussian likelihood than Yule-Walker models of the same order. The goal of the generalized versions is to extend some of these desirable properties to the generalized setting.

3.2 Subset versions of three lattice algorithms

Since $\mathbf{x}_t$ was defined to be zero outside the time interval $[1, n]$, the empirical prediction errors $\hat{\boldsymbol{\varepsilon}}_J(t)$ and $\hat{\boldsymbol{\eta}}_{J^*}(t - k_m)$, are not particularly meaningful for t close to 1 and $n + k_m$, respectively. This suggests replacing the lower and upper limits of summation in (10) by $k_m + 1$ and n, respectively. We will adopt this truncated summation in all the lattice algorithms proposed in this section, the introduction of which first necessitates the definition of the following matrices:

$$\hat{\Omega}_{\epsilon\epsilon} = \frac{1}{n - k_m} \sum_{t=k_m+1}^{n} \hat{\boldsymbol{\varepsilon}}_J(t) \hat{\boldsymbol{\varepsilon}}_J(t)', \qquad \hat{\Omega}_{\epsilon\eta} = \frac{1}{n - k_m} \sum_{t=k_m+1}^{n} \hat{\boldsymbol{\varepsilon}}_J(t) \hat{\boldsymbol{\eta}}_{J^*}(t - k_m)',$$

and

$$\hat{\Omega}_{\eta\eta} = \frac{1}{n - k_m} \sum_{t=k_m+1}^{n} \hat{\boldsymbol{\eta}}_{J^*}(t - k_m) \hat{\boldsymbol{\eta}}_{J^*}(t - k_m)'.$$

Note that $\hat{\Omega}_{\varepsilon\varepsilon}$ and $\hat{U}_J$ are two different estimates of U_J, while $\hat{\Omega}_{\eta\eta}$ and $\hat{V}_{J^*}$ are estimates of V_{J^*}. Roughly speaking, it is the combination of these different estimates together with the truncation of the sum that leads to the three modifications of the algorithm proposed below. In Appendix A.2 we show that the resulting algorithms can be obtained as the minimizers of certain prediction errors. The first modification of Algorithm 2 reduces to the widely used algorithm of Morf, Vieira, Lee, and Kailath (1978), in the case when $K = \{1, \ldots, m\}$.

Algorithm 3 (Subset Vieira-Morf)
Replace (10) in Algorithm 2 by

$$\hat{\Phi}_K(k_m) = \hat{U}_J^{1/2} \hat{R} \hat{V}_{J^*}^{-1/2}, \tag{16}$$

where

$$\hat{R} = \hat{\Omega}_{\epsilon\epsilon}^{-1/2} \hat{\Omega}_{\epsilon\eta} \hat{\Omega}_{\eta\eta}^{-1/2}. \tag{17}$$

A similar modification was given by Strand (1977) (and independently by A.H. Nuttall and others). The following algorithm reduces to the Nuttall-Strand algorithm in the case $K = \{1, \ldots, m\}$.

Algorithm 4 (Subset Nuttall-Strand)
Replace (10) in Algorithm 2 by

$$\hat{\Phi}_K(k_m) = \hat{U}_J^{1/2} \hat{R} \hat{V}_{J^*}^{-1/2}, \tag{18}$$

where,

$$\mathit{vec}\, \hat{R} = 2 \left[I_d \otimes \hat{U}_J^{-1/2} \hat{\Omega}_{\epsilon\epsilon} \hat{U}_J^{-1/2} + \hat{V}_{J^*}^{-1/2} \hat{\Omega}_{\eta\eta} \hat{V}_{J^*}^{-1/2} \otimes I_d \right]^{-1} \times \mathit{vec} \left[\hat{U}_J^{-1/2} \hat{\Omega}_{\epsilon\eta} \hat{V}_{J^*}^{-1/2} \right]. \tag{19}$$

A computationally more convenient form of the same algorithm is given by

$$\hat{\Phi}_K(k_m) = \hat{\Delta} \hat{V}_{J^*}^{-1}, \tag{20}$$

where

$$\text{vec}\, \hat{\Delta} = 2 \left[I_d \otimes \hat{\Omega}_{\epsilon\epsilon} \hat{U}_J^{-1} + \hat{\Omega}_{\eta\eta} \hat{V}_{J^*}^{-1} \otimes I_d \right]^{-1} \text{vec}\, \hat{\Omega}_{\epsilon\eta}. \tag{21}$$

(The equality of both solutions is proved in Appendix A.2.) Since the term in the square brackets of (21) is close to $2 I_d \otimes I_d$, the right hand side of (20) is close, apart from the range of summation, to that of (10). An advantage of the Nutall-Strand algorithm however, is that it leads to a causal solution *in the full-subset case.*

In Burg's (1968) method for fitting full-subset autoregressions, the reflection coefficients were chosen so as to minimize the sums of squares of the forward and backward prediction errors. A natural multivariate subset generalization is to select $\hat{\Phi}_K(k_m)$ so as to minimize,

$$S_K(\hat{\Phi}_K(k_m)) = \sum_{t=k_m+1}^{n} \left[\hat{\boldsymbol{\varepsilon}}_K(t)' \hat{\boldsymbol{\varepsilon}}_K(t) + \hat{\boldsymbol{\eta}}_{K^*}(t-k_m)' \hat{\boldsymbol{\eta}}_{K^*}(t-k_m) \right], \tag{22}$$

with respect to $\hat{\Phi}_K(k_m)$, where $\hat{\boldsymbol{\varepsilon}}_K(t)$ and $\hat{\boldsymbol{\eta}}_{K^*}(t-k_m)$ are given by (14) and (15), respectively. With the aid of (11), $S_K(\hat{\Phi}_K(k_m))$ can then be expressed as a function of $\hat{\Phi}_K(k_m)$ and estimates computed at earlier stages of the recursive modeling procedure. This allows the minimizing value of $\hat{\Phi}_K(k_m)$ to be obtained explicitly (see Appendix A.2), leading to our last multivariate subset algorithm.

Algorithm 5 (Subset Burg)
Replace (10) in Algorithm 2 by

$$vec\, \hat{\Phi}_K(k_m) = \left[\hat{\Omega}_{\eta\eta} \otimes I_d + \hat{V}_{J^*}^2 \otimes \hat{U}_J^{-1}\hat{\Omega}_{\epsilon\epsilon}\hat{U}_J^{-1}\right]^{-1} vec\left[\hat{\Omega}_{\epsilon\eta} + \hat{U}_J^{-1}\hat{\Omega}_{\epsilon\eta}\hat{V}_{J^*}\right]. \tag{23}$$

Remark. Although minimization of the sum of squares of the forward and backward prediction errors (Burg's algorithm) leads in the univariate full-subset case to models with generally higher Gaussian likelihood than the empirical Yule-Walker equations, the asymmetry of multivariate subset modeling suggests that a weighted average of the forward and backward prediction errors might be more appropriate. In Appendix A.2, we prove that the subset Nuttall-Strand algorithm is obtained by minimizing the weighted sums of squares of prediction errors

$$\sum_{t=k_{m+1}}^{n} \left[\hat{\boldsymbol{\varepsilon}}_K(t)'\hat{U}_J^{-1}\,\hat{\boldsymbol{\varepsilon}}_K(t) + \hat{\boldsymbol{\eta}}_{K^*}(t-k_m)'\hat{V}_{J^*}^{-1}\,\hat{\boldsymbol{\eta}}_{K^*}(t-k_m)\right],$$

with respect to $\hat{\Phi}_K(k_m)$, where $\hat{\boldsymbol{\varepsilon}}_K(t)$ and $\hat{\boldsymbol{\eta}}_{K^*}(t-k_m)$ are given by (14) and (15), respectively. We also prove that the subset Vieira-Morf algorithm can be viewed as the solution of a minimization problem in which, with a natural but slightly different standardization of the error vectors, we obtain parameter estimates with the very desirable property that they simultaneously minimize both forward and backward error criteria (and their sum). From a theoretical perspective these algorithms therefore lead to more convincing estimators.

4 Comparing the Performance of the Algorithms

Since the primary aim of the four versions of Algorithm 2 is to provide fast and simple methods for the fitting of SVAR models with high likelihoods, it is of considerable interest to compare the actual likelihoods achieved by each. In this section we present such a comparison, by simulating realizations from a variety of univariate and bivariate SVAR models with independent Gaussian noise.

4.1 Preliminaries

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a realization of the causal SVAR process $\{\mathbf{X}_t\}$ defined by

$$\mathbf{X}_t = \Phi_K(k_1)\mathbf{X}_{t-k_1} + \cdots + \Phi_K(k_m)\mathbf{X}_{t-k_m} + \mathbf{Z}_t, \quad \{\mathbf{Z}_t\} \sim \text{IID N}(\mathbf{0}, \Sigma). \quad (24)$$

The likelihood, $\mathcal{L}(\Theta)$, where $\Theta = \{\Phi_K(k_1), \ldots, \Phi_K(k_m), \Sigma\}$, is a function of $md^2 + (d^2 + d)/2$ scalar parameters. For each realization simulated from a particular model, we obtain the Yule-Walker, Vieira-Morf, Nuttall-Strand, and Burg estimators, and compute the respective values of $-2\log\mathcal{L}(\hat{\Theta}_{AL})$ (the generic subscript AL signifying that the estimators are obtained via one of these four algorithms). The maximum likelihood estimator (MLE) is also obtained, and the corresponding value of $-2\log\mathcal{L}(\hat{\Theta}_{ML})$ (the subscript ML signifying that the estimators are obtained via maximization of the likelihood) subtracted from those obtained from each of the four algorithms, to give, for each algorithm, a value of

$$\text{NL} := -2\log\mathcal{L}(\hat{\Theta}_{AL}) + 2\log\mathcal{L}(\hat{\Theta}_{ML}).$$

(The minimization of $-2\log\mathcal{L}$ was carried out with a version of the algorithm of Hooke and Jeeves (1961). For the univariate examples it is straightforward to calculate the likelihood exactly. For the bivariate examples the likelihoods were computed by using the truncated series expansion of the autocovariances, $\Gamma(h) \approx \sum_{j=0}^{100} \Psi_{h+j}\Sigma\Psi'_j$, where the matrices Ψ_j are the coefficients in the representation $\mathbf{X}_t = \sum_{j=0}^{\infty} \Psi_j \mathbf{Z}_{t-j}$ and can be computed recursively as described in Brockwell and Davis, 1991, sec. 11.3.)

The *characteristic polynomial* of the SVAR model (24) is

$$P(z) = \det\left[I_d - \Phi_K(k_1)z^{k_1} - \cdots - \Phi_K(k_m)z^{k_m}\right].$$

The model is causal if the zeroes of its characteristic polynomial are all greater than one in magnitude. It is well-known that in the univariate full-subset case, the YW estimators can be severely biased if the roots of the AR characteristic polynomial are close to the unit circle. To allow for the expected dependence of performance on the location of the zeroes of $P(z)$, we simulate from models with a variety of configurations of these zeroes.

4.2 Univariate case

For each univariate model we generated 1,000 realizations, each of length 100, with $\{Z_t\} \sim$ IID N$(0, 1)$.

Example 1

$$(1+0.5B)(1-(0.1-0.3i)B)(1-(0.1+0.3i)B)X_t = Z_t.$$

This is the causal subset model, $X_t + 0.30X_{t-1} + 0.05X_{t-3} = Z_t$*, with* $K = \{1,3\}$ *and characteristic roots (moduli)* -2, $1 \pm 3i$ (3.16).

Example 2

$$(1+0.98B)(1-0.98B)(1+0.98iB)(1-0.98iB)X_t = Z_t.$$

This is the causal subset model, $X_t - 0.92X_{t-4} = Z_t$*, with* $K = \{4\}$ *and characteristic roots* ± 1.0204, $\pm 1.0204i$.

Example 3

$$(1+0.98B)(1-0.95B^3)X_t = Z_t.$$

This is the causal subset model $X_t + 0.98X_{t-1} - 0.95X_{t-3} - 0.93X_{t-4} = Z_t$*, with* $K = \{1,3,4\}$ *and characteristic roots (moduli)* $-0.5086 \pm 0.8809i$ (1.0172), 1.0172, -1.0204.

Example 4

$$(1-0.95B^2)(1+0.98B)(1-0.98B)X_t = Z_t.$$

This is the causal subset model $X_t - 1.91X_{t-2} + 0.91X_{t-4} = Z_t$*, with* $K = \{2,4\}$ *and characteristic roots* ± 1.0204, ± 1.0260.

The means, medians, and standard deviations of the values of NL are shown in Table 1, along with the percentage of realizations for which each method scored the lowest value. Figure 1 displays boxplots of the values of NL for the 1,000 realizations of each example. The performance of the subset Yule-Walker estimators is particularly poor compared with the other three estimators when the autoregressive roots are close to the unit circle. Overall, the Burg, Nutall-Strand and Vieira-Morf estimators give consistently higher likelihoods with less variability between realizations than the subset Yule-Walker estimators. Note that although different in general, the Burg and Nuttall-Strand solutions coincide in Examples 2 and 4. This is due to the particular configuration of the lags in the sets K (see remark below).

Remark. (Coincidence of Burg and Nuttall-Strand solutions). First note that in the univariate case all quantities are scalars. If we differentiate between the Burg and Nuttall-Strand solutions by topping the corresponding estimators with a tilde (˜) and breve (˘), respectively, then (23) and (18) reduce to

$$\tilde{\Phi}_K(k_m) = \frac{\tilde{U}_J(\tilde{U}_J + \tilde{V}_{J^*})\tilde{\Omega}_{\epsilon\eta}}{\tilde{V}_{J^*}^2\tilde{\Omega}_{\epsilon\epsilon} + \tilde{U}_J^2\tilde{\Omega}_{\eta\eta}}, \quad \text{and} \quad \breve{\Phi}_K(k_m) = \frac{2\breve{U}_J\breve{\Omega}_{\epsilon\eta}}{\breve{V}_{J^*}\breve{\Omega}_{\epsilon\epsilon} + \breve{U}_J\breve{\Omega}_{\eta\eta}}.$$

It follows immediately that if

(i) $\tilde{U}_J = \tilde{V}_{J^*}$, $\breve{U}_J = \breve{V}_{J^*}$, *and*

(ii) all remaining Burg quantities from the previous iteration steps are equal to their Nuttall-Strand counterparts,

then $\tilde{\Phi}_K(k_m) = 2\tilde{\Omega}_{\epsilon\eta}/(\tilde{\Omega}_{\epsilon\epsilon} + \tilde{\Omega}_{\eta\eta}) = \breve{\Phi}_K(k_m)$, so that the two solutions coincide. This happens trivially whenever K is comprised of just one lag (Example 2), since $J = \{\emptyset\} = J^*$. If K is comprised of one lag then $K = K^*$ and by (11) also $\tilde{\Psi}_{K^*}(k_m) = \tilde{\Phi}_K(k_m) = \breve{\Phi}_K(k_m) = \breve{\Psi}_{K^*}(k_m)$ and $\tilde{U}_K = \tilde{V}_{K^*} = \breve{U}_K = \breve{V}_{K^*}$. When $K = \{k_1, k_2\}$, a sufficient condition for coincidence of the solutions is that $k_2 = 2k_1$, since this implies $J = J^*$ (Example 4, $J = \{2\} = J^*$). In Example 1, $J = \{1\} \neq \{2\} = J^*$, so that (i) is violated. To obtain the estimators for the model of Example 3, we first need those of the intermediate model with $J = \{1, 3\}$. But this is now the situation of Example 1, so that $\tilde{\Phi}_J(j_{m-1}) \neq \breve{\Phi}_J(j_{m-1})$ (where j_{m-1} is the largest integer of J) which ultimately implies $\tilde{\Phi}_K(k_m) \neq \breve{\Phi}_K(k_m)$ also when $K = \{1, 3, 4\}$.

4.3 Bivariate case

Due to the difficulties involved in finding maximum likelihood estimators in the multivariate setting, we concentrate on bivariate models with subset size one. 200 realizations are simulated from each model, each of sample size 100, with $\mathbf{Z}_t \sim \mathrm{N}_2(\mathbf{0}, I_2)$, and configurations of roots of the characteristic polynomial that mimic those of the univariate examples.

Example 5
The causal bivariate subset VAR(2) model

$$\mathbf{X}_t - \begin{bmatrix} 0.547 & -0.300 \\ 0.700 & -0.457 \end{bmatrix} \mathbf{X}_{t-2} = \mathbf{Z}_t,$$

with characteristic polynomial,

$$P(z) = (1 - 0.25z^2)(1 + 0.16z^2).$$

Roots of characteristic polynomial: ± 2, $\pm 2.5i$.

Table 1: Summary statistics by method for the data of Examples 1-4

Example	Method	Mean of NL	Median of NL	Std. Dev. of NL	Frequency of lowest NL (%)
1	Yule-Walker	0.011	0.002	0.027	34.0
	Vieira-Morf	0.003	0.001	0.007	29.1
	Nuttall-Strand	0.003	0.001	0.007	12.3
	Burg	0.003	0.001	0.007	24.6
2	Yule-Walker	1.629	0.994	1.84	14.3
	Vieira-Morf	0.108	0.052	0.16	47.2
	Burg and Nuttall-Strand	0.111	0.053	0.17	38.5
3	Yule-Walker	6.019	3.710	6.603	6.1
	Vieira-Morf	0.504	0.285	0.770	28.1
	Nuttall-Strand	0.507	0.285	0.769	22.2
	Burg	0.505	0.284	0.767	43.6
4	Yule-Walker	200.18	200.802	48.83	0.0
	Vieira-Morf	0.32	0.133	0.64	50.6
	Burg and Nuttall-Strand	0.38	0.139	0.80	49.4

Example 6
The bivariate causal subset VAR(2) model

$$\mathbf{X}_t - \begin{bmatrix} 1.0091 & -0.3000 \\ 0.7000 & -1.0670 \end{bmatrix} \mathbf{X}_{t-2} = \mathbf{Z}_t,$$

with characteristic polynomial,

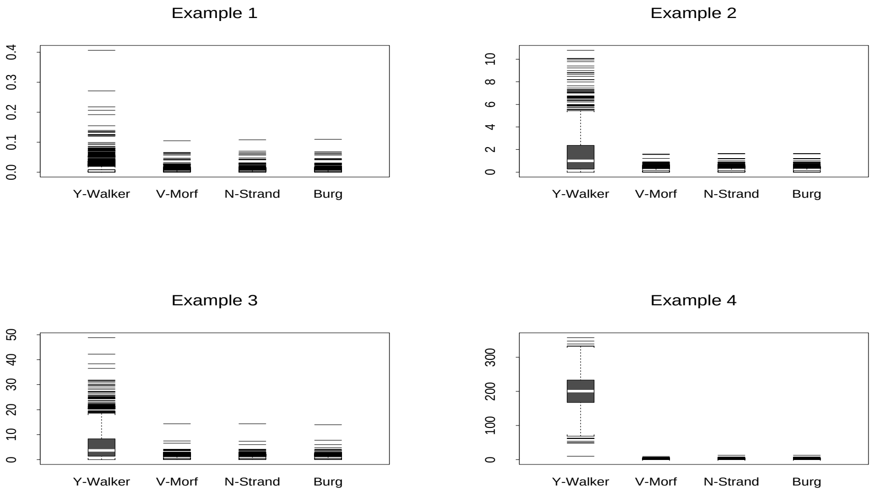
$$P(z) = (1 + 0.98^2 z^2)(1 - 0.95^2 z^2).$$

Roots of characteristic polynomial: ± 1.0526, $\pm 1.0204i$.

Example 7
The bivariate causal subset VAR(2) model

$$\mathbf{X}_t - \begin{bmatrix} 0.4 & -1.2 \\ 0.9 & -0.4 \end{bmatrix} \mathbf{X}_{t-2} = \mathbf{Z}_t,$$

Figure 1: Boxplots of NL's for the data of Examples 1-4.



with characteristic polynomial,

$$P(z) = (1 + 0.92z^4).$$

Roots (modulus) of characteristic polynomial: $\pm 0.722 \pm 0.722i$ (1.0211).

Example 8
The bivariate causal subset VAR(2) model

$$\mathbf{X}_t - \begin{bmatrix} 1.4135 & -0.3000 \\ 0.7000 & 0.4969 \end{bmatrix} \mathbf{X}_{t-2} = \mathbf{Z}_t,$$

with characteristic polynomial,

$$P(z) = (1 - 0.98^2 z^2)(1 - 0.95^2 z^2).$$

Roots of characteristic polynomial: ± 1.0204, ± 1.0260. *(239 realizations were simulated here; in 39 of these the Burg white noise covariance matrix estimate was negative definite, and the likelihood of the resulting model could not be computed. These 39 realizations were omitted.)*

The results are visually presented in Figure 2, and summarized in Table 2. As in the univariate examples, the performance of the Yule-Walker estimators is inferior to that of the three new lattice algorithms.

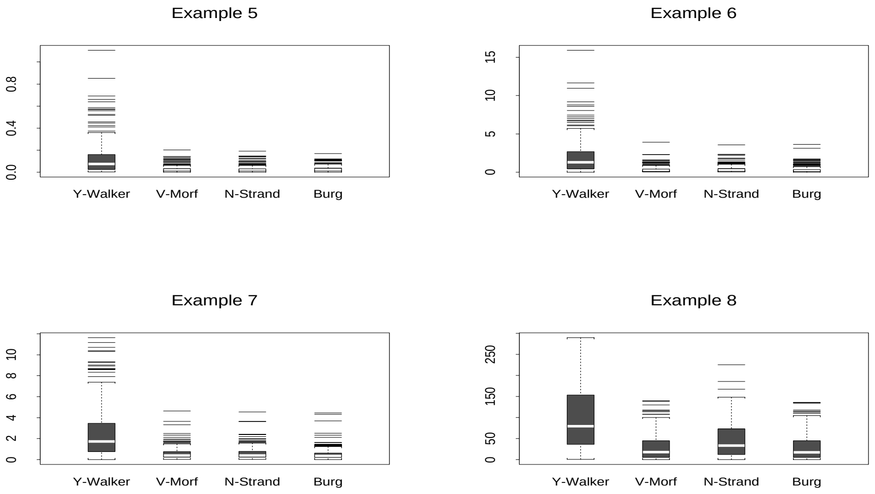
Table 2: Summary statistics by method for the data of Examples 5-8.

Example	Method	Mean of NL	Median of NL	Std. Dev. of NL	Frequency of lowest NL (%)
5	Yule-Walker	0.137	0.076	0.168	12.5
	Vieira-Morf	0.028	0.018	0.029	32.0
	Nuttall-Strand	0.028	0.020	0.029	30.0
	Burg	0.030	0.021	0.027	25.5
6	Yule-Walker	2.07	1.29	2.39	10.0
	Vieira-Morf	0.37	0.22	0.45	26.0
	Nuttall-Strand	0.40	0.26	0.46	11.0
	Burg	0.33	0.20	0.45	53.0
7	Yule-Walker	2.551	1.744	2.527	10.0
	Vieira-Morf	0.610	0.393	0.630	20.0
	Nuttall-Strand	0.608	0.387	0.635	14.0
	Burg	0.538	0.339	0.617	56.0
8	Yule-Walker	97.7	79.5	72.7	15.5
	Vieira-Morf	29.8	18.1	32.2	48.0
	Nuttall-Strand	46.9	33.1	42.3	2.0
	Burg	29.9	17.1	32.5	34.5

5 Conclusions

We have developed several new recursive algorithms, based on empirical forward and backward prediction errors, for estimating parameters of d-variate subset autoregressions, and compared their performance with the multivariate subset version of the Levinson-Durbin algorithm. Like the latter, the new algorithms require the manipulation of matrices of dimension $d \times d$ only. They are very fast compared with maximum likelihood estimation, and are found, in a range of simulations of Gaussian subset models, to give consistently higher likelihoods than the Yule-Walker models produced by the Levinson-Durbin algorithm. The Vieira-Morf algorithm has the additional

Figure 2: Boxplots of NL's for the data of Examples 5-8.



attractive property of simultaneously minimizing the weighted forward and backward prediction errors. These observations are consistent with comparable results obained also in the univariate case.

Acknowledgments

This work was partially supported by NSF Grant DMS-9972015 (PB, AT), and a von Neumann Guest Professorship in the Zentrum Mathematik, Technische Universität, München (PB).

A Appendix

A.1 Multivariate Projections

In this section we state two theorems on multivariate projection which play a key role in the proofs of Propositions 1 and 2, respectively. These theorems appear in various forms in the literature. For a proof in the form given here see Brockwell and Dahlhaus (2002).

Theorem A.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*

$$\hat{\mathbf{X}}(\mathbf{Y}, \mathbf{Z}) = A_1\mathbf{Y} + A_2\mathbf{Z}, \tag{25}$$

where

$$\begin{aligned} A_2 &= < \mathbf{X} - B\mathbf{Y}, \mathbf{Z} > v_{\mathbf{Z}|\mathbf{Y}}^{-1} =< \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}), \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}) > v_{\mathbf{Z}|\mathbf{Y}}^{-1}, && (26) \\ A_1 &= B - A_2C, && (27) \end{aligned}$$

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*

$$v_{\mathbf{X}|\mathbf{Y},\mathbf{Z}} = v_{\mathbf{X}|\mathbf{Y}} - A_2 v_{\mathbf{Z}|\mathbf{Y}} A_2'. \tag{28}$$

Interchanging the roles of $\mathbf{X}$ and $\mathbf{Z}$ in Theorem A.1 gives the following Corollary.

Corollary A.1
Under the conditions of Theorem A.1,

$$\hat{\mathbf{Z}}(\mathbf{Y}, \mathbf{X}) = D_1\mathbf{Y} + D_2\mathbf{X}, \tag{29}$$

where

$$D_2 =< \mathbf{Z} - C\mathbf{Y}, \mathbf{X} > v_{\mathbf{X}|\mathbf{Y}}^{-1} =< \mathbf{Z} - \hat{\mathbf{Z}}(\mathbf{Y}), \mathbf{X} - \hat{\mathbf{X}}(\mathbf{Y}) > v_{\mathbf{X}|\mathbf{Y}}^{-1}, \tag{30}$$

$$\begin{aligned} D_1 &= C - D_2B, && (31) \\ v_{\mathbf{Z}|\mathbf{X},\mathbf{Y}} &= v_{\mathbf{Z}|\mathbf{Y}} - D_2 v_{\mathbf{X}|\mathbf{Y}} D_2', && (32) \end{aligned}$$

and $v_{\mathbf{X}|\mathbf{Y}}^{-1}$ *is any generalized inverse of* $v_{\mathbf{X}|\mathbf{Y}}$.

The proof of Theorem A.1 and its corollary makes no use of the particular inner product, $E(X_iY_j)$, of the components X_i and Y_j of $\mathbf{X}$ and $\mathbf{Y}$. We can therefore express the results in the following more general form.

Theorem A.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 $\mathcal{S}$. For any two such vectors, $\mathbf{X}$ 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 (25)–(28) as in Theorem A.1.

Corollary A.2
Equations (29)–(32) remain valid in the context of Theorem A.2.

A.2 Minmization of Sums of Weighted Forward and Backward Prediction Errors

In this section we show how the three algorithms of Section 3.2 can be obtained as the minimizers of weighted sums of squares of the forward and backward prediction errors. Starting from the *standardized residuals*

$$\tilde{\boldsymbol{\varepsilon}}_t \equiv \hat{U}_J^{-1/2} \hat{\boldsymbol{\varepsilon}}_J(t), \quad \tilde{\boldsymbol{\eta}}_t \equiv \hat{V}_{J^*}^{-1/2} \hat{\boldsymbol{\eta}}_{J^*}(t - k_m),$$

and defining the matrices, $\tilde{A} \equiv \hat{U}_J^{1/2} A \hat{U}_J^{1/2}$, and $\tilde{B} \equiv \hat{V}_{J^*}^{1/2} B \hat{V}_{J^*}^{1/2}$, for some positive definite symmetric matrices A and B (to be specified later), we consider instead of (22), the more general weighted minimization problem:

$$\sum_{t=k_{m+1}}^{n} \left[\hat{\boldsymbol{\varepsilon}}_K(t)' A \, \hat{\boldsymbol{\varepsilon}}_K(t) + \hat{\boldsymbol{\eta}}_{K^*}(t - k_m)' B \, \hat{\boldsymbol{\eta}}_{K^*}(t - k_m) \right] , \qquad (33)$$

with $\hat{\boldsymbol{\varepsilon}}_K(t)$ and $\hat{\boldsymbol{\eta}}_{K^*}(t - k_m)$ given by (14) and (15), respectively. Letting $R = \hat{U}_J^{-1/2} \hat{\Phi}_K(k_m) \hat{V}_{J^*}^{1/2}$, (11) gives $\Psi_{K^*}(k_m) = \hat{V}_{J^*}^{1/2} R' \hat{U}_J^{-1/2}$. With these definitions, (33) becomes

$$\sum_{t=k_m+1}^{n} \left[(\tilde{\boldsymbol{\varepsilon}}_t - R\tilde{\boldsymbol{\eta}}_t)'\tilde{A}(\tilde{\boldsymbol{\varepsilon}}_t - R\tilde{\boldsymbol{\eta}}_t) + (\tilde{\boldsymbol{\eta}}_t - R'\tilde{\boldsymbol{\varepsilon}}_t)'\tilde{B}(\tilde{\boldsymbol{\eta}}_t - R'\tilde{\boldsymbol{\varepsilon}}_t) \right]. \tag{34}$$

Minimization with respect to $\hat{\Phi}_K(k_m)$, is now equivalent to minimization with respect to R. Noting that $\frac{\partial tr(CR)}{\partial R} = C'$, taking the first derivative of (34) and setting it to zero, leads to the equation

$$\sum_{t=k_m+1}^{n} \left[\tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\varepsilon}}_t' R\tilde{B} + \tilde{A}R\tilde{\boldsymbol{\eta}}_t\tilde{\boldsymbol{\eta}}_t' \right] = \sum_{t=k_m+1}^{n} \left[\tilde{A}\tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\eta}}_t' + \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\eta}}_t'\tilde{B} \right]. \tag{35}$$

Noting that $\operatorname{vec}(CRD) = (D' \otimes E)\operatorname{vec} R$, we vec both sides of (35) to obtain

$$\begin{aligned} &\left[\tilde{B} \otimes \left(\sum_t \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\varepsilon}}_t' \right) + \left(\sum_t \tilde{\boldsymbol{\eta}}_t\tilde{\boldsymbol{\eta}}_t' \right) \otimes \tilde{A} \right] \operatorname{vec} R \\ &\qquad = \operatorname{vec}\left[\tilde{A}\left(\sum_t \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\eta}}_t' \right) + \left(\sum_t \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\eta}}_t' \right)\tilde{B} \right], \end{aligned} \tag{36}$$

which gives finally

$$\begin{aligned} \operatorname{vec} R = &\left[\tilde{B} \otimes \left(\sum_t \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\varepsilon}}_t' \right) + \left(\sum_t \tilde{\boldsymbol{\eta}}_t\tilde{\boldsymbol{\eta}}_t' \right) \otimes \tilde{A} \right]^{-1} \\ &\qquad \times \operatorname{vec}\left[\tilde{A}\left(\sum_t \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\eta}}_t' \right) + \left(\sum_t \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\eta}}_t' \right)\tilde{B} \right]. \end{aligned} \tag{37}$$

It can be shown that the second derivative of (34) with respect to $\operatorname{vec} R$ is equal to

$$2\left[\tilde{B} \otimes \left(\sum_t \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\varepsilon}}_t' \right) + \left(\sum_t \tilde{\boldsymbol{\eta}}_t\tilde{\boldsymbol{\eta}}_t' \right) \otimes \tilde{A} \right].$$

Each of the two summands is positive definite almost surely, implying that the above minimum is unique almost surely.

Different choices of A and B now lead to different algorithms:

Subset Nutall-Strand Algorithm. Setting $\tilde{A} = \tilde{B} = I_d$, gives, with $\sum_t \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\varepsilon}}_t' = \hat{U}_J^{-1/2}\hat{\Omega}_{\epsilon\epsilon}\hat{U}_J^{-1/2}$, and $\sum_t \tilde{\boldsymbol{\eta}}_t\tilde{\boldsymbol{\eta}}_t' = \hat{V}_{J^*}^{-1/2}\hat{\Omega}_{\eta\eta}\hat{V}_{J^*}^{-1/2}$, the solution (19), and as a consequence, (18) for $\hat{\Phi}_K(k_m)$. Thus, the subset

Nutall-Strand algorithm is obtained by minimizing a *weighted* sum of forward and backward prediction errors, where the weights are chosen to be the inverse of the covariance matrix of the prediction errors in the previous step.

The computationally more convenient form (20)-(21) of the same algorithm, is obtained by setting $\tilde{A} = \tilde{B} = I_d$ in (35), and multiplying the equation from the left by $\hat{U}_J^{1/2}$, and from the right by $\hat{V}_{J^*}^{1/2}$. Setting $\Delta \equiv \hat{U}_J^{1/2} R \hat{V}_{J^*}^{1/2}$, leads to

$$\hat{\Omega}_{\epsilon\epsilon}\hat{U}_J^{-1}\Delta + \Delta\hat{V}_{J^*}^{-1}\hat{\Omega}_{\eta\eta} = 2\hat{\Omega}_{\epsilon\eta},$$

which ultimately gives (21).

Subset Burg Algorithm. Setting $A = B = I_d$ gives the solution (23) of the subset Burg algorithm. This is obtained by setting $\tilde{A} = \hat{U}_J$ and $\tilde{B} = \hat{V}_{J^*}$ in (35), and multiplying the equation from the left by $\hat{U}_J^{-1/2}$, and from the right by $\hat{V}_{J^*}^{1/2}$. With $\hat{\Phi}_K(k_m) = \hat{U}_J^{1/2} R \hat{V}_{J^*}^{-1/2}$, this leads to

$$\hat{U}_J^{-1}\hat{\Omega}_{\epsilon\epsilon}\hat{U}_J^{-1}\hat{\Phi}_K(k_m)\hat{V}_{J^*}^2 + \hat{\Phi}_K(k_m)\hat{\Omega}_{\eta\eta} = \hat{\Omega}_{\epsilon\eta} + \hat{U}_J^{-1}\hat{\Omega}_{\epsilon\eta}\hat{V}_{J^*},$$

which implies (23).

Subset Vieira-Morf Algorithm. The subset Vieira-Morf algorithm is obtained, *not* as a special case of the above, but as the solution of a similar minimization problem. If we standardize the residuals by their empirical standard deviations, i.e. if we set

$$\tilde{\boldsymbol{\varepsilon}}_t = \hat{\Omega}_{\epsilon\epsilon}^{-1/2}\hat{\boldsymbol{\varepsilon}}_J(t), \quad \text{and} \quad \tilde{\boldsymbol{\eta}}_t = \hat{\Omega}_{\eta\eta}^{-1/2}\hat{\boldsymbol{\eta}}_{J^*}(t - k_m),$$

we obtain $(\sum_t \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\varepsilon}}_t') = 1 = (\sum_t \tilde{\boldsymbol{\eta}}_t\tilde{\boldsymbol{\eta}}_t')$. If we set $\tilde{A} = \tilde{B} = I_d$, minimization of (34) with respect to R now gives, as a special case of (37), the solution

$$R = \frac{1}{n - k_m}\sum_{t=k_{m+1}}^{n} \tilde{\boldsymbol{\varepsilon}}_t\tilde{\boldsymbol{\eta}}_t' = \hat{\Omega}_{\epsilon\epsilon}^{-1/2}\hat{\Omega}_{\epsilon\eta}\hat{\Omega}_{\eta\eta}^{-1/2},$$

which is exactly (17). The same solution is obtained if we minimize either the forward prediction error, by setting $\tilde{A} = I_d$ and $\tilde{B} = 0$, or the backward prediction error, by setting $\tilde{A} = 0$ and $\tilde{B} = I_d$. This is a particularly nice property which sets the Vieira-Morf algorithm apart from the other two.

References

[1] Brockwell P.J., and Dahlhaus R. (2002), "Generalized Levinson-Durbin and Burg Algorithms", *Journal of Econometrics*, to appear.

[2] Brockwell P.J., and Davis R.A. (1991), *Time Series: Theory and Methods*, 2nd ed., New York: Springer-Verlag.

[3] Burg, J.P. (1968), "A New Analysis Technique for Time Series Data", in *Modern Spectrum Analysis*, (1978), D.G. Childers (ed.), NATO Advanced Study Institute of Signal Processing with emphasis on Underwater Acoustics, New York: IEEE Press.

[4] Haykin, S. (1996), *Adaptive Filter Theory*, New York: Prentice-Hall.

[5] Hooke, R. and Jeeves T. (1961), "A direct search solution of numerical and statistical problems", *Journal of Association for Computing Machinery*, 8, 212-229.

[6] Itakura, F. and Saito, S. (1971), "Digital filtering techniques for speech analysis and synthesis", *Proceedings of the 7th International Congress on Acoustics*, Budapest, Paper 25-C-1, 261-264.

[7] Jones, R.H. (1978), "Multivariate autoregression estimation using residuals", in *Applied Time Series Analysis*, (1978), D.F. Findley (ed.), Proceedings of the First Applied Time Series Symposium, Tulsa, Okla., 1976, New York: Academic Press.

[8] McClave, J. (1975), "Subset Autoregression", *Technometrics* 17, 213-220.

[9] Makhoul, J. (1977), "Stable and efficient lattice methods for linear prediction", *IEEE Transactions on Acoustics, Speech and Signal processing* 25, 423-428.

[10] Morf, M., Vieira, A., Lee, D.T.L., and Kailath, T. (1978), "Recursive Multichannel Maximum Entropy Spectral Estimation", *IEEE Transactions on Geoscience and Electronics*, GE-16, 85-94.

[11] Penm, J.H.W. and Terrell, R.D. (1982), "On the recursive fitting of subset autoregressions", *Journal of Time Series Analysis*, 3, 43-59.

[12] Sarkar, A. and Sharma, K.M.S. (1997), "An Approach to Direct Selection of Best Subset AR Model", *Journal of Statistical Computation and Simulation*, 56, 273-291.

[13] Strand, O.N. (1977), "Multichannel complex maximum entropy (autoregressive) spectral analysis", *IEEE Trans. Automat. Control.*, 22, 634-640.

[14] Trindade, A.A. (2001), "Implementing Modified Burg Algorithms in Multivariate Subset Autoregressions", Dept. of Statistics Technical Report 2001-002, University of Florida.

[15] Zhang, X. and Terrell, R.D. (1997), "Projection Modulus: A New Direction for Selecting Best Subset Autoregressive Models", *Journal of Time Series Analysis*, 18, 195-212.