A Note on Generation, Estimation and Prediction of Stationary Processes

Michael A. Hauser[1], Wolfgang Hörmann[1], Robert M. Kunst[2] and Jörg Lenneis[1]
[1]University of Economics and Business Administration,
Department of Statistics, Vienna
[2]Institute for Advanced Studies, Vienna.

is published in:

Dutter, R. and Grossman, W.(eds.) (1994):
COMPSTAT 1994, Proceedings in Computational Statistics,
Physica-Verlag, Heidelberg
pp. 323-328

Unfortunately there is an error in the printed version in equation (9).
The following

The variance of the linear predictor is, however,

$$\mathrm{V}[Y_\tau|Y_T] = L_{\tau\tau} H_\tau L'_{\tau\tau} \tag{9}$$

with $H_{T+\tau} = \begin{bmatrix} H_T & 0 \\ 0 & H_\tau \end{bmatrix}$.

should be read as

The variance of the linear predictor is, however,

$$\mathrm{V}[Y_\tau|Y_T] = L_{\tau\tau} H_{\tau|Y_T} L'_{\tau\tau} \tag{9}$$

with $H_{T+\tau} = \begin{bmatrix} H_T & 0 \\ 0 & H_\tau \end{bmatrix}$ and $H_{\tau|Y_T} = \mathrm{E}[H_\tau|Y_T]$.

The enclosed copy is corrected for that error.

A Note on Generation, Estimation and Prediction of Stationary Processes

Michael A. Hauser[1], Wolfgang Hörmann[1], Robert M. Kunst[2] and Jörg Lenneis[1]
[1]University of Economics and Business Administration,
Department of Statistics, Vienna
[2]Institute for Advanced Studies, Vienna.

Keywords. Cholesky decomposition, Toeplitz matrices, fractional integration

1 Introduction

Some recently discussed stationary processes like fractionally integrated processes cannot be described by low order autoregressive or moving average (ARMA) models rendering the common algorithms for generation estimation and prediction partly very misleading [cf. Hosking(1981,1984), Sowell(1992), Ray(1993)]. We offer an unified approach based on the Cholesky decomposition of the covariance matrix which makes these problems exactly solvable in an efficient way.

Our starting point are stationary processes with a Wold representation of the form

$$y_t - \mu = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}, \tag{1}$$

where ϵ_t is uncorrelated noise with mean zero. The ψ_i are quadratic summable and the (unconditional) variance of the noise, σ_ϵ^2, is greater than zero. We assume for simplicity of the presentation that $\mu = 0$. Y_T denotes the vector $(y_1, \ldots, y_T)'$ and $E_T = (\epsilon_1, \ldots, \epsilon_T)'$. The covariance matrix of Y_T, Σ_T, is positive definite, symmetric and Toeplitz, and thus persymmetric. It may by factorized according to the Cholesky decomposition.

$$\Sigma_T = L_T L_T'. \tag{2}$$

L_T is a lower triangular matrix.

One possibility for the *generation* of a sample of length T of a given process which possesses exactly the same covariance structure is to use the relation

$$Y_T = L_T E_T. \tag{3}$$

Under the assumption of normal distributed noise *estimation* may be performed by maximizing the Gaussian likelihood

$$f(Y_T; \mu, \Sigma_T) = (2\pi)^{-T/2} |\Sigma_T|^{-1/2} \exp[-(Y_T - \mu)' \Sigma_T^{-1} (Y_T - \mu)/2]. \tag{4}$$

For ARMA models there exist computationally simpler presentations of the likelihood. For fractionally integrated models, however, this is the only known exact form [Li and McLeod(1986) or Sowell(1992)].

The implicit *noise* vector may be obtained by

$$E_T = L_T^{-1} Y_T. \tag{5}$$

The *linear prediction* for one step to τ steps ahead may simply be performed by extending the above equation to $T+\tau$ and replacing the future noises by their expectation which is zero. This is

$$Y_{T+\tau} = \begin{bmatrix} Y_T \\ Y_\tau \end{bmatrix} = L_{T+\tau} E_{T+\tau} = \begin{bmatrix} L_T & 0 \\ L_{\tau T} & L_{\tau\tau} \end{bmatrix} \begin{bmatrix} E_T \\ E_\tau \end{bmatrix} \text{ and}$$

$$\mathrm{E}[Y_\tau | Y_T] = \begin{bmatrix} L_{\tau T} & L_{\tau\tau} \end{bmatrix} \begin{bmatrix} E_T \\ 0 \end{bmatrix} = L_{\tau T} E_T. \tag{6}$$

The *variance of the linear forecast* Y_τ given Y_T, E_T respectively, is given by means of the covariance matrix $\Sigma_{T+\tau}$, with $\Sigma_{T+\tau} = L_{T+\tau} L'_{T+\tau}$,

$$\mathrm{V}[Y_\tau | Y_T] = \mathrm{E}[(Y_\tau - \mathrm{E}[Y_\tau|Y_T])(Y_\tau - \mathrm{E}[Y_\tau|Y_T])' | Y_T] = L_{\tau\tau} L'_{\tau\tau}. \tag{7}$$

If the *innovations* are *conditional heteroscedastic* and Gaussian - i.e. ϵ_t are uncorrelated and normal with non-constant variances, which depend on the past - the process likelihood is given by (4) by replacing the covariance matrix Σ_T by a process dependent covariance matrix [see Hauser and Kunst (1993)]

$$\Sigma_T = L_T H_T L'_T \tag{8}$$

where H_T is diagonal and contains the conditional variances of the normalized ϵ_t. In case of homoscedasticity the H_T matrix reduces to I_T.
Generation and linear prediction is analogous to the homoscedastic case once the heteroscedastic innovations are given. The variance of the linear predictor is, however,

$$\mathrm{V}[Y_\tau | Y_T] = L_{\tau\tau} H_{\tau|Y_T} L'_{\tau\tau} \tag{9}$$

with $H_{T+\tau} = \begin{bmatrix} H_T & 0 \\ 0 & H_\tau \end{bmatrix}$ and $H_{\tau|Y_T} = \mathrm{E}[H_\tau | Y_T]$.

The numerical problems addressed above can be summarized as follows: Generation and prediction require the calculation of the Cholesky factor, the inverse of the Cholesky factor, and the repeated multiplication of the Cholesky factor with an arbitrary vector. Estimation, i.e. the inversion of the covariance matrix, may be implemented by factorizing Σ_T^{-1} in a MDM', M a lower triangular matrix with ones in the diagonal, D a diagonal matrix, via the Levinson algorithm. The determinant of the covariance matrix is then equal $|D|$. The calculation of the variance of the predictor may be obtained

by calculating only the lower right $\tau \times \tau$ part of the Cholesky matrix.
How the necessary operations can be performed in an efficient way is discussed below.

2 The multiplication of the Cholesky factor with an arbitrary vector

Notation and some properties of Toeplitz matrices:

$$\Sigma_{T+1} = \begin{bmatrix} \Sigma_T & \Sigma'_{1T} \\ \Sigma_{1T} & \Sigma_{11} \end{bmatrix}, L_{T+1} = \begin{bmatrix} L_T & 0 \\ L_{1T} & L_{11} \end{bmatrix}, R_{T+1} = \begin{bmatrix} R_T & Er \\ (Er)' & 1 \end{bmatrix}$$

E is a square matrix with ones in the secondary diagonal and zeros else. It holds that $EE = I, E^{-1} = E$. R_T is the correlation matrix, $\Sigma_T = \sigma_y^2 R_T$. It is symmetric and Toeplitz, so that $ER_TE = R_T$ and $ER_T = R_TE$ holds. R_T^{-1} is also symmetric and persymmetric.

Lemma 1: [Brockwell and Davis(1991, p.168)]
The best linear 1-step ahead predictor of $\hat{y}_{T+1}$ of y_{T+1} in terms of Y_T and its mean squared error are

$$\hat{y}_{T+1} = \Sigma_{1T}\Sigma_T^{-1}Y_T, \qquad v_T = \Sigma_{11} - \Sigma_{1T}\Sigma_T^{-1}\Sigma'_{1T}. \tag{10}$$

In case of multivariate normal distributed Y_{T+1} this is identical to the moments given by the conditional normal distribution [Johnson(1987, p.50)]. The coefficient in front of y_1 may be interpreted as the T-th partial autocovariance.

Proposition 1:
The best linear 1-step ahead predictor of $\hat{y}_{T+1}$ of y_{T+1} in terms of the Cholesky factors and past innovation vector E_T and its mean squared error are

$$\hat{y}_{T+1} = L_{1T}E_T, \qquad v_T = L_{11}L_{11}. \tag{11}$$

Proof:
This may be easily seen by using $\Sigma_{T+1} = L_{T+1}L'_{T+1}$ in the partition representation as given above, multiplying out, and replacing the Σ-matrices by the corresponding expressions in terms of the L-matrices in (10). For Y_T use $Y_T = L_TE_T$. □

The predictor is given by the multiplication of the last line of the Cholesky matrix by the vector $(E_T, 0)'$.

For the generation of samples of a process with given true covariance matrix the best linear predictor can be easily used recursively in the following way starting at $T = 0$ with $v_0 = \sigma_y^2$ [cf. Hosking(1984, p.1900)]:

$$y_{T+1} = \Sigma_{1T}\Sigma_T^{-1}Y_T + \sqrt{v_T}\epsilon_{T+1}, \tag{12}$$

where the ϵ_t are an (possibly heteroscedastic) innovation sequence.

In notation of the Cholesky matrix this amounts to

$$y_{T+1} = L_{1T}E_T + L_{11}\epsilon_{T+1}. \tag{13}$$

This is the multiplication of the last line of L_{T+1} with E_{T+1}, or more compactly for the whole vector Y_{T+1}, $Y_{T+1} = L_{T+1}E_{T+1}$.

An efficient algorithm to compute the best linear predictor and its mean squared error is the Durbin-Levinson algorithm [Brockwell and Davis(1991, p.169)]. Thus the Durbin-Levinson algorithm does multiply the Cholesky matrix with the vector E_{T+1} by requiring $O(T^2)$ flops and $O(T)$ storage. More generally, this algorithm performs the multiplication of the Cholesky matrix of a symmetric Toeplitz matrix with any arbitrary vector. This is remarkable, since there is no procedure known for the simply structured Toeplitz matrices to compute the Cholesky matrix with less than $O(T^3)$ flops and $O(T^2)$ storage. Below we will give a derivation of an equivalent algorithm based on matrix computations and the use of the Durbin algorithm which solves the Yule-Walker equations [Golub and VanLoan(1989, p.185)].

Derivation of the algorithm:
The idea for the algorithm is identical to the first step of the recursion of the Trench algorithm as presented in Golub and VanLoan(1989, p.188). For simplicity we reformulate the problem in correlations instead of covariances, which implies $\sigma_y^2 = 1$, $\Sigma_T = \sigma_y^2 R_T$ respectively. The first two moments of y_{T+1} as given in (10) simplify to $(Er)'R_T^{-1}Y_T$ and $1 - r'R_T^{-1}r$ using the properties of Toeplitz matrices and the matrix E and the notation given above.
$R_{T+1}^{-1} = \begin{bmatrix} R_T & Er \\ (Er)' & 1 \end{bmatrix}^{-1} = \begin{bmatrix} B & v \\ v' & \gamma \end{bmatrix}$. This implies that
$\begin{bmatrix} R_T & Er \\ (Er)' & 1 \end{bmatrix} \begin{bmatrix} v \\ \gamma \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.
Solving this system for v and γ yields $R_T v = -\gamma Er$ from the first equation. So v can be expressed via the solution y of the Yule-Walker equations, $R_T y = -r$, $y = -R_T^{-1}r$ and $v = \gamma Ey$.
By replacing v in the second equation γ can be expressed as $\gamma = 1/(1+r'y) = 1/(1 - r'R_T^{-1}r)$.
The first two moments of y_{T+1} can be then expressed in terms of y. That is: $(Er)'R_T^{-1}Y_T = -(Ey)'Y_T$ and $1 - r'R_T^{-1}r = 1/\gamma = 1 + r'y$. □

The algorithm gives the multiplication of the Cholesky factor with an arbitrary vector. The storage requirements are those of Durbin algorithm which are linear. Its number of flops are $O(T^2)$ which increases by two vector multiplications.

3 The inverse of the Cholesky factor

Proposition 2:
The inverse Cholesky matrix is related to the Cholesky matrix of the inverse by transposing with respect to the secondary diagonal.

Proof:
Σ is positive definite, symmetric and persymmetric. The Cholesky decompositions of Σ and its inverse, which is also symmetric and persymmetric, are $\Sigma = AA'$ and $\Sigma^{-1} = BB'$.
The inversion of the first decomposition is $\Sigma^{-1} = (A')^{-1}A^{-1}$. A, A^{-1} and B are lower triangular matrices. So there is a lower triangular decomposition and an upper triangular decomposition of the same matrix.
$\Sigma^{-1} = E\Sigma^{-1}E = E(A')^{-1}EEA^{-1}E = (E(A')^{-1}E)(EA^{-1}E) = BB'$. Since the Cholesky decomposition is well defined $EA^{-1}E = B'$ and, thus, $A^{-1} = EB'E$ follows. $\square$

4 Computations

As given above *generation* of samples of the process (y_t) may be obtained efficiently in linear storage requirements, once the autocorrelation function is given. [For the calculation of the autocovariance function of fractional integrated processes see Sowell(1992).]

If the *estimation* is performed via the likelihood function given in (4) the Levinson algorithm [see Marple(1987, p.87)] may be used to calculate the Cholesky decomposition of Σ_T^{-1}, $\Sigma_T^{-1} = MDM'$, and thus also the required determinant. This algorithm is $O(T^2)$ in storage and $O(T^2)$ in flops.
The resulting *innovations* may be calculated using the Cholesky decomposition of the last iteration of the optimization procedure, Proposition 2 and (5).

The linear 1- to τ-step *prediction* (forecast) vector given Y_T may be calculated via the (estimated) residual vector and (6) - linear in storage and quadratic in flops - using the (estimated) autocovariance function.

Especially in case of calculating the *variance of the linear predictor*, (7), Proposition 1 is very helpful since τ is typically small. Multiplying $L_{T+\tau}$ by a vector with zeros and a 1 in position $(T+j)$ picks out exactly the $(T+j)$-

th column which is the column j in $L_{\tau\tau}$. Without storing the intermediate results of the multiplication of $L_{T+\tau}$ with the first T zeros the number of flops is $O(\tau T^2)$. The storage is linear if the diagonal elements are needed only.
The procedure can be easily generalized for heteroscedastic innovations. The 1's have to be replaced by the square root of the conditional variances.

5 Summary

An efficient algorithm - $O(T)$ in storage and $O(T^2)$ in flops - for multiplying the Cholesky factor by an arbitrary vector is presented. It may be used for generation of linear processes, linear prediction and calculation of the predictor variance.
It is shown that the Cholesky factor of an inverse symmetric Toeplitz matrix is a simple function of the inverse Cholesky factor of the Toeplitz matrix itself. Thus, given the Cholesky factor of the inverse covariance matrix the noise vector may be easily obtained.
We have outlined that for the simulation of stationary processes, for estimation and prediction two different algorithms are sufficient: the Levinson algorithm for calculating the Cholesky decomposition of the inverse covariance matrix and the algorithm giving a multiplication of a vector with the Cholesky matrix of the covariance matrix. Moreover this way is also very efficient.

References

Brockwell, P.J. and Davis, R.A., 1991, Time series: Theory and methods (Springer, New York).
Golub, G.H. and Van Loan, Ch.F., 1989, Matrix computations, (John Hopkins University Press, Baltimore).
Hauser, M.A. and Kunst, R.M., 1993, Fractionally Integrated Models with ARCH Errors, Paper presented at ESEM, Uppsala.
Hosking, J.R.M., 1981, Fractional differencing, Biometrika, 68, 165-176.
Hosking, J.R.M., 1984, Modelling persistence in hydrological time series using fractional differencing, Water Resources Research, 20, 1898-1908.
Johnson, M.E. 1987, Multivariate Statistical Simulation (John Wiley, New York).
Li, W.K. and McLeod, A.I., 1986, Fractional time series modelling, Biometrika, 73, 217-221.
Marple, S.L. Jr., 1987, Digital Spectral Analysis (Prentice Hall, Englewood Cliffs).
Ray, B.K., 1993, Modeling long-memory processes for optimal long-range prediction, Journal of Time Series Analysis, 14, 511-526.
Sowell, F., 1992, Maximum likelihood estimation of stationary univariate fractionally integrated time series models, Journal of Econometrics, 53, 165-188.