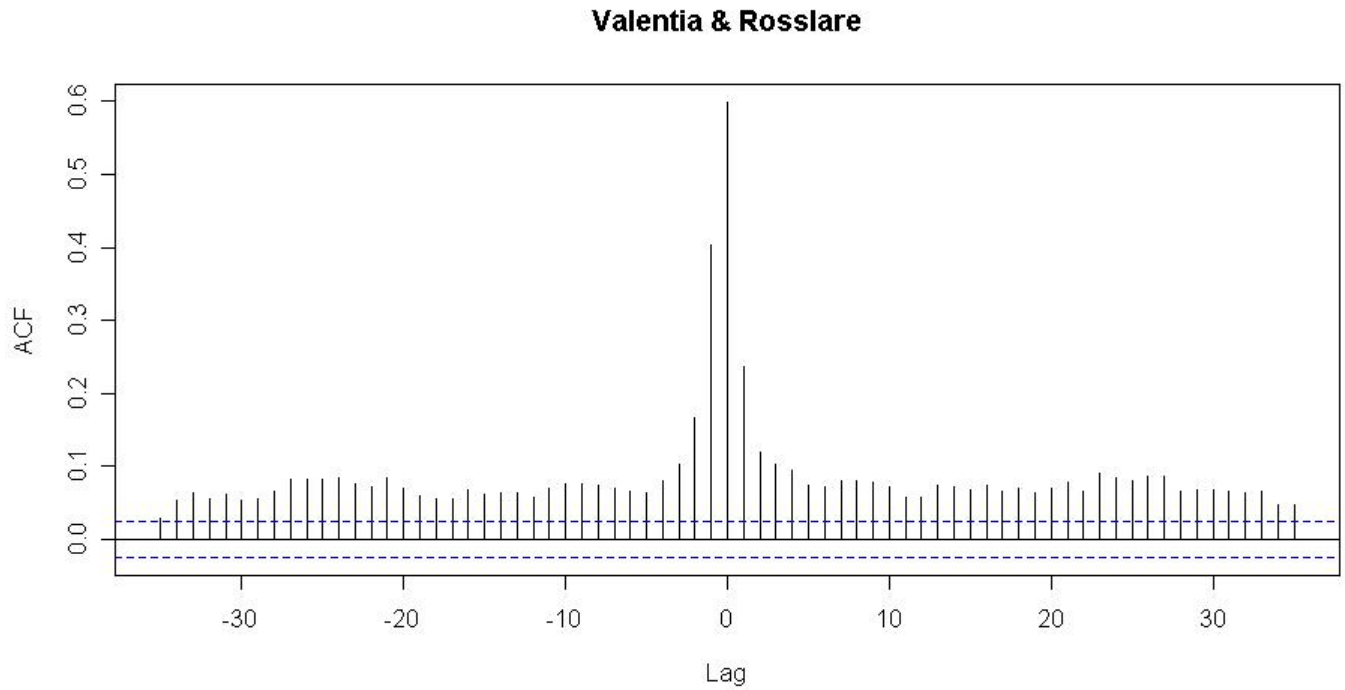
Computationally Efficient Gaussian Maximum Likelihood Methods for Vector ARFIMA Models

Rebecca J. Sela
Stern School of Business, New York University
Joint work with Clifford Hurvich

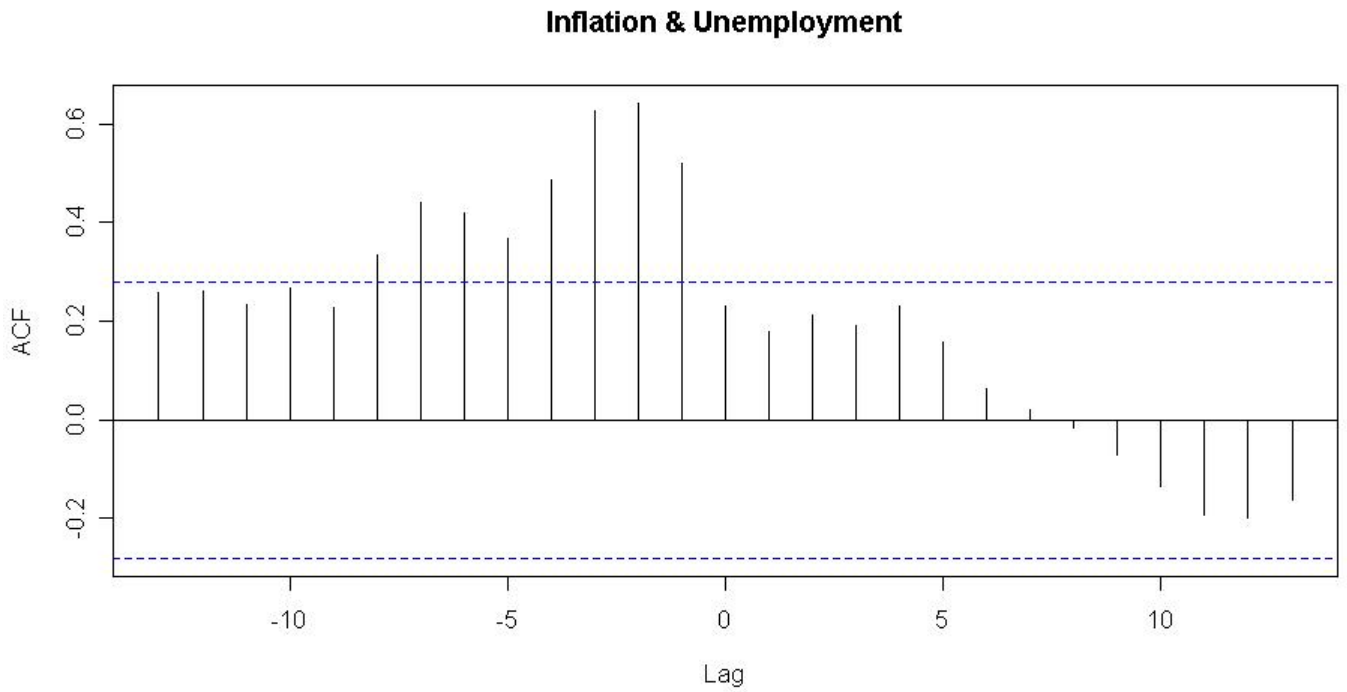
May 9, 2008

Slowly decaying cross-covariances appear in meteorological data.



The estimated cross-correlation function of daily wind speeds in Valentia and Rosslare, Ireland, from 1961 to 1978.

They can also appear in macroeconomic data.



The estimated cross-correlation function of the unemployment rate and the inflation rate, based on annual data from 1948 to 1996.

The most commonly used multivariate model is a vector autoregression.

$$\begin{aligned} A(L)X_t &= \epsilon_t \\ \epsilon_t &\sim Normal(0, \Sigma) \end{aligned}$$

All the roots of $|A(L)|$ must be outside the unit circle for the model to be stationary.

$$X_t - A_1 X_{t-1} = \epsilon_t$$

All the singular values of A_1 must be less than one for the model to be stationary.

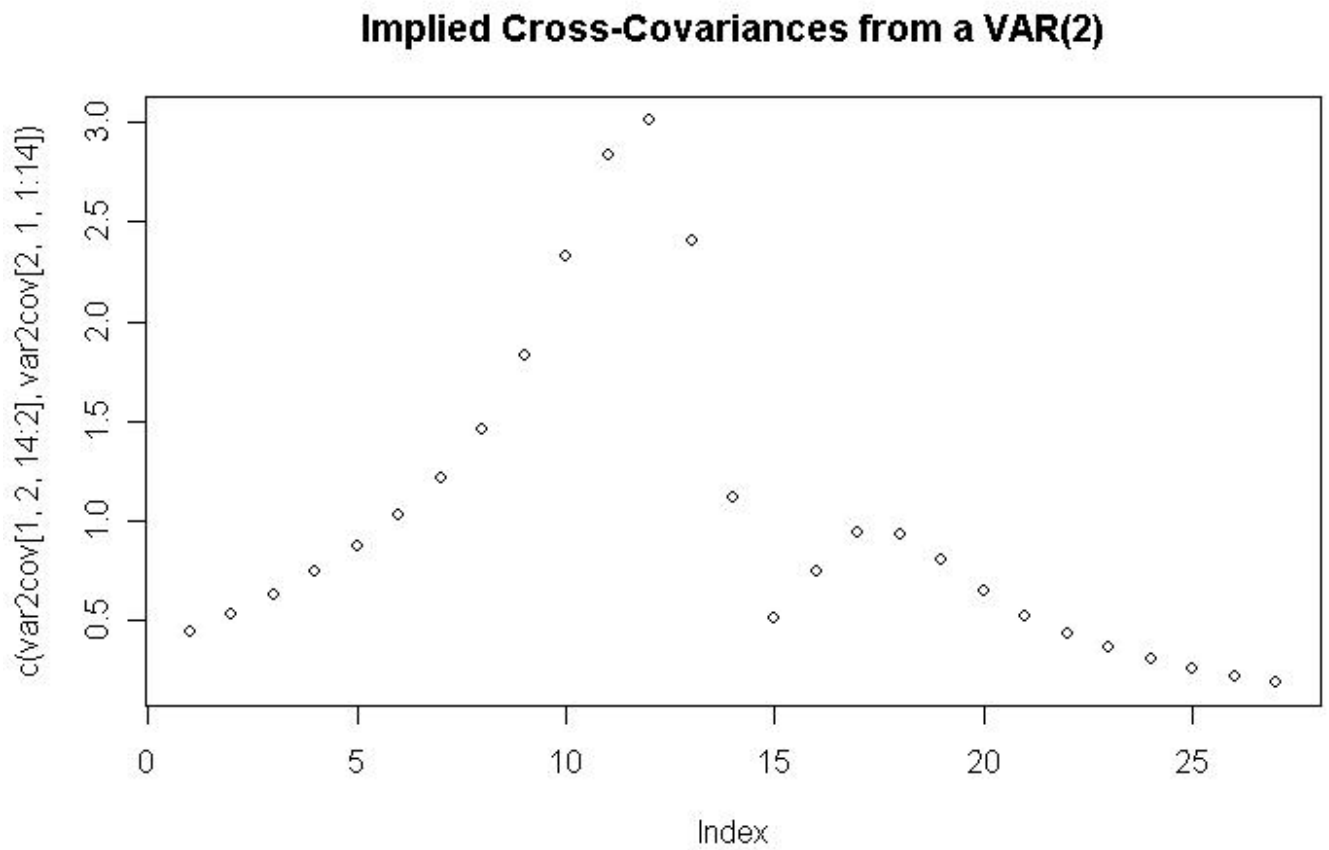
The covariances of a VAR decay at an exponential rate.

For a $VAR(1)$, with $A(L) = I - A_1 L$,

$$\begin{aligned} vec(\text{Var}(X_t)) &= (I_{K^2} - A_1 \otimes A_1)^{-1} vec(\Sigma) \\ \text{Cov}(X_t, X_{t-r}) &= A_1^r \text{Var}(X_t), r > 0 \\ \text{Cov}(X_t, X_{t+r}) &= \text{Cov}(X_t, X_{t-r})' \end{aligned}$$

Since any VAR can be written as a VAR(1), the covariances of any VAR decay exponentially fast.

Vector autoregressive models cannot capture slowly decaying cross-covariances.



Long memory models allow for slowly-decaying autocorrelations.

- Long memory models are defined in part by their differencing parameter, d.
- If $d = 0$, the models do not have long memory. Autoregressive models are one type of model with $d = 0$.
- If $0 < |d| < \frac{1}{2}$, then the model is stationary, with autocovariances which of the form, $\omega(r) \sim C|r|^{2d-1}$.

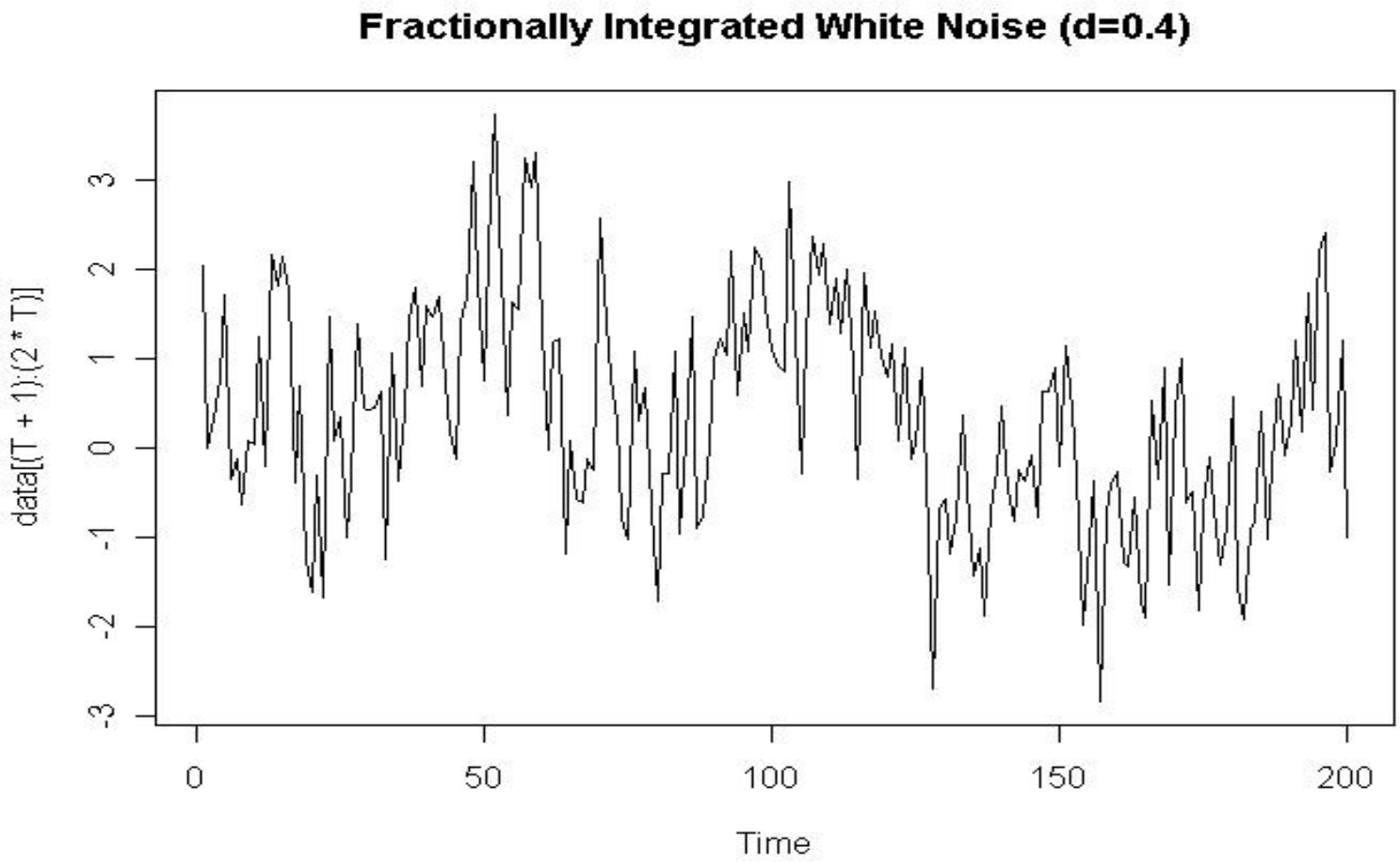
The simplest example of a univariate long memory model is fractionally integrated white noise.

$$
\begin{aligned}
(1-L)^d y_t &= \epsilon_t \\
\epsilon_t &\sim Normal(0, \sigma^2) \\
(1-L)^d &= \sum_{j=0}^{\infty} (-1)^j \binom{d}{j} L^j \\
\binom{d}{j} &= \frac{d(d-1)\cdots(d-j+1)}{j!}
\end{aligned}
$$

where $0 < |d| < \frac{1}{2}$. The spectral density is:

$$\frac{\sigma^2}{2\pi} |1 - e^{-i\lambda}|^{-2d}$$

What fractionally integrated white noise looks like:



We can combine fractionally integrated white noise and ARMA models to create ARFIMA models.

We can have an ARMA model driven by fractionally integrated white noise:

$$a(L)x_t = b(L)[(1-L)^{-d}\epsilon_t]$$

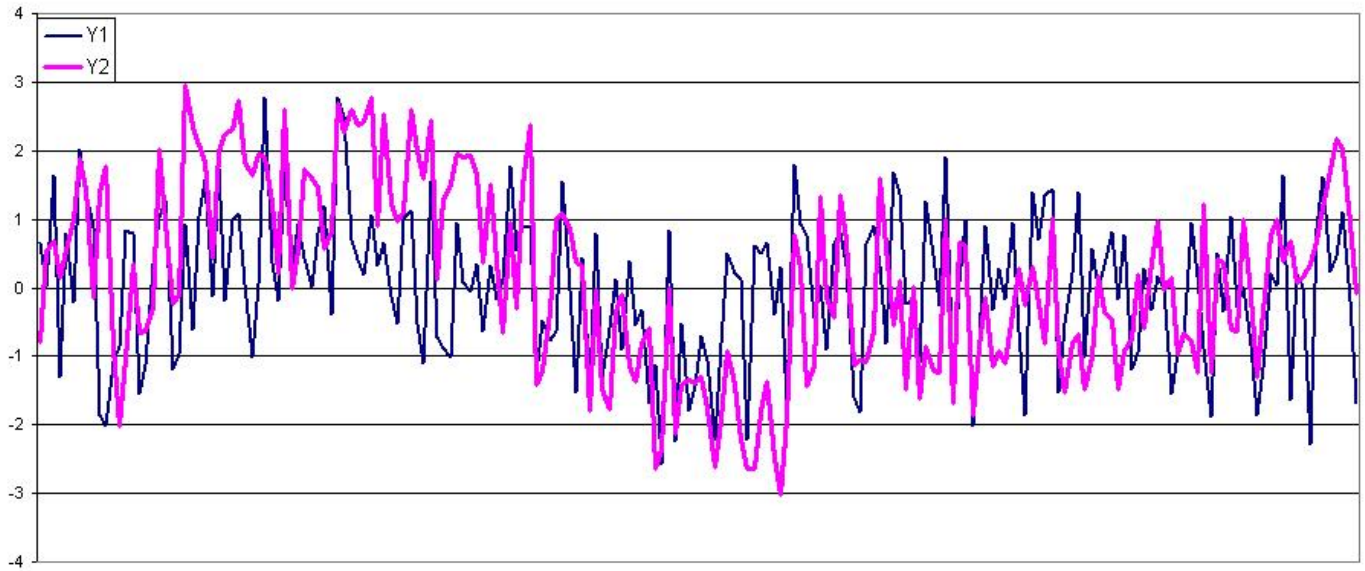
Equivalently, we can have an ARMA model which is then fractionally integrated:

$$x_t = (1-L)^{-d}\left(\frac{b(L)}{a(L)}\epsilon_t\right)$$

Now, we define multivariate fractionally integrated white noise.

$$
\begin{aligned}
D(L)Y_t &= \epsilon_t \\
\epsilon_t &\sim Normal(0, \Sigma) \\
D(L) &= \begin{pmatrix} (1-L)^{d_1} & 0 & \dots & 0 \\ 0 & (1-L)^{d_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (1-L)^{d_K} \end{pmatrix}
\end{aligned}
$$

What fractionally integrated multivariate white noise looks like:



$$d = (0.1, 0.4), \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

Now, there are two possible combinations of fractionally integrated white noise and VAR models.

We can have a VAR model driven by fractionally integrated white noise (VAR-FI):

$$A(L)X_t = D(L)^{-1}\epsilon_t$$

We can also have a VAR model which is then fractionally integrated (FI-VAR):

$$X_t = D(L)^{-1}\left(A(L)^{-1}\epsilon_t\right)$$

These two models are not equivalent.
(The difference between the two types of models was first noted by Lobato, 1997.)

These models have different properties.

- FIVAR: $A(L)D(L)X_t = \epsilon_t$
  - The k^{th} time series is integrated of order d_k.
  - There is no cointegration.
- VARFI: $D(L)A(L)X_t = \epsilon_t$
  - It is possible that each time series is integrated of order $\max(\vec{d})$.
  - It is often possible to find a linear combination of present and past values of the elements of X_t, with exponentially decaying weights, which is integrated of order less than $\max(\vec{d})$.

Now that we have two models, we might want to:

- Compute the autocovariance and cross-covariance sequences.
- Simulate from it.
- Estimate it with Gaussian maximum likelihood, which will require us to:
  - Compute a quadratic form.
  - Compute a determinant.

For FIVAR processes, there are three possible ways to compute the autocovariance sequence:

- We could use the relationship between the spectral density and the autocovariances and apply numerical integration methods:

$$Cov(X_t, X_{t-h}) = \int_{-\pi}^{\pi} e^{ih\lambda} f_X(\lambda) d\lambda$$

- Sowell (1989b) worked out an exact expression for the autocovariances of the FIVAR process, which depends on the hypergeometric function.

Both of these methods are slow.

We will generalize the "splitting method" of Bertelli and Caporin (2002) instead.

Define

$$
\begin{aligned}
X_t &= D(L)^{-1} Z_t = \sum_{j=0}^{\infty} \Psi_j Z_{t-j} \\
Z_t &= A(L)^{-1} \epsilon_t \\
\mathrm{Cov}(X_t, X_{t-h}) &= \mathrm{Cov}\left(\sum_{i=0}^{\infty} \Psi_i Z_{t-i}, \sum_{j=0}^{\infty} \Psi_j Z_{t-j-h} \right) \\
&= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \Psi_i \mathrm{Cov}(Z_{t-i}, Z_{t-h-j}) \Psi_j'
\end{aligned}
$$

Since the Ψ_j are diagonal matrices, we may write the (k, l) element of the sum as:

$$\begin{aligned}
\text{Cov}(X_{k,t}, X_{l,t-h}) &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\psi(i,d_k)\psi(j,d_l)\text{Cov}(Z_{k,t}, Z_{l,t-h-j+i}) \\
&= \sum_{m=0}^{\infty}\sum_{j=m}^{\infty}\psi(j-m,d_k)\psi(j,d_l)\text{Cov}(Z_{k,t}, Z_{l,t-h-m}) \\
&= \sum_{m=0}^{\infty}\text{Cov}(Z_{k,t}, Z_{l,t-h-m})\left(\sum_{j=m}^{\infty}\psi(j,d_l)\psi(j-m,d_k)\right)
\end{aligned}$$

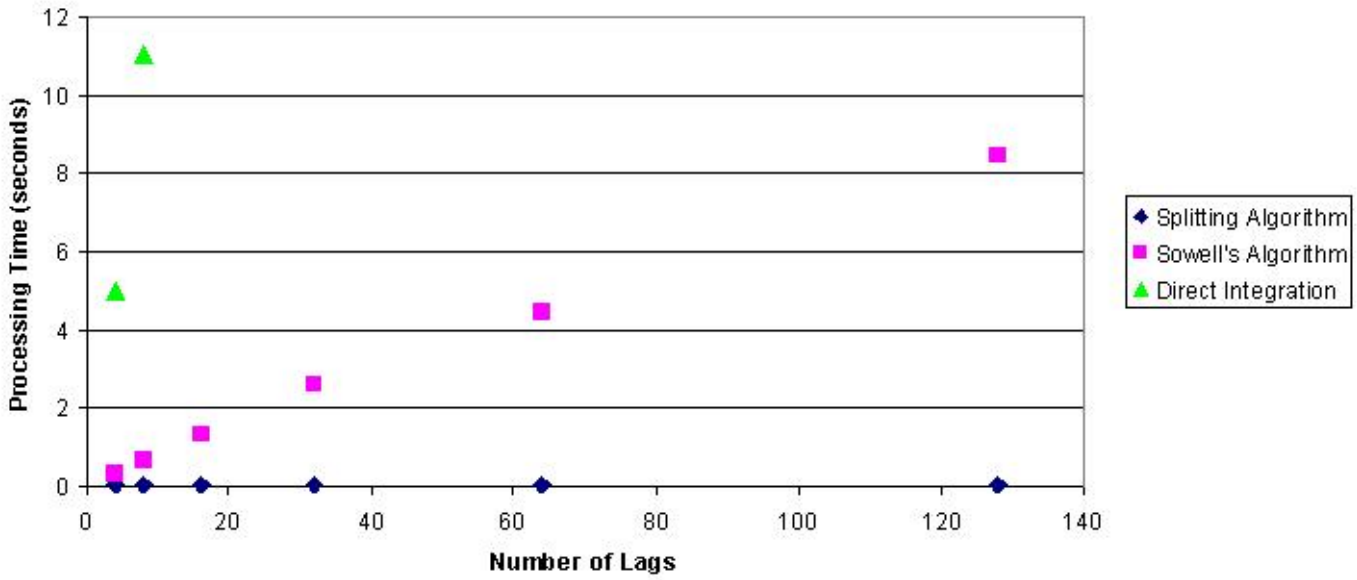
Sowell (1989b) gives a closed form for the second sum in terms of the gamma function.

We now approximate the infinite convolution by a convolution of finitely many terms.

$$\mathrm{Cov}(X_{k,t}, X_{l,t-h}) = \sum_{m=0}^{\infty} \mathrm{Cov}(Z_{k,t}, Z_{l,t-h-m}) \left(\sum_{j=m}^{\infty} \psi(j, d_l)\psi(j-m, d_k) \right)$$

- Since Z_t is a vector autoregressive process, its autocovariances decay exponentially quickly.
- That means we may choose M such that their sum after beyond lag M is less than any given error tolerance.
- We then truncate the sum by setting $\mathrm{Cov}(Z_{k,t}, Z_{l,t-h}) = 0$ for all $|h| > M$.

The splitting method dramatically reduces computing time.



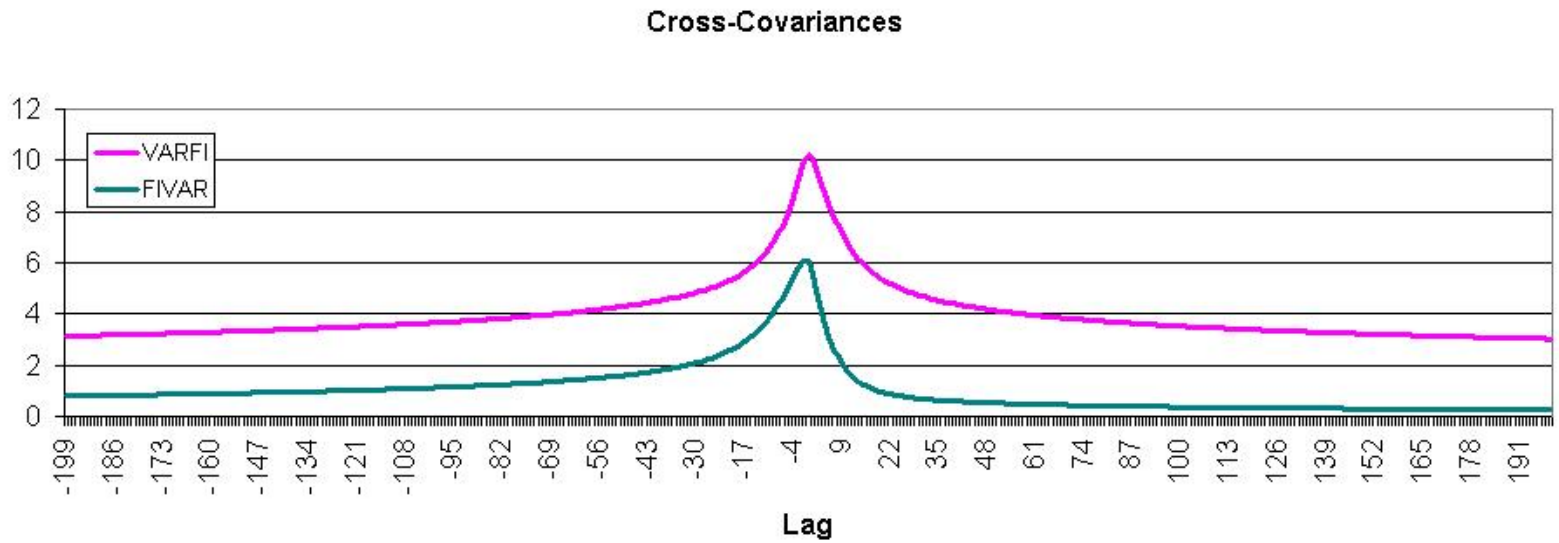
Processing time needed to compute the autocovariances of a FIVAR process with $d = (0.1, 0.4)$, $A_1 = (0.7, 0.1, 0.2, 0.6)$, and $\Sigma = (1, .5, .5, 2)$, using the Sowell method, the integral-based method and the splitting method.

- The resulting computed values match to at least 8 decimal places; more precision could be achieved by increasing M.

A different computation must be used for VARFI processes.

- Using the spectral density, we may write the autocovariances of a VARFI process as a convolution of terms that depend on the eigenvalues of the autoregressive matrix and on the fractionally integrated white noise process.
- As before, we truncate the convolution.
- Again, computations using this splitting method are much faster than computations using the integral definition directly.

Using our algorithms, we can compute the cross-covariances of FIVAR and VARFI processes.



$$d = (0.1, 0.4), \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, A_1 = \begin{pmatrix} 0.7 & 0.1 \\ 0.2 & 0.6 \end{pmatrix}$$

- Compute the autocovariance and cross-covariance sequences.
- Simulate from it.
- Estimate it with Gaussian maximum likelihood, which will require us to:
  - Compute a quadratic form.
  - Compute a determinant.

For multivariate models, there are two typical simulation algorithms.

- Find an autoregressive representation of X_t, and then simulate X_{t+1} by adding a random error on to a linear combination of past values.
  - This would require truncation for long memory processes.
- Find a matrix, A, such that $AA^* = \mathrm{Var}(X)$.
  - This matrix can be computed in $O(T^2)$ steps using the Durbin-Levinson algorithm (also presented by Sowell, 1989a).
  - We present an alternative algorithm which requires only $O(T \log T)$ steps.

Our algorithm use the structure of the covariance matrix for all of the observations.

- Let $X = (X_{1\cdot}, ..., X_{K\cdot})'$, where $X_{k\cdot} = (X_{k1}, ..., X_{kT})'$.
- We know that $\Omega = \mathrm{Var}(X)$ has a block Toeplitz structure.
- We take advantage of this structure for simulation.

This idea was suggested by Davies and Harte (1987) for univariate time series and extended by Wood and Chan (1994) to spatial time series.

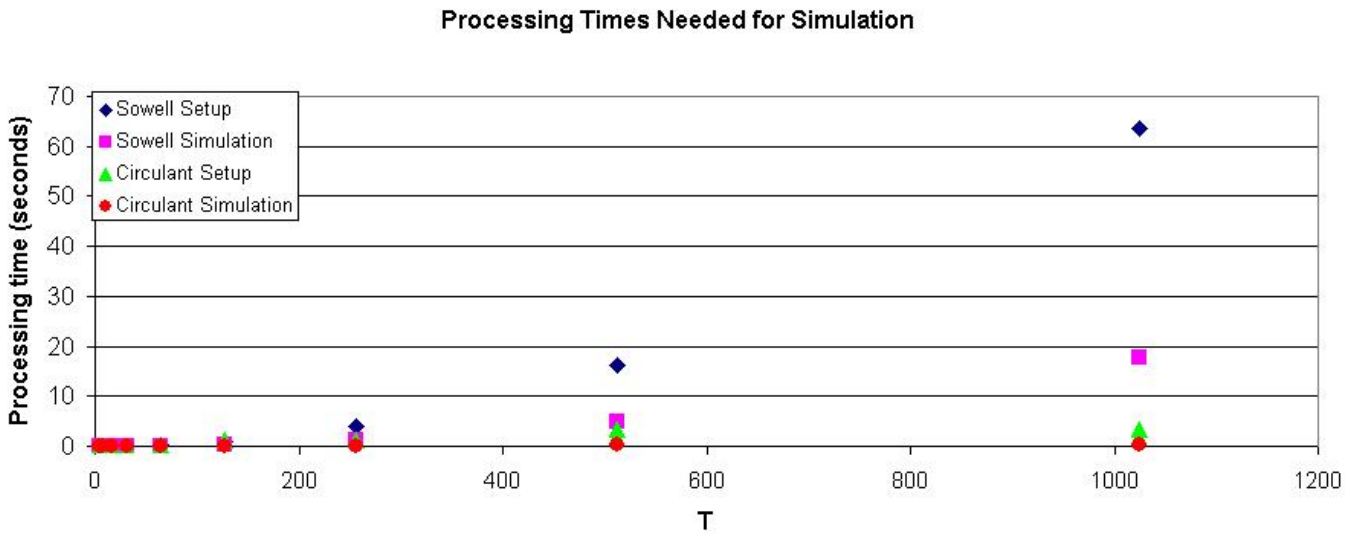
We can embed a Toeplitz matrix in a circulant matrix.

$$\begin{pmatrix} c_0 & c_1 & \cdots & c_{T-1} & c_{1-T} & \cdots & c_{-1} \\ c_{-1} & c_0 & \cdots & c_{T-2} & c_{T-1} & \cdots & c_{-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_{1-T} & c_{2-T} & \cdots & c_0 & c_1 & \cdots & c_{T-1} \\ c_{T-1} & c_{1-T} & \cdots & c_{-1} & c_0 & \cdots & c_{T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & \cdots & c_{1-T} & c_{2-T} & \cdots & c_0 \end{pmatrix}$$

Using circulant embedding on each block of the block Toeplitz matrix yields a block circulant matrix.

- If the resulting block circulant matrix is positive definite, then it is a covariance matrix, C.
- The properties of circulant matrices can be used to create efficient algorithms to compute A such that $AA^* = C$.
- We simulate using a representation for A and then extract the observations that correspond to the original process.

This new method is fast to initialize and fast for each additional simulation.



- Compute the autocovariance and cross-covariance sequences.
- Simulate from it.
- Estimate it with Gaussian maximum likelihood, which will require us to:
  - Compute a quadratic form.
  - Compute a determinant.

Likelihood computations have two parts.

$$l(\theta|X) = -\frac{1}{2}\log|\Omega(\theta)| - \frac{1}{2}X'\Omega(\theta)^{-1}X$$

- The second term can be computed using a version of the preconditioned conjugate gradient algorithm given by Chan and Olkin (1994).
- We present two approximations to the first term: one that is commonly used and one which works much better.

We start with a few facts about the determinant.

- According to the Durbin-Levinson algorithm, $|\Omega(T)| = \prod_{r=0}^{T-1} |v(r)|$, where $v(r) = \text{Var}(X_t | X_{t-1}, ..., X_{t-r})$.
- $|v(r)|$ is decreasing and bounded below by the innovation variance, $|\Sigma|$.
- It is conjectured that $|v(r)| \sim |\Sigma| \exp\left(\frac{\alpha}{r-1}\right)$, where α is a constant that depends on $\vec{d}$ and other aspects of the process.

This graph plots $|v(r)|$ as a function of r for FIVAR and VARFI processes.

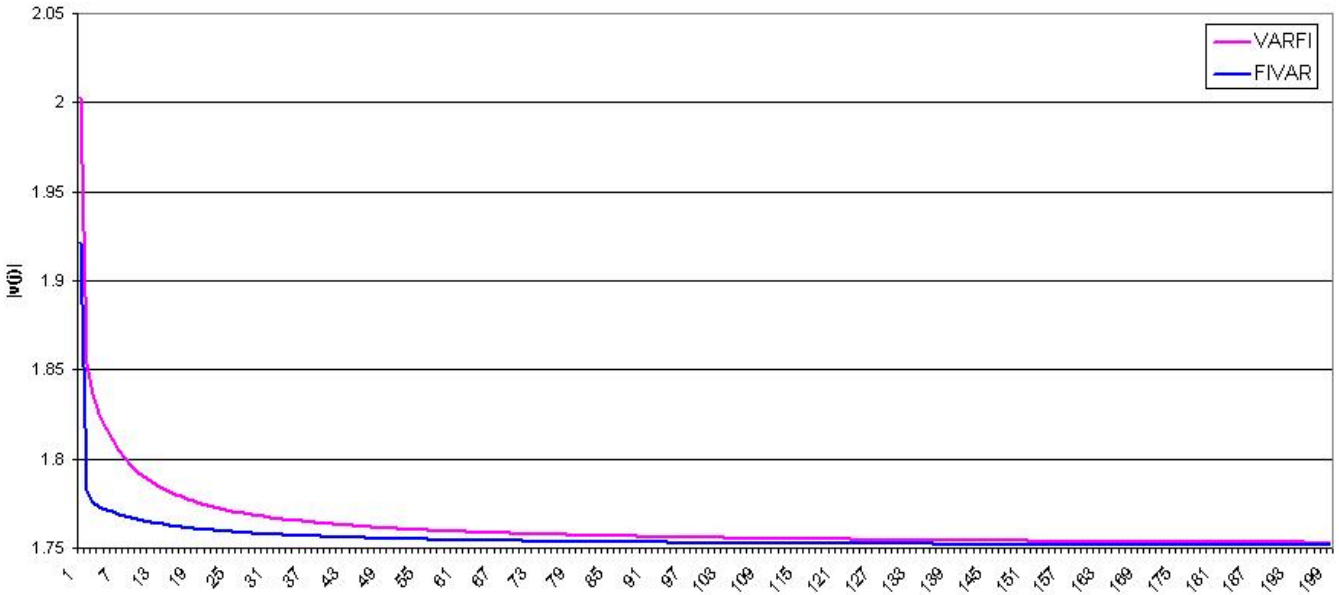


Figure: $|v(r)|$ for VARFI and FIVAR processes for r ranging from 1 to 199.

The asymptotic behavior is more clear if we exclude the first few lags.

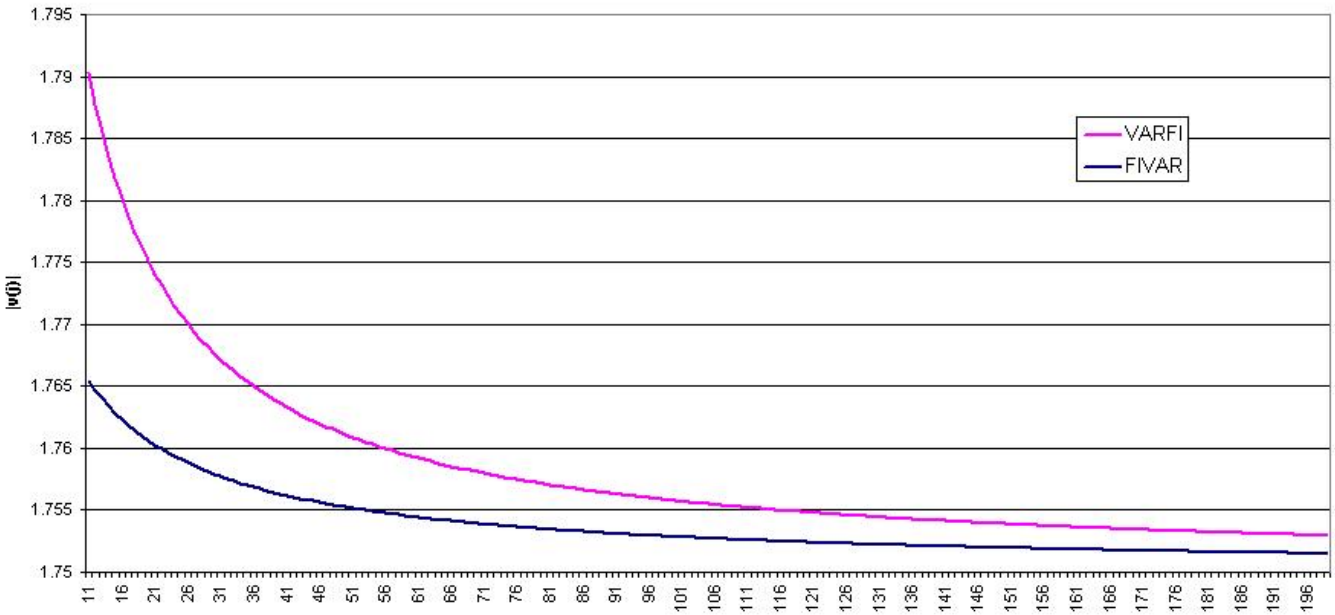


Figure: $|v(r)|$ for VARFI and FIVAR processes for r ranging from 11 to 199.

A simple, widely used approximation is $|\Omega| = |\Sigma|^T$.

- This approximation is equivalent to setting $|v(r)| = |\Sigma|$ for all r.
- As we saw on the last few slides, this approximation is quite wrong, especially for small values of r.
- In addition, how wrong the approximation is depends on the parameter values.
- As we will see in the Monte Carlo results, this approximation does not work well in estimation.

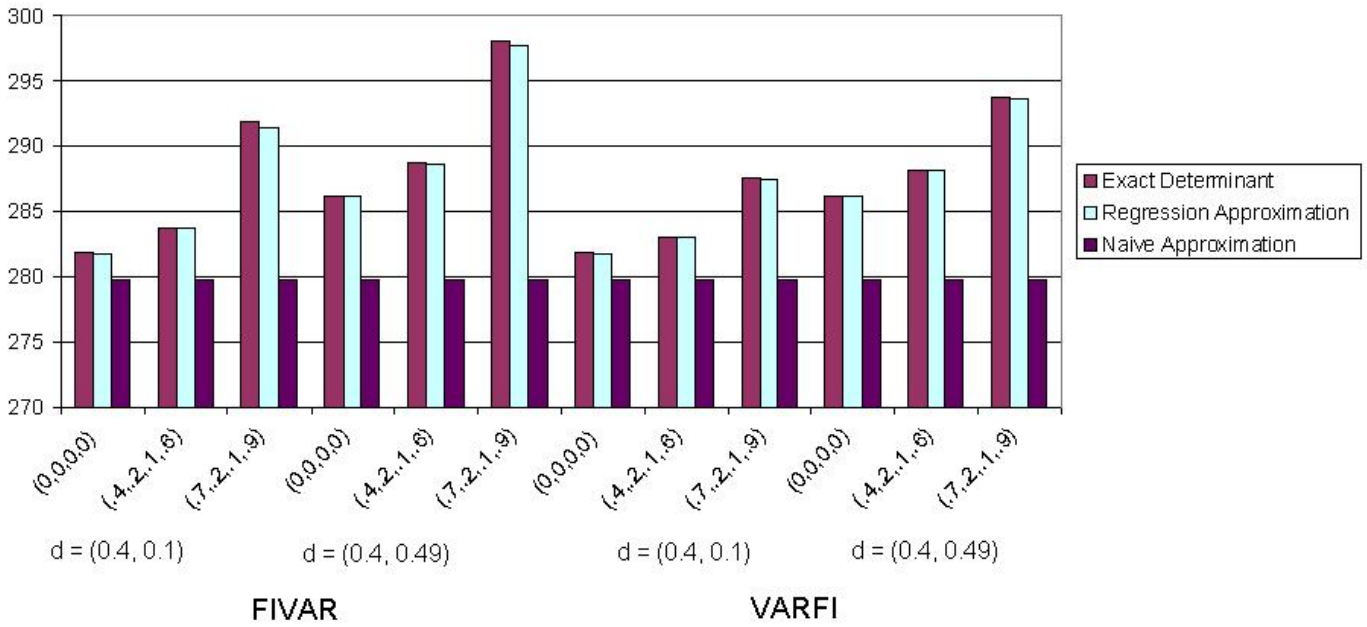
We suggest an alternative approximation.

- Use the Durbin-Levinson algorithm to compute $|v(0)|, ..., |v(S)|$ for a small S.
- Use the preconditioned conjugate gradient algorithm to compute $|v(T-1)|$.
- Fit the relationship:

$$\begin{aligned} r\sqrt{|v(r)|} &= \alpha + \beta r \\ |v(r)| &= \beta^2 + \frac{2\alpha\beta}{r} + \frac{\alpha^2}{r^2} \end{aligned}$$

- Use the fitted values to approximate $|v(r)|$ for $r = S+1, ..., T-2$.

The regression approximation is much closer for a variety of parameter configurations.



$$T = 500, \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

We now use Monte Carlo to check the performance of our estimators.

- We will compare the difference in estimates based on the two approximations for the determinant.
- Then, we will compare the performance of the Gaussian maximum likelihood estimator to the Whittle estimator.

A good approximation leads to closer estimates.

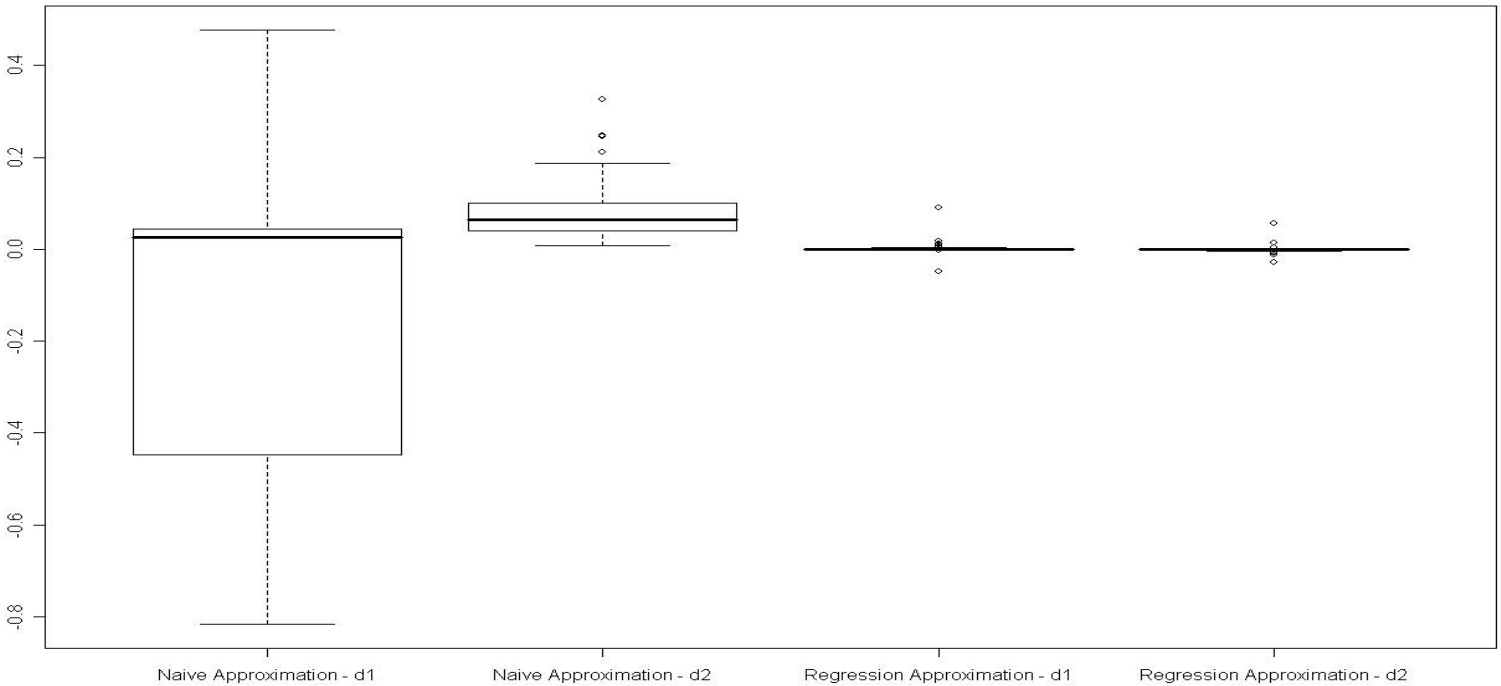


Figure: Boxplots of the differences between exact ML estimates and the two approximations for $|\Omega|$ in a FIVAR process.

The regression approximation is superior to the naive approximation.

- The difference between the estimates from the naive approximation and the estimates from exact Gaussian maximum likelihood are quite variable.
- In a significant fraction of the trials, the estimated values of Σ from the naive approximation were over 100.
- The difference between the estimates from the regression approximation and the estimates from exact maximum likelihood were generally small and had mean zero.
- These results held for both FIVAR and VARFI processes.

In a second Monte Carlo, we look at the performance of our estimator and the Whittle estimator.

The Whittle approximation to the log likelihood is:

$$-\frac{T}{2}\log|\Sigma| - \frac{1}{2}\sum_{j=1}^{T} tr\left(f^{-1}\left(\frac{2\pi j}{T}\right) I\left(\frac{2\pi j}{T}\right)\right)$$

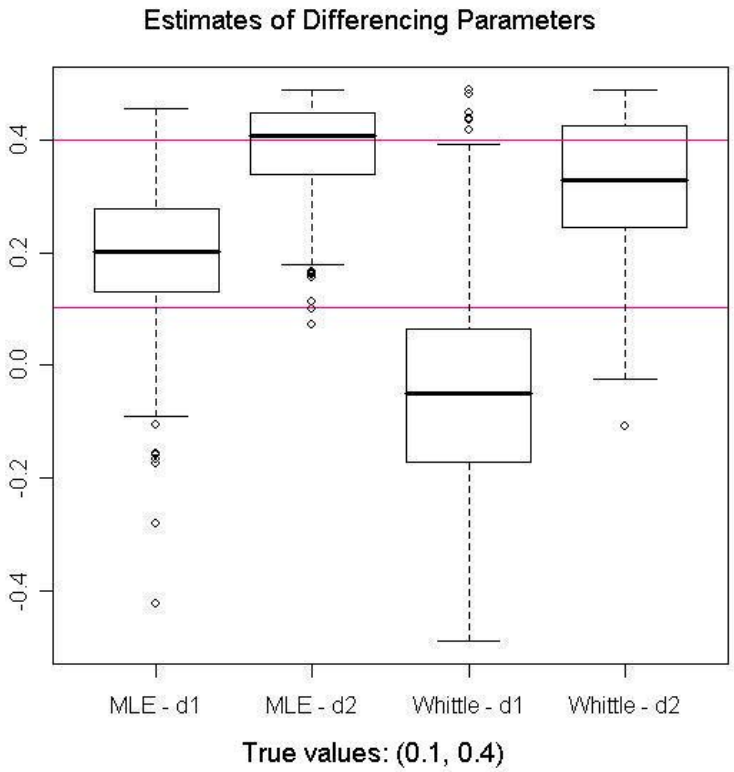
where

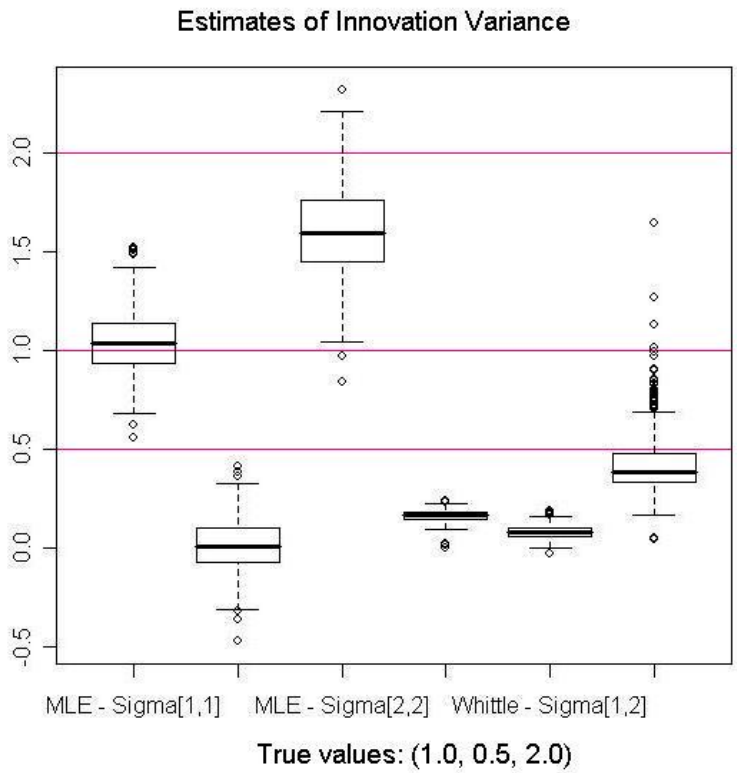
- $I(\cdot)$ is the cross periodogram:

$$I(\lambda) = \frac{1}{2\pi T}\sum_{t=1}^{T}\sum_{s=1}^{T} X_t X_s' \exp(i\lambda(t-s))$$

- f is the spectral density of the process.

We now compare the performance of the two estimators.





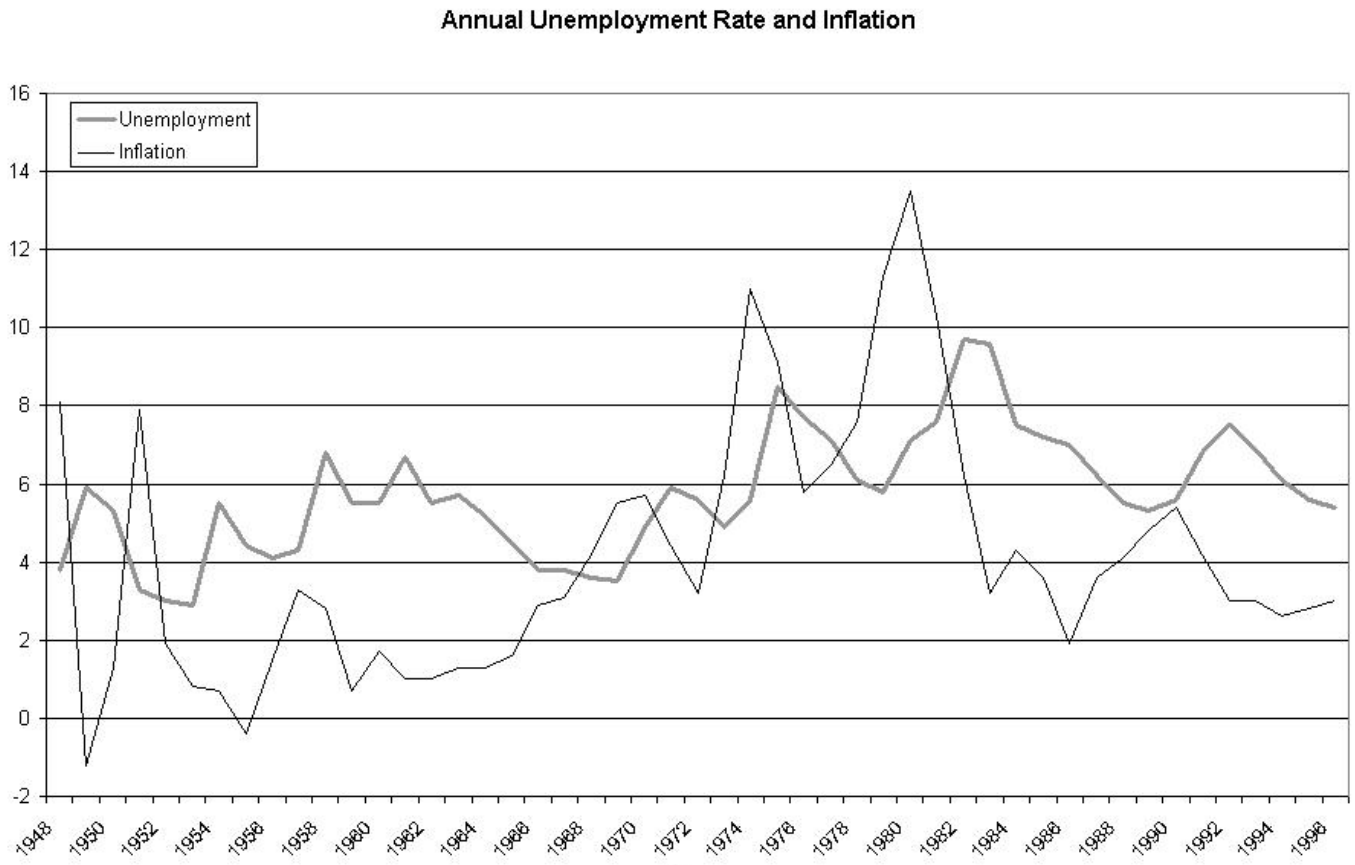
Preliminary Monte Carlo results are mixed.

- The Whittle estimates of Σ appear to be biased toward 0.
- The maximum likelihood estimates for d generally have a slightly smaller root mean squared error than the estimates from the Whittle estimator.
- The off-diagonal elements of A_1 are generally estimated better by the Whittle estimator.
- With nine parameters being estimated, there is no clear winner yet.

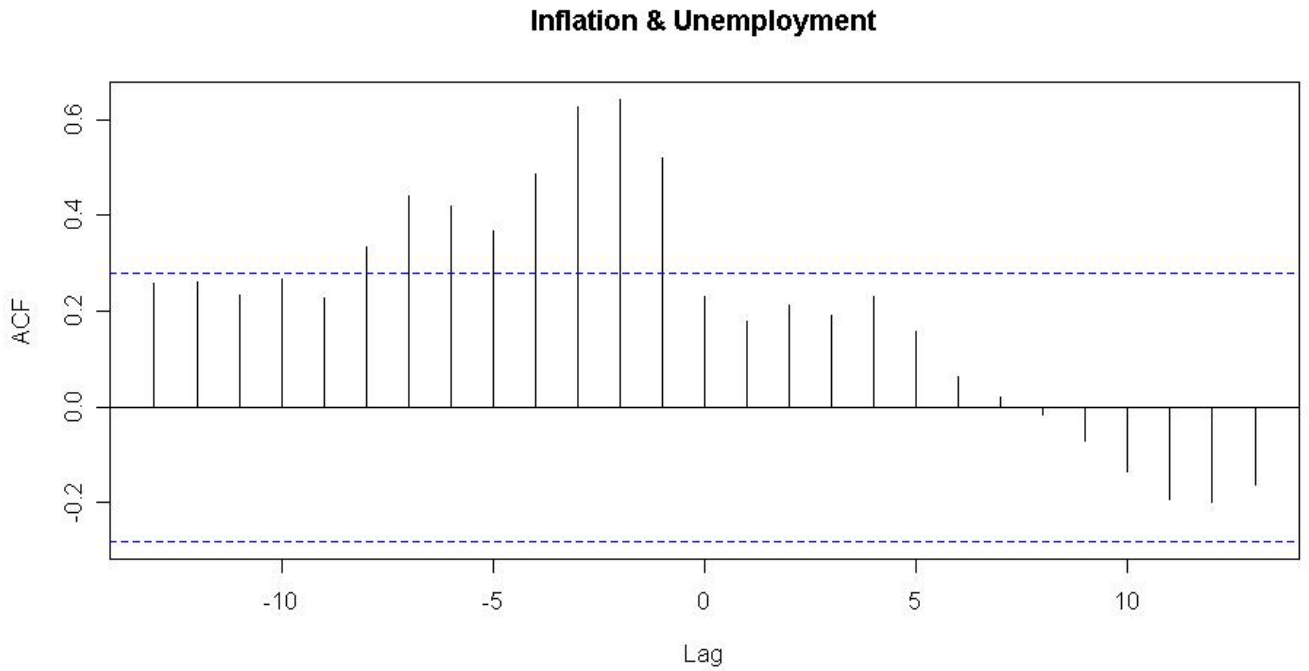
The Phillips curve describes the relationship between unemployment and inflation.

- The traditional Phillips curve implies that an increase in the unemployment rate leads to a decrease in inflation.
- Because the inflation rate is persistent, most models suggest that inflation would decrease relative to expected inflation.
- If expected inflation equaled inflation from the previous period, this would suggest a unit root in inflation.

However, both the unemployment rate and inflation seem to be persistent but mean-reverting.



Since both series seem to have long memory, a FIVAR or VARFI model might be useful.



The empirical cross-correlation function of the unemployment rate and the inflation rate, based on annual data from 1948 to 1996.

Comparing the log likelihoods at the maximum likelihood estimates, a VARFI model is preferred.

$$
\begin{aligned}
unemp(t) &= 0.223 unemp(t-1) - 0.045 infl(t-1) + u_{1t} \\
infl(t) &= -0.902 unemp(t-1) - 0.360 infl(t-1) + u_{2t} \\
\begin{pmatrix} (1-L)^{0.448} u_{1t} \\ (1-L)^{0.240} u_{2t} \end{pmatrix} &\sim Normal\left(0, \begin{pmatrix} 2.225 & -1.473 \\ -1.473 & 4.965 \end{pmatrix}\right)
\end{aligned}
$$

- As predicted by the Phillips curve, low unemployment in one period is associated with higher inflation in the next period.

We have discussed two multivariate generalizations of the univariate ARFIMA model.

- FIVAR and VARFI models can both be used to model long memory time series, but they have different implications.
- We have discussed efficient algorithms for calculating their autocovariances and simulating from the models.
- We have introduced an approximation to the determinant which is preferable to the traditional determinant approximation, $|\Sigma|^T$.
- We have fit the models to macroeconomic data.

There is much future work to be done.

- As I speak, I am running extensive simulations to further understand the properties of the Whittle and maximum likelihood estimators for these models.
- While the regression approximation to the determinant is effective, it could be refined.
- There have been many applications of univariate long memory models to data – there are also many potential applications for FIVAR and VARFI models.

References I

Bertelli, Stefano, and Massimiliano Caporin (2002), "A note on calculating autocovariances of long memory processes," *Journal of Time Series Analysis*, 23(5): 503–508.

Chan, Tony F., and Julia A. Olkin (1994), "Circulant preconditioners for Toeplitz-block matrices," *Numerical Algorithms* 6:89–101.

Davies, R.B., and D.S. Harte (1987), "Tests for the Hurst effect," *Biometrika* 74:95-101.

Lobato, Ignacio N. (1997), "Consistency of the averaged cross-periodogram in long memory time series," *Journal of Time Series Analysis* 18(2):137-155.

References II

Sowell, Fallaw (1989a), "A decomposition of block Toeplitz matrices with applications to vector time series."
Sowell, Fallaw (1989b), "Maximum likelihood estimation of fractionally integrated time series models."
Wood, Andrew T.A., and Grace Chan (1994), "Simulation of stationary gaussian processes in $[0,1]^d$," *Journal of Computational and Graphical Statistics* 3(4):409-432.