

Department of Economics

A Stochastic Variance Factor Model for Large Datasets and an Application to S&P Data

Andrea Cipollini and George Kapetanios

Working Paper No. 506

February 2004

ISSN 1473-0278



Queen Mary
University of London

A Stochastic Variance Factor Model for Large Datasets and an Application to S&P data

A. Cipollini* and G. Kapetanios†
Queen Mary, University of London.

February 10, 2004

Abstract

The aim of this paper is to consider multivariate stochastic volatility models for large dimensional datasets. We suggest use of the principal component methodology of Stock and Watson (2002) for the stochastic volatility factor model discussed by Harvey, Ruiz, and Shephard (1994). The method is simple and computationally tractable for very large datasets. We provide theoretical results on this method and apply it to S&P data.

JEL Codes: C32, C33, G12

Keywords: Stochastic Volatility, Factor Models, Principal Components

1 Introduction

The aim of this paper is to consider multivariate stochastic volatility models for large dimensional datasets. It is by now a well known fact that many financial time series exhibit time varying volatilities and cross correlations. Serial correlation in these changes is a well known feature and has been studied extensively starting with the work of Engle (1982) on ARCH models and

*Department of Economics, Queen Mary, University of London, Mile End Rd., London E1 4NS. Email: A.Cipollini@qmul.ac.uk

†Department of Economics, Queen Mary, University of London, Mile End Rd., London E1 4NS. Email: G.Kapetanios@qmul.ac.uk

their extension to GARCH models by Bollerslev (1986). Following these seminal papers, a huge numbers of alternative models have been suggested in the literature. A review of them is well beyond the scope of a single paper. The main characteristic of these models is the dependence of current volatility on past model errors and volatilities.

Within the GARCH framework, increasing attention has been devoted to model multivariate volatility models for large dimensional dataset. The models suggested by Diebold and Nerlove (1989), King, Sentana, and Wadhvani (1994) and more recently by Dungey, Martin, and Pagan (2000) to retrieve latent factor which exhibit GARCH behaviour are intractable when applied to large datasets. This is due to the large number of parameters to be estimated and the complicated constraints on the parameter space. The recent studies of Alexander (2000) and Engle (2002) allow for modelling large dimensional conditional heteroscedastic asset returns. Specifically, Alexander (op. cit.) proposes to split the analysis in two stages by, first, modelling the asset returns as a linear combination of few of their orthogonal principal components. In the second stage it is possible to obtain time varying conditional covariances of the original series modelling the conditional variances of the principal components as time varying (for instance, using a GARCH(1,1) process). In his study, Engle (op. cit) suggests to estimate consistently the parameters of a large dimensional multivariate GARCH model in two stages. First, the estimation of univariate GARCH process allows to retrieve the conditional volatilities of each return. In the second stage, the remaining two coefficients (indexing the time varying correlation matrix) are estimated. For a recent application of the Engle methodology, see Engle and Sheppard (2001) who focus on 132 asset returns.

A large class of models, alternative to the GARCH framework, considers volatilities to depend on unobserved processes which can be modelled in a variety of ways. These models are collectively known as stochastic volatility models. A seminal paper on the modelling of stochastic volatility models is Harvey, Ruiz, and Shephard (1994). This paper provides a first attempt at modelling the volatility of multivariate time series. The current paper follows directly from their work. The state space approach advocated in Harvey, Ruiz, and Shephard (1994) is powerful and intuitive but not able to deal with very large datasets due to computational constraints.

Recently, Chib, Nardari, and Shephard (2002) have suggested a Bayesian estimation methodology (which relies on Markov Chain Monte Carlo) to estimate a large dimensional multivariate stochastic volatility model (the performance of their method is based upon the simulation of 40 series and for an application of their method, see Nardari and Scruggs (2003) who focus on 17 asset returns). We argue that possible computational alternatives, such as this, based on Bayesian methods require specialist knowledge both of computational algorithms such as Markov Chain Monte Carlo and entail selecting a number of parameters such as prior distributions which may be controversial.

An alternative to state space factor models that by now is well established for modelling the first moments of large datasets is the approach advocated by Stock and Watson (2002) based on principal components. We suggest use of this methodology for the problem analysed using state space models by Harvey, Ruiz, and Shephard (1994). The method is simple and computationally tractable for very large datasets as we will also show in the empirical application which is based on S&P data.

The paper is structured as follows: Section 2 discusses the stochastic volatility model. Section 3 gives details of, and theoretical justification for, the application of principal components to the problem at hand. Section 4 discusses our empirical application. Finally, Section 5 concludes.

2 The Stochastic Volatility Factor Model

Let $y_t = (y_{1,t}, \dots, y_{N,t})'$ be an n -dimensional vector of observations, at time t , with elements given by

$$y_{i,t} = \epsilon_{i,t}(e^{h_{i,t}})^{1/2} \quad (1)$$

where $\epsilon_t = (\epsilon_{1,t}, \dots, \epsilon_{N,t})'$ is a multivariate noise vector with mean zero and covariance matrix Σ where Σ has diagonal elements equal to unity. $h_{i,t}$ is an unobserved random process whose properties we will specify in what follows. Denote $h_t = (h_{1,t}, \dots, h_{N,t})'$. Then, using the standard logarithmic transformation we have that $w_t = (\ln(y_{1,t}^2), \dots, \ln(y_{N,t}^2))'$ can be written as

$$w_t = \mu + h_t + \xi_t \quad (2)$$

where

$$\xi_t = (\xi_{1,t}, \dots, \xi_{N,t})' = (\ln(\epsilon_{1,t}^2) - E(\ln(\epsilon_{1,t}^2)), \dots, \ln(\epsilon_{N,t}^2) - E(\ln(\epsilon_{N,t}^2)))'$$

and

$$\mu = (E(\ln(\epsilon_{1,t}^2)), \dots, E(\ln(\epsilon_{N,t}^2)))'$$

This forms a general class of models for studying time varying volatilities. The properties of particular models depend on the assumptions made about h_t . Harvey, Ruiz, and Shephard (1994) have made two suggestions. The first is simply to consider

$$h_t = h_{t-1} + \eta_t \tag{3}$$

i.e. h_t is a multivariate random walk. The second suggestion is the one we consider in more detail in the current paper. They suggest that

$$h_t = Af_t \tag{4}$$

and

$$f_t = f_{t-1} + \eta_t \tag{5}$$

where f_t is a k dimensional multivariate random walk where $k < N$. This is a factor model in volatilities. The first thing to note about the model is the nature of the factor representation. This representation may appear restrictive but by appropriate redefinition of the factor vector more complicated dynamic processes may be accommodated. Harvey, Ruiz, and Shephard (1994) estimate this model using the Kalman filter and assuming normality for ϵ_t . We use this setup as a platform for discussing the application of principal components in the next section.

3 Principal Components and Factors in Stochastic Volatility Models

In a seminal paper Stock and Watson (2002) have introduced a new method of analysing large datasets and sparked a large literature on empirical and theoretical work on factor extraction, (see e.g. Forni and Reichlin (1996, 1998), Forni, Hallin, Lippi, and Reichlin (2000, 2004) and Kapetanios and Marcellino (2003)). The canonical model for factor analysis used by Stock and Watson (2002) is

$$x_t = Af_t + v_t \tag{6}$$

where $x_t = (x_{1,t}, \dots, x_{N,t})'$ is a multivariate series, f_t is a $k \times 1$ factor vector and $v_t = (v_{1,t}, \dots, v_{N,t})'$ is a vector of idiosyncratic errors. Clearly, this model is suitable for analysing stochastic volatility data. Principal components minimise

$$V(r) = \min_{A,f} (NT)^{-1} \sum_{i=1}^N \sum_{t=1}^T (x_{i,t} - a_i' f_t)^2$$

where $f = (f_1, \dots, f_T)'$, to obtain estimates of the factors and factor loadings. Note that extra restrictions are needed to identify the factors and factor loadings separately since, for any nonsingular matrix, H , $a_i' f_t = a_i' H H^{-1} f_t = a_i^{*'} f_t^*$. Defining $X = (x_1, \dots, x_T)'$, it can be shown that, under the implicit restriction $f' f / T = I$, the estimate of f , denoted \hat{f} is given by the eigenvectors corresponding to the largest k eigenvalues of the matrix XX' . Bai (2003) has shown that application of principal components leads to consistent estimation of f in the sense that for each t

$$\|\hat{f}_t - H f_t\| = O_p(\min\{\sqrt{N}, \sqrt{T}\}) \quad (7)$$

where $\|\cdot\|$ denotes matrix norm and H denotes a nonsingular matrix. Note that since f_t cannot be identified without further restrictions, consistency holds only up to all $k \times k$ nonsingular transformations.

For this result to hold a number of mild conditions on f_t and v_t need to be applied: In particular the following set of assumptions are sufficient for consistency to hold.

Assumption 1 $E\|f_t\|^4 \leq M < \infty$, $T^{-1} \sum_{t=1}^T f_t f_t' \xrightarrow{p} \Sigma$ for some $k \times k$ positive definite matrix Σ .

Assumption 2 $\|a_i\| \leq \bar{a} < \infty$

Assumption 3 $E(v_{i,t}) = 0$, $E|v_{i,t}|^8 \leq M$

Assumption 4 For $\tau_{i,j,t,s} \equiv E(v_{i,t} v_{j,s})$ the following hold

- $(NT)^{-1} \sum_{s=1}^T \sum_{t=1}^T |\sum_{i=1}^N \tau_{i,i,t,s}| \leq M$
- $|\sum_{i=1}^N \tau_{i,i,s,s}| \leq M$ for all s
- $N^{-1} \sum_{i=1}^N \sum_{j=1}^N |\tau_{i,j,s,s}| \leq M$

- $(NT)^{-1} \sum_{s=1}^T \sum_{t=1}^T \sum_{i=1}^N \sum_{j=1}^N |\tau_{i,j,t,s}| \leq M$
- For every (t, s) , $E|(N)^{-1/2} \sum_{i=1}^N (v_{i,s}v_{i,t} - \tau_{i,i,s,t})|^4 \leq M$

Assumption 5 f_t and $v_{i,t}$ are independent

Similar results are shown to hold in Bai (2004) if the factor is a random walk process. In fact, as we see from (7), these assumptions are sufficient not only for consistency but also for determining rates of convergence for \hat{f}_t . Anticipating the analysis of the empirical section we provide a small extension of the results of Bai (2003) in the following theorem

Theorem 1 *The results of Bai (2003) on consistency of factor estimates using principal components are valid if the factor processes are stationary long memory ARFIMA(p, d, q) processes driven by shocks with finite fourth moment.*

Long memory processes form a large class of time series processes characterised by the slow decline of the autocorrelation function which asymptotically declines hyperbolically rather than exponentially as in the more usual case of short memory processes such as ARMA processes. A large class of the processes can be characterised by the Fractional ARIMA (ARFIMA) model

$$(1 - L)^d f_t = u_t \tag{8}$$

where u_t is a finite ARMA process, d is a real number and $(1 - L)^d$ is defined in terms of its binomial expansion as

$$(1 - L)^d = \sum_{i=0}^{\infty} \frac{\Gamma(i - d)}{\Gamma(i + 1)\Gamma(-d)} L^i = \sum_{i=0}^{\infty} b_i L^i$$

For $0 < d < 0.5$, f_t is stationary with $\sum_{i=0}^{\infty} b_i^2 < \infty$. As long as the fourth moment of u_t exists it then follows that the fourth moment of f_t exists. We assume that the process driving the ARMA process, u_t , has a finite fourth moment and hence u_t has a finite fourth moment. Further, Hosking (1982) has shown that a law of large numbers holds for $1/T \sum_{t=1}^T f_t^2$ which converges to its expected value. These results combined provide Assumption 1 of the sufficient list of assumptions given above for consistency of factor estimates and hence factor estimation is consistent when the factors are stationary long

memory processes.

Our study is in line with the findings of long memory in volatility in the financial instruments returns. Specifically, using a parallel with ARMA and ARFIMA processes Baillie, Bollerslev, and Mikkelsen (1996) suggested a Fractionally Integrated GARCH process (FIGARCH) for the asset return conditional variance, allowing the integration coefficient to vary in the range $[0,1]$.

From the above discussion it is clear that our model is a generalisation of the model suggested by Harvey, Ruiz, and Shephard (1994), given by (2), (4) and (5) since we entertain a much more general process for f_t than (5). We simply require that f_t satisfies Assumption 1 and $\xi_{i,t}$ satisfies Assumptions 3-4. Once f_t has been estimated one can model it using a variety of models. Following Bai (2003), the error that arises, in modelling, from the fact that f_t is estimated rather than known is negligible if $\sqrt{T}/N \rightarrow 0$. Such models can then provide forecasts for f_t and hence for $w_{i,t}$.

However, the factor model is still not general enough to capture important aspects of the data as reported in various empirical studies. For example, time varying correlations are not captured by this model. Nevertheless, a number of extensions can be envisaged to enable modelling of such features. We suggest the following extension to (5)

$$h_{i,t} = a'_i f_t + u_{i,t} \tag{9}$$

Then

$$w_{i,t} = \mu_i + a'_i f_t + u_{i,t} + \xi_{i,t} = \mu_i + a'_i f_t + \zeta_{i,t} \tag{10}$$

where $\zeta_{i,t} = u_{i,t} + \xi_{i,t}$. As long as $\zeta_{i,t}$ satisfies Assumptions 3-4, the factor can be still estimated consistently. Then, one can model the estimate of $\zeta_{i,t}$ using univariate stochastic volatility representations. This two step approach is very flexible and can capture a wide variety of volatility features. For example, the proportion of $w_{i,t}$ explained by $a'_i f_t$ and $u_{i,t}$ respectively, conditioning on the past, can vary over time giving rise to time varying correlations.

4 Empirical Application

We apply our suggested method of analysing stochastic volatility data to two large datasets. These are the S&P500 and S&P100 datasets. Data, obtained from Datastream are daily returns and span the period 01/01/1995-13/01/2004 comprising 2356 observations. We choose to consider only companies for which data are available throughout the period leading us to have $N = 412$ for the S&P500 dataset and $N = 93$ for the S&P100 dataset. Once all periods when markets were closed are dropped from the dataset the number of observations is 2275.

We first demean daily returns, denoted y_t , to get $\tilde{y}_t = y_t - 1/T \sum_{t=1}^T y_t$. Then, we transform the data to get $w_{i,t} = \ln(\tilde{y}_{i,t}^2)$. Finally, we demean the transformed data to get $\tilde{w}_t = w_t - 1/T \sum_{t=1}^T w_t$. We apply principal components to \tilde{w}_t .

The exponent of the first factor for both datasets is plotted in Figure 1. The two factors are highly correlated with correlation equal to 0.956. We then calculate the average R^2 across all \tilde{w}_t and report it, cumulatively, for the first 10 factors in Table 1. It is clear that whereas the first factor explains 10% and 11% of the variation in the two datasets respectively, further factors can add only marginally to the explanatory power of the set of factors. We therefore conclude that one factor captures a large common component of the stochastic volatility of the two datasets. It is worth noting that although 10% may appear to be a small part of the variation, it is one series that can explain on average 10% for 412 series which is a considerable achievement. We concentrate our analysis on the first factor of the two datasets. We leave to future work a full analysis of the volatilities of the dataset with common and idiosyncratic components as suggested in the previous section. We simply wish to provide an illustrative example of the new method.

Next, we examine the dynamic properties of the factors. Firstly we apply an *AR* model to them using the Akaike information criterion with maximum lag order equal to 24. For both factors a lag order of 11 is chosen. Coefficient estimates, t-statistics and the maximum absolute eigenvalue of the *AR* polynomial are reported in Table 2. Clearly, both factors are highly persistent and only admit an *AR* representation with long lags. We carry out the Lee, White, and Granger (1993) nonlinearity test on our chosen *AR* specification.

Probability values of 0.12 and 0.37 for the two factors respectively indicates little evidence for neglected nonlinearity in the specification.

Further insight is obtained by plotting the autocorrelation functions of the factors. This is done in Figure 2. We also plot the upper 95% bound of the confidence interval of the null hypothesis that the process is white noise. Clearly, the autocorrelation function declines very slowly. Even after 400-500 periods the autocorrelations are positive and significant. This points towards long memory models whose autocorrelation function declines hyperbolically.

We therefore fit $ARFI(p, d)$ to the factors. As we only allow for an AR component in the model we choose a large maximum lag order of 24 lags. We choose the lag order by minimising the Akaike information criterion. As Theorem 1 states, factor estimation is still valid in the presence of stationary long memory. Table 3 reports the results. Clearly there is evidence of long memory as the estimated long memory parameter is significantly different from zero. Nevertheless, both factors appear to be stationary since \hat{d} in both cases is around 0.45.

Further insight into the factor may be obtained by looking at the impulse responses of the fitted long memory model. The impulse responses up to horizon 30 are plotted in Figure 3. Clearly they decline much more slowly than the exponential rate of short memory ARMA processes. Further analysis of the persistence of the process can be had by looking at its half life. This is defined as the time it takes for half the effect of the shock to disappear. There is a large literature on measures of half life in macroeconomics. Usually these measures are obtained from fitted AR(1) models. Clearly this is not appropriate here and would lead to incorrect conclusions. Recently, Chortareas and Kapetanios (2004) have provided a new half life measure which defines half life as the point at which half the cumulative impulse response effect has been experienced. This measure juxtaposes itself from the usual measure which is defined as the point in time at which the instantaneous response to the shock is half compared to the instantaneous response when the shock impacts on the process. More details may be found in Chortareas and Kapetanios (2004). Using this new measure we get that, for the S&P 500 dataset, the half life of a shock is 23 days compared to 22 days for the S&P 100 dataset. Interestingly, according to the standard definition of half life both series have a rather short half life of one day.

5 Conclusion

In the past twenty years there has been an explosion of research interest in univariate and multivariate volatility modelling. Out of many approaches a factor based modelling approach seems to have a large numbers of desirable characteristics chief among which is parsimony. However, existing methods are mostly based on the state space representation of factor models and the need to use iterative techniques for estimation may be problematic in large datasets.

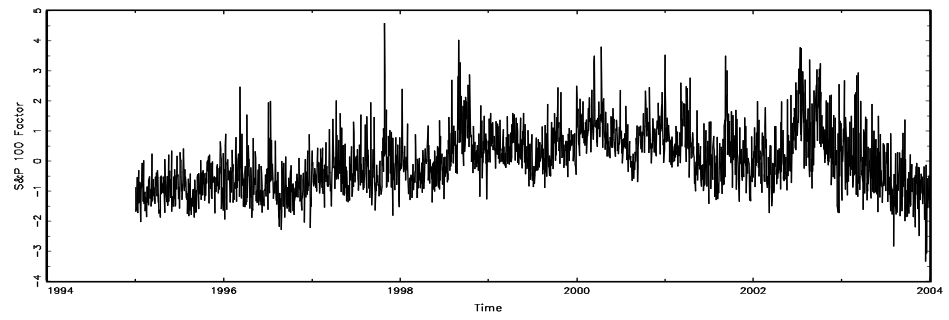
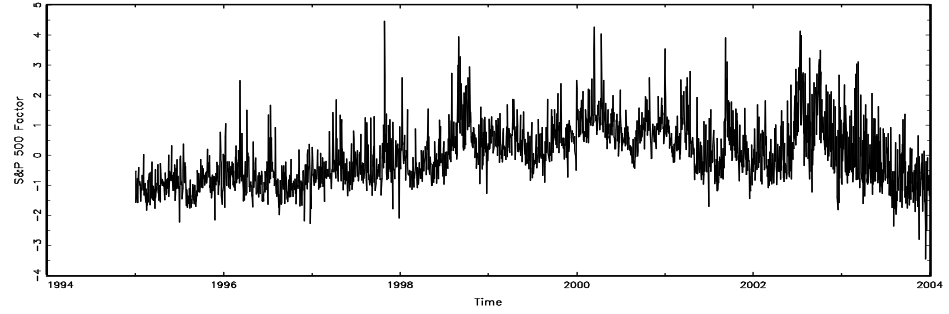
This paper has suggested the use of principal components as advocated by Stock and Watson (2002) to complement stochastic volatility modelling of multivariate time series. The theoretical properties of the new approach have been discussed. The method has been extended to highly persistent stationary data which exhibit long memory behaviour. The method has been applied to the S&P 100 and S&P 500 datasets with interesting results. More specifically, we have found that a single factor can explain a large proportion of volatility in the datasets. This factor is highly persistent and exhibits long memory. A fruitful avenue for future research appears to be the combination of our multivariate factor methodology with univariate stochastic volatility representations for individual idiosyncratic components.

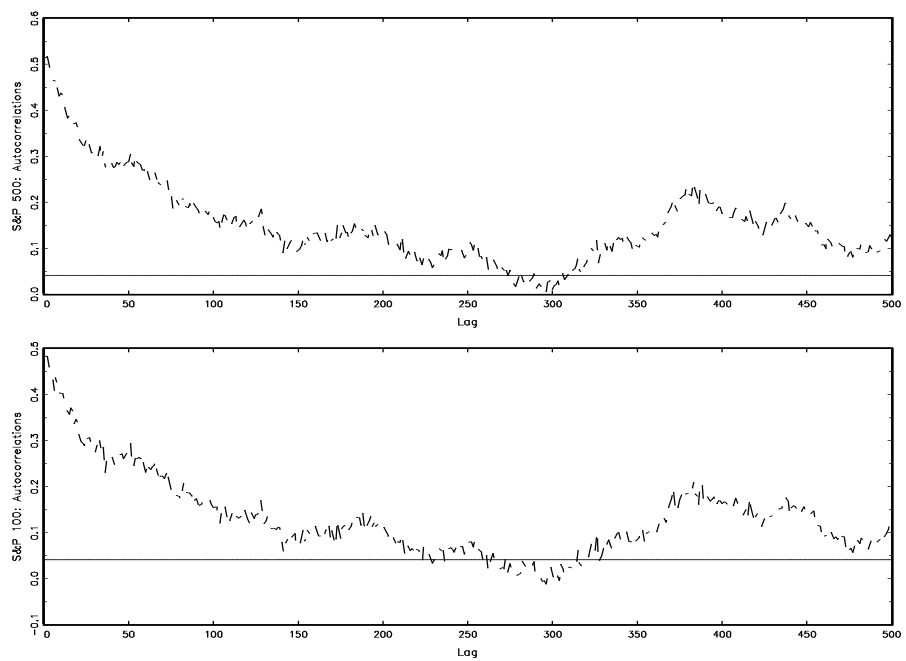
References

- ALEXANDER, C. (2000): “A primer on the orthogonal GARCH model,” *University of Reading discussion paper*.
- BAI, J. (2003): “Inferential Theory for Factor Models of Large Dimensions,” *Econometrica*, 71, 135–173.
- (2004): “Estimating Cross-Section Common Stochastic Trends in Nonstationary Panel Data,” *Journal of Econometrics*, Forthcoming.
- BAILLIE, R., T. BOLLERSLEV, AND H. MIKKELSEN (1996): “Fractionally Integrated Generalised Autoregressive Conditional Heteroscedasticity,” *Journal of Econometrics*, 74, 3–30.
- BOLLERSLEV, T. (1986): “Generalised Autoregressive Conditional Heteroscedasticity,” *Journal of Econometrics*, 51, 307–327.

- CHIB, S., F. NARDARI, AND N. SHEPHARD (2002): “Analysis of high dimensional multivariate stochastic volatility models,” *Washington University in St. Louis working paper*.
- CHORTAREAS, G., AND G. KAPETANIOS (2004): “How Puzzling is the PPP Puzzle,” *Queen Mary, University of London Mimeo*.
- DIEBOLD, F., AND M. H. NERLOVE (1989): “The dynamics of exchange rate volatility: a multivariate latent factor ARCH model,” *Journal of Applied Econometrics*, 4, 1–21.
- DUNGEY, M., V. MARTIN, AND A. PAGAN (2000): “A Multivariate Latent Factor Decomposition of International Bond Yield Spreads,” *Journal of Applied Econometrics*, 16, 697–715.
- ENGLE, R. (2002): “Dynamic conditional correlation: a simple class of multivariate GARCH models,” *Journal of Business and Economics Statistics*, 20, 339–350.
- ENGLE, R., AND K. SHEPPARD (2001): “Theoretical and empirical properties of Dynamic conditional correlation multivariate GARCH,” *UCSD discussion paper 2001-15*.
- ENGLE, R. F. (1982): “Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation,” *Econometrica*, 50(1), 987–1007.
- FORNI, M., M. HALLIN, M. LIPPI, AND L. REICHLIN (2000): “The Generalised Factor Model: Identification and Estimation,” *Review of Economics and Statistics*, 82, 540–554.
- (2004): “The Generalised Factor Model: Consistency and Rates,” *Journal of Econometrics*, Forthcoming.
- FORNI, M., AND L. REICHLIN (1996): “Dynamic Common Factors in Large Cross Sections,” *Empirical Economics*, 21, 27–42.
- (1998): “Let’s Get Real: A Dynamic Factor Analytical Approach to the Disaggregated Business Cycle,” *Review of Economic Studies*, 65, 453–474.

- HARVEY, A. C., E. RUIZ, AND N. SHEPHARD (1994): “Multivariate Stochastic Variance Models,” *Review of Economic Studies*, 61, 247–264.
- HOSKING, J. R. M. (1982): “Modelling Persistence in Hydrological Time Series Using Fractional Differencing,” *Water Resources Research*, 20, 1898–1908.
- KAPETANIOS, G., AND M. MARCELLINO (2003): “A Comparison of Estimation Methods for Dynamic Factor Models of Large Dimensions,” *Queen Mary, University of London Working Paper No. 489*.
- KING, M., E. SENTANA, AND S. WADHWANI (1994): “Volatility and links between national stock markets,” *Econometrica*, 62, 901–933.
- LEE, T. H., H. WHITE, AND C. W. J. GRANGER (1993): “Testing for Neglected Nonlinearity in Time Series Models: A Comparison of Neural Network Methods and Alternative Tests,” *Journal of Econometrics*, 56, 269–290.
- NARDARI, F., AND J. SCRUGGS (2003): “Analysis of linear factor Models with multivariate stochastic volatility models for stocks and bond returns,” *W.P. Carey School of Business working paper*.
- STOCK, J. H., AND M. W. WATSON (2002): “Macroeconomic Forecasting Using difucions Indices,” *Journal of Business and Economic Statistics*, 20, 147–162.





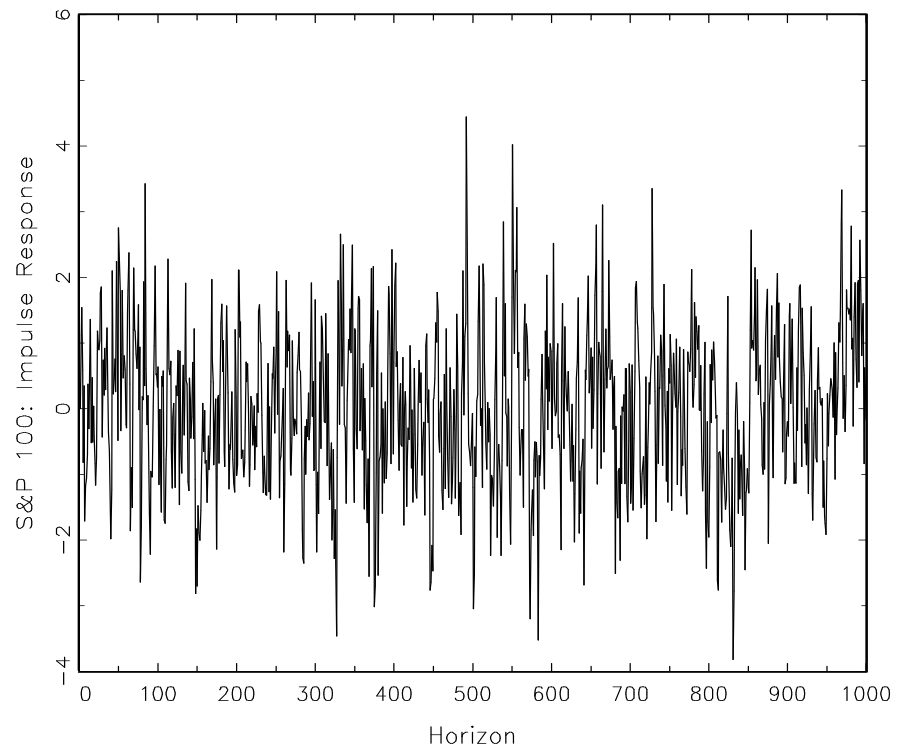


Table 1: Cumulative Explained Variation		
No. of Factors	S&P500	S&P100
1	0.096	0.112
2	0.109	0.132
3	0.122	0.150
4	0.131	0.166
5	0.139	0.181
6	0.143	0.195
7	0.147	0.208
8	0.151	0.224
9	0.156	0.237
10	0.161	0.250
11	0.166	0.264
12	0.169	0.277
13	0.173	0.290
14	0.177	0.302
15	0.181	0.314
16	0.184	0.327
17	0.188	0.338
18	0.192	0.350
19	0.195	0.362
20	0.199	0.374

	S&P500		S&P100	
	Coeff	<i>t</i> -stat	Coeff	<i>t</i> -stat
AR(1)	0.181	8.581	0.172	8.174
AR(2)	0.160	7.497	0.151	7.084
AR(3)	0.107	4.966	0.091	4.198
AR(4)	0.094	4.341	0.084	3.868
AR(5)	0.085	3.932	0.080	3.688
AR(6)	0.036	1.651	0.027	1.249
AR(7)	0.077	3.555	0.089	4.128
AR(8)	0.041	1.881	0.040	1.851
AR(9)	0.023	1.064	0.031	1.415
AR(10)	0.050	2.332	0.047	2.199
AR(11)	0.060	2.866	0.054	2.500
$\max \lambda_i $	0.98	0	0.98	0

	S&P500	S&P100
p	1	1
\hat{d}	0.416	0.398
$std(\hat{d})$	0.0197	0.0198
95% $CI(\hat{d})$	0.456	0.437
5% $CI(\hat{d})$	0.376	0.358

**This working paper has been produced by
the Department of Economics at
Queen Mary, University of London**

**Copyright © 2004 Andrea Cipollini and George Kapetanios
All rights reserved.**

**Department of Economics
Queen Mary, University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 5096 or Fax: +44 (0)20 8983 3580
Email: j.conner@qmul.ac.uk
Website: www.econ.qmul.ac.uk/papers/wp.htm**