SCHOOL OF ACCOUNTING, FINANCE AND MANAGEMENT

Essex Finance Centre

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

A. Cipollini
University of Essex

G. Kapetanios
Queen Mary, University of London

Discussion Paper No. DP 07/05

November, 2007

¥

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

A. Cipollini*
University of Essex

G. Kapetanios†
Queen Mary, University of London.

October 2, 2007

Abstract

The aim of this paper is to consider multivariate stochastic volatility models for large dimensional datasets. We suggest the use of the principal component methodology of Stock and Watson (2002) for the stochastic volatility factor model discussed by Harvey, Ruiz, and Shephard (1994). We provide theoretical and Monte Carlo 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. For this purpose we use a common factor approach along the lines of Harvey, Ruiz, and Shephard (1994). More recently, Bayesian estimation methods, relying on Markov Chain Monte Carlo, have been put forward by Chib, Nardari, and Shephard (2006) to estimate relatively large multivariate stochastic volatility models. However, computational constraints can be binding when dealing with very large datasets such as, e.g, S&P 500 constituents. For instance, the Bayesian modelling approach put forward by Chib, Nardari, and Shephard (2006) is illustrated by modelling a dataset of only 20 series of stock returns. Recently, Stock and Watson (2002) have shown that principal component estimates of the common factor underlying large datasets can be used successfully in forecasting conditional means. We propose the use of principal component estimation for the volatility processes of large datasets. A Monte Carlo study and an application to the modelling of the volatilities of the S&P constituents illustrate the usefulness of our approach.

*Department of Accounting, Finance and Management, University of Essex, Wivenhoe Park, C04 3SQ, London. Email acipol@essex.ac.uk

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

2 The Stochastic Volatility Factor Model

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

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

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

$$w_t = \mu + h_t + \xi_t, \tag{2}$$

where $\xi_t = (\xi_{1,t}, \ldots, \xi_{N,t})' = (ln(\epsilon_{1,t}^2) - E(ln(\epsilon_{1,t}^2)), \ldots, ln(\epsilon_{N,t}^2) - E(ln(\epsilon_{N,t}^2))'$ and $\mu = (E(ln(\epsilon_{1,t}^2)), \ldots, E(ln(\epsilon_{N,t}^2)))'$. 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. In line with Harvey, Ruiz, and Shephard (1994) we model h_t through common factors. We have

$$h_t = Af_t, \tag{3}$$

where f_t is a $k \times 1$ vector of factor processes. Given the computational constraints in estimating state space model representations of the common factors (underlying the large dimensional dataset of stochastic volatilities) via either Maximum Likelihood as in Harvey, Ruiz, and Shephard (1994), or via Bayesian estimation methods put forward by Chib, Nardari, and Shephard (2006), we estimate the common factor through applying principal components to w_t. Following Bai (2003), the error that arises, in modelling, from the fact that f_t is estimated rather than known is negligible if $\sqrt{T}/N \to 0$. In order to forecast w_t we need to introduce dynamics in f_t. For the particular application of the factor model to S&P 500 constituents, considered in Section 4, the slow and hyperbolic decline in the autocorrelation function of the factors suggests the presence of long memory. So, we fit an ARFIMA model of the form $(1-L)^d f_t = u_t$, where u_t is a finite order ARMA process, i.e. $A(L)u_t = B(L)\eta_t$, where $A(L)$ and $B(L)$ are lag polynomials and $A(L)$ has its roots outside the unit circle; d is a real number and $(1-L)^{-d}$ is defined in terms of its binomial expansion as $(1-L)^{-d} = \sum_{i=0}^{\infty} \frac{\Gamma(1-d)}{\Gamma(i+1)\Gamma(-d-i+1)}(-1)^i L^i = \sum_{i=0}^{\infty} b_i L^i$. For $0 < d < 0.5$, f_t is stationary with $\sum_{i=0}^{\infty} b_i^2 < \infty$. In order to obtain consistent estimates of the factors through principal components, the regularity conditions of Bai (2003) must hold. In particular, the existence of a finite fourth moment for f_t is needed. It is straightforward to show (and this is shown in Cipollini and Kapetanios (2004)) that finiteness of the fourth moment of η_t is sufficient for these regularity conditions to hold for a long memory and stationary f_t. The common factor modelling per se is not general enough to capture important aspects of the data as

reported in various empirical studies. So we suggest the following extension to (3):

$$h_{i,t} = a_i' f_t + \psi_{i,t} \tag{4}$$

where $\psi_t = (\psi_{1,t}, \ldots, \psi_{N,t})'$ is a vector of idiosyncratic errors. Then

$$w_{i,t} = \mu_i + a_i' f_t + \psi_{i,t} + \xi_{i,t} = \mu_i + a_i' f_t + \zeta_{i,t} \tag{5}$$

where $\zeta_{i,t} = \psi_{i,t} + \xi_{i,t}$. As long as $\psi_{i,t}$ satisfies the regularity condition of Bai (2003), f_t can be consistently estimated and $\zeta_{i,t}$ can be modelled, as a residual, by fitting individual state space stochastic volatility models. Specifically, the estimated model for each $\zeta_{i,t}$, is of the form

$$\zeta_{i,t} = \delta_i \varphi_{i,t} + \xi_{i,t} \tag{6}$$

$$\varphi_{i,t} = \rho_i \varphi_{i,t-1} + \chi_{i,t} \tag{7}$$

(6) is estimated by maximum likelihood. For this estimation we can set $E(\chi_{i,t}^2) = 1$, $E(\varphi_{i,0}) = 0$ and $E(\varphi_{i,0}^2) = \frac{1}{1-\rho_i^2}$. As shown in Harvey, Ruiz, and Shephard (1994) Gaussian ML estimation is consistent. 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 $\psi_{i,t}$ respectively, conditioning on the past, can vary over time giving rise to time varying covariances. To see this note that

$$E(y_{i,t} y_{j,t} | t-1) = \sigma_{ij} E\left(e^{0.5(h_{i,t}+h_{j,t})} \middle| t-1\right) = \sigma_{ij} E\left(e^{0.5(a_i' f_t + a_j' f_t + \psi_{i,t} + \psi_{j,t})} \middle| t-1\right) \tag{8}$$

3 Monte Carlo Analysis

The model we consider is given by

$$y_{i,t} = \epsilon_{i,t} (e^{h_{i,t}})^{1/2} \tag{9}$$

$$h_t = A f_t \tag{10}$$

We consider two alternative data generation processes for the factor. The first is an $AR(1)$ model given by $f_t = \rho f_{t-1} + \eta_t$ where we set $k = 1$. The second is an $ARFIMA(1, d, 0)$ given by $(1 - \rho L)(1 - L)^d f_t = \eta_t$. Throughout $\epsilon_{i,t}, \eta_t \sim i.i.d.N(0, 1)$. We consider $N = 50, 100, 200$ and $T = 200, 500, 1000, 2000$. For the $AR(1)$ factor model, $\rho = 0.1, 0.5, 0.9$. For the $ARFIMA(1, d, 0)$ model, $\rho = 0.5$ and $d = 0.2, 0.4$. Estimation of the ARFIMA model is carried out by minimising the conditional sum of squares as discussed in Baillie, Chung, and Tieslau (1996). For every experiment we carry out 1000 replications. We report two performance indicators: (i) the average absolute correlation of the true and estimated factor over replications and (ii) the proportion of the series variance explained by the estimated

factor compared to the proportion of the series variance explained by the true factor averaged over both series and replications. Results for the $AR(1)$ factor model are reported in Table 1A and for the $ARFIMA(1, d, 0)$ model in Table 1B.

The estimation method works well. The average absolute correlation between true and estimated factor never drops below 0.95. It improves with N and with higher ρ. As discussed in Bai (2003), the performance of the method depends on the minimum of $\sqrt{N}$ and T. Since we consider $N < T$ the fact that performance does not improve with T is intuitive. Moving on to the proportion of series variance explained, we see that the estimated factor does as well or even better compared to the true factor.

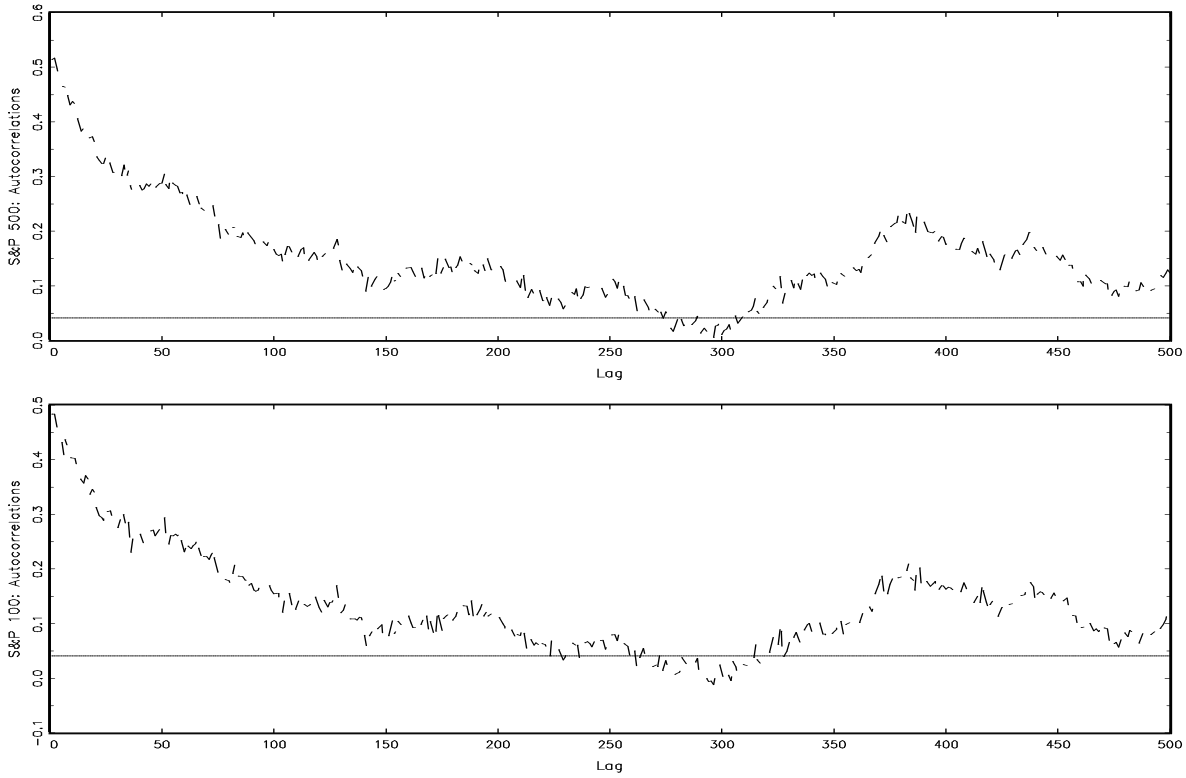
4 Empirical analysis

We apply our suggested method of analysing stochastic volatility data to large datasets given by the constituents of S&P500 and S&P100. Data, obtained from Datastream are daily returns and span the period 01/01/1995-13/01/2004 comprising 2356 observations. We consider only companies for which data are available throughout the period leading us to have $N = 438$ for the S&P500 dataset and $N = 93$ for the S&P 100 dataset. Once all periods when markets were closed are dropped from the datasets 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 $\tilde{w}_t = w_t - 1/T \sum_{t=1}^{T} w_t$, and we apply principal components to $\tilde{w}_t$. In Table 2, we report the cumulative average R^2 across all $\tilde{w}_t$ for the first 20 factors. It is clear that whereas the first factor explains about 10% of the variation in the datasets, further factors can add only marginally to the explanatory power of the set of factors. Therefore we conclude that one factor captures a large common component of the stochastic volatility of these large datasets. Further insight is obtained by plotting the autocorrelation functions (with the upper 95% bound of the confidence interval of the null hypothesis that the process is white noise) of the factors in Figure 1. The autocorrelation functions decline very slowly. This points towards long memory models whose autocorrelation function declines hyperbolically. Consequently, we fit an $ARFIMA(p, d, 0)$ to the factors. The results in Table 3 show evidence of stationary long memory.

We, next, consider stochastic simulations (using 1000 replications) to generate the density forecast for the an equally weighted portfolio return, whose constituents are those from the S&P500. The density forecasts are produced out of sample, using recursive estimation of the parameters estimates of the model and the forecast evaluation periods is made of the last 100 observations.

The density forecast of the factor stochastic volatility model are obtained by simulating the common factor using its estimated long memory representation. Then, we couple the

Figure 1: Factor Autocorrelation Function



realisations for the artificial generation of the factor together with stochastic simulations for the idiosyncratic components obtained by simulating the state space model in (6). This gives the artificial generated paths for $\tilde{w}_t$ and combining this with (1) we obtain the stochastic simulations for the demeaned returns which are then added up to provide the forecast of the portfolio returns under alternative scenarios. All error terms are set to be $N(0,1)$ random variables. We compare the accuracy of the density forecast for the factor stochastic volatilities with the one associated with a model using a separate stochastic volatility specification (INDIV), an Orthogonal GARCH (Alexander (2000)), a multivariate EWMA specification (J. P. Morgan (1996)) and a constant covariance model (CCOV)[1]. In the case of INDIV we simply remove the effect of the factors in retrieving the stochastic volatility estimate. In order to obtain portfolio simulations from the OGARCH model we, use random draws from a

[1]We attempted to estimate the DCC model developed by Engle (2002) for the S&P500 dataset using the MATLAB routine developed by Kevin Sheppard. However, we were not able to use the available routine. As pointed out by the author of the routine, for such a large dataset, the system would require approximately 36GB of memory available for MATLAB.

$N(0,1)$ to stochastically simulate the first principal component of the vector of (de-meaned) stock returns as a GARCH(1,1). For the artificial generation of portfolio returns obtained from the CCOV and from the multivariate EWMA, we simulate $z_{t+1}H_{t+1}^{0.5}$, where the N-dimensional vector of the errors, z, is drawn from a $N(0,I)$ distribution. Specifically, as for the CCOV model, H_{t+1} is set to the constant sample covariance matrix; as for the EWMA, both the volatilities and also the cross products in H_{t+1}have an IGARCH specification with weights equal to 0.94 and 0.06, for the GARCH and ARCH component, respectively.

We consider two methods of evaluating the predictive densities we obtain. We, first, consider the Kolmogorov-Smirnov (KS) test to test the null of i.i.d. uniformity in the probability integral transform $z_t = \int_{-\infty}^{y_t} p_t(u)du$,where y_t is portfolio return realisation and $p_t(u)$ is conditional prediction of the portfolio return associated with scenario u. Then, we consider the Berkowitz (2001) Likelihood Ratio test for the null of normality and serial independence in the series $\Phi^{-1}(z_t)$,where $\Phi^{-1}(.)$ is the inverse normal cumulative density function. The probability values for the KS test and the Berkowitz (2001) test are respectively 0.30 and 0.14. Conversely, for the case of the INDIV model the relevant probability values are 0.06 and 0.004, whereas for both OGARCH models, EWMA and CCOV both probability values are 0.

We also consider the performance of each model in producing the probability forecast of an event characterised by negative portfolio returns. The probability forecast estimation for each time period that belongs to the forecast evaluation period is obtained by counting the number of scenarios for which the equally weighted portfolio return are negative and dividing this by the number of replications. The performance measure used is the Kuipers score (see Granger and Pesaran (2000)), which is defined as the difference between the proportion of negative returns events that were correctly forecasted, and we use 0.5 as the cut-off value to call a negative return event via probability forecast. Kuipers scores above zero mean that the model generates proportionally more correct forecasts than false alarms. Looking at the scores for the four models we consider we get the following scores: 0.117 (FACTOR), 0.003 (INDIV), -0.082 (OGARCH1), -0.132(CCOV) and 0.038 (EWMA). Clearly, the factor model is to be preferred over the other specifications according to this criterion.

5 Conclusion

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 method has been extended to highly persistent stationary data which exhibit long memory behaviour. A small Monte Carlo analysis has been undertaken. The method has been applied to the S&P 500 constituent dataset with very encouraging results.

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 Dinensions," *Econometrica*, 71, 135–173.

Baillie, R. T., C. F. Chung, and M. A. Tieslau (1996): "Analysing Inflation by the Fractionally Integrated ARFIMA-GARCH Model," *Journal of Applied Econometrics*, 11, 23–40.

Berkowitz, J. (2001): "Testing Density Forecasts with Applications to Risk Management," *Journal of Business and Economic Statistics*, 19, 465–474.

Chib, S., F. Nardari, and N. Shephard (2006): "Analysis of high dimensional multivariate stochastic volatility models," *Journal of Econometrics*, 134, 317–341.

Cipollini, A., and G. Kapetanios (2004): "A Stochastic Variance Factor Model for Large Datasets and an Application to S&P data," *Queen Mary, University of London Working Paper 506*.

Engle, R. (2002): "Dynamic conditional correlation: a simple class of multivariate GARCH models," *Journal of Business and Economics Statistics*, 20, 339–350.

Granger, C. W. J., and M. H. Pesaran (2000): "Economic and statistical measures of forecast accuracy," *Journal of Forecasting*, 19, 537–560.

Harvey, A. C., E. Ruiz, and N. Shephard (1994): "Multivariate Stochastic Variance Models," *Review of Economic Studies*, 61, 247–264.

J. P. Morgan (1996): "Riskmetrics," *Technical Documents, 4th ed. New York*.

Stock, J. H., and M. W. Watson (2002): "Macroeconomic Forecasting Using Diffusion Indices," *Journal of Business and Economic Statistics*, 20, 147–162.

Table 1A: Results for $AR(1)$					
ρ	N/T	200	500	1000	2000
Average Absolute Correlation					
0.1	50	0.953	0.953	0.954	0.954
	100	0.975	0.976	0.976	0.976
	200	0.988	0.988	0.988	0.988
0.5	50	0.963	0.964	0.965	0.965
	100	0.981	0.982	0.982	0.982
	200	0.990	0.991	0.991	0.991
0.9	50	0.988	0.990	0.990	0.990
	100	0.994	0.995	0.995	0.995
	200	0.997	0.997	0.998	0.998
Average Relative Explained Variance					
0.1	50	1.062	1.048	1.048	1.048
	100	1.035	1.025	1.024	1.024
	200	1.026	1.013	1.012	1.012
0.5	50	1.073	1.043	1.038	1.037
	100	1.057	1.026	1.020	1.019
	200	1.066	1.015	1.010	1.010
0.9	50	1.081	1.023	1.012	1.010
	100	1.056	1.019	1.008	1.005
	200	1.055	1.015	1.006	1.003

Table 1B: Results for $ARFIMA(1, d, 0)$					
d	N/T	200	500	1000	2000
Average Absolute Correlation					
0.2	50	0.974	0.975	0.976	0.976
	100	0.987	0.987	0.988	0.988
	200	0.993	0.994	0.994	0.994
0.4	50	0.985	0.988	0.989	0.989
	100	0.993	0.994	0.994	0.994
	200	0.996	0.997	0.997	0.997
Average Relative Explained Variance					
0.2	50	1.315	1.095	1.055	1.032
	100	1.194	1.077	1.033	1.018
	200	1.188	1.075	1.024	1.014
0.4	50	1.275	1.124	1.058	1.026
	100	1.252	1.101	1.041	1.021
	200	1.218	1.101	1.041	1.018

Table 2: 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

Table 3: Long Memory Model		
	S&P500	S&P100
p	1	1
$\hat{d}$	0.416	0.398
$std(\hat{d})$	0.0197	0.0198
$95\%CI(\hat{d})$	(0.37, 0.46)	(0.36, 0.44)
CI: Confidence Interval		