ARTICLE IN PRESS

Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

High-dimensional copula-based distributions with mixed frequency data[☆]

Dong Hwan Oh[a], Andrew J. Patton[b,*]

[a] Quantitative Risk Analysis Section, Federal Reserve Board, Washington DC 20551, United States

[b] Department of Economics, Duke University, Box 90097, Durham NC 27708, United States

ARTICLE INFO

Article history:
Available online xxxx

JEL classification:
C32
C51
C58

Keywords:
High frequency data
Forecasting
Composite likelihood
Nonlinear dependence

ABSTRACT

This paper proposes a new model for high-dimensional distributions of asset returns that utilizes mixed frequency data and copulas. The dependence between returns is decomposed into linear and nonlinear components, enabling the use of high frequency data to accurately forecast linear dependence, and a new class of copulas designed to capture nonlinear dependence among the resulting uncorrelated, low frequency, residuals. Estimation of the new class of copulas is conducted using composite likelihood, facilitating applications involving hundreds of variables. In- and out-of-sample tests confirm the superiority of the proposed models applied to daily returns on constituents of the S&P 100 index.

1. Introduction

A model for the multivariate distribution of the returns on large collections of financial assets is a crucial component in modern risk management and asset allocation. Modelling high-dimensional distributions, however, is not an easy task and only a few models are typically used in high dimensions, most notably the Normal distribution, which is still widely used in practice and academia despite its notorious limits, for example, thin tails and zero tail dependence.

This paper provides a new approach for constructing and estimating high-dimensional distribution models. Our approach builds on two active areas of recent research in financial econometrics. First, high frequency data has been shown to be superior to daily data for measuring and forecasting variances and covariances, see Andersen et al. (2006) for a survey of this very active area of research. This implies that there are gains to be had by modelling linear dependence, as captured by covariances, using high frequency data. Second, copula methods have been shown to be useful for constructing flexible distribution models in high dimensions, see Christoffersen et al. (2013), Oh and Patton (2016) and Creal and Tsay (2014). These two findings naturally lead to the question of whether high frequency data and copula methods can be combined to improve the modelling and forecasting of high-dimensional return distributions.

Exploiting high frequency data in a lower frequency copula-based model is not straightforward as, unlike variances and covariances, the copula of low frequency (say daily) returns is not generally a known function of the copula of high frequency returns. Thus the link between high frequency volatility measures (e.g., realized variance and covariance) and their low frequency counterparts cannot generally be exploited when considering dependence via the copula function. We overcome this hurdle by decomposing the dependence structure of low frequency asset returns into linear and nonlinear components. We then use high frequency data to accurately model the linear dependence, as measured by covariances, and a new class of copulas to capture the remaining dependence in the low frequency standardized residuals.

The difficulty in specifying a copula-based model for standardized, uncorrelated, residuals, is that the distribution of the residuals must imply an identity correlation matrix. Independence is only sufficient for uncorrelatedness, and we wish to allow for

☆ We thank the guest editor (Eric Ghysels), two anonymous referees, and Tim Bollerslev, Federico Bugni, Jia Li, Oliver Linton, Bruno Rémillard, Enrique Sentana, Neil Shephard, and George Tauchen as well as seminar participants at the Federal Reserve Board, Rutgers University, SUNY-Stony Brook, Toulouse School of Economics, University of Cambridge, and University of Montreal for their insightful comments. We also benefited from data mainly constructed by Sophia Li and Ben Zhao. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Federal Reserve Board.

* Corresponding author.
E-mail addresses: donghwan.oh@frb.gov (D.H. Oh), andrew.patton@duke.edu (A.J. Patton).

http://dx.doi.org/10.1016/j.jeconom.2016.04.011
0304-4076/© 2016 Elsevier B.V. All rights reserved.

Please cite this article in press as: Oh, D.H., Patton, A.J., High-dimensional copula-based distributions with mixed frequency data. Journal of Econometrics (2016), http://dx.doi.org/10.1016/j.jeconom.2016.04.011

possible nonlinear dependence between these linearly unrelated variables. Among existing work, only the multivariate Student's t distribution has been used for this purpose, as an identity correlation matrix can be directly imposed on this distribution. We dramatically increase the set of possible models for uncorrelated residuals by proposing methods for generating "jointly symmetric" copulas. These copulas can be constructed from any given (possibly asymmetric) copula, and when combined with any collection of (possibly heterogeneous) symmetric marginal distributions they guarantee an identity correlation matrix. Evaluation of the density of our jointly symmetric copulas turns out to be computationally difficult in high dimensions, but we show that composite likelihood methods (see Varin et al., 2011 for a review) may be used to estimate the model parameters and undertake model selection tests.

This paper makes four main contributions. Firstly, we propose a new class of "jointly symmetric" copulas, which are useful in multivariate density models that contain a covariance matrix model (e.g., GARCH–DCC, HAR, stochastic volatility, etc.) as a component. Second, we show that composite likelihood methods may be used to estimate the parameters of these new copulas, and in an extensive simulation study we verify that these methods have good finite-sample properties. Third, we propose a new and simple model for high-dimensional covariance matrices drawing on ideas from the HAR model of Corsi (2009) and the DCC model of Engle (2002), and we show that this model outperforms the familiar DCC model empirically. Finally, we present a detailed empirical application of our model to 104 individual U.S. equity returns, showing that our proposed approach significantly outperforms existing approaches both in-sample and out-of-sample.

Our methods and application are related to several existing papers. Most closely related is the work of Lee and Long (2009), who also consider the decomposition into linear and nonlinear dependence, and use copula-based models for the nonlinear component. However, Lee and Long (2009) focus only on bivariate applications, and their approach, which we describe in more detail in Section 2, is computationally infeasible in high dimensions. Our methods are also clearly related to copula-based density models, some examples of which are cited above, however in those approaches only the variances are modelled prior to the copula stage, meaning that the copula model must capture both the linear and nonlinear components of dependence. This makes it difficult to incorporate high frequency data into the dependence model. Papers that employ models for the joint distribution of returns that include a covariance modelling step include Chiriac and Voev (2011), Jondeau and Rockinger (2012), Hautsch et al. (2015), and Jin and Maheu (2013). As models for the standardized residuals, those papers use the Normal or Student's t distributions, both of which are nested in our class of jointly symmetric models, and which we show are significantly beaten in our application to U.S. equity returns.

The paper is organized as follows. Section 2 presents our approach for modelling high-dimensional distributions. Section 3 presents multi-stage, composite likelihood methods for model estimation and comparison, which are studied via simulations in Section 4. Section 5 applies our model to daily equity returns and compares it with existing approaches. Section 6 concludes. An appendix contains all proofs, and a web appendix contains additional details, tables and figures (see Appendix A).

2. Models of linear and nonlinear dependence

We construct a model for the conditional distribution of the N-vector $\mathbf{r}_t$ as follows:

$$\mathbf{r}_t = \boldsymbol{\mu}_t + \mathbf{H}_t^{1/2}\mathbf{e}_t \qquad (1)$$

$$\text{where } \mathbf{e}_t \sim iid\, \mathbf{F}(\cdot;\boldsymbol{\eta}) \qquad (2)$$

where $\mathbf{F}(\cdot;\boldsymbol{\eta})$ is a joint distribution with zero mean, identity covariance matrix and "shape" parameter $\boldsymbol{\eta}$, and $\boldsymbol{\mu}_t = E[\mathbf{r}_t|\mathcal{F}_{t-1}]$, $\mathbf{H}_t = V[\mathbf{r}_t|\mathcal{F}_{t-1}]$, $\mathcal{F}_t = \sigma(\mathbf{Y}_t, \mathbf{Y}_{t-1}, \ldots)$, and $\mathbf{Y}_t$ includes $\mathbf{r}_t$ and possibly other time t observables, such as realized variances and covariances. To obtain $\mathbf{H}_t^{1/2}$, we suggest using the spectral decomposition due to its invariance to the order of the variables. Note that by assuming that $\mathbf{e}_t$ is iid, we impose that all dynamics in the conditional joint distribution of $\mathbf{r}_t$ are driven by the conditional mean and (co)variance. This common, and clearly strong, assumption goes some way towards addressing the curse of dimensionality faced when N is large.

In existing approaches, see Chiriac and Voev (2011), Jondeau and Rockinger (2012), Hautsch et al. (2015), and Jin and Maheu (2013) for example, $\mathbf{F}$ would be assumed multivariate Normal (which reduces to independence, given that $\mathbf{e}_t$ has identity covariance matrix) or Student's t, and the model would be complete. Instead, we consider the decomposition of the joint distribution $\mathbf{F}$ into marginal distributions F_i and copula $\mathbf{C}$ using Sklar's (1959) theorem:

$$\mathbf{e}_t \sim \mathbf{F}(\cdot;\boldsymbol{\eta}) = \mathbf{C}(F_1(\cdot;\boldsymbol{\eta}), \ldots, F_N(\cdot;\boldsymbol{\eta});\boldsymbol{\eta}). \qquad (3)$$

Note that the elements of $\mathbf{e}_t$ are uncorrelated but may still exhibit cross-sectional dependence, which is completely captured by the copula $\mathbf{C}$. Combining Eqs. (1)–(3) we obtain the following density for the distribution of returns:

$$\mathbf{f}_t(\mathbf{r}_t) = \det\left(\mathbf{H}_t^{-1/2}\right) \times \mathbf{c}(F_1(e_{1t}), \ldots, F_N(e_{Nt})) \times \prod_{i=1}^{N} f_i(e_{it}). \qquad (4)$$

Thus this approach naturally reveals two kinds of dependence between returns: "linear dependence," captured by conditional covariance matrix $\mathbf{H}_t$, and any "nonlinear dependence" remaining in the uncorrelated residuals $\mathbf{e}_t$, captured by the copula $\mathbf{C}$. There are two important advantages in decomposing a joint distribution of returns in this way. First, it allows the researcher to draw on the large literature on measuring, modelling and forecasting conditional covariance matrix $\mathbf{H}_t$ with low and high frequency data. For example, GARCH-type models such as the multivariate GARCH model of Bollerslev et al. (1988), the BEKK model of Engle and Kroner (1995), and the dynamic conditional correlation (DCC) model of Engle (2002) naturally fit in Eqs. (1) and (2). The increasing availability of high frequency data also enables us to use more accurate models for the conditional covariance matrix, see, for example, Bauer and Vorkink (2011), Chiriac and Voev (2011), and Noureldin et al. (2012), and those models are also naturally accommodated by Eqs. (1)–(2).[1] Second, the model specified by Eqs. (1)–(3) is easily extended to high-dimensional applications given that multi-stage separate estimation of the conditional mean of the returns, the conditional covariance matrix of the returns, the marginal distributions of the standardized residuals, and finally the copula of the standardized residuals is possible. Of course, multi-stage estimation is less efficient than one-stage estimation, however the main difficulty in high-dimensional applications is the proliferation of parameters and the growing computational burden as the dimension increases. By allowing for multi-stage estimation we overcome this obstacle.

[1] In the part of our empirical work that uses realized covariance matrices, we take these as given, and do not take a stand on the specific continuous-time process that generates the returns and realized covariances. This means that, unlike a DCC-type model, for example, which only considers daily returns, or a case where the continuous-time process was specified, we cannot simulate or generate multi-step ahead predictions from these models.

Please cite this article in press as: Oh, D.H., Patton, A.J., High-dimensional copula-based distributions with mixed frequency data. Journal of Econometrics (2016), http://dx.doi.org/10.1016/j.jeconom.2016.04.011