Journal of Econometrics 201 (2017) 384–399

Contents lists available at ScienceDirect

Journal of Econometrics

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

Using principal component analysis to estimate a high dimensional factor model with high-frequency data☆

Yacine Aït-Sahalia [a,*], Dacheng Xiu [b]

[a] Department of Economics, Princeton University and NBER, 26 Prospect Avenue, Princeton, NJ 08540, USA
[b] Booth School of Business, University of Chicago, 5807 S Woodlawn Avenue, Chicago, IL 60637, USA

ARTICLE INFO

Article history:
Available online 19 August 2017

JEL classification:
C13
C14
C55
C58
G01

Keywords:
High-dimensional data
High-frequency data
Latent factor model
Principal components
Portfolio optimization

ABSTRACT

This paper constructs an estimator for the number of common factors in a setting where both the sampling frequency and the number of variables increase. Empirically, we document that the covariance matrix of a large portfolio of US equities is well represented by a low rank common structure with sparse residual matrix. When employed for out-of-sample portfolio allocation, the proposed estimator largely outperforms the sample covariance estimator.

1. Introduction

This paper proposes an estimator, using high frequency data, for the number of common factors in a large-dimensional dataset. The estimator relies on principal component analysis (PCA) and novel joint asymptotics where both the sampling frequency and the dimension of the covariance matrix increase. One by-product of the estimation method is a well-behaved estimator of the increasingly large covariance matrix itself, including a split between its systematic and idiosyncratic matrix components.

Principal component analysis (PCA) and factor models represent two of the main methods at our disposal to estimate large covariance matrices. If nonparametric PCA determines that a common structure is present, then a parametric or semiparametric factor model becomes a natural choice to represent the data. Prominent examples of this approach include the arbitrage pricing theory (APT) of Ross (1976) and the intertemporal capital asset pricing model (ICAPM) of Merton (1973), which provide an economic rationale for the presence of a factor structure in asset returns. Chamberlain and Rothschild (1983) extend the APT strict factor model to an approximate factor model, in which the residual covariances are not necessarily diagonal, hence allowing for comovement that is unrelated to the systematic risk factors. Based on this model, Connor and Korajczyk (1993), Bai and Ng (2002), Amengual and Watson (2007), Onatski (2010) and Kapetanios (2010) propose statistical methodologies to determine the number of factors, while Bai (2003) provides tools to conduct statistical inference on the common factors and their loadings. Connor and Korajczyk (1988) use PCA to test the APT.

In parallel, much effort has been devoted to searching for observable empirical proxies for the latent factors. The three-factor model by Fama and French (1993) and its many extensions are widely used examples, with factors constructed using portfolios returns often formed by sorting firm characteristics. Chen et al. (1986) propose macroeconomic variables as factors, including inflation, output growth gap, interest rate, risk premia, and term premia. Estimators of the covariance matrix based on observable factors are proposed by Fan et al. (2008) in the case of a strict factor model and Fan et al. (2011) in the case of an approximate factor model. A factor model can serve as the reference point for shrinkage estimation (see Ledoit and Wolf (2012) and Ledoit and

☆ We are benefited from the very helpful comments of the Editor and two anonymous referees, as well as extensive discussions with Jianqing Fan, Alex Furger, Chris Hansen, Jean Jacod, Yuan Liao, Nour Meddahi, Markus Pelger, and Weichen Wang, as well as seminar and conference participants at CEMFI, Duke University, the 6th French Econometrics Conference in Honor of Christian Gouriéroux, the 8th Annual SoFiE Conference, the 2015 IMS-China International Conference on Statistics and Probability, and the 11th World Congress of the Econometric Society. We are also grateful to Chaoxing Dai for excellent research assistance.

* Corresponding author.
E-mail addresses: yacine@princeton.edu (Y. Aït-Sahalia), dacheng.xiu@chicagobooth.edu (D. Xiu).

http://dx.doi.org/10.1016/j.jeconom.2017.08.015
0304-4076/

Wolf (2004)). Alternative methods rely on various forms of thresholding (Bickel and Levina, 2008a,b; Cai and Liu, 2011; Fryzlewicz, 2013; Zhou et al., 2014) whereas the estimator in Fan et al. (2013) is designed for latent factor models.

The above factor models are static, as opposed to the dynamic factor models introduced in Gouriéroux and Jasiak (2001) to represent stochastic means and volatilities, extreme risks, liquidity and moral hazard in insurance analysis. Dynamic factor models are developed in Forni et al. (2000), Forni and Lippi (2001), Forni et al. (2004) and Doz et al. (2011), in which the lagged values of the unobserved factors may also affect the observed dependent variables; see Croux et al. (2004) for a discussion. Forni et al. (2009) adapt structural vector autoregression analysis to dynamic factor models.

Both static and dynamic factor models in the literature have typically been cast in discrete time. By contrast, this paper provides methods to estimate continuous-time factor models, where the observed variables are continuous Itô semimartingales. The literature dealing with continuous-time factor models has mainly focused on models with observable explanatory variables in a low dimensional setting. For example, Mykland and Zhang (2006) develop tools to conduct analysis of variance as well as univariate regression, while Todorov and Bollerslev (2010) add a jump component in the univariate regression setting and Aït-Sahalia et al. (2014) extend the factor model further to allow for multivariate regressors and time-varying coefficients.

When the factors are latent, however, PCA becomes the main tool at our disposal. Aït-Sahalia and Xiu (2015) extend PCA from its discrete-time low frequency roots to the setting of general continuous-time models sampled at high frequency. The present paper complements it by using PCA to construct estimators for the number of common factors, and exploiting the factor structure to build estimators of the covariance matrix in an increasing dimension setting, without requiring that a set of observable common factors be pre-specified. The analysis is based on a general continuous-time semiparametric approximate factor model, which allows for stochastic variation in volatilities as well as correlations. Independently, Pelger (2015a, b) propose an alternative estimator for the number of factors and factor loadings, with a distributional theory that is entry-wise, whereas the present paper concentrates on the matrix-wise asymptotic properties of the covariance matrix and its inverse.

This paper shares some theoretical insights with the existing literature of approximate factor models in discrete time, in terms of the strategy for estimating the number of factors. However, there are several distinctions, which require a different treatment in our setting. For instance, the identification restrictions we impose differ from those given by e.g., Bai (2003), Doz et al. (2011) and Fan et al. (2013), due to the prevalent presence of heteroscedasticity in high frequency data. Also, the discrete-time literature on determining the number of factors relies on random matrix theory for i.i.d. data (see, e.g., Bai and Ng, 2002; Onatski, 2010; Ahn and Horenstein, 2013; Trapani, 2017), which is not available for semimartingales.

The methods in this paper, including the focus on the inverse of the covariance matrix, can be useful in the context of portfolio optimization when the investable universe consists of a large number of assets. For example, in the Markowitz model of mean–variance optimization, an unconstrained covariance matrix with d assets necessitates the estimation of $d(d+1)/2$ elements, which quickly becomes unmanageable as d grows, and even if feasible would often result in optimal asset allocation weights that have undesirable properties, such as extreme long and short positions. Various approaches have been proposed in the literature to deal with this problem. The first approach consists in imposing some further structure on the covariance matrix to reduce the number of parameters to be estimated, typically in the form of a factor model along the lines discussed above, although Green and Hollifield (1992) argue that the dominance of a single factor in equity returns can lead empirically to extreme portfolio weights. The second approach consists in imposing constraints on the portfolio weights (Jagannathan and Ma, 2003; Pesaran and Zaffaroni, 2008; DeMiguel et al., 2009a; El Karoui, 2010; Fan et al., 2012; Gandy and Veraart, 2013) or penalties Brodie et al. (2009). The third set of approaches are Bayesian and consist in shrinkage of the covariance estimates (Ledoit and Wolf, 2003), assuming a prior distribution for expected returns and covariances and reformulating the Markowitz problem as a stochastic optimization one (Lai et al., 2011), or simulating to select among competing models of predictable returns and maximize expected utility (Jacquier and Polson, 2010). A fourth approach consists in modeling directly the portfolio weights in the spirit of Aït-Sahalia and Brandt (2001) as a function of the asset's characteristics (Brandt et al., 2009). A fifth and final approach consists in abandoning mean–variance optimization altogether and replacing it with a simple equally-weighted portfolio, which may in fact outperform the Markowitz solution in practice (DeMiguel et al., 2009b).

An alternative approach to estimating covariance matrices using high-frequency data is fully nonparametric, i.e., without assuming any underlying factor structure, strict or approximate, latent or not. Two issues have attracted much attention in this part of the literature, namely the potential presence of market microstructure noise in high frequency observations and the potential asynchronicity of the observations: see Aït-Sahalia and Jacod (2014) for an introduction. Various methods are available, including Hayashi and Yoshida (2005), Aït-Sahalia et al. (2010), Christensen et al. (2010), Barndorff-Nielsen et al. (2011), Zhang (2011), Shephard and Xiu (2012) and Bibinger et al. (2014). However, when the dimension of the asset universe increases to a few hundreds, the number of synchronized observations is bound to drop, which requires severe downsampling and hence much longer time series to be maintained. Dealing with an increased dimensionality without a factor structure typically requires the additional assumption that the population covariance matrix itself is sparse (see, e.g., Tao et al., 2011, 2013b, a). Fan et al. (2016) assume a factor model but with factors that are observable.

The rest of the paper is organized as follows. Section 2 sets up the model and assumptions. Section 3 describes the proposed estimators and their properties. We show that both the factor-driven and the residual components of the sample covariance matrix are identifiable, as the cross-sectional dimension increases. The proposed PCA-based estimator is consistent, invertible and well-conditioned. Additionally, based on the eigenvalues of the sample covariance matrix, we provide a new estimator for the number of latent factors. Section 4 provides Monte Carlo simulation evidence.

Section 5 implements the estimator on a large portfolio of stocks. We find a clear block-diagonal pattern in the residual correlations of equity returns, after sorting the stocks by their firms' global industrial classification standard (GICS) codes. This suggests that the covariance matrix can be approximated by a low-rank component representing exposure to some common factors, plus a sparse component, which reflects their sector/industry specific exposure. Empirically, we find that the factors uncovered by PCA explain a larger fraction of the total variation of asset returns than that explained by observable portfolio factors such as the market portfolio, the Fama–French portfolios, as well as the industry-specific ETF portfolios. Also, the residual covariance matrix based on PCA is sparser than that based on observable factors, with both exhibiting a clear block-diagonal pattern. Finally, we find that the PCA-based estimator outperforms the sample covariance estimator in out-of-sample portfolio allocation. Section 6 concludes. Mathematical proofs are in the appendix.