Extracting Statistical Factors When Betas Are Time-Varying [*]

Patrick Gagliardini[†] and Hao Ma[‡]

First Draft: January 2019
This Version: September 2020

Abstract

This paper deals with identification and inference on the unobservable *conditional* factor space and its dimension in large unbalanced panels of asset returns. The model specification is nonparametric regarding the way the loadings vary in time as functions of common shocks and individual characteristics. The number of active factors can also be time-varying as an effect of the changing macroeconomic environment. The method deploys Instrumental Variables (IV) which have full-rank covariation with the factor betas in the cross-section. It allows for a large dimension of the vector generating the conditioning information by machine learning techniques. In an empirical application, we infer the conditional factor space in the panel of monthly returns of individual stocks in the CRSP dataset between January 1971 and December 2017.

Keywords: Large Panel, Unobservable Factors, Conditioning Information, Instrumental Variables, Machine Learning, Post-Lasso, Artificial Neural Networks

[*] We thank T. Andersen, F. Carlini, R. Engle, Y. Li, M. Pelger, M. Rubin, O. Scaillet, K. Singleton, V. Todorov, F. Trojani, P. Zaffaroni and participants at the SoFiE 2019 Annual Meeting in Shanghai, the CFE 2019 conference in London, the EFA 2020 virtual conference and a seminar at the University of Geneva for useful comments. We gratefully acknowledge financial support from the Swiss National Science Foundation (SNSF) Grant 105218 186233.

[†] Università della Svizzera italiana (USI, Lugano) and Swiss Finance Institute (SFI). E-mail address: patrick.gagliardini@usi.ch

[‡] Università della Svizzera italiana (USI, Lugano) and Swiss Finance Institute (SFI). E-mail address: hao.ma@usi.ch

1 Introduction

This paper studies empirical asset pricing models with unobservable risk factors and time-varying factor sensitivities. The exposures of assets to various forms of systematic risk can be time-varying, as an effect of the changing macroeconomic environment and individual asset characteristics. The empirical asset pricing literature deploying time-varying beta specifications has mostly focused on models with observable factors (see e.g. Connor and Korajczyk (1989), Shanken (1990), Cochrane (1996), Ferson and Schadt (1996), Ferson and Harvey (1991, 1999), Lettau and Ludvigson (2001), Petkova and Zhang (2005) using portfolios, and Gagliardini, Ossola and Scaillet (2016, 2019) using individual stocks). In such models, the risk factors are a-priori identified with observable economic variables. However, empirical research shows that there is a great latitude in the choice of the economic factors explaining equity portfolios returns, see in particular the emerging literature on the "factor zoo" (e.g. Cochrane (2011), Feng, Giglio and Xiu (2019)). Moreover, it is not clear whether the same risk factors are relevant for individual stocks as well.

The difficulty in identifying economic risk factors can be solved empirically by using latent factor models. In recent years, there has been a growing interest for large panel models with unobservable factors. However, this literature has been confined mostly to a framework with time-invariant factor sensitivities (see e.g. the pioneering work of Bai and Ng (2002), Bai (2003), Stock and Watson (2002), and the more recent developments in e.g. Onatski (2009), Ahn and Horenstein (2013)). Allowing for general time-variation of the coefficients as functions of the conditioning information is key when addressing the questions on how many, and which ones, are the systematic risk factors. In fact, a model with one risk factor, say, and time-varying risk exposures can be confused with a model with several risk factors, when the exposures are erroneously assumed time-invariant. The aim of this paper is to overcome these difficulties by simultaneously accommodating the unobservability of the factor values and the time-variation of the factor sensitivities. This task requires to develop new econometric methodologies, because the traditional approach for inference in linear latent factor models relying on Principal Component Analysis (PCA) cannot be applied with time-varying loadings.

The contributions of this paper are twofold. First, we develop a novel methodology to extract statistical (i.e. unobservable) factors in linear cross-sectional factor models with time-varying coefficients. Our modeling framework is both coherent with the implications of the absence of arbitrage opportunities in large economies with conditioning information, and very general regarding the assumptions on the dynamics of factor betas. In fact, our approach is essentially nonparametric with respect to the beta dynamics, and focuses on valid inference on the conditional factor structure - both regarding the dimension and the span of the conditional factor space. A major distinctive feature of our work compared to existing literature is that the dimension of the conditional factor space can be time-varying, as an effect of the systematic changes in factor exposures. To the best of our knowledge, this paper is the first to propose inference on the number of latent factors with conditional betas. Our methodology relies on two key inputs, namely availability of a sufficiently large number of instruments, and knowledge of the variables spanning the filtration of common shocks. Instruments are variables assumed to be *cross-sectionally* uncorrelated with idiosyncratic errors but full-rank correlated with factor betas, and are used to construct cross-sectional averages from which factors can be extracted, in a similar vein as in Gagliardini and Gourieroux (2017). Given the conditional setting adopted in this paper, the factor space at a given date is defined in terms of innovations which are unpredictable w.r.t. the sigma-field of aggregate shocks at the previous date. This implies that factor extraction requires finding the residuals of the projection of the cross-sectional averages on the sigma-field of aggregate shocks at the previous date. Being the latter sigma-field generated by a potentially very large number of economic and financial variables, we deploy a machine learning approach via post-Lasso estimators (see e.g. Belloni et al. (2012)) or artificial neural networks. We investigate the small- and large-sample properties of our estimators and inferential procedures. We show the consistency of the estimator of the (time-varying) number of factors in a large-n-large-T asymptotics [1].

Second, our empirical contribution consists in estimating the conditional factor space from the unbalanced panel of monthly returns for individual stocks in the CRSP dataset in the period between January 1971 and December 2017. We select 22 characteristics for each firm as our instrumental variables from Freyberger, Neuhierl, and Weber (2017) and the portfolio weights of some observable factor returns. We include 34 different monthly variables to generate the information set of aggregate shocks. Our methodology provides a first evidence that the number of latent conditional factors is time-varying and small, ranging between 1 and 2 in most months in our sample. Our estimator selects only one conditional factor in almost 75% of the months and especially often after 2000. Some finite-sample volatility of the estimator makes challenging the exact determination of the regime shifts in the number of conditional factors. We also investigate to which extent the conditional factor space coincides with that of traditional specifications with observable factors. Considering a model with only 1 conditional latent factor, we estimate the conditional canonical correlation between our latent factor and the market factor. The results show that the latent factor has a conditional correlation over 0.5 with the market factor in most months in the period 1980-2000, and that it tends to correlate more just before or during recessions.

[1] Zaffaroni (2019) considers inferential theory with unobserved factors in a setting with n large and T fixed.

When we let the number of conditional factors be time-varying, we see large improvements in the first conditional canonical correlation of the latent factors with Fama-French factors, compared with the single factor case. Such increases mean that not only the market factor but also e.g. the SMB and HML factors help explain our conditional latent factor space. However, the estimated second conditional correlation is constantly rather small, suggesting that Fama-French factors do not span the entire conditional factor space.

Some recent papers allow time variation in betas in unobserved factor models via instruments. Connor, Linton and Hagmann (2012) estimate nonparametrically characteristic-based factor models in which the betas are individual-independent functions of observed regressors. An extension of their base specification allows for time-variation of the regressors. Fan, Liao and Wang (2016) develop the Projected PCA method which accommodates an additive individual-specific component in the characteristic-based betas but in a time-invariant framework. Pelger and Xiong (2018) consider kernel estimation of betas which are functions of a small vector of observable state variables. Gagliardini, Ossola and Scaillet (2019) develop a diagnostic criterion to select the number of omitted unconditional latent factors in a large approximate conditional model with observed factors. Recent studies in the same topic as ours include Kelly, Pruitt and Su (2017, 2019). In their pioneering work, these authors introduce Instrumented Principal Component Analysis (IPCA) for estimating both the time-varying beta loadings and the latent factors. The method makes it possible to disentangle the dynamics of beta loadings from latent factors by assuming a constant mapping matrix Γ, which links linearly "instrumental" variables to factor loadings. The estimation method consists in solving a least squares problem with respect to matrix Γ and sample factor values in a similar vein as panel fixed effects estimators. Both our and Kelly, Pruitt and Su (2017, 2019) papers follow similar ideas as PCA but under a dynamic setting to extract statistical latent factors from a large dataset. In order to do so, we both use instumental variables as additional information for identifying factors. However, our paper is distinct from theirs in several aspects, and the two methodologies are complementary in some respect. Firstly, we estimate the conditional factor space in a more general setting without specifying any particular model that links the factor loadings to instruments. The framework of IPCA is useful instead as it allows to estimate the mapping of factor loadings to instrumental variables at the same time. Secondly, we introduce a consistent selection procedure for the number of conditional factors and allow the latter to be time-varying to better accommodate reality. Thirdly, our methodology allows to estimate non-parametrically the time-varying risk premium associated with the conditional latent factors, and the associated Stochastic Discount Factor (SDF). Some recent work also takes advantage of machine learning methods to achieve greater flexibility in modeling beta dynamics and accommodate large conditioning information sets. Among others, Gu, Kelly and Xiu (2019) use autoencoder artificial neural networks. Finally, Ait-Sahalia and Xiu (2017), Liao and Yang (2018), Pelger (2019 a,b), Li, Todorov and Tauchen (2019) consider inference in large-dimensional models with unobserved factors using high-frequency data.

The rest of the paper is structured as follows. Section 2 introduces the model framework. Section 3 presents identification of the conditional factor space and its dimension. Section 4 provides estimates of the factor space and a procedure to infer the number of conditional factors. Section 5 discusses the economic identification of the factor space. Section 6 provides the empirical analysis on the CRSP dataset. Technical developments and a Monte Carlo analysis are relegated to the Appendix.

2 Conditional Factor Model

We consider the following linear factor model for asset returns with time-varying coefficients:

$$y_{i,t} = a_{i,t-1} + b'_{i,t-1} f_t + \varepsilon_{i,t}, \tag{1}$$

where $y_{i,t}$ denotes the excess return on asset i in period t, for $i = 1, \cdots, n$ and $t = 1, \cdots, T$. Coefficient $a_{i,t-1}$ is a $\mathcal{F}_{i,t-1}$-measurable scalar, and $b_{i,t-1}$ is a $\mathcal{F}_{i,t-1}$-measurable $k \times 1$ vector, where $\mathcal{F}_{i,t}$ for t varying is the filtration of the relevant information for asset i. The sigma-field $\mathcal{F}_{i,t}$ is such that $\mathcal{F}_{i,t} = \mathcal{F}_t \vee \mathcal{G}_{i,t}$, where $\mathcal{F}_t$ is the information set of common shocks (see Section 3.1 for a formal definition) and $\mathcal{G}_{i,t}$ is an additional component that is both asset- and time-dependent. The $k \times 1$ vector f_t is $\mathcal{F}_t$-measurable and represents the systematic risk factors. It is unobservable to the econometrician. The number of factors k is assumed unknown and has to be estimated from data. In the time-varying setting the number of active factors with non-vanishing loadings can be smaller than k at some dates t, is $\mathcal{F}_{t-1}$ measurable, and has also to be determined. The idiosyncratic error terms $\varepsilon_{i,t}$ are such that $E(\varepsilon_{i,t}|\mathcal{F}_{i,t-1}) = 0$ and $Cov(\varepsilon_{i,t}, f_t|\mathcal{F}_{i,t-1}) = 0$, which implies that $a_{i,t-1}$ and $b_{i,t-1}$ are conditional alphas and betas for asset i, respectively. Moreover, the error terms are weakly cross-sectionally correlated as in an approximate factor structure à la Chamberlain and Rothschild (1983) under assumptions introduced below.

Gagliardini, Ossola and Scaillet (2016), henceforth referred as GOS, study the implications of Arbitrage Pricing Theory (APT) in conditional factor models for large economies. They build on the framework of Al-Najjar (1995, 1998, 1999) with a continuum of assets, and extend it to a setting with conditional information and random draws of the indices of the n assets in the sample.

Under the Assumptions APR.1-4 in GOS, the following pricing restriction holds:

$$a_{i,t-1} = b'_{i,t-1}\nu_t, \tag{2}$$

for all i, where vector ν_t is $\mathcal{F}_{t-1}$-measurable. By inserting equation (2) into (1), we get:

$$y_{i,t} = b'_{i,t-1}g_t + \varepsilon_{i,t}, \tag{3}$$

where:

$$g_t = \nu_t + f_t. \tag{4}$$

Then, the vector of equity risk premia is:

$$\lambda_t = E[g_t|\mathcal{F}_{t-1}] = \nu_t + E[f_t|\mathcal{F}_{t-1}], \tag{5}$$

where $E[\cdot|\mathcal{F}_t]$ denotes conditional expectation given $\mathcal{F}_t$.

The focus of this paper is on identification and statistical inference for the conditional factor space, including its dimension, in a framework that is nonparametric regarding the dynamics of factor loadings.

Definition 1. *The conditional factor space is identifiable if the stochastic process f_t can be identified from observable data up to linear affine conditional transformations mapping f_t into $R_{t-1}f_t + r_{t-1}$, where the non-singular $k \times k$ matrix R_{t-1} and the $k \times 1$ vector r_{t-1} are $\mathcal{F}_{t-1}$-measurable.*

In fact, latent factor models are invariant to one-to-one affine transformations of the unobservable factors. In our conditional setting, the transformation can be time-varying and predetermined, i.e. function of the information in $\mathcal{F}_{t-1}$. [2] Under suitable normalization restrictions on the conditional factor space introduced below, a unique “representer" can be identified. The data available to the econometrician consists in a large panel of asset returns, the variables generating the information set $\mathcal{F}_t$ of common shocks, and a set of instrumental variables in the filtrations $\mathcal{F}_{i,t}$ to be defined below. For expository purpose, we present the identification and estimation strategies in the framework of a balanced panel of asset returns. The extension of the methodology to accommodate unbalanced panels with missing-at-random data is simple and is discussed later.

[2] Specifically, the model can be written in terms of the transformed factor $f_t^* = R_{t-1}f_t + r_{t-1}$ and transformed loading $b_{i,t-1}^* = (R'_{t-1})^{-1}b_{i,t-1}$, with $\nu_t^* = R_{t-1}\nu_t - r_{t-1}$ and $\lambda_t^* = R_{t-1}\lambda_t$.

3 Identification

Our identification strategy heavily relies on cross-sectional averages of the available data. The next subsection sets the ground to ensure that the probability limits of such cross-sectional averages in large samples exist and are measurable w.r.t. suitable information sets.

3.1 Probability limits of cross-sectional averages

Let $W_{i,t}$ denote a random vector measurable w.r.t. information set $\mathcal{F}_{i,t}$. [3]

Assumption 1. *(i) For any $W_{i,t}$ such that $\sup_{i \geq 1} E[\|W_{i,t}\|^{\beta}] < \infty$ for some $\beta > 1$, the probability limit*

$$\mathbb{E}^c[W_{\cdot,t}] := \operatorname*{plim}_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} W_{i,t}$$

exists, for any t, and is measurable w.r.t. the sigma-field $\mathcal{F}_t$. (ii) The variables generating $\mathcal{F}_t$ are observable and known to the econometrician.

We work with Assumption 1 (i) as a convenient high-level regularity condition that accommodates various forms of cross-sectional and time-series dependence. With cross-sectionally i.i.d. data, Assumption 1 (i) corresponds to the standard Law of Large Numbers (LLN) and the sigma field $\mathcal{F}_t$ is trivial. If the data $W_{i,t}$, for $i = 1, 2, ...$, are exchangeable, these data are i.i.d. conditionally on the sigma-field generated by symmetric functions, i.e. $\mathcal{F}_t = \bigcap_{N=1}^{\infty} \mathcal{F}_t^N$, where $\mathcal{F}_t^N$ denotes the sigma-field generated by N-symmetric functions of the variables $W_{i,t}$, for $i = 1, 2, ...$ (see e.g. Andrews (2005), Hall and Heyde (1980), Chapter 7). Assumption 1 (i) clarifies why sigma-field $\mathcal{F}_t$ corresponds to common - or systematic - shocks in the economy, that is, the shocks which are not eliminated by diversification across a large number of assets. Assumption 1 (ii) implies that conditional expectations of observable quantities given $\mathcal{F}_t$ are identifiable for the econometrician. [4]

3.2 Identification with time-invariant number of factors

Suppose first that the true number of factors k, unknown to the econometrician, is time-invariant. We start our identification strategy by imposing the existence of $K \times 1$ instrumental variables $w_{i,t}$

[3] This condition covers the case in which $W_{i,t} = (x'_{i,1}, ..., x'_{i,t})'$ consists of the history of variable x up to time t.

[4] Note that Assumption 1 (ii) does not conflict with the latent character of the factors, since the factors are unknown functions of the variables generating the sigma-field $\mathcal{F}_t$.

with $K \geq k$, i.e. we assume a known upper bound on the number of unobservable factors and the availability of at least as many asset-level observable variables which satisfy the next assumption.

Assumption 2. *There exists a $K \times 1$ vector of instrumental variables $w_{i,t-1}$ measurable w.r.t. $\mathcal{F}_{i,t-1}$ such that:*

$$(i)\, \mathbb{E}^c[w_{\cdot,t-1}\varepsilon_{\cdot,t}] = 0,$$

$$(ii)\, \Gamma_{t-1} := \mathbb{E}^c[w_{\cdot,t-1}b'_{\cdot,t-1}] \textit{ is a } K \times k \textit{ full-rank matrix, P-a.s.,}$$

for any $t \geq 1$.

Hence, instruments are predetermined variables which feature zero correlation with error terms and full-rank covariance with the factor loadings *cross-sectionally.* Lagged values of asset characteristics such as firm size, book-to-market ratio, earnings per share, etc. can provide valid instrumental variables. [5]

Define the limit cross-sectional average:

$$\xi_t = \mathbb{E}^c[w_{\cdot,t-1}y_{\cdot,t}], \tag{6}$$

which is measurable w.r.t. $\mathcal{F}_t$ under Assumption 1 and is identifiable being a function of the observed data distribution. Vector ξ_t corresponds to the returns of K (well-diversifies) portfolios. Then, equation (3) and Assumption 2 imply:

$$\xi_t = \Gamma_{t-1}\, g_t, \tag{7}$$

where Γ_{t-1} is the $K \times k$ full-rank, $\mathcal{F}_{t-1}$-measurable matrix defined in Assumption 2 (ii). By computing the conditional variance of vector ξ_t given information set $\mathcal{F}_{t-1}$, we get the $K \times K$ symmetric matrix

$$V(\xi_t|\mathcal{F}_{t-1}) = \Gamma_t\, V(g_t|\mathcal{F}_{t-1})\, \Gamma'_t \tag{8}$$

which has rank k. Thus, the true number of latent factors k is identifiable by the rank of the symmetric matrix $V(\xi_t|\mathcal{F}_{t-1})$, or equivalently, the number of its non-zero eigenvalues.

Let us now come to the identification of process g_t. From equation (8) the eigenvectors of conditional variance-covariance matrix $V(\xi_t|\mathcal{F}_{t-1})$ associated to the non-zero eigenvalues span the range of matrix Γ_{t-1}, which is therefore identifiable. Then, from equation (7) vector g_t is identifiable

[5] Assumption 2 does not require instrumental variables to be necessarily time-varying. In the Appendix, we show that we can generate time-invariant instrumental variables from return data and prove that they satisfy the identification conditions above.

up to a one-to-one linear transformation that is $\mathcal{F}_{t-1}$ measurable. To make the identification argument more operational in view of the estimation step, we introduce a normalization of the conditional factor space to fix this conditional transformation uniquely. Let J_{t-1} be the $K \times k$ matrix of normalized eigenvectors associated to the non-zero eigenvalues of $V(\xi_t|\mathcal{F}_{t-1})$. This matrix is $\mathcal{F}_{t-1}$-measurable, full-rank, and identifiable up to sign changes if the eigenvalues of $V(f_t|\mathcal{F}_{t-1})$ are distinct, P-a.s.

Assumption 3. *Without loss of generality, we assume that the following normalization restriction holds for the latent factors:*

$$\Gamma_{t-1} = J_{t-1},$$

for any t.

Assumption 3 is not a restriction on the Data Generating Process (DGP) since this normalization can always be achieved by a conditional transformation of the factors. Indeed, for any normalization we have $\Gamma_{t-1} = J_{t-1}R_{t-1}$, where R_{t-1} is a $k \times k$ invertible matrix, measurable w.r.t. $\mathcal{F}_{t-1}$. The conditional factor structure is observationally invariant under the transformation from f_t to the new factor $f_t^* = A_{t-1}f_t$, say, with loadings $b_{i,t-1}^* = (A_{t-1}^{-1})'b_{i,t-1}$, where A_{t-1} is a non-singular $k \times k$ matrix that is measurable w.r.t. the information set $\mathcal{F}_{t-1}$. Then, for the rotated factors we have:

$$\begin{aligned}\Gamma_{t-1}^* &= \mathbb{E}^c[w_{\cdot,t-1}b_{\cdot,t-1}^{*\prime}] \\ &= \mathbb{E}^c[w_{\cdot,t-1}b_{\cdot,t-1}']A_{t-1}^{-1} = \Gamma_{t-1}A_{t-1}^{-1} = J_{t-1}R_{t-1}A_{t-1}^{-1} = J_{t-1},\end{aligned}$$

if we choose $A_{t-1} = R_{t-1}$. Assumption 3 is a normalization of the conditional factors, which is alternative to e.g. imposing a conditional variance of the factor vector equal to the identity matrix (plus additional restrictions on the factor loadings). We prefer the former approach over the latter one since it simplifies the identification argument and the derivation of the estimators. In fact, under Assumption 3 equation (8) implies:

$$V(f_t|\mathcal{F}_{t-1}) = \Lambda_t, \tag{9}$$

where Λ_t is the diagonal $k \times k$ matrix of the non-zero eigenvalues of matrix $V(\xi_t|\mathcal{F}_{t-1})$.

From equation (7) we deduce that under Assumption 3 vector g_t is given by:

$$g_t = J_{t-1}'\xi_t, \tag{10}$$

i.e., the population *conditional* principal component of vector ξ_t. The identified value g_t depends on the normalization implied by Assumption 3. We can interpret this identification strategy as if we were using the exactly identified set of instrumental variables $J'_{t-1}w_{i,t-1}$, and take the limit cross-sectional averages of cross-products of these new asset characteristics and returns: $g_t = \mathbb{E}^c[(J'_{t-1}w_{\cdot,t-1})\, y_{\cdot,t}]$.

For the purpose of identification of the vector f_t of factor values, we need the next assumption.

Assumption 4. *Without loss of generality, we assume $E[f_t|\mathcal{F}_{t-1}] = 0$.*

This assumption can always be imposed on the DGP by a conditional shift of the factor. [6] Under Assumption 4 the risk premium vector is:

$$\lambda_t = \nu_t = E[g_t|\mathcal{F}_{t-1}] = (J_{t-1})' E[\xi_t|\mathcal{F}_{t-1}]. \tag{11}$$

From (4) and (11) we can identify the vector of unobservable factors:

$$f_t = g_t - \lambda_t = (J_{t-1})'(\xi_t - E[\xi_t|\mathcal{F}_{t-1}]). \tag{12}$$

We have proved the following result.

Proposition 1. *Under Assumptions 1-4, the number of latent conditional factors k, the conditional factor values f_t, and the conditional risk premium vector λ_t, are identifiable, for all t.*

In particular, for the identification of the conditional factor space we need the observability of the sigma-field $\mathcal{F}_t$ of aggregate shocks, but not of the whole $\mathcal{F}_{i,t}$, nor the specification of the conditional betas dynamics.

It is instructive to compare our identification strategy with other recent contributions in the theory of large-dimensional latent factor models. Bai and Ng (2002), Bai (2003), Onatski (2008), and Ahn and Horenstein (2013) among others propose methods for estimating the number of factors, and the factor space, in static latent factor models. These methods rely on the eigenvalue-eigenvector decomposition of the sample variance-covariance matrix of the data under the assumption of time-invariant loadings. In our setting with time-varying loadings, we rely instead on the time-series conditional variance-covariance matrix of a vector of cross-sectional averages, which is shown to have reduced rank equal to the number of unobservable conditional factors.

The use of instrumental variables to model time-varying betas is a central idea in the IPCA methodology of Kelly, Pruitt and Su (2017, 2019). IPCA assumes that the number of factors is known and constant through time, and that the cross-sectional regression coefficient matrix

[6] Indeed, suppose $E[f_t|\mathcal{F}_{t-1}] \neq 0$. We can rewrite our model (1) in an observationally equivalent way as $y_{i,t} = a^*_{i,t-1} + b'_{i,t-1}f^*_t + \varepsilon_{i,t}$, where $a^*_{i,t-1} = a_{i,t-1} + b'_{i,t-1}E(f_t|\mathcal{F}_{t-1})$ and $f^*_t = f_t - E(f_t|\mathcal{F}_{t-1})$. The new factor f^*_t matches Assumption 4.

$(\mathbb{E}^c[w_{\cdot,t-1}w'_{\cdot,t-1}])^{-1}\Gamma_{t-1}$ is time-variant. Further, IPCA does not impose the no-arbitrage restriction (2).

3.3 Identification with time-varying number of factors

We can extend our identification strategy to treat the case where the true number of factors is time-varying. In model (1) time variation of the number of factors is implied by the possibly changing dimension of the space spanned by the loadings (i.e., the column space of the loadings matrix for a large number of assets) during different economic phases. The number of common factors at date t is denoted k_t and is defined as the rank of the cross-sectional variance-covariance matrix of the loadings:

$$k_t = \text{Rank } \mathbb{E}^c\left[(b_{\cdot,t-1} - \mathbb{E}^c[b_{\cdot,t-1}])(b_{\cdot,t-1} - \mathbb{E}^c[b_{\cdot,t-1}])'\right], \tag{13}$$

and is $\mathcal{F}_{t-1}$ measurable. We have $k_t < k$ at date t, if the loadings of $k - k_t$ factors are zero for most assets at that date, or more generally if $(k - k_t)$ linear combinations - with $\mathcal{F}_{t-1}$-measurable weights - of the components of vector $b_{i,t-1}$ are zero for most assets.

Identification of factor dimension k_t relies on the following assumption, which is version of Assumption 2 adapted to cover a time-varying (TV) number of factors.

Assumption 2.TV *There exists a $K \times 1$ vector of instrumental variables $w_{i,t-1}$ measurable w.r.t. $\mathcal{F}_{i,t-1}$ such that:*

$$(i)\, \mathbb{E}^c[w_{\cdot,t-1}\varepsilon_{\cdot,t}] = 0,$$

$$(ii)\, \Gamma_{t-1} := \mathbb{E}^c[w_{\cdot,t-1}b'_{\cdot,t-1}] \textit{ is a } K \times k \textit{ matrix of column rank } k_t,$$

for any t, where $k_t \leq k$ is $\mathcal{F}_{t-1}$-measurable.

By the argument in Section 3.2 we identify k_t from the rank of matrix $V(\xi_t|\mathcal{F}_{t-1}) = \Gamma_{t-1}\, V(g_t|\mathcal{F}_{t-1})\, \Gamma'_{t-1}$. Hence, the time-varying number of factors is identifiable at each date under Assumption 2.TV.

For the identification of vector g_t we mimic the strategy in Section 3.2. First, let J_{t-1} be the $K \times k_t$ matrix having as columns the normalized eigenvectors of matrix $V(\xi_t|\mathcal{F}_{t-1})$ associated with the k_t non-zero eigenvalues.

Assumption 3.TV *Without loss of generality, the following normalization restriction holds for the latent factors:*

$$\Gamma_{t-1} = [J_{t-1} \;:\; 0_{K\times(k-k_t)}],$$

for any t.

The interpretation of this normalization restriction when $k_t < k$ is that we are adopting a transformation of the unobservable factors such that the loadings of the last $k - k_t$ components are zero for almost all assets at date t. This yields the right block of zeros $0_{K\times(k-k_t)}$ in matrix Γ_t. The normalization of the left block of Γ_{t-1} to get J_{t-1} is achieved by an additional transformation of the first k_t components of the factor vector in analogy to Section 3.2.

From equation (7) we deduce that under Assumption 3.TV:

$$\bar{g}_t := J'_{t-1}\xi_t \tag{14}$$

identifies the k_t-dimensional sub-block of the vector g_t corresponding to the rotated factors with non-vanishing loadings at date t. The corresponding factor values $\bar{f}_t$ and risk premia vector $\bar{\lambda}_t$ are identified as before under Assumption 4 as

$$\bar{f}_t = \bar{g}_t - \bar{\lambda}_t = (J_{t-1})'(\xi_t - E[\xi_t|\mathcal{F}_{t-1}]), \qquad \bar{\lambda}_t = E[\bar{g}_t|\mathcal{F}_{t-1}] = (J_{t-1})'E[\xi_t|\mathcal{F}_{t-1}]. \tag{15}$$

The remaining $k - k_t$ factor values and risk premia are not identifiable at date t.

We summarize our results in the next proposition.

Proposition 2. *Under Assumptions 1, 2.TV, 3.TV and 4, the number of latent conditional factors k_t, the conditional factor values $\bar{f}_t$, and the conditional risk premium vector $\bar{\lambda}_t$, are identifiable, for all t.*

4 Estimation

The identification strategy developed in the previous section naturally leads to an estimation methodology by the plug-in approach consisting in replacing population quantities with their sample analogues. Conditional expectations $E(\cdot|\mathcal{F}_{t-1})$ given the information set of common shocks are involved in several intermediate steps of the identification strategy. We estimate these conditional expectations by nonparametric methods. To cope with the possibly high-dimensional framework and the implied curse of dimensionality, we adopt machine learning approaches that we present in the next subsection.

4.1 Estimating conditional expectations by machine learning

To start with, we need to specify the variables generating the information set $\mathcal{F}_{t-1}$.

Assumption 5. *The information set $\mathcal{F}_t$ is generated by the d-dimensional observable vector Markov process Z_t.*

This assumption implies:

$$E[\zeta_t|\mathcal{F}_{t-1}] = E[\zeta_t|Z_{t-1}] =: \psi^\zeta(Z_{t-1})$$

for a function $\psi^\zeta(\cdot)$ and any random vector ζ_t. We adopt the notation $\psi_l^\zeta(Z_{t-1}) = E[\zeta_{tl}|Z_{t-1}]$, for $l = 1, \cdots, L$, where ζ_{tl} is the lth component of the L-dimensional vector ζ_t, and ψ_l^ζ is the lth component of the vector function ψ^ζ. The estimation of the conditional expectation $E[\zeta_{tl}|\mathcal{F}_{t-1}]$ amounts to estimation of function ψ_l^ζ in the nonparametric regression model:

$$\zeta_{tl} = \psi_l^\zeta(Z_{t-1}) + u_{tl}, \quad E[u_{tl}|Z_{t-1}] = 0,$$

for $l = 1, \cdots, L$. Nonparametric regression estimators popular in the econometrics literature include kernel smoothing and Sieve (e.g. series) estimators (e.g. Haerdle and Linton (1994), Chen (2007)). Nonparametric regression estimators suffer from the curse of dimensionality, i.e. the convergence rate deteriorates as dimension d grows, implying unreliable estimates in samples of realistic size when d is large. [7] Imposing functional restrictions on the regression function via e.g. additive specifications or index models is a way to cope with the curse of dimensionality. Here we consider two methods in machine learning which gained popularity in econometrics for estimating high-dimensional conditional expectations.

i) Lasso and Post-Lasso estimators

We firstly adopt the post-Lasso estimator used e.g. in Belloni et al. (2012). Suppose that there are p known functions of Z_{t-1} that we collect in the vector $h(Z_{t-1}) = (h_1(Z_{t-1}), \cdots, h_p(Z_{t-1}))'$ to be used in estimation of conditional expectation functions $\psi_l^\zeta(Z_{t-1})$, $l = 1, \cdots, L$. The list $h(Z_{t-1})$ can consist of series terms with respect to the components of vector Z_{t-1}, such as orthogonal polynomials, B-splines or other function bases used in Sieve estimation, and their number p is possibly much larger than the sample size T.

Sparsity condition. *Each conditional expectation function $\psi_l^\zeta(Z_{t-1})$ is well-approximated by a*

[7] For e.g. a tensor Sieve estimator with polynomial basis, the number of parameters to approximate a function in d dimensions using polynomials of order M grow like $O(M^d)$. This leads to a convergence rate of the Sieve estimator $O_p(T^{-\frac{m}{2p+m}})$ to estimate a function in the Holder class of degree m.

list of $s \geq 1$ unknown common variables functions:

$$\psi_l^{\zeta}(Z_{t-1}) = h(Z_{t-1})'\gamma_{l0} + e_l(Z_{t-1}), \quad l = 1, \cdots, L,$$
$$\underset{1\leq l\leq L}{max} \|\gamma_{l0}\|_0 \leq s = o(T), \quad \underset{1\leq l\leq L}{max} [\mathbb{E}_T e_l(Z_{t-1})^2]^{1/2} \leq c_s = O_p(\sqrt{s/T}),$$

where $\mathbb{E}_T$ denotes sample average across the T observations of variable Z_t, $e_l(Z_{t-1})$ is the approximation error, and $\|\gamma\|_0$ denotes the number of non-zero elements of vector γ.

The above sparsity assumption, which is Condition AS in Belloni et al. (2012), imposes that only $s \ll T$ functions in the list - with unknown identity - are relevant for approximating the conditional expectation functions of interest.

The Lasso estimator of vector γ_l is defined as a solution of the convex optimization problem

$$\widehat{\gamma}_l = \arg \min_{\gamma \in \mathbb{R}^p} \widehat{Q}_l(\gamma) + \frac{\theta}{T} \|\widehat{\Upsilon}_l \gamma\|_1$$

where $\widehat{Q}_l(\gamma) = \mathbb{E}_T[(\zeta_{tl} - h(Z_{t-1})'\gamma)^2]$ is the least square criterion function using ζ_{tl} as the dependent variable and $h(Z_{t-1})$ as regressors, θ is the penalty level, $\|z\|_1 = \sum_{l=1}^{p} |z_l|$ is the L^1-norm of vector z in $\mathbb{R}^p$, and $\widehat{\Upsilon}_l = diag(\widehat{\upsilon}_{l1}, \cdots, \widehat{\upsilon}_{lp})$ is a diagonal matrix specifying the penalty loadings. [8]

The L^1-penalty in the Lasso criterion implies that the elements of the estimate vector $\hat{\gamma}_{l,i}$ are different from zero only for the indices i in a subset $\hat{S}$ of $\{1, ..., p\}$. The post-Lasso estimator $\widehat{\widehat{\gamma}}_l$ is obtained by performing Ordinary Least Squares (OLS) on the variables that are selected by Lasso in a first step, i.e.

$$(\widehat{\widehat{\gamma}}_l)_{\hat{S}} = \mathbb{E}_T[h(Z_{t-1})_{\hat{S}} h(Z_{t-1})'_{\hat{S}}]^{-1} \mathbb{E}_T[h(Z_{t-1})_{\hat{S}} \zeta_{tl}]$$

and $(\widehat{\widehat{\gamma}}_l)_{\hat{S}^c} = 0$, where $(\gamma)_S$ denotes the subvector of γ for the components with indices $i \in S$, set $\hat{S}^c$ is the complement of $\hat{S}$ in $\{1, ..., p\}$, and $\hat{s} = |\hat{S}|$ is the cardinality of $\hat{S}$. Then, the post-Lasso estimator of the conditional expectation $\psi_l^{\zeta}(Z_{t-1})$ takes the form $\widehat{\psi_l^{\zeta}}(Z_{t-1}) = h(Z_{t-1})'\widehat{\widehat{\gamma}}_l$, for $l = 1, \cdots, L$.

Under the above Sparsity condition and other assumptions, Belloni et al. (2012) show that the post-Lasso estimator $\widehat{\psi_l^{\zeta}}$ has a convergence rate $O_p\left(\sqrt{\frac{s \log(s \vee T)}{T}}\right)$ in the empirical L^2 norm. The convergence rate depend on s but not on d directly.

[8] Belloni et al. (2012) characterize the optimal penalty loadings. In order to implement them, in the empirical application we use the same algorithm as in the Appendix of Belloni et al. (2012).

ii) Neural Networks

The other nonparametric estimation method that we adopt for our analysis is the artificial neural network. In the mathematical theory of artificial neural networks, the universal approximation theorem states that a feed-forward network with a single hidden layer containing a large number of neurons can approximate any continuous function (Gallant, White (1988), Cybenko (1989), Hornik et al. (1991), Hornik (1991)).

Universal Approximation Theorem. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a non-constant, bounded and continuous function, and let $I \subset \mathbb{R}^d$ be a compact set. The set of functions*

$$ANN_d = \left\{ \psi(Z) = \sum_{m=1}^{M} v_m \phi(w'_m Z + b_m),\ v_m, b_m \in \mathbb{R},\ w_m \in \mathbb{R}^d,\ m = 1, ..., M,\ M \in \mathbb{N} \right\}$$

is dense in the set $C^d(I)$ of real-valued continuous functions on I.

We base our analysis on feed-forward networks, which consist of an input layer, one or more hidden layers, and an output layer. Figure 1 shows a graphical representation in our case with one hidden layer that contains M neurons. Each neuron m firstly forms a linear combination of the d input predictors in vector Z_{t-1}, and then applies the nonlinear activation function ϕ to its aggregated value before sending its output $H_{l,m,t-1}$ for component l to the next layer. At last, the prediction for ζ_{lt} is modeled as a linear combination of the outputs $H_{l,m,t-1}$ for the different neurons m. In formulas:

$$H_{l,m,t-1} = \phi(b_{l,m} + w'_{l,m} Z_{t-1}),\ m = 1, ..., M,$$

$$\psi_l^{\zeta}(Z_{t-1}; \gamma_l) = v_{l,m} + \sum_{m=1}^{M} v_{l,m} H_{l,m,t-1},$$

where vector γ_l consists of the network parameters b, w, v for component l. The estimator $\widehat{\psi_l^{\zeta}}(Z_{t-1}) = \psi_l^{\zeta}(Z_{t-1}; \hat{\gamma}_l)$ is obtained by replacing vector γ_l with an estimate from a penalized nonlinear least square criterion:

$$\hat{\gamma}_l = \arg\min_{\gamma \in \mathbb{R}^p} \mathbb{E}_T \left[(\zeta_{tl} - \psi_l^{\zeta}(Z_{t-1}; \gamma))^2 \right] + \frac{\theta}{T} \text{pen}(\gamma).$$

As activation function we use $\phi(x) = x_+$, where $x_+ = x$ if $x \geq 0$, and $= 0$ otherwise, i.e. the positive part of x. [9] Several features of the the neural network's architecture influence its approximation quality, such as the number of hidden layers ("width"), the number of neurons in

[9] This function is commonly refered to as rectified linear unit in the artificial neural network literature.

each layer ("depth"), and which units are connected. We discuss our choice in the Monte Carlo and empirical application sections. We refer to e.g. Hastie et al. (2009), Chapter 11, for more details. Chen and Shen (1998) and Chen and White (1999) prove that, if $\psi_l^\zeta(\cdot)$ belongs to a smothness class defined by an integrability condition for the Fourier transform, then a single-layer feed-forward ANN estimator has convergence rate $O_p\left((T/\log T)^{-\frac{1}{4}\frac{1+2/(d+1)}{1+1/(d+1)}}\right)$ in L^2 norm. The convergence rate is faster than $T^{-1/4}$ for any $d \geq 1$. Gu, Kelly, and Xiu (2018) and Chen, Pelger and Zhu (2019) provide recent applications of neural networks and other machine learning methods in empirical asset pricing.

Equipped with these machine learning methods to estimate conditional expectations, we denote $\widehat{E}(\zeta_t|\mathcal{F}_{t-1}) := (\widehat{\psi_1^\zeta}(Z_{t-1}), \cdots, \widehat{\psi_L^\zeta}(Z_{t-1}))'$ the vector of estimates. In some cases vector ζ_t is unobservable and has to be estimated by a consistent estimator $\hat{\zeta}_t$, say. In those cases, we apply the Lasso methodology to $\hat{\zeta}_t$ and denote as $\hat{E}(\hat{\zeta}_t|\mathcal{F}_{t-1}) = \widehat{\psi^{\hat{\zeta}}}(Z_{t-1})$ the estimated conditional expectation function. Moreover, when ζ_t is a random matrix, the machine learning estimator is defined for each element.

We now turn to the estimation of the conditional factor space and the determination of its dimension.

4.2 Estimation of the conditional factor space

i) Time-invariant number of factors

Let us first assume that the number of latent factors k is time-invariant and known to the econometrician, and suppose that a possibly overidentified set of $K \geq k$ instruments is available (this assumption is relaxes afterwards). We start with the estimation of process ξ_t defined in equation (6) by means of a cross-sectional average:

$$\hat{\xi}_t = \frac{1}{n}\sum_{i=1}^{n} w_{i,t-1} y_{i,t}. \tag{16}$$

This estimator is consistent as $n \to \infty$ under Assumption 1.

Next, to estimate vector g_t using equation (10), the conditional variance-covariance matrix $V(\xi_t|\mathcal{F}_{t-1})$ is estimated by the machine learning methods introduced in Subsection 4.1 applied to the estimated process $\hat{\xi}_t$ in (16), namely:

$$\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1}) := \widehat{\psi^{\hat{\xi}\hat{\xi}'}}(Z_{t-1}) - \widehat{\psi^{\hat{\xi}}}(Z_{t-1})[\widehat{\psi^{\hat{\xi}}}(Z_{t-1})]'. \tag{17}$$

Since the machine learning estimator is computed elementwise, we regularize the estimate $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$ to ensure positive semi-definiteness. The eigenvectors associated with the k largest eigenvalues of the regularized matrix yield the columns of matrix $\hat{J}_{t-1}$. The signs are chosen recursively such that $\hat{J}'_{t,j}\hat{J}_{t-1,j} \geq 0$ for all columns $j = 1, \cdots, k$ and dates $t = 2, \cdots, T$. Then, the estimator of g_t is

$$\hat{g}_t = (\hat{J}_{t-1})'\hat{\xi}_t, \tag{18}$$

and corresponds to the vector of the k first sample *conditional* principal components of $\hat{\xi}_t$.

Finally, building on equations (11) and (12), the estimators for the factor values and risk premia vectors are:

$$\hat{f}_t = \hat{g}_t - \hat{\lambda}_t = (\hat{J}_{t-1})'(\hat{\xi}_t - E[\hat{\xi}_t|\mathcal{F}_{t-1}]), \qquad \hat{\lambda}_t = (\hat{J}_{t-1})'E[\hat{\xi}_t|\mathcal{F}_{t-1}]. \tag{19}$$

The estimator of the conditional factor space is in closed-form up to the conditional expectation estimate by machine learning deployed for getting $E[\hat{\xi}_t|\mathcal{F}_{t-1}]$ and $\hat{J}_{t-1}$. Moreover, the methodology relies on cross-sectional averaging for estimating $\hat{g}_t$, which implies that it readily applies to an unbalanced panel under the missing-at-random assumption.

ii) Time-varying number of factors

When the number of factors k_t is time-varying, the estimators (18) and (19) apply, with $\hat{J}_{t-1}$ being the $K \times k_t$ matrix of the standardized eigenvectors associated with the k_t largest eigenvalues of $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$. In this case estimates $\hat{\hat{g}}_t = \hat{J}'_{t-1}\hat{\xi}_t$, $\hat{\hat{f}}_t = (\hat{J}_{t-1})'(\hat{\xi}_t - E[\hat{\xi}_t|\mathcal{F}_{t-1}])$ and $\hat{\hat{\lambda}}_t = (\hat{J}_{t-1})'E[\hat{\xi}_t|\mathcal{F}_{t-1}]$ correspond to the k_t-dimensional conditional rotation of the factor vector with non-vanishing loadings at date t.

4.3 Estimation of the number of factors

In this section we consider the problem of estimating the true number of conditional factors. Recall that, under Assumption 2 (resp., Assumption 2.TV), the factor dimension k (resp. k_t) equals the rank of the conditional variance-covariance matrix $V(\xi_t|\mathcal{F}_{t-1}) = \Gamma_{t-1}\, V(g_t|\mathcal{F}_{t-1})\, \Gamma'_{t-1}$. Hence, inference on the factor space dimension at date t is tantamount to inference on the rank of matrix $V(\xi_t|\mathcal{F}_{t-1})$, or, equivalently, on the number of non-zero eigenvalues of that matrix. We build on the insight of Ahn and Horenstein (2013) and adapt their eigenvalue-ratio selection principle to our setting with conditional factors. The idea is that, if the eigenvalues of the empirical counterpart of matrix $V(\xi_t|\mathcal{F}_{t-1})$ feature a scree-plot decay behavior, the number of conditional factors can be estimated by the integer r such that the ratio between the rth and $(r+1)$th largest eigenvalues is maximal. [10] Next we detail the selection procedure in the cases of constant and time-varying

[10] Statistical tests for the rank of a matrix V are developed e.g. by Cragg and Donald (1996), Robin and Smith (2000), Kleibergen and Paap (2006), Al-Sadoon (2017). These tests rely on the asymptotic normality of the

number of conditional factors, respectively.

i) Time-invariant case

Let us first assume that the true number of conditional factors k is constant through time, and $k \leq q$ for some known upper bound q. Let us denote $\delta_r(A)$ the rth largest eigenvalue of symmetric matrix A. With this notation, let $\hat{\rho}_{r,t} = \delta_r[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]/\delta_{r+1}[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]$ be the ratio of two consecutive eigenvalues, and $\hat{\rho}_r = \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\hat{\rho}_{r,t}$ the average across time of such eigenvalue ratios, for any integer $r \leq q$, where the estimator $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$ is defined in (17). The index set $\mathcal{T}_1 = \left\{t \leq T \ : \ \delta_{q+1}[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})] \geq \sigma^a\right\}$ with cardinality T_1, where $a > 1$ is a constant and $\sigma \downarrow 0$ is a positive sequence shrinking to zero as n and T increase, is used as a trimming device to avoid too small (or negative) eigenvalue estimates in the denominator of the eigenvalue ratio. Then, the estimator of the number of conditional factors is:

$$\hat{k} = \arg\max_{1\leq r\leq q} \hat{\rho}_r, \tag{20}$$

i.e. the location of the largest average eigenvalue ratio.

ii) Time-varying case

Let us now allow for the number of conditional factors k_t being time-varying as in Subsection 3.3. Then, the estimator of the number of conditional factors at date t is:

$$\hat{k}_t = \arg\max_{1\leq r\leq q} \hat{\rho}_{r,t}, \tag{21}$$

for $t \in \mathcal{T}_1$, i.e. the eigenvalue ratio is maximized at each date t after trimming.

We stress that the consistency of estimators $\hat{k}$ and $\hat{k}_t$ does not follow from the theory developed in Ahn, Horenstein (2013). Indeed, matrix $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$ is a time-series *conditional* variance-covariance matrix of a vector of cross-sectional averages, and not a sample unconditional variance-covariance matrix of a large panel. We show the consistency of estimators $\hat{k}$, $\hat{k}_t$ is Section 4.4. Moreover, other estimators could possibly be developed building on the approaches of e.g. Bai, Ng (2002) and Onatski (2008). Our estimators $\hat{k}$, $\hat{k}_t$ based on Ahn, Horenstein (2013) eigenvalue ratio principle have to advantage not to require the choice of a tuning parameter for the scale of the eigenvalues.

estimator of matrix V. Such distributional results are available for kernel regression and Sieve estimators, but are much more scarce for machine learning estimators to the best of our knowledge. Moreover, when the matrix of interest V is a variance-covariance matrix, the results for standard tests may fail, e.g. Donald, Fortuna and Pipiras (2007, 2014). We avoid these difficulties by adopting a model selection approach for the inference on the number of conditional latent factors.

4.4 Large sample results

We provide regularity conditions for proving the consistency, and establishing the convergence rates, of the estimators of Sections 4.2 and 4.3 in the double asymptotics with $n, T \to \infty$. Let us denote by $\|h(Z_t)\|_{p,T} = (\mathbb{E}_T[\|h(Z_t)\|^p])^{1/p}$ the sample L^p norm, for $p > 0$, and $\|h(Z_t)\|_{\infty,T} = \sup\limits_{1 \leq t \leq T} \|h(Z_t)\|$ the sample sup norm. Moreover, $a_T \downarrow 0$ denotes a deterministic sequence such that $a_T > 0$ and $a_T = o(1)$, and $X = O_p^{-1}(1)$ means $X^{-1} = O_p(1)$ for a non-zero random variable X.

i) Factor values and risk premia estimates

We start by establishing consistency and convergence rates for the estimators of factor values and risk premia under our normalization.

Assumption 6. *(i) The estimator $\hat{\xi}_t$ is such that $\hat{\xi}_t = \xi_t + \frac{1}{\sqrt{n}} u_t$ for all t, where $\|u_t\|_{4,T} = O_p(1)$. (ii) It holds $\|f_t\|_{\infty,T} = O_p(\tau_T)$ and $\|\nu_t\|_{\infty,T} = O_p(\tau_T)$, where $\tau_T = O([\log T]^b)$ for $b > 0$. (iii) The k_t non-zero eigenvalues of matrix $V(f_t|\mathcal{F}_{t-1})$ are distinct, and $\varrho_t := \max\limits_{r=1,\ldots,k_t} \sum_{j=1, j\neq r}^{k_t} |\mu_{r,t} - \mu_{j,t}|^{-1}$, where $\mu_{j,t} = \delta_j[V(f_t|\mathcal{F}_{t-1})]$, is such that $\|\varrho_t\|_{\infty,T} = O_p(\tau_T)$.*

Assumption 7. *The machine learning regression estimator is such that: (i) $\|\widehat{\psi^X}(Z_{t-1}) - \psi^X(Z_{t-1})\|_{2,T} = O_p(a_T)$, and (ii) $\|\widehat{\psi^{X+u}}(Z_{t-1}) - \widehat{\psi^X}(Z_{t-1})\|_{2,T} = O_p(\|u_t\|_{2,T} + b_T)$, where X_t is either ξ_t, $vech(\xi_t \xi_t')$, or g_t, for any process u_t and some rates $a_T \downarrow 0$ and $b_T \downarrow 0$.*

Assumption 6 (i) requires that the cross-sectional estimator $\hat{\xi}_t$ has convergence rate $\sqrt{n}$. The bounds in Assumption 6 (ii) are implied by tail conditions on the stationary distribution of processes f_t and ν_t. Assumption 6 (iii) is a lower bound on the separation among the eigenvalues of matrix $V(f_t|\mathcal{F}_{t-1})$. This condition is used to control how the estimation error on matrix $V(\xi_{t-1}|\mathcal{F}_{t-1})$ propagates to the estimation error on its eigenvectors J_{t-1}. Assumption 7 concerns the machine learning estimator of conditional expectations. Part (i) implies the convergence rate a_T in the sample root mean square error (RMSE). Part (ii) upper bounds the effect on the regression estimator of a small "perturbation" on the regressand process, by means of the sample L^2-norm of the perturbation u_t and rate b_T. Such condition is required to control the effect of replacing ξ_t with $\hat{\xi}_t$, and g_t with $\hat{g}_t$, in the meaching learning estimator. Assumption 7 is stated in general form to cover different machine learning methodologies. In the Appendix we detail primitive regularity conditions to ensure Assumption 7 for the Lasso estimator and we give the corresponding rates a_T and b_T.

The next proposition establishes the consistency of factor values and risk premia estimates in the $\| \cdot \|_{2,T}$ norm (i.e., in RMSE) and provides upper bounds for the convergence rates. Let $\hat{\Phi}_t = \hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1}) - V(\xi_t|\mathcal{F}_{t-1})$ be the estimation error on the conditional variance.

Proposition 3. *(a) Under Assumption 1-6 we have* $\|\hat{g}_t - g_t\|_{2,T} = O_p\left(\tau_T^2 \|\hat{\Phi}_t\|_{2,T} + \frac{1}{\sqrt{n}}\right)$. *(b) Under Assumptions 1-7 we have* $\|\hat{\Phi}_t\|_{2,T} = O_p(a_T + \frac{1}{\sqrt{n}}\tau_T + b_T)$ *and*

$$\|\hat{f}_t - f_t\|_{2,T} \;=\; O_p[\tau_T^2(a_T + \frac{1}{\sqrt{n}}\tau_T + b_T)], \quad \|\hat{\lambda}_t - \lambda_t\|_{2,T} = O_p[\tau_T^2(a_T + \frac{1}{\sqrt{n}}\tau_T + b_T)].$$

The nonparametric convergence rate implied by a_T is slower than the parametric rate $\sqrt{T}$. Thus, if $T = O(n)$ as expected in our empirical application with individual stock return data, the error from cross-sectional estimation of $\hat{\xi}_t$ is asymptotically negligible compared to the machine learning estimation of conditional expectations.

ii) Selection of the number of conditional factors

Let us now turn to the consistency of estimators $\hat{k}$ and $\hat{k}_t$. Define $\hat{\Psi}_t = \hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1}) - \Gamma_{t-1}\hat{V}(g_t|\mathcal{F}_{t-1})\Gamma'_{t-1}$, that is the sum of the estimation error $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1}) - \hat{V}(\xi_t|\mathcal{F}_{t-1})$ induced by replacing ξ_t with $\hat{\xi}_t$, plus the estimation error $\hat{V}(\xi_t|\mathcal{F}_{t-1}) - \Gamma_{t-1}\hat{V}(g_t|\mathcal{F}_{t-1})\Gamma'_{t-1}$ induced by the time-variation of Γ_{t-1}. Then, we can write:

$$\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1}) = \Gamma_{t-1}\hat{V}(g_t|\mathcal{F}_{t-1})\Gamma'_{t-1} + \hat{\Psi}_t. \tag{22}$$

We use equation (22) and a perturbation theory argument in the Appendix to derive the asymptotic behavior of the eigenvalues of $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$. The first matrix on the RHS of equation (22) has reduced rank k_t and drives the behaviour of the k_t largest eigenvalues of $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$. The second term in the RHS converges to zero in sample RMSE under our assumptions, and drives the behaviour of the eigenvalues after the k_tth one.

Assumption 8. *We have (i)* $\|\hat{\Phi}_t\|_{\infty,T} = o_p(1)$, *(ii)* $\|\hat{\Psi}_t\|^2_{2,T} = o(\sigma^a)$, *and*

$$\begin{aligned} &\textit{either } (iii) \quad \frac{1}{T}\sum_{t=1}^{T} 1\{\|\hat{\Psi}_t\| \leq \epsilon\sqrt{\sigma},\ \sigma \leq \delta_{q-k}(\Pi'_{t-1}\hat{\Psi}_{t-1}\Pi_{t-1})\} = O_p^{-1}(1), \\ &\textit{or } (iv) \quad \mathbb{P}\left(\|\hat{\Psi}_t\| \leq \epsilon\sqrt{\sigma},\ \sigma \leq \delta_{q-k_t}(\Pi'_{t-1}\hat{\Psi}_t\Pi_{t-1}),\ \forall t \in \mathcal{T}\right) \to 1, \end{aligned}$$

for a subset $\mathcal{T}$ *of* $\{1, ..., T\}$, *a sequence* $\sigma \downarrow 0$, *and constants* $a > 1$ *and* $\epsilon > 0$ *small, where the columns of the* $K \times (K - k_t)$ *matrix* Π_{t-1} *are orthonormal vectors which span the orthogonal complement to the range of matrix* J_{t-1}.

For the maximum eigenvalue-ratio principle to work, we need that the eigenvalues of $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$ for orders larger than k_t are small and bounded away from zero from below by a small positive quantity (depending on T), with probability close to 1, so that the eigenvalue ratios for orders larger than k_t are not the dominating ones. The leading terms in the asymptotic expansions for these eigenvalues after the k_tth one are determined by the eigenvalues of matrix $\Pi'_{t-1}\hat{\Psi}_t\Pi_{t-1}$. This remark explains the conditions in Assumptions 8 (iii) and (iv), which apply for the time-invariant, and the time-variyng case, respectively.

Proposition 4. *(a) Let the number of conditional factors be time-invariant. Then, under Assumptions 1-7 and 8 (i)-(iii), estimator $\hat{k}$ is consistent: $\hat{k} = k$ w.p.a. 1. (b) With a possibly time-varying number of conditional factors, under Assumptions 1-7 and 8 (i), (ii), (iv) the estimator $\hat{k}_t$ is uniformly consistent over $\mathcal{T}$, i.e.* $\mathbb{P}\left(\hat{k}_t = k_t,\ \forall\ t \in \mathcal{T}\right) \to 1$.

As a plausibility remark, we notice that in Assumptions 8 (iii) and (iv) the norm of $\hat{\Psi}_t$ is required to be $O(\sqrt{\sigma})$ while the eigenvalues of the transformation $\Pi'_{t-1}\hat{\Psi}_{t-1}\Pi_{t-1}$ have to be bigger than σ, i.e. a threshold that is smaller that $\sqrt{\sigma}$ when σ tends to 0. While possible for kernel regression and Sieve estimators, deriving primitive conditions implying Assumptions 8 (iii) and (iv) for Lasso, artificial neural networks or other machine learning estimators require to develop the theory considerably beyond the currently available results.

4.5 Smoothing the estimated number of conditional factors

i) Unsupervised learning

With time-varying number of conditional factors, the estimates $\hat{k}_t$ may be noisy in samples of moderate size T as in our empirical application. We can use unsupervised learning approaches to smooth the estimated paths. To simplify let us focus on the problem of distinguishing between two regimes: $k_t = 1$ (i.e. a single conditional factor) vs. $k_t \geq 2$ (i.e. at least two conditional factors). Further let us assume that these regimes are persistent, i.e. the transitions occur infrequently compared to the monthly data sampling frequency. Let $\gamma(t)$ denote the indicator of the first regime. We build on the ideas in Horenko (2010 a, b) to regularize the regime transitions by a Total Variation (TV) bound. We solve the constrained optimization problem:

$$\max_{\{\gamma(t),\ t=1,\ldots,T\}} \sum_{t=1}^{T} \Delta\hat{\rho}_t \gamma(t) \quad \text{s.t. } 0 \leq \gamma(t) \leq 1,\ \forall t,\ \text{and } \sum_{t=2}^{T} |\gamma(t) - \gamma(t-1)| \leq C,$$

where $\Delta\hat{\rho}_t = \hat{\rho}_{1,t} - \max_{j=2,\ldots,q} \hat{\rho}_{j,t}$. Without the TV constraint, the objective function is maximized by assigning $\gamma(t) = 1$ to the dates when $\Delta\hat{\rho}_t > 0$, i.e. the largest eigenvalue ratio is the first one, and $\gamma(t) = 0$ otherwise. The constant $C > 0$ defining the TV bound can be interpreted as a the number of allowed regime changes. If we set $C = \infty$ we get $\gamma(t) = 1\{\Delta\hat{\rho}_t > 0\} = 1\{\hat{k}_t = 1\}$ i.e. the unregularized time-varying estimator, while $C = 0$ yields $\gamma(t) = 1\{\sum_{t=1}^{T} \Delta\hat{\rho}_t > 0\} = 1\{\hat{k} = 1\}$ i.e. the time-invariant estimator. [11]

ii) Supervised learning

5 Economic Identification of the Conditional Factor Space

Since the latent conditional factor space is identified only up to a conditional transformation (fixed by the normalization restrictions in Assumptions 3 and 4), the estimated factor values and risk premia do not admit a direct economic interpretation. In an unconditional setting, Giglio and Xiu (2017) estimate the risk premia of the projection of the latent factors onto observable factors. In this section we consider identifiable features of the conditional factor space, which can be interpreted in economic terms.

5.1 Conditional canonical correlations with observed factors

Let us consider a vector f_t^O of K^O observed factors, such as the vector of three Fama-French factors with $K^O = 3$. We want to measure to extent to which the spaces spanned by the latent and observed factors f_t and f_t^O coincide *conditionally* on $\mathcal{F}_{t-1}$. For this purpose we estimate the conditional canonical correlations between these vectors $\rho_{r,t}$, $r = 1, 2, \cdots, \underline{K}_t$, where $\underline{K}_t = \min\{K^O, k_t\}$. Specifically, the first conditional canonical correlation is defined by:

$$\begin{aligned} \rho_{1,t} \quad &= \max_{a_1 \in \mathbb{R}^{K^O},\ b_1 \in \mathbb{R}^{k_t}} \ Cov(a_1' f_t^O, b_1' \bar{f}_t | \mathcal{F}_{t-1}) \\ &\quad \text{s.t. } V(a_1' f_t^O | \mathcal{F}_{t-1}) = 1, \quad V(b_1' \bar{f}_t | \mathcal{F}_{t-1}) = 1. \end{aligned}$$

[11] Alternatively, we can solve the regularized log-likelihood maximization problem:

$$\max_{\{p(t),\ t=1,\ldots,T\}} \sum_{t=1}^{T} 1\{\hat{k}_t = 1\} \log p(t) + 1\{\hat{k}_t \geq 2\} \log[1 - p(t)], \quad \text{s.t. } 0 \leq p(t) \leq 1,\ \forall t,\ \text{and } \sum_{t=2}^{T} |p(t) - p(t-1)| \leq C,$$

where $p(t)$ can be interpreted as the probability to be in the first regime at time t.

It is a stochastic process with $\mathcal{F}_{t-1}$ measurable values in $[0,1]$. [12] The second, third, etc. conditional canonical correlations are defined recursively by:

$$\begin{aligned}
\rho_{r,t} \quad &= \quad \max_{a_r \in \mathbb{R}^{K^O},\ b_r \in \mathbb{R}^{k_t}} \quad Cov(a_r' f_t^O, b_r' \bar{f}_t | \mathcal{F}_{t-1}) \\
& \qquad \text{s.t.} \quad V(a_r' f_t^O | \mathcal{F}_{t-1}) = 1, \quad V(b_r' \bar{f}_t | \mathcal{F}_{t-1}) = 1, \\
& \qquad \qquad Cov(a_r' f_t^O, a_j' f_t^O | \mathcal{F}_{t-1}) = 0, \quad Cov(b_r' \bar{f}_t, b_j' \bar{f}_t | \mathcal{F}_{t-1}) = 0, \quad j = 1, \cdots, r-1,
\end{aligned}$$

for $r = 2, ..., \underline{K}_t$. By Proposition 2 the conditional canonical correlations are identifiable under Assumptions 1, 2.TV, 3.TV and 4. By analogy with the unconditional setting (see e.g. Anderson (2003)), the squared conditional canonical correlations are the $\underline{K}_t$ largest eigenvalues of matrix $\mathscr{R}_{t-1} = V(f_t^O | \mathcal{F}_{t-1})^{-1} Cov(f_t^O, \bar{f}_t | \mathcal{F}_{t-1}) V(\bar{f}_t | \mathcal{F}_{t-1})^{-1} Cov(\bar{f}_t, f_t^O | \mathcal{F}_{t-1})$. By equations (14) and (15) we have $V(\bar{f}_t | \mathcal{F}_{t-1}) = \Lambda_{t-1}$ and $Cov(f_t^O, \bar{f}_t | \mathcal{F}_{t-1}) = Cov(f_t^O, \xi_t | \mathcal{F}_{t-1}) J_{t-1}$, where Λ_{t-1} is the diagonal matrix of the k_t non-zero eigenvalues of $V(\xi_t | \mathcal{F}_{t-1})$. Thus:

$$\mathscr{R}_{t-1} = V(f_t^O | \mathcal{F}_{t-1})^{-1} Cov(f_t^O, \xi_t | \mathcal{F}_{t-1}) V(\xi_t | \mathcal{F}_{t-1})^{\dagger} Cov(\xi_t, f_t^O | \mathcal{F}_{t-1}),$$

where $V(\xi_t | \mathcal{F}_{t-1})^{\dagger} = J_{t-1} (\Lambda_{t-1})^{-1} (J_{t-1})'$ is the pseudo-inverse of matrix $V(\xi_t | \mathcal{F}_{t-1})$ based on Singular Value Decomposition (SVD).

We estimate the conditional canonical correlations at each date t by plug-in of the estimators of Section 4. Specifically, they are the square roots of the $\hat{\underline{K}}_t = \min\{\hat{k}_t, K^O\}$ largest eigenvalues of matrix

$$\hat{\mathscr{R}}_{t-1} = \hat{V}(f_t^O | Z_{t-1})^{-1} \widehat{Cov}(f_t^O, \hat{\xi}_t | Z_{t-1}) \hat{V}(\hat{\xi}_t | Z_{t-1})^{\dagger} \widehat{Cov}(\hat{\xi}_t, f_t^O | Z_{t-1}),$$

where $\hat{\xi}_t$ is the cross-sectional average in equation (16), $\hat{V}(\cdot | Z_{t-1})$ and $\widehat{Cov}(\cdot, \cdot | Z_{t-1})$ denote the estimators of the conditional variance and covariance based on machine learning methods of Section 4.1, and $\hat{V}(\hat{\xi}_t | Z_{t-1})^{\dagger} = \hat{J}_{t-1} (\hat{\Lambda}_{t-1})^{-1} (\hat{J}_{t-1})'$ and $\hat{\Lambda}_{t-1} = \text{diag}(\delta_j [\hat{V}_{t-1}(\hat{\xi}_t)], j = 1, ..., \hat{k}_t)$ as defined in Section 4.2, and $\hat{k}_t$ is the estimated number of conditional latent factors from Section 4.3.

5.2 Double Machine Learning (DML) estimation of average conditional canonical correlations

Let $\theta = (\psi_1^{\zeta}(\cdot), ..., \psi_L^{\zeta}(\cdot))'$, where $\psi_l^{\zeta}(Z_{t-1}) = E(\zeta_{tl} | Z_{t-1})$ for $l = 1, ..., L$, and the vector $\zeta_t = (f_t^{O\prime}, vech(f_t^O f_t^{O\prime})', \xi_t', vech(\xi_t \xi_t')', vec(f_t^O \xi_t')')'$ stacks the elements of vectors f_t^O and ξ_t, and their cross-products, whose conditional expectations build the variance and covariance matrices $V(f_t^O | Z_{t-1})$,

[12] The vectors a_1 and b_1 which solve the minimization problem are the first conditional canonical directions and are also $\mathcal{F}_{t-1}$ measurable. Their time dependence is not made explicit to ease notation.

$V(\xi_t|Z_{t-1})$ and $Cov(f_t^O, \xi_t|Z_{t-1})$. Then, we can write the rth conditional canonical correlation as $\rho_{r,t} = \delta_r[\mathscr{R}(\theta(Z_{t-1}))]^{1/2} \equiv \rho_r[\theta(Z_{t-1})]$, where the matrix-valued function $\mathscr{R}(\cdot)$ is such that $\mathscr{R}(\theta(Z_{t-1})) = V(f_t^O|Z_{t-1})^{-1}Cov(f_t^O, \xi_t|Z_{t-1})V(\xi_t|Z_{t-1})^{\dagger}Cov(\xi_t, f_t^O, |Z_{t-1})$. We consider the finite-dimensional parameter

$$c(\theta_0) = E(W_t\rho_{r,t}) = E[W_t\rho_r(\theta_0(Z_{t-1}))],$$

where $W_t = W(Z_{t-1})$ is a given scalar function of Z_{t-1}. This class of functionals, and simple nonlinear transformations thereof, include e.g. average conditional canonical correlations (when $W_t = 1$) and the regression coefficients of conditional canonical correlations onto functions of Z_{t-1}. It also includes fixed-bandwidth kernel regression of $\rho_{k,t}$ onto a vector $Z_{t-1}^* = f(Z_{t-1})$, with $W_t = K((Z_{t-1}^* - z^*)/h)/E[K((Z_{t-1}^* - z^*)/h)]$, for given z^* and $h > 0$.

The classical plug-in principle of semi-parametric econometrics suggests an estimator of $c(\theta_0)$ that is asymptotically normal under some assumptions (e.g., Chen and Shen (1998), Chen and White (2000)). The more recent literature on Double Machine Learning (DML, see e.g. Chernozhukov et al. (2018a, b), Chernozhukov, Newey and Robins (2018), Chernozhukov, Newey, Singh (2019)) shows that, if the estimator is modified by using a locally robust orthogonality restriction and sample splitting, then a more basic set of regularity conditions can be invoqued, which may apply in more general high-dimensional settings. More specifically, the modified orthogonality restriction for DML is:

$$E[W_t\rho_r(\theta(Z_{t-1})) - c + \alpha(Z_{t-1})'(\zeta_t - \theta(Z_{t-1}))] = 0, \tag{23}$$

with scalar parameter c and functional parameters θ, α. Here, the true value $\alpha_0(\cdot)$ of vector function $\alpha(\cdot)$ is the L^2 Rietz representer of the Gateaux derivative of functional $c(\theta)$, i.e.

$$\frac{\partial c}{\partial \theta}[\theta - \theta_0] := \lim_{\tau \to 0} \frac{c(\theta + \tau(\theta - \theta_0)) - c(\theta_0)}{\tau} = \langle \alpha_0, \theta - \theta_0 \rangle,$$

for any θ in a neighborhood of θ_0, where $\langle \theta, \vartheta \rangle = \int \theta(z)'\vartheta(z)dP_0(z)$ denotes the vector L^2 scalar product w.r.t. the true stationary distribution P_0 of Z_{t-1}. The orthogonality restriction in (23) is locally robust in the sense that, for $c = c_0$, the orthogonality restriction holds for *any* α and $\theta = \theta_0$, and for *any* θ and $\alpha = \alpha_0$ at first-order. From $\frac{\partial c}{\partial \theta}[\theta - \theta_0] = E[W_t\nabla\rho_r(\theta_0(Z_{t-1}))'(\theta(Z_{t-1}) - \theta_0(Z_{t-1}))]$ we get:

$$\alpha_0(Z_{t-1}) = W_t\nabla\rho_r(\theta_0(Z_{t-1})), \tag{24}$$

where $\nabla\rho_r(\cdot)$ is the gradient of function $\rho_r(\cdot)$ (see the Appendix for its expression).

Let I_s for $s = 1, 2$ be two subintervals yielding a splitting of the sample along the time dimension.

[13] The DML estimator of c_0 is:

$$\hat{c} = \frac{1}{T}\sum_{s=1}^{2}\sum_{t\in I_s}\Big(W_t\rho_r(\hat{\theta}_s(Z_{t-1})) + \hat{\alpha}_s(Z_{t-1})'(\hat{\zeta}_t - \hat{\theta}_s(Z_{t-1})\Big) \tag{25}$$

where $\hat{\alpha}_s(\cdot)$ and $\hat{\theta}_s(\cdot)$ are estimators of the vectors of Rietz representer and conditional expectations obtained from the sample excluding dates in I_s. It is obtained from the orthogonality restriction (23) replacing ζ_t with $\hat{\zeta}_t$, plugging-in estimates $\hat{\alpha}_s(\cdot)$ and $\hat{\theta}_s(\cdot)$ and using sample splitting. We can use $\hat{\alpha}_s = W(\cdot)\nabla\rho_r(\hat{\theta}_s(\cdot))$.

We establish asymptotic normality of estimator $\hat{c}$ under the next assumptions.

Assumption 9. *(i) Functions $\alpha_0(\cdot)$ and $\Sigma_u(\cdot) = V(u_t|Z_{t-1} = \cdot)$ are bounded. (ii) $E[W_t^2] < \infty$. (iii) We have $\int W(z)^2[\rho_r(\theta(z)) - \rho_r(\theta_0(z))]^2 dP_0(z) \leq C\|\theta - \theta_0\|^{\bar{a}}$, for $\bar{a} > 0$, and (iv) $|\int W(z)[(\rho_r(\theta(z)) - \rho_r(\theta_0(z)) - \nabla\rho_r(\theta_0(z))'(\theta(z) - \theta_0(z)))]dP_0(z)| \leq C\|\theta - \theta_0\|^2$, for θ in a neighborhoud of θ_0, and a constant $C > 0$.*

Assumption 10. *We have: (i) $\|\hat{\theta}_s - \theta_0\| = o_p(T^{-1/4})$, (ii) $\|\hat{\alpha}_s - \alpha_0\| = o_p(1)$ and (iii) $\sqrt{T}\|\hat{\theta}_s - \theta_0\|\|\hat{\alpha}_s - \alpha_0\| = o_p(1)$, for $s = 1, 2$.*

Assumption 11. *We have $E[\|\hat{\zeta}_t - \zeta_t\|^2]^{1/2} = O(1/\sqrt{n})$.*

Assumption 12. *The process $\{Y_t = (Z_t, y_{i,t}, w_{i,t}', i = 1, ..., n)'\}$ is beta-mixing, with beta-mixing coefficient*

$$\beta(j) = \sup_{n\geq 1}\sup_{t\geq 1} E\left[\sup\left\{|P(B|\mathscr{Y}_{-\infty}^{t}) - P(B)| : \ B \in \mathscr{Y}_{t+j}^{\infty}\right\}\right]$$

such that $\beta(j) \leq Cj^{-\bar{m}}$, for $\bar{m} > 3$ and $C > 0$, where $\mathscr{Y}_t^s = \sigma(Y_u : t \leq u \leq s)$.

Assumption 13. *The orthogonality vector $\psi_t = W_t\rho_r(\theta_0(Z_{t-1})) + \alpha_0(Z_{t-1})'(\zeta_t - \theta_0(Z_{t-1})) = W_t[\rho_r(\theta_0(Z_{t-1})) + \nabla\rho_r(\theta_0(Z_{t-1}))'u_t]$, where $u_t = \zeta_t - \theta_0(Z_{t-1})$, is such that $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\psi_t \Rightarrow N(0, \sigma_1^2 + \sigma_2^2)$ as $T \to \infty$, where*

$$\sigma_1^2 = \sum_{h=-\infty}^{\infty} E[W_tW_{t-h}\nabla\rho_r(\theta_0(Z_{t-1}))'u_tu_{t-h}'\nabla\rho_r(\theta_0(Z_{t-h}))], \tag{26}$$

and

$$\sigma_2^2 = \sum_{h=-\infty}^{\infty} Cov\left(W_t\rho_r(\theta_0(Z_{t-1})), W_{t-h}\rho_r(\theta_0(Z_{t-h}))\right). \tag{27}$$

Then, we have the next asymptotic normality result.

[13] We could consider sample splitting in more than two subintervals. However this leads to computing estimators on samples with non-consecutive dates, which breaks the serial dependence structure.

Proposition 5. *Under Assumptions 9-13 and regularity conditions, if $n, T \to \infty$ such that $T/n = o(1)$, we have $\sqrt{T}(\hat{c} - c_0) \Rightarrow N(0, \sigma_1^2 + \sigma_2^2)$, where σ_1^2 and σ_2^2 are given in (26) and (27).*

If the errors u_t are a conditional martingale difference sequence given Z_t, namely $E(u_t | \underline{u_{t-1}}, \underline{Z_t}) = 0$, then the asymptotic variance component σ_1^2 becomes:

$$\sigma_1^2 = E[W(Z_{t-1})^2 \nabla \rho_k(\theta_0(Z_{t-1}))' \Sigma_u(Z_{t-1}) \nabla \rho_k(\theta_0(Z_{t-1}))] \tag{28}$$

where $\Sigma_u(Z_{t-1}) = E(u_t u_t' | Z_{t-1})$.

5.3 Stochastic discount factor implied by the conditional factors

It is well-known that a (conditional) linear multi-factor model for returns corresponds to a specification of the Stochastic Discount Factor (SDF) which is (conditionally) linear in the factors. In our conditional framework, equations (3)-(5) and the factor normalization in Assumption 4 imply that the conditional expected excess returns are:

$$E(y_{i,t} | \mathcal{F}_{t-1}) = b_{i,t-1}' \lambda_t.$$

By using $b_{i,t-1} = V(f_t | \mathcal{F}_{t-1})^{-1} Cov(f_t, y_{i,t} | \mathcal{F}_{t-1})$ and rearranging terms, we get $E\left[y_{i,t}\left(1 - f_t' V(f_t | \mathcal{F}_{t-1})^{-1} \lambda_t\right) | \mathcal{F}_{t-1}\right] = 0$. When the number of factors is time-varying, the same equation holds after replacing f_t and λ_t with $\bar{f}_t$ and $\bar{\lambda}_t$. Thus, we get:

$$E\left(m_{t-1,t} y_{i,t} | \mathcal{F}_{t-1}\right) = 0, \tag{29}$$

where the SDF between dates $t-1$ and t is given by:

$$m_{t-1,t} = R_{f,t}^{-1}\left(1 - \bar{f}_t' V(\bar{f}_t | \mathcal{F}_{t-1})^{-1} \bar{\lambda}_t\right), \tag{30}$$

and $R_{f,t} = E\left(m_{t-1,t} | \mathcal{F}_{t-1}\right)^{-1}$ is the gross return of the conditionally risk-free asset. [14] Note that, while the vectors of factor values and risk premia are identifiable only up to conditional transformations, the quantity defining the SDF is invariant to such conditional transformations.

In an asset pricing model defined by a SDF, the cross-section of conditional expected excess returns is spanned by the assets conditional betas with respect to a single factor (see e.g. Singleton (2006), Chapter 11, and references therein). In fact, equation (29) can be rewritten as:

$$E\left(y_{i,t} | \mathcal{F}_{t-1}\right) = \beta_{i,t-1}^* \lambda_t^*, \tag{31}$$

[14] The SDF in (30) does not necessarily meet the positivity condition.

where $\beta^*_{i,t-1} = V\left(r^*_t|\mathcal{F}_{t-1}\right)^{-1} Cov\left(r^*_t, y_{i,t}|\mathcal{F}_{t-1}\right)$,

$$r^*_t = \frac{m_{t-1,t}}{E\left(m^2_{t-1,t}|\mathcal{F}_{t-1}\right)}, \tag{32}$$

and $\lambda^*_t = -R_{f,t}V\left(m_{t-1,t}|\mathcal{F}_{t-1}\right)/E\left(m^2_{t-1,t}|\mathcal{F}_{t-1}\right)$. The single factor r^*_t is the gross return of an asset with payoff equal to the SDF $m_{t-1,t}$. Equivalently, r^*_t is the minimum conditional second-moment gross return. If this asset is tradable, $\lambda^*_t = E\left(r^*_t|\mathcal{F}_{t-1}\right) - R_{f,t}$ equals its expected excess return.

Our methodology yields the non-parametric identification for the SDF and the minimum conditional second-moment return provided by equations (30) and (32). It only relies on a linear conditional linear factor structure for a large cross-section of assets. It is very general for what concerns the observability of the factors by the econometrician, their possibly time-varying number, and the dynamics of betas.

Finally, two remarks are in order. First, the single-factor structure in expected excess returns does not imply that the return conditional factor space itself is one-dimensional. In fact, the implied error terms $y_{i,t} - \beta^*_{i,t-1}r^*_t$ are not necessarily weakly correlated across assets. Second, we stress that we have deployed the conditional linear factor structure to derive the no-arbitrage restrictions by the results in GOS. Otherwise, in our conditional economy with an infinity (continuum) of assets, the derivation of a SDF without assuming a factor structure for returns is unknown to us.

6 Empirical Analysis

We conduct our empirical analysis on U.S. equities. The dependent variable is excess return over risk-free rate on each individual stock in our sample. We proxy the risk-free rate with the monthly 30-day T-bill beginning-of-month yield. To cope with an overidentified case we choose 22 firm-level characteristics listed below as our instruments.

6.1 Data Description

We get monthly stock returns from CRSP and quarterly firm characteristics from COMPUSTAT. The sample begins in January 1971 and ends in December 2017, which consists of 564 months.

For each firm we select 15 characteristics as our instrumental variables $w_{i,t}$ from Freyberger, Neuhierl, and Weber (2017). The FNW characteristics are grouped into four categories: (i) investment-related characteristics, which are change in total assets (investment), change in book equity (Δceq), change in shares outstanding (Δshrout), (ii) profitability-related characteristics, which are earnings per share (eps), gross profit over book equity (prof), return on asset (roa), return on equity (roe), (iii) value-related characteristics, which are book to market ratio (beme), cash flow to total liabilities (c2a), sales growth (sales_g), market capitalization (me), and (iv) past return based variables, which are lagged 1 month return (r_{2-1}), return from 6 to 2 months before current period (r_{6-1}), return from 12 to 2 months before current period (r_{12-2}) and return from 12 to 7 months before current period (r_{12-7}). For a detailed description of the 15 variables, see Freyberger, Neuhierl, and Weber (2017).

Next, we obtain the estimates of ξ_t by cross-sectionally averaging the products of characteristics $w_{i,t-1}$ and excess return $y_{i,t}$ of the same stock between consecutive dates as in (16). Notice that if we rescale $w_{i,t-1}$ to have $\sum_{i=1}^{n} w_{i,t-1} = 1$, then ξ_t can be simply seen as a portfolio return. One may recall that any tradable return factors, such as SMB or HML, are essentially returns of portfolios in which different weights for stocks are assigned according to some firm-level characteristics. Therefore, our ξ_t can be interpreted as a vector of portfolio returns based on some functions of certain stock characteristics. As a result, any existing tradable return factors such as Fama-French factors can be regarded as a candidate for our ξ_t using some function of firm-level characteristics as weights. Thus, we include 7 well known return factors as additional components of vector $\hat{\xi}_t$. The 7 factors are the Fama-French 5 factors MKT, SMB, HML, RMW, CMA, together with Momentum (MOM) and Betting Against Beta (BAB).

Our Markov process Z_t, which generates the information set $\mathcal{F}_t$ of aggregate shocks, consists of 18 different monthly variables. We group those state variables into 4 categories: (i) financial indicators as proposed in Goyal and Welch (2008), which are dividend-price ratio (DP), earnings-price ratio (EP), book-to-market ratio (BM), net equity expansion (NTIS), Treasury-bill rate (TBL), term spread (TMS), default spread (DFY) and stock variance (SVAR), (ii) risk factors, which are the 5 factors in Fama-French 5-factor model (ExMKT, SMB, HML, RMW, CMA), BAB and MOM, (iii) macroeconomic variables, which are aggregate consumption growth (CONSr), nominal inflation rate (CPIr) and unemployment rate (UNEMP), and (iv) past values of our cross-sectional weighted returns ξ_{t-1}.

6.2 Data preparation and parameter settings

We only retain the stock×month observations for which all 22 characteristics are non-missing and convert all quarterly firm characteristics into monthly basis simply by assuming a constant value across the quarter. For each instrumental variable, we standardize its values cross-sectionally at each time period, in order to mitigate influence from outliers as well as to make variables more comparable. Specifically, at each month we calculate every individual stock's ranks for each of the 22 characteristics, divide the ranks by the number of observations, multiply by 2 and subtract 1. The values of the new characteristics after remapping are in the $[-1, 1]$ interval. Remember that the definition of $\hat{\xi}_t$ is equivalent to building 22 portfolios using the values of the 22 remapped characteristics at month $t-1$ as stocks' weights in each portfolio. We calculate the value-weighted portfolios by multiplying the market capitalization weight of each stock.

We compute the cross-sectional average $\hat{\xi}_t$ month by month. The sample start is chosen such that the estimate $\hat{\xi}_t$ is obtained by averaging returns times characteristics over more than 2000 stocks at each date, which should imply that $\hat{\xi}_t$ is close to the cross-sectional limit ξ_t as in Assumption 1. Figure 5 shows the evolution of the cross-sectional sample size over time. The number of observations steadily increases from about 2000 to close to 9000 from 1971 to 1998. After the internet bubble, there is a decrease but the sample size still keeps above 5000. Moreover, we standardize each element in $\hat{\xi}_t$ across time to make sure that all elements are of the same order of magnitude.

For estimating conditional expectations $E[\zeta_t|Z_{t-1}]$, we can adopt the machine learning methods discussed in Subsection 4.1. In this empirical application, we choose to use neural networks to estimate the conditional expectations element by element. Specifically, we use a feed-forward network with single hidden layer, of which the number of neurons is equal to the number of variables in our common information vector Z_t. Because the time-series sample size and the dimension of conditioning variables are not very large (compared e.g. with image recognition problem), we will not deploy the widely-used *stochastic gradient descent* algorithm which optimizes on random sub-samples for the purpose of efficiency. Instead, we are able to efficiently optimize our objective function over the whole sample.

To improve generalization and mitigate overfitting of our neural networks, we adopt two methods: the Bayesian Regularization training function, in conjunction with averaging across multiple parallel neural networks.

We take the *Bayesian Regularization backpropagation* from Matlab as our optimization algorithm. Bayesian regularization backpropagation is a network training function that updates the weight and bias values according to Levenberg-Marquardt optimization algorithm. It minimizes a combination of squared errors and weights, and then determines the correct combination so as to produce a network that generalizes well. The biggest advantage of this algorithm is that it determines the optimal regularization parameters in an automated fashion.

The network averaging method is adopted because the mean squared error for the average output is likely to be lower than most of the individual neural network performances. In our case, for each estimate we train 48 parallel networks at the same time using different initial values and then take the average as the final output.

6.3 Inference on the number of conditional factors

According to Subsection 4.3, the conditional factor space dimension k_t equals the rank of the conditional variance-covariance matrix $V(\xi_t|Z_{t-1})$, which can be estimated by:

$$\hat{V}(\hat{\xi}_t|Z_{t-1}) = \hat{E}(\hat{\xi}_t\hat{\xi}_t{}'|Z_{t-1}) - \hat{E}(\hat{\xi}_t|Z_{t-1})\hat{E}(\hat{\xi}_t|Z_{t-1})'. \tag{33}$$

The estimated dimension $\hat{k}_t$ is obtained by maximizing the eigenvalue-ratio criterion introduced in Subsection 4.3. Moreover, here we also propose several other eigenvalue-based ratios as alternative estimation methods for k_t. By carrying out the eigen-decomposition of the estimated conditional variance $\hat{V}(\hat{\xi}_t|Z_{t-1})$, we are able to infer the conditional factor space dimension.

Figure 6 displays four time series of eigenvalue ratios, averaged by quarters. For most quarters, the average ratio of the first to the second eigenvalues is the largest one. This shows the presence of a dominating first factor in the conditional factor space. Figure 7 reports the time series of incremental explanatory power ratios, averaged by quarter. These quantities are defined as the ratios of the rth largest eigenvalue to the sum of the first r eigenvalues, for any integer r. Their interpretation is supported by the decomposition of the conditional variability of ξ_t as $Tr[V(\xi_t|\mathcal{F}_{t-1})] = \sum_{r=1}^{K} \delta_r[V(\xi_t|\mathcal{F}_{t-1})] = \sum_{r=1}^{k_t} V(f_{r,t}|\mathcal{F}_{t-1})$. For ranks r larger than 4, these incremental explanatory power ratios are smaller than 10% for almost all factors. In Figure 8 we display the time series of accumulative explanatory power ratios, averaged by quarter. The explanatory power ratio for rank r is defined as the ratio of the sum of the first r eigenvalues to the sum of the first $k_{\max}$ eigenvalues. We set $k_{\max} = 4$ following Figure 7, which suggests that the contributions of the remaining factors is small. The first factor has an explanatory power around or above 50%. The accumulative explanatory power of the first three factors is around 90%. The yellow vertical bars denote periods in which there is an increase of the accumulative explanatory power ratios, e.g. the one of the first factor reaches $60-70\%$, and the one of the first three factors is above 90%. Figure 9 reports the time series of accumulative explanatory power ratios computed using squared eigenvalues. This reflects the contribution of conditional covariation in addition to conditional variation, similarly as in Fiorentini and Sentana (2015). These ratios are larger than those in Figure 8, highlighting the importance of conditional factors in explaining conditional covariation.

6.4 Estimation of the conditional factor space

Once we estimated the number of conditional factors, we are able to adopt the method in Subsection 4.2 to estimate the values of the latent conditional factors. We start by assuming that the number of conditional factors is time-invariant. We choose the number of factors to be 1, given the dominance of the first eigenvalue ratio in Figure 6.

To conduct an economically-meaningful analysis, we estimate the conditional canonical correlation between our latent factor and some observable factors as well as state variables following steps in Subsection 5.1. Next, we calculate the in-sample averages of those conditional correlations, and then rank them in a descending order and display them in Figure 10. Not surprisingly, the first conditional factor can be mostly explained by the excess market return, with average conditional correlation close to 70%. The next most powerful variables are HML and CMA, with average conditional correlation around 60%. The results are not surprising as well, since the two factors are highly correlated in the first place according to Fama and French (2015).

Based on the analysis above, the market factor explains most of the factor space of our first conditional latent factor. For the second latent factor, we follow the same procedures and provide the bar charts in Figure 11. A value around 45% for the conditional correlation indicates that SMB is likely to be the most important factor in spanning our second conditional factor space.

Together with the previous results for the first conditional factor, we may conclude that the factor space of our two latent factors are mostly driven by two observable static factors: the market factor and the SMB factor.

TO BE CONTINUED...

7 References

Ait-Sahalia, Y., and Xiu, D. (2017). Using Principal Component Analysis to Estimate a High Dimensional Factor Model with High-Frequency Data. *Journal of Econometrics*, 201, 384-399.

Al-Najjar, N. (1995). Decomposition and Characterization of Risk With a Continuum of Random Variables. *Econometrica*, 63 (5), 1195-1224.

Al-Najjar, N. (1998). Factor Analysis and Arbitrage Pricing in Large Asset Economies. *Journal of Economic Theory*, 78 (2), 231-262.

Al-Najjar, N. (1999). Decomposition and Characterization of Risk With a Continuum of Random Variables: Corrigendum. *Econometrica*, 67 (4), 919-920.

Ahn, S. C., and Horenstein, A. R. (2013). Eigenvalue Ratio Test for the Number of Factors. *Econometrica*, 81(3), 1203-1227.

Al-Sadoon, M. M. (2017): A Unifying Theory of Tests of Rank, *Journal of Econometrics*, 199, 49-62.

Andreou, E., Gagliardini, P., Ghysels, E., and Rubin. M., (2019): Inference in Group Factor Models, with an Application to Mixed Frequency Data, forthcoming in *Econometrica*.

Andrews, D. W. (2005). Cross-section Regression with Common Shocks. *Econometrica*, 73(5), 1551-1585.

Bai, J., and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1), 191-221.

Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71, 135-171.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica*, 80(6), 2369-2429.

Bernstein, D. S. (2009). Matrix Mathematics: Theory, Facts, and Formulas. *Princeton University Press.*

Chamberlain, G., and Rothschild, M. (1983). Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica*, 51 (5), 1281-1304.

Chen, L., Pelger, M., & Zhu, J. (2019). Deep learning in asset pricing. *Working Paper*

Chen, X. (2007). Large Sample Sieve Estimation of Semi-Nonparametric Models. In *Handbook of Econometrics*, Volume 6, Chapter 76, 5549-5632.

Cochrane, J. H. (1996). A Cross-sectional Test of an Investment-based Asset Pricing Model. *Journal of Political Economy*, 104(3), 572-621.

Cochrane, J. H. (2011). Presidential Address: Discount Rates. *Journal of Finance*, 66(4), 1047-1108.

Connor, G., Hagmann, M. and Linton, O. (2012): Efficient Semiparametric Estimation of the Fama-French Model and Extensions, *Econometrica*, 80, 713-754.

Connor, G., and Korajczyk, R. A. (1989). An Intertemporal Equilibrium Beta Pricing Model. *The Review of Financial Studies*, 2(3), 373-392.

Cragg, J. G. and Donald, S.G. (1996). On the Asymptotic Properties of LDU-Based Tests of the Rank of a Matrix, *Journal of the American Statistical Association*, 91, 1301-1309.

Cragg, J. G. and Donald, S.G. (1997). Inferring the Rank of a Matrix, *Journal of Econometrics*, 76, 223-250.

Cybenko, G. (1989). Approximation by Superposition of a Sigmoidal Function. *Mathematics of Control, Signals and Systems*, 2, 303-314.

Donald, S. G., Fortuna, N. and Pipiras, V. (2007): On Rank Estimation in Symmetric Matrices: The Case of Indefinite Matrix Estimators, *Econometric Theory*, 23, 1103-1123.

Donald, S. G., Fortuna, N. and Pipiras, V. (2014): On Estimating the Rank of Semidefinite Matrix, Working Paper.

Fama, E. F., and French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1), 3-56.

Fan, J., Liao, Y., and Wang, W. (2016): Projected Principal Component Analysis in Factor Models, *Annals of Statistics*, 44, 219-254.

Feng, G., Giglio, S., and Xiu, D. (2017). Taming the Factor Zoo: A Test of New Factors. Forthcoming in *Journal of Finance.*

Ferson, W. E., and Harvey, C. R. (1991). The Variation of Economic Risk Premiums. *Journal of Political Economy*, 99(2), 385-415.

Ferson, W. E., and Harvey, C. R. (1999). Conditioning Variables and the Cross Section of Stock Returns. *The Journal of Finance*, 54(4), 1325-1360.

Ferson, W. E., and Schadt, R. W. (1996). Measuring Fund Strategy and Performance in Changing Economic Conditions. *The Journal of Finance*, 51(2), 425-461.

Fiorentini, G., and Sentana, E. (2015): Tests for serial dependence in static, non-Gaussian factor models. In: Koopman, S.J., Shephard, N. (Eds.), Unobserved Components and Time Series Econometrics. Oxford University Press, Oxford.

Freyberger, J., Neuhierl, A., and Weber, M. (2017). Dissecting Characteristics Nonparametrically, Working Paper. *National Bureau of Economic Research.*

Gagliardini, P., and Gourieroux, C. (2017). Double Instrumental Variable Estimation of Interaction Models with Big Data. *Journal of Econometrics*, 201(2), 176-197.

Gagliardini, P., Ossola, E., and Scaillet, O. (2016). Time-Varying Risk Premium in Large Cross-Sectional Equity Data Sets. *Econometrica*, 84(3), 985-1046.

Gagliardini, P., Ossola, E., and Scaillet, O. (2019). A Diagnostic Criterion for Approximate Factor Structure. *Journal of Econometrics, forthcoming.*

Gallant, A. R., and White, H. (1988). There Exists a Neural Network that Does not Make Avoidable Mistakes. In *IEEE Second International Conference on Neural Networks*, I, 657-664.

Giglio, S., and Xiu, D. (2017). Asset Pricing with Omitted Factors. Working Paper.

Goyal, A. and Welch, I. (2007). A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21(4), 1455-1508.

Gu, S., Kelly, B. T., and Xiu, D. (2018). Empirical Asset Pricing via Machine Learning. Working Paper.

Gu, S., Kelly, B. T., and Xiu, D. (2019). Autoencoder Asset Pricing Models. Working Paper.

Haerdle, W., and Linton, O. (1994). Applied Nonparametric Methods. In *Handbook of Econometrics*, R. Engle and D. McFadden eds., Volume 4, Chapter 38, 2295-2339.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2, 359-366.

Hornik, K. (1991). Approximation Capabilities of Multilayer Feed-Forward Networks. *Neural Networks*, 4(2), 251-257.

Kelly, B., Pruitt, S., and Su, Y. (2017): Instrumented Principal Component Analysis. Working Paper.

Kelly, B., Pruitt, S., and Su, Y. (2019): Characteristics are Covariances: A Unified Model of Risk and Return, *Journal of Financial Economics, forthcoming.*

Kleibergen, F., and Paap, R. (2006): Generalized Reduced Rank Tests Using the Singular Value Decomposition, *Journal of Econometrics*, 133, 97-126.

Kuersteiner, G. M., and Prucha, I. R. (2013). Limit Theory for Panel Data Models with Cross Sectional Dependence and Sequential Exogeneity. *Journal of Econometrics*, 174(2), 107-126.

Lettau, M., and Ludvigson, S. (2001). Consumption, Aggregate Wealth, and Expected Stock Returns. *The Journal of Finance*, 56(3), 815-849.

Li, J., Todorov, V., and Tauchen, V. (2019). Jump Factor Models in Large Cross-Sections. *Quantitative Economics*, 10, 419-456.

Liao, Y., and Yang, X. (2017). Uniform Inference for Characteristics Effect of Large Continuous-Time Linear Models. Working Paper.

Onatski, A. (2009). Testing Hypotheses about the Number of Factors in Large Factor Models. *Econometrica*, 77(5), 1447-1479.

Pelger, M. (2019). Large-Dimensional Factor Modeling Based on High-Frequency Observations. *Journal of Econometrics*, 4, 23-42.

Pelger, M. (2019). Understanding Systematic Risk: A High-Frequency Approach. *Journal of Finance, forthcoming.*

Pelger, M., and Xiong, R. (2018): State-Varying Factor Models of Large Dimensions, Working Paper.

Petkova, R., and Zhang, L. (2005). Is Value Riskier than Growth?. *Journal of Financial Economics*, 78(1), 187-202.

Robin, J. M., and Smith, R. (2000): Tests of Rank, *Econometric Theory*, 16, 151-175.

Shanken, J. (1990). Intertemporal Asset Pricing: An Empirical Investigation. *Journal of Econometrics*, 45(1-2), 99-120.

Singleton, K. (2006). Empirical Dynamic Asset Pricing. Model Specification and Econometric Assessment. Princeton University Press.

Stock, J. H., and Watson, M. W. (2002). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460), 1167-1179.

Tao, T. (2012). Topics in Random Matrix Theory Graduate Studies in Mathematics. *American Mathematical Society.*

Zaffaroni, P. (2019) Factor Models for Asset Pricing. Working Paper.

APPENDIX

A Proof of Proposition 3

(a) With identity weighting matrix we have $\hat{\bar{g}}_t = \hat{J}_t' \hat{\xi}_t$. First, let us show the RMSE convergence of $\hat{J}_t$ to J_t. We use a result in perturbation theory providing the approximation for the eigenvectors of a matrix with explicit characterization of the Lipshitz constant (see Proposition 7 in Carlini and Gagliardini (2018), which is an extension of Theorem 3 in Izenman (1975)).

Lemma 1. *Let A be a symmetric $n \times n$ matrix of rank $k \leq n$, with distinct non-zero eigenvalues $\mu_1 > \mu_2 > ... > \mu_k > 0$, and associated standardized eigenvectors v_1, v_2, ... v_k (hence, the null eigenvalue $\mu_0 = 0$ has eigenspace of dimension $n - k$). Let $\hat{A}$ be a symmetric $n \times n$ matrix (a "perturbation" of A), and let $\hat{\mu}_1$, ... , $\hat{\mu}_k$ be its k largest eigenvalues and $\hat{v}_1$, ... , v_k the associated standardized eigenvectors. Then:*

$$\|\hat{v}_j - v_j\| \leq c\rho \|\hat{A} - A\|$$

for $j = 1, ..., k$, where $\rho := \max_{j:j=1,...,k} \sum_{l=0, l\neq j}^{k} |\mu_j - \mu_l|^{-1}$ and c is a universal constant (that can be chosen equal to $c = 6 + 5\sqrt{2}$).

Write $J_t = [J_{1,t} : \cdots : J_{k_t,t}]$ and $\hat{J}_t = [\hat{J}_{1,t} : \cdots : \hat{J}_{k_t,t}]$ the matrices of standardized eigenvectors to the first k_t eigenvalues of matrices $V(\xi_t|\mathcal{F}_{t-1})$ and $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$. Recall that matrix $V(\xi_t|\mathcal{F}_{t-1})$ has rank k_t, under Assumption 4.TV its k_t non-zero eigenvalues are the non-zero diagonal elements of diagonal matrix $V(f_t|\mathcal{F}_{t-1})$, and the latter are distinct $\mu_{1,t} > \mu_{2,t} > ... > \mu_{k_t,t} > 0$ under Assumption 6 (iii). Moreover, $\sum_{l=0,l\neq j}^{k_k} |\mu_{j,t} - \mu_{l,t}|^{-1} \leq \varrho_t$, for all $j = 1, ..., k_t$ and all $t = 1, ..., T$ (where $\mu_{0,t} \equiv 0$). Then, from Lemma 1 we have:

$$\|\hat{J}_{j,t} - J_{j,t}\| \leq c\varrho_t \|\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1}) - V(\xi_t|\mathcal{F}_{t-1})\| = c\varrho_t \|\hat{\Phi}_t\|,$$

for all $j = 1, ..., k_t$ and all $t = 1, ..., T$, where constant c is independent of j and t. Thus, from Assumption 6 (iii) we get:

$$\|\hat{J}_t - J_t\|_{2,T} = O_p(\tau_T \|\hat{\Phi}_t\|_{2,T}). \tag{A.1}$$

Let us now show the RMSE convergence of $\hat{\bar{g}}_t$. We use $\hat{\xi}_t = \hat{J}_t \bar{g}_t - (\hat{J}_t - J_t)\bar{g}_t + \frac{1}{\sqrt{n}} u_t$, which implies

$$\hat{\bar{g}}_t = \bar{g}_t - \hat{J}_t'(\hat{J}_t - J_t)\bar{g}_t + \frac{1}{\sqrt{n}} \hat{J}_t' u_t. \tag{A.2}$$

We use $\|\hat{J}_t\| = 1$. Then, from Assumption 6 (ii) we get

$$\|\hat{\bar{g}}_t - \bar{g}_t\| \leq C_1 \tau_T \|\hat{J}_t - J_t\| + \frac{1}{\sqrt{n}} \|u_t\|, \tag{A.3}$$

w.p.a. 1, for constant C_1. Then, from bound (A.1) and Assumption 6 we get:

$$\|\hat{g}_t - g_t\|_{2,T} = O_p\left(\tau_T^2 \|\hat{\Phi}_t\|_{2,T} + \frac{1}{\sqrt{n}}\right). \tag{A.4}$$

(b) First, let us show the convergence of $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$ to $V(\xi_t|\mathcal{F}_{t-1})$ in RMSE. We use $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1}) = \widehat{\psi^{\hat{\xi}\hat{\xi}'}}(Z_{t-1}) - \widehat{\psi^{\hat{\xi}}}(Z_{t-1})\widehat{\psi^{\hat{\xi}}}(Z_{t-1})'$, $\hat{V}(\xi_t|\mathcal{F}_{t-1}) = \widehat{\psi^{\xi\xi'}}(Z_{t-1}) - \widehat{\psi^{\xi}}(Z_{t-1})\widehat{\psi^{\xi}}(Z_{t-1})'$ and $V(\xi_t|\mathcal{F}_{t-1}) = \psi^{\xi\xi'}(Z_{t-1}) - \psi^{\xi}(Z_{t-1})\psi^{\xi}(Z_{t-1})'$, and the equations $\hat{\xi}_t = \xi_t + \frac{1}{\sqrt{n}}u_t$ and $\hat{\xi}_t\hat{\xi}_t' = \xi_t\xi_t' + \frac{1}{\sqrt{n}}(u_t\xi_t' + \xi_t u_t') + \frac{1}{n}u_t u_t'$ for all t. From Assumption 6 (ii) process ξ_t is $O_p(\tau_T)$ uniformly in t. Then, from Assumptions 7 (ii) and 6 (i)-(ii) we have $\|\widehat{\psi^{\hat{\xi}}}(Z_{t-1}) - \widehat{\psi^{\xi}}(Z_{t-1})\|_{2,T} = O_p(\frac{1}{\sqrt{n}} + b_T)$ and $\|\widehat{\psi^{\hat{\xi}\hat{\xi}'}}(Z_{t-1}) - \widehat{\psi^{\xi\xi'}}(Z_{t-1})\|_{2,T} = O_p(\frac{1}{\sqrt{n}}\tau_T + b_T)$. Therefore $\|\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1}) - \hat{V}(\xi_t|\mathcal{F}_{t-1})\|_{2,T} = O_p(\frac{1}{\sqrt{n}}\tau_T + b_T)$. Moreover, from Assumption 7 (i) we have $\|\hat{V}(\xi_t|\mathcal{F}_{t-1}) - V(\xi_t|\mathcal{F}_{t-1})\|_{2,T} = O_p(a_T)$. Thus, we get:

$$\|\hat{\Phi}_t\|_{2,T} = \|\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1}) - V(\xi_t|\mathcal{F}_{t-1})\|_{2,T} = O_p(a_T + \frac{1}{\sqrt{n}}\tau_T + b_T). \tag{A.5}$$

From (A.4) this implies

$$\|\hat{g}_t - g_t\|_{2,T} = O_p[\tau_T^2(a_T + \frac{1}{\sqrt{n}}\tau_T + b_T)]. \tag{A.6}$$

Finally, we can prove the RMSE convergence of $\hat{\lambda}_t$ and $\hat{f}_t$. We have $\hat{\lambda}_t = \widehat{\psi^{\hat{g}}}(Z_{t-1})$ and $\lambda_t = \psi^{g}(Z_{t-1})$. From Assumption 7 (ii) and bound (A.6) we have $\|\widehat{\psi^{\hat{g}}}(Z_{t-1}) - \widehat{\psi^{g}}(Z_{t-1})\|_{2,T} = O_p[\tau_T^2(a_T + \frac{1}{\sqrt{n}}\tau_T + b_T)]$, and from Assumption 7 (i) we have $\|\widehat{\psi^{g}}(Z_{t-1}) - \psi^{g}(Z_{t-1})\|_{2,T} = O_p(a_T)$. Hence, $\|\hat{\lambda}_t - \lambda_t\|_{2,T} = O_p[\tau_T^2(a_T + \frac{1}{\sqrt{n}}\tau_T + b_T)]$. From the latter bound and (A.6), and using $\hat{f}_t = \hat{g}_t - \hat{\lambda}_t$, we get $\|\hat{f}_t - f_t\|_{2,T} = O_p[\tau_T^2(a_T + \frac{1}{\sqrt{n}}\tau_T + b_T)]$.

B Proof of Proposition 4

In this appendix we prove the consistency of the rank test on the number of conditional factors based on the eigenvalue-ratio principle.

B.1 Perturbation theory for the eigenvalues of matrix $\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$

Let $\hat{V}_t := \hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})$ denote an estimator of matrix $V_t := V(\xi_t|\mathcal{F}_{t-1}) = \Gamma_t V(g_t|\mathcal{F}_{t-1})\Gamma_t'$. Under Assumption 8 (i), we have an asymptotic expansion of the form:

$$\hat{V}_t = \tilde{V}_t + \hat{\Psi}_t, \tag{B.1}$$

where $\tilde{V}_t := \Gamma_t \hat{V}(g_t|\mathcal{F}_{t-1})\Gamma_t'$ and $\hat{V}(g_t|\mathcal{F}_{t-1})$ is a (generally infeasible) estimator of $V(g_t|\mathcal{F}_{t-1})$, the estimation error term $\hat{\Psi}_t$ is such that $\hat{\Psi}_t = O_p(R)$, and $R = R_{n,T}$ is a positive rate tending to zero as $n, T \to \infty$.

We derive an asymptotic expansion for the $K - k_t$ smallest eigenvalues of matrix $\hat{V}_t$ using equation (B.1) and perturbation theory. Recall that under Assumptions 2.TV and 4.TV, we have $\Gamma_t = [J_t : 0_{K \times (k-k_t)}]$, where the columns of matrix J_t are the normalized eigenvectors of V_t associated to the k_t non-zero eigenvalues. The $K - k_t$ smallest eigenvalues of both V_t and $\tilde{V}_t$ are equal to zero. The corresponding eigenspace is the orthogonal complement of the column space of matrix J_t. Let Π_t denote a $K \times (K - k_t)$ full column rank matrix, whose columns span this eigenspace and are normalized to have length 1. Then we have $\Pi_t'\Gamma_t = 0$ and $\Pi_t'\Pi_t = I_{K-k_t}$. Let $\hat{W}_t$ denote the $K \times (K - k_t)$ matrix whose columns are the orthonormalized eigenvectors of $\hat{V}_t$ associated to the $K - k_t$ smallest eigenvalues, and let $\hat{\Lambda}_t$ be the diagonal matrix of these eigenvalues. They solve the eigenvalue-eigenvector equation:

$$\hat{V}_t \hat{W}_t = \hat{W}_t \hat{\Lambda}_t. \tag{B.2}$$

Since the columns of the orthogonal matrix $[J_t \ : \ \Pi_t]$ span $\mathbb{R}^K$, we can write

$$\hat{W}_t = (\Pi_t + J_t \hat{\alpha}_t)\mathcal{U}_t, \tag{B.3}$$

for some matrices $\hat{\alpha}_t$ and $\mathcal{U}_t$, where matrix $\mathcal{U}_t$ is non-singular. By perturbation theory, matrices $\hat{\alpha}_t$ and $\hat{\Lambda}_t$ converge to zero as $n, T \to \infty$ (see below).

The asymptotic expansions of the eigenvalues of matrix $\hat{V}_t$ are provided in the next lemma.

Lemma 2. *We have:*

$$|\delta_j(\hat{V}_t) - \delta_j(V_t)| \leq \|\hat{\Psi}_t\| + \|\hat{V}(g_t|\mathcal{F}_{t-1}) - V(g_t|\mathcal{F}_{t-1})\|, \tag{B.4}$$

for $j = 1, ..., k_t$, and:

$$|\delta_j(\hat{V}_t) - \delta_{j-k_t}(\Pi_t'\hat{\Psi}_t\Pi_t)| \leq C\|\hat{\Psi}_t\|^2, \tag{B.5}$$

for $j = k_t + 1, ..., K$, where C is a universal constant.

Proof of Lemma 2: We plug (B.1) and (B.3) into the eigenvalue-eigenvector equation (B.2) to get:

$$\tilde{V}_t J_t \hat{\alpha}_t \mathcal{U}_t + \hat{\Psi}_t \Pi_t \mathcal{U}_t + \hat{\Psi}_t J_t \hat{\alpha}_t \mathcal{U}_t = \Pi_t \mathcal{U}_t \hat{\Lambda}_t + J_t \hat{\alpha}_t \mathcal{U}_t \hat{\Lambda}_t.$$

Pre-multiplying this equation by Π_t', and by J_t', we get:

$$\Pi_t'\hat{\Psi}_t\Pi_t\mathcal{U}_t + \Pi_t'\hat{\Psi}_t J_t \hat{\alpha}_t \mathcal{U}_t = \mathcal{U}_t\hat{\Lambda}_t, \tag{B.6}$$

and:

$$J_t'\tilde{V}_t J_t \hat{\alpha}_t \mathcal{U}_t + J_t'\hat{\Psi}_t \Pi_t \mathcal{U}_t + J_t'\hat{\Psi}_t J_t \hat{\alpha}_t \mathcal{U}_t = \hat{\alpha}_t \mathcal{U}_t \hat{\Lambda}_t, \tag{B.7}$$

respectively, using $\Pi_t' J_t = 0$, $\Pi_t'\Pi_t = I_{K-k_t}$ and $J_t' J_t = I_{k_t}$. From equation (B.6) we get:

$$\hat{\Lambda}_t = \mathcal{U}_t^{-1}\left(\Pi_t'\hat{\Psi}_t\Pi_t + \Pi_t'\hat{\Psi}_t J_t\hat{\alpha}_t\right)\mathcal{U}_t. \tag{B.8}$$

By plugging (B.8) into (B.7), and multiplying times $\mathcal{U}_t^{-1}$ from the right, we get:

$$J_t'\tilde{V}_t J_t \hat{\alpha}_t + J_t'\hat{\Psi}_t \Pi_t + J_t'\hat{\Psi}_t J_t \hat{\alpha}_t = \hat{\alpha}_t\left(\Pi_t'\hat{\Psi}_t\Pi_t + \Pi_t'\hat{\Psi}_t J_t\hat{\alpha}_t\right).$$

By using $J_t'\tilde{V}_t J_t = \hat{V}(\bar{g}_t|\mathcal{F}_{t-1})$, this yields:

$$\hat{\alpha}_t = \hat{V}(\bar{g}_t|\mathcal{F}_{t-1})^{-1}\left\{-J_t'\hat{\Psi}_t\Pi_t - J_t'\hat{\Psi}_t J_t\hat{\alpha}_t + \hat{\alpha}_t\left[\Pi_t'\hat{\Psi}_t\Pi_t + \Pi_t'\hat{\Psi}_t J_t\hat{\alpha}_t\right]\right\}. \tag{B.9}$$

We use this equation to upper bound the norm of matrix $\hat{\alpha}_t$. For this purpose, we deploy $\|\hat{V}(\bar{g}_t|\mathcal{F}_{t-1})^{-1}\| = O_p(1)$. By computing the norms on both sides of (B.9) we get:

$$\|\hat{\alpha}_t\| \leq c_1 + c_2\|\hat{\alpha}_t\| + c_3\|\hat{\alpha}_t\|^2, \tag{B.10}$$

where $c_1 = O_p(\|\hat{\Psi}_t\|)$, $c_2 = O_p(\|\hat{\Psi}_t\|)$ and $c_3 = O_p(\|\hat{\Psi}_t\|)$. Hence, $\|\hat{\alpha}_t\|$ satisfies an inequality of second order, either $\|\hat{\alpha}_t\| \leq \frac{1-c_2-\sqrt{(1-c_2)^2-4c_1c_3}}{2c_3}$ or $\|\hat{\alpha}_t\| \geq \frac{1-c_2+\sqrt{(1-c_2)^2-4c_1c_3}}{2c_3}$. The second case is not admissible, since $\frac{1-c_2+\sqrt{(1-c_2)^2-4c_1c_3}}{2c_3} \simeq c_3^{-1} \xrightarrow{p} \infty$ (...). Hence we have $\|\hat{\alpha}_t\| \leq \frac{1-c_2-\sqrt{(1-c_2)^2-4c_1c_3}}{2c_3}$. Using that $\frac{1-c_2-\sqrt{(1-c_2)^2-4c_1c_3}}{2c_3} \simeq \frac{2c_1c_3}{2c_3} = c_1$, we get:

$$\|\hat{\alpha}_t\| = O_p(c_1) = O_p(\|\hat{\Psi}_t\|). \tag{B.11}$$

From equation (B.8) we have:

$$\begin{aligned}\delta_j(\hat{\Lambda}_t) &= \delta_j\left(\mathcal{U}_t^{-1}(\Pi_t'\hat{\Psi}_t\Pi_t + \Pi_t'\hat{\Psi}_t J_t\hat{\alpha}_t)\mathcal{U}_t\right)\\ &= \delta_j\left(\Pi_t'\hat{\Psi}_t\Pi_t + \Pi_t'\hat{\Psi}_t J_t\hat{\alpha}_t\right).\end{aligned}$$

If we can apply a Weyl's inequality argument, from (B.11) we deduce:

$$\begin{aligned}\delta_j(\hat{\Lambda}_t) &= \delta_j(\Pi_t'\hat{\Psi}_t\Pi_t) + O_p(\|\Pi_t'\hat{\Psi}_t J_t\hat{\alpha}_t\|) = \delta_j(\Pi_t'\hat{\Psi}_t\Pi_t) + O_p(\|\hat{\Psi}_t\|\|\hat{\alpha}_t\|)\\ &= \delta_j(\Pi_t'\hat{\Psi}_t\Pi_t) + O_p(\|\hat{\Psi}_t\|^2).\end{aligned}$$

The bound in (B.5) follows by using $\delta_{k_t+j}(\hat{V}_t) = \delta_j(\hat{\Lambda}_t)$.

Note: The application of the Weyl's inequality is not straightforward, and we check this here in more detail. The Weyl's inequalities for two *symmetric* matrices A and B state that $\delta_{i+j-1}(A +$

$B) \leq \delta_i(A) + \delta_j(B)$. In particular, for $j = 1$ we get: $\delta_i(A+B) \leq \delta_i(A) + \delta_1(B)$. Now, this inequality also implies $\delta_i(A) \leq \delta_i(A+B) + \delta_1(-B) \leq \delta_i(A+B) - \delta_m(B)$, if m is the dimension of B. Thus, we have: $\delta_i(A+B) \geq \delta_i(A) + \delta_m(B) \geq \delta_i(A) - \max\limits_{1 \leq j \leq m} |\delta_j(B)|$. Using that for a symmetric matrix B we have $\|B\| = \max\limits_{1 \leq j \leq m} |\delta_j(B)|$, we have just shown that:

$$|\delta_i(A+B) - \delta_i(A)| \leq \|B\|, \tag{B.12}$$

(this inequality corresponds to Fact 9.12.4 in Bernstein (2009) [15]). The above application of this inequality was not fully justified because one of the two matrices is not symmetric. Equation (B.3) implies by the orthonormality of the eigenvectors in $\hat{W}_t$:

$$I_{K-k_t} = \mathcal{U}_t' \, (I_{K-k_t} + \hat{\alpha}_t'\hat{\alpha}_t) \, \mathcal{U}_t. \tag{B.13}$$

Hence:

$$\mathcal{U}_t^{-1} = \mathcal{U}_t' + \mathcal{U}_t'\hat{\alpha}_t'\hat{\alpha}_t. \tag{B.14}$$

By plugging into the expression of $\hat{\Lambda}_t$ given in (B.8), we have:

$$\hat{\Lambda}_t = \mathcal{U}_t'\Pi_t'\hat{\Psi}_t\Pi_t\mathcal{U}_t + \mathcal{U}_t' \left\{ \hat{\alpha}_t'\hat{\alpha}_t\Pi_t'\hat{\Psi}_t\Pi_t + (I_{K-k_t} + \hat{\alpha}_t'\hat{\alpha}_t)\Pi_t'\hat{\Psi}_t J_t \hat{\alpha}_t \right\} \mathcal{U}_t.$$

Moreover, from (B.13) we have $I_{K-k_t} = \mathcal{U}_t'\mathcal{U}_t + \mathcal{U}_t'\hat{\alpha}_t'\hat{\alpha}_t\mathcal{U}_t$, which implies $I_{K-k_t} = (\mathcal{U}_t'\mathcal{U}_t)^{-1/2} + C$, where $C := I_{K-k_t} - (I_{K-k_t} - \mathcal{U}_t'\hat{\alpha}_t'\hat{\alpha}_t\mathcal{U}_t)^{-1/2}$. Then we can write $\hat{\Lambda}_t = A + B$, where

$$\begin{aligned} A &= \bar{\mathcal{U}}_t'\Pi_t'\hat{\Psi}_t\Pi_t\bar{\mathcal{U}}_t, \\ B &= C\mathcal{U}_t'\Pi_t'\hat{\Psi}_t\Pi_t\bar{\mathcal{U}}_t + \bar{\mathcal{U}}_t'\Pi_t'\hat{\Psi}_t\Pi_t\mathcal{U}_t C + C\mathcal{U}_t'\Pi_t'\hat{\Psi}_t\Pi_t\mathcal{U}_t C' \\ &\quad + \mathcal{U}_t' \left\{ \hat{\alpha}_t'\hat{\alpha}_t\Pi_t'\hat{\Psi}_t\Pi_t + (I_{K-k_t} + \hat{\alpha}_t'\hat{\alpha}_t)\Pi_t'\hat{\Psi}_t J_t \hat{\alpha}_t \right\} \mathcal{U}_t \end{aligned}$$

where $\bar{\mathcal{U}}_t := \mathcal{U}_t(\mathcal{U}_t'\mathcal{U}_t)^{-1/2}$ is an orthogonal matrix. Matrix A and therefore also $B = \hat{\Lambda}_t - A$ are symmetric. Hence, inequality (B.12) implies $\delta_j(\hat{\Lambda}_t) = \delta_j(A) + O_p(\|B\|)$, uniformly in $j = 1, ..., K - k_t$. Now, $\delta_j(A) = \delta_j(\bar{\mathcal{U}}_t^{-1}\Pi_t'\hat{\Psi}_t\Pi_t\bar{\mathcal{U}}_t) = \delta_j(\Pi_t'\hat{\Psi}_t\Pi_t)$ because $\bar{\mathcal{U}}_t$ is orthogonal. To bound the operator norm of matrix B, we use that $\|\mathcal{U}_t\| = \delta_1(\mathcal{U}_t'\mathcal{U}_t)^{1/2} = \delta_1(I_{K-k_t} - \mathcal{U}_t'\hat{\alpha}_t'\hat{\alpha}_t\mathcal{U}_t)^{1/2} \leq 1$. Using the series expansion of the inverse square-root matrix function and the bound in (B.11), we get $\|C\| = O_p(\|\hat{\Psi}_t\|^2)$. Thus we get $\|B\| = O_p(\|\hat{\Psi}_t\|^2)$. We conclude

$$\delta_j(\hat{\Lambda}_t) = \delta_j(\Pi_t'\hat{\Psi}_t\Pi_t) + O_p(\|\hat{\Psi}_t\|^2). \tag{B.15}$$

Then, (B.5) follows from $\delta_{k_t+j}(\hat{V}_t) = \delta_j(\hat{\Lambda}_t)$.

The bound in (B.4) is a straightforward implication of equation $\hat{V}_t = \tilde{V}_t + \hat{\Psi}_t = V_t + \hat{\Psi}_t + \Gamma_t(\hat{V}(g_t|\mathcal{F}_{t-1}) - V(g_t|\mathcal{F}_{t-1}))\Gamma_t'$, the result in (B.12) and triangular inequality. Q.E.D.

[15] A related result is the Weilandt-Hoffmann inequality: $\sum_{j=1}^{m} |\delta_j(A+B) - \delta_j(A)|^2 \leq \|B\|_F^2$, see Tao (2012), p.137.

B.2 Consistency of the eigenvalue-ratio test

B.2.1 Constant number of conditional factors

We first focus on the case where the number of conditional factors k is constant.

(i) Let $j \leq k-1$. We have

$$\frac{\delta_j(\hat{V}_t)}{\delta_{j+1}(\hat{V}_t)} \leq \frac{\delta_j(V_t)}{\delta_{j+1}(V_t)} \frac{1+x}{1-y},$$

where $x := |\delta_j(\hat{V}_t) - \delta_j(V_t)|/\delta_j(V_t)$ and $y := |\delta_{j+1}(\hat{V}_t) - \delta_{j+1}(V_t)|/\delta_{j+1}(V_t)$. From Lemma 1 we have $|\delta_j(\hat{V}_t) - \delta_j(V_t)| \leq \|\hat{\Phi}_t\|$ for all j and t, where $\|\hat{\Phi}_t\| = \|\hat{\Psi}_t\| + \|\hat{V}(g_t|\mathcal{F}_{t-1}) - V(g_t|\mathcal{F}_{t-1})\|$. Moreover $\bar{c} \geq \delta_j(V_t) \geq \underline{c}$, for all j, t. Define $\mathcal{T} = \{\|\hat{\Phi}_t\| \leq \epsilon,\ t = 1, ..., T\}$ where $\epsilon = \rho\underline{c}$ and $\rho \in (0,1)$. Further we use that $\frac{1+x}{1-y} \leq 1 + \frac{1}{1-\rho}(x+y)$ for all $x, y \in [0, \rho]$. Thus we get on $\mathcal{T}$:

$$\frac{\delta_j(\hat{V}_t)}{\delta_{j+1}(\hat{V}_t)} \leq \frac{\delta_j(V_t)}{\delta_{j+1}(V_t)} + C\|\hat{\Phi}_t\|, \tag{B.16}$$

for universal constant $C = 2(\bar{c}/\underline{c}^2)(1-\rho)^{-1}$.

(ii) Let $j = k$. From Lemma 1 we have $|\delta_j(\hat{V}_t) - \delta_j(V_t)| \leq \|\hat{\Phi}_t\|$ and $z := |\delta_{j+1}(\hat{V}_t) - \delta_1(\Pi_t'\hat{\Psi}_t\Pi_t)| \leq C\|\hat{\Psi}_t\|^2$. Then on $\mathcal{T}$ we have:

$$\begin{aligned} \frac{\delta_j(\hat{V}_t)}{\delta_{j+1}(\hat{V}_t)} &\geq \frac{\delta_j(V_t)(1-x)}{\delta_1(\Pi_t'\hat{\Psi}_t\Pi_t) + z} \\ &\geq \frac{\underline{c}(1-x)}{\delta_1(\Pi_t'\hat{\Psi}_t\Pi_t) + C\|\hat{\Psi}_t\|^2} \geq \frac{\underline{c}(1-\rho)}{(1+C\epsilon)\|\hat{\Psi}_t\|}. \end{aligned}$$

(iii) Let $j \geq k+1$ and $j \leq q-1$. Let us define the sets of time indices:

$$\begin{aligned} \mathcal{T}_1 &= \left\{t : \ 1 \leq t \leq T,\ \delta_q(\hat{V}_t) \geq \sigma^a\right\}, \\ \mathcal{T}^* &= \left\{t : \ 1 \leq t \leq T,\ \|\hat{\Psi}_t\| \leq \epsilon\sqrt{\sigma},\ \sigma \leq \delta_{q-k}(\Pi_t'\hat{\Psi}_t\Pi_t)\right\}, \end{aligned}$$

with $a > 1$ and $\sigma \downarrow 0$. We have $\mathcal{T}^* \subset \mathcal{T}_1$, for ϵ small. Indeed, from Lemma 1 we have for $t \in \mathcal{T}^*$:

$$\begin{aligned} \delta_q(\hat{V}_t) &\geq \delta_{q-k}(\Pi_t'\hat{\Psi}_t\Pi_t) - C\|\hat{\Psi}_t\|^2 \\ &\geq (1 - C\epsilon^2)\sigma \geq \sigma^a, \end{aligned}$$

for $\epsilon \leq \sqrt{1/(2C)}$ and $\sigma \leq (1/2)^{1/(a-1)}$. For $t \in \mathcal{T}_1$ and on $\mathcal{T}$ we have:

$$\begin{aligned}\frac{\delta_j(\hat{V}_t)}{\delta_{j+1}(\hat{V}_t)} &\leq \frac{\delta_{j-k}(\Pi_t'\hat{\Psi}_t\Pi_t) + C\|\hat{\Psi}_t\|^2}{\sigma^a} \\ &\leq \frac{(1+C\epsilon)\|\hat{\Psi}_t\|}{\sigma^a}.\end{aligned}$$

We summarize the previous findings. For $j \leq k-1$ and on $\mathcal{T}$ we have:

$$\frac{\delta_j(\hat{V}_t)}{\delta_{j+1}(\hat{V}_t)} \leq \frac{\delta_j(V_t)}{\delta_{j+1}(V_t)} + C\|\hat{\Phi}_t\|,$$

for $j = k$ and on $\mathcal{T}$ we have:

$$\frac{\delta_j(\hat{V}_t)}{\delta_{j+1}(\hat{V}_t)} \geq C\|\hat{\Psi}_t\|^{-1},$$

and finally for $j \geq k+1$ and $j \leq q-1$ on $\mathcal{T}$ and for $t \in \mathcal{T}_1$ we have:

$$\frac{\delta_j(\hat{V}_t)}{\delta_{j+1}(\hat{V}_t)} \leq C\frac{\|\hat{\Psi}_t\|}{\sigma^a},$$

where C denotes a generic positive constant (not necessarily equal in all instances).

Let $\hat{\rho}_j := \frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\frac{\delta_j(\hat{V}_t)}{\delta_{j+1}(\hat{V}_t)}$ with $T_1 = |\mathcal{T}_1|$. We have for $j \leq k-1$ and on $\mathcal{T}$:

$$\hat{\rho}_j \leq C + C\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\|\hat{\Phi}_t\| \leq C(1+\epsilon),$$

for $j = k$ on $\mathcal{T}$:

$$\begin{aligned}\hat{\rho}_j &\geq C\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\|\hat{\Psi}_t\|^{-1} \geq C\left(\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\|\hat{\Psi}_t\|\right)^{-1} \\ &\geq C(T^*/T)\left(\frac{1}{T}\sum_{t=1}^{T}\|\hat{\Psi}_t\|\right)^{-1},\end{aligned}$$

where $T^* = |\mathcal{T}^*| \leq T_1$ and using Jensen inequality, and finally for $j \geq k+1$ and $j \leq q-1$ on $\mathcal{T}$:

$$\hat{\rho}_j \leq \frac{C}{\sigma^a}\frac{1}{T_1}\sum_{t\in\mathcal{T}_1}\|\hat{\Psi}_t\| \leq \frac{C}{\sigma^a}(T^*/T)^{-1}\frac{1}{T}\sum_{t=1}^{T}\|\hat{\Psi}_t\|.$$

Let $\hat{k} = \underset{1\leq j\leq q-1}{\arg\max}\hat{\rho}_j$. Then, using the Cauchy-Schwarz inequality, we have $\hat{k} = k$ w.p.a. 1, if $\mathbb{P}(\mathcal{T}) \to 1$, $T^*/T = O_p^{-1}(1)$, i.e. T^*/T is bounded away from zero w.p.a. 1, and $\frac{1}{T}\sum_{t=1}^{T}\|\hat{\Psi}_t\|^2 \ll \sigma^a$.

Lemma 3. *Let us assume that*

$$
\begin{aligned}
&i) \qquad \mathbb{P}\left(\sup_{1\leq t\leq T}\|\hat{\Phi}_t\| \leq \epsilon\right) \to 1,\\
&ii) \qquad \frac{1}{T}\sum_{t=1}^{T} 1\{\|\hat{\Psi}_t\| \leq \epsilon\sqrt{\sigma},\ \sigma \leq \delta_{q-k}(\Pi_t'\hat{\Psi}_t\Pi_t)\} = O_p^{-1}(1),\\
&iii) \qquad \frac{1}{T}\sum_{t=1}^{T}\|\hat{\Psi}_t\|^2 = o(\sigma^a),
\end{aligned}
$$

for $\sigma \downarrow 0$*,* $a > 1$*, and* $\epsilon > 0$ *small. Then* $\hat{k} = k$ *w.p.a.* 1.

B.2.2 Time-varying number of factors

Let us now consider the case with time-varying number of factors k_t. We assume that $1 \leq k_t \leq q$. Let $\hat{\rho}_{j,t} = \delta_j(\hat{V}_t)/\delta_{j+1}(\hat{V}_t)$, and $\hat{k}_t := \max_{j:1\leq j\leq q} \hat{\rho}_{j,t}$. Let t be such that $\|\hat{\Phi}_t\| \leq \epsilon$, for $\epsilon > 0$ small. From the arguments in the previous subsection we have:

(i) $\hat{\rho}_{j,t} \leq C(1+\epsilon)$, for $j < k_t$, and

(ii) $\hat{\rho}_{j,t} \geq \frac{C}{1+C\epsilon}\|\hat{\Psi}_t\|^{-1}$, for $j = k_t$, where C is a generic universal constant. Moreover:

(iii) Suppose that date t is in the set $\{t \ :\ \|\hat{\Psi}_t\| \leq \epsilon\sqrt{\sigma},\ \sigma \leq \delta_{q-k_t}(\Pi_t'\hat{\Psi}_t\Pi_t)\}$. Then, we have $\delta_q(\hat{V}_t) \geq (1 - C\epsilon^2)\sigma$, and thus $\hat{\rho}_{j,t} \leq \frac{(1+C\epsilon)\|\hat{\Psi}_t\|}{(1-C\epsilon^2)\sigma}$, for $j > k_t$.

Therefore, at a date t in set $\{t \ :\ \|\hat{\Phi}_t\| \leq \epsilon,\ \|\hat{\Psi}_t\| \leq \epsilon\sqrt{\sigma},\ \sigma \leq \delta_{q-k_t}(\Pi_t'\hat{\Psi}_t\Pi_t)\}$, we have $\hat{k}_t = k_t$ if :

$$
\begin{aligned}
\frac{C}{1+C\epsilon}\|\hat{\Psi}_t\|^{-1} > C(1+\epsilon) \qquad &\Leftrightarrow \qquad \|\hat{\Psi}_t\| < \frac{1}{(1+\epsilon)(1+C\epsilon)},\\
\frac{C}{1+C\epsilon}\|\hat{\Psi}_t\|^{-1} > \frac{(1+C\epsilon)\|\hat{\Psi}_t\|}{(1-C\epsilon^2)\sigma} \qquad &\Leftrightarrow \qquad \|\hat{\Psi}_t\|^2 < \frac{C(1-C\epsilon^2)}{(1+C\epsilon)^2}\sigma,
\end{aligned}
$$

and the latter inequalities hold true if ϵ is small enough.

Lemma 4. *For any subset* $\mathcal{T}_1$ *of* $\{1, ..., T\}$*, constant* $\epsilon > 0$ *small enough, and* $\sigma \downarrow 0$*, we have:*

$$
\mathbb{P}\left(\hat{k}_t = k_t,\ \forall t \in \mathcal{T}_1\right) \geq \mathbb{P}\left(\|\hat{\Phi}_t\| \leq \epsilon,\ \|\hat{\Psi}_t\| \leq \epsilon\sqrt{\sigma},\ \sigma \leq \delta_{q-k_t}(\Pi_t'\hat{\Psi}_t\Pi_t),\ \forall t \in \mathcal{T}_1\right).
$$

In particular, if the probability in the RHS tends to 1*, then* $\hat{k}_t = k_t$ *for all* $t \in \mathcal{T}_1$ *w.p.a.* 1.

C Finding instrumental variables

Time-invariant instrumental variables w_i are generated by time averages:

$$w_i = E_i[\varphi(W)] = \operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \varphi(W_{i,t}), \tag{C.1}$$

where $\varphi(\cdot)$ denotes a generic (integrable) function of variable W. From Assumption 1 variable w_i is measurable w.r.t. sigma-field $\mathcal{H}_i$.

Assumption 14. *For any variable w_i which is measurable w.r.t. $\mathcal{H}_i$, we have:*

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} w_i \varepsilon_{i,t} = 0.$$

Under Assumption 14 the asset-specific and time-invariant characteristics are *cross-sectionally* uncorrelated with the idiosyncratic errors. Then, time-invariant instrument w_i satisfies Assumption 1 (i).

As an example, suppose the factor loadings vector $b_{i,t-1}$ is such that:

$$b_{i,t-1} = b(Z_{t-1}, \alpha_i), \tag{C.2}$$

where Z_{t-1} generates the common information set $\mathcal{F}_{t-1}$ at time $t-1$, and α_i contains all the firm-specific characteristics for asset i that generate $\mathcal{H}_i$. Taking (C.2) into (C.1) with $W_{i,t} = y_{i,t}$, we have:

$$\begin{aligned} w_i &= \operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \varphi\left[b(Z_{t-1}, \alpha_i)' g_t + \varepsilon_{i,t}\right] \\ &= E[\varphi(b(Z_{t-1}, \alpha_i)' g_t + \varepsilon_{i,t})|\alpha_i] \\ &= \int \varphi(b(Z_{t-1}, \alpha_i)' g_t + \varepsilon_{i,t}) dP_i(Z_{t-1}, g_t, \varepsilon_{i,t}) \equiv U(\alpha_i), \end{aligned}$$

where $E[\cdot|\alpha_i]$ represents the expectation across time for given asset i, and $P_i(\cdot)$ denotes the conditional probability law of $Z_{t-1}, g_t, \varepsilon_{i,t}$ given α_i, for asset i. As a result, $U(\alpha_i)$ is a time-invariant function, which only depends on firm-specific characteristics α_i. Assumption 14 (or equivalently Assumption 2 (i)) is satisfied if the α_i and the $\varepsilon_{i,t}$ are cross-sectionally independent at any date. Assumption 2 (ii) is satisfied if matrix $E_t^c[w_i b_{i,t-1}'] = \int U(\alpha_i) b(Z_{t-1}, \alpha_i)' dP(\alpha_i)$ is full-rank, P-a.s.

D Proof of equation (9)

Matrices Λ_t and J_t are the diagonal matrix of the k non-zero eigenvalues of $V(\xi_t|\mathcal{F}_{t-1})$, and the matrix of the associated standardized eigenvectors, respectively. Thus, $V(\xi_t|\mathcal{F}_{t-1})J_t = J_t\Lambda_t$. From equation (8) and Assumption 3 we get:

$$\begin{aligned} J_t\Lambda_t &= V(\xi_t|\mathcal{F}_{t-1})J_t = J_t V(g_t|\mathcal{F}_{t-1})J_t' J_t \\ &= J_t V(f_t|\mathcal{F}_{t-1}), \end{aligned}$$

where we use $J_t'J_t = I_k$ and $V(g_t|\mathcal{F}_{t-1}) = V(f_t|\mathcal{F}_{t-1})$. Then, since matrix J_t is full column-rank, equation (9) follows.

E. Monte Carlo Simulations

In this section, we report the results of our Monte Carlo simulation study to investigate the finite sample properties of our estimators for the factor space and its dimension.

E.1 DGPs with time-invariant number of factors

E.1.1 Data generating process

We simulate the excess returns from the conditional factor model (3)-(4) consistent with the no-arbitrage restrictions, where f_t is a $k \times 1$ vector with distribution $i.i.d.N(\mathbf{0}, I_k)$ and $\varepsilon_{i,t} \sim i.i.d.N(0,1)$, mutually independent. The factor loadings $b_{i,t-1}$ and the vector ν_t are generated by:

$$\begin{aligned} b_{i,t-1} &= b_{i,0} + b_{i,1}Z_{t-1} \\ \nu_t &= \nu_0 + \nu_1 Z_{t-1}, \end{aligned} \tag{C.3}$$

where Z_t is a $d \times 1$ vector with distribution $i.i.d.N(\mathbf{0}, I_d)$, $b_{i,0}$ is a $k \times 1$ vector with distribution $i.i.d.N(\mathbf{1_k}, I_k)$, where $\mathbf{1_k}$ denotes a $k \times 1$ vector with unit elements, and $b_{i,1}$ is a $k \times d$ matrix with each element following $i.i.d.N(0,1)$, ν_0 is a $k \times 1$ constant vector and ν_1 is a $k \times d$ constant matrix. All random variables are assumed mutually independent.

The $K \times 1$ vector of instrumental variables $w_{i,t}$ is generated as follows:

$$w_{i,t} = \begin{bmatrix} SA_t(Cb_{i,t} + u_{i,t}) \\ Qu_{i,t}^W \end{bmatrix} \tag{C.4}$$

and S is a $(k+k_R) \times k$ matrix with $S = \begin{bmatrix} I_k \\ R \end{bmatrix} U\Lambda^{-1/2}$, where R is a full-rank $k_R \times k$ matrix, U is the matrix of normalized eigenvectors of $I_k + R'R$ and Λ is the diagonal matrix of eigenvalues. Therefore, the columns of matrix S are orthonormal

$$S'S = \Lambda^{-1/2}U'(I_k + R'R)U\Lambda^{-1/2} = I_k. \tag{C.5}$$

Moreover, C is a $k \times k$ diagonal matrix with $\rho_1, ..., \rho_k$ on its diagonal, Q is a $(K-k-k_R) \times k_W$ matrix, $u_{i,t} \sim i.i.d.N(\mathbf{0}, I_k)$ and $u^W_{i,t} \sim i.i.d.N(\mathbf{0}, I_{k_W})$ mutually independent. In Appendix D we show that with the choice

$$A_t = \left[(1+\|Z_t\|^2)I_k + \mathbf{1_k}\mathbf{1_k}'\right]^{-1} C^{-1} \tag{C.6}$$

Assumptions 2 and 4 are satisfied. When $k_R = k_W = 0$ we have $S = I_k$, the setting is exactly identified and matrix C controls for the correlation of the instruments with the loadings. Otherwise, matrix S is designed to create k useful and k_R redundant instruments, and matrix Q generates $K-k-k_R$ useless instruments from k_W pure noise variables.

We simulate 1000 panel datasets for each of the following DGP settings:

DGP 1 : $k = 1$, $k_R = k_W = 0$, $d = 1$, $\nu_0 = 1$, $\nu_1 = 1$

DGP 2 : $k = 2$, $k_R = k_W = 0$, $d = 1$, $\nu_0 = [1\ \ -1]'$, $\nu_1 = [1\ \ 1]'$

DGP 3 : $k = 2$, $k_R = k_W = 0$, $d = 10$, $\nu_0 = [1\ \ -1]'$, ν_1 is a 2×10 matrix of ones

DGP 4 : $k = 2$, $k_R = k_W = 0$, $d = 20$, $\nu_0 = [1\ \ -1]'$, ν_1 is a 2×20 matrix given by:

$$\nu_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & \cdots 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & \cdots 0 \end{bmatrix}$$

DGP 5 : $k = 2$, $k_R = 1$, $k_W = 3$, $d = 10$, $\nu_0 = \begin{bmatrix} 1 & -1 \end{bmatrix}'$,

$$\begin{aligned} \nu_1 &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & \cdots 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & \cdots 0 \end{bmatrix}, \\ R &= \begin{bmatrix} -1 & 1 \end{bmatrix},\ Q = I_3,\ C = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}, \end{aligned}$$

where ν_1 is a 2×10 matrix.

DGP1 and DGPs 2-4 are exactly identified with 1 and 2 factors, respectively, and different numbers of common variables. DGP 5 is overidentified with 1 redundant and 3 useless factors.

E.1.2 Simulation Results

We start with exactly identified settings (DGP 1-4). In order to understand how our estimation approach works for different finite samples, we perform simulation exercises combining different values of the cross-sectional dimension n and the time series dimension T. Tables 1 to 4 in the Appendix show percentiles of root mean square error (RMSE) of our estimators $\hat{g}_t$, $\hat{\nu}_t$ and $\hat{f}_t$ of different sample sizes under corresponding DGP processes. RMSE is calculated via the following formulas:

$$RMSE(l) = \sqrt{\frac{1}{kT} \sum_{j=1}^{k} \sum_{t=1}^{T} (\widehat{l_{jt}} - l_{jt})^2}, \tag{C.7}$$

where $l = g, \nu, f$.

Since we estimate g_t based on a pure cross-sectional average at each time t, we expect that the RMSE of $\hat{g}_t$ is stable with respect to time series dimension T and decreases with cross-sectional dimension n. RMSEs of $\hat{g}_t$ in Tables 1-4 verify this property: for each column, the RMSE stays approximately the same with constant n but different T, while it is decreasing when n increases.

The RMSEs of $\hat{\nu}_t$ will show an opposite property since we estimate $\hat{\nu}_t$ by using conditional expectation model, which is basically a time-series weighted average. However, since we replace g_t with $\hat{g}_t$ when estimating ν_t, we may also expect RMSEs to decrease with cross-sectional dimension n. From the panels corresponding to $\hat{\nu}_t$ in each table, we observe that the RMSEs of $\hat{\nu}_t$ are relatively stable as n increases, while clearly decreasing with T. We deduce that the estimation error from replacing g_t with $\hat{g}_t$ can be neglected compared to the error from time series regression.

As $\hat{f}_t$ is simply given by $\hat{g}_t - \hat{\nu}_t$, we expect the RMSEs of $\hat{f}_t$ to decrease with both sample dimensions, which we indeed observe from the panel regarding $\hat{f}_t$ in each table.

To summarize, in the exactly identified case, the root mean square errors of our estimates are rather small for realistic sample sizes, showing that our methodology has good finite-sample properties for the considered DGPs. Moreover, the root mean square errors decrease with the sample sizes N and T as predicted by theory.

We now move to the overidentified setting with DGP 5. Table 5 shows percentages of selection of $\hat{k}$ among all repetitions of different sample sizes. For the smallest considered sample sizes, namely $n = 1000$ and $T = 250$, the estimate $\hat{k}$ selects the correct number of factors $k = 2$ in 76% of the cases, and underestimates k by one unit in the rest of the cases. When the time dimension increases to $T = 500$, which is similar to our empirical application, the percentage of correct selection increases to 95%. As the sample sizes continue to increase to $n \geq 5000$, the simulation results reach a 100% rate of correct selection, showing that the estimator $\hat{k}$ behaves very well in the time-invariant case.

In order to assess the accuracy of the estimator of the number of factors $\hat{k}_t$ at the given date t, in Figure 2 we display histograms of $\hat{k}_t$ across dates $t = 1, ..., T$ and 100 Monte Carlo replications for different combinations of n and T. As expected, the histograms become more peaked at the true number of factors $k = 2$ as n and T increase.

In order to highlight the importance of accounting for the conditional nature of the factors, we compare our results with those obtained by selecting the number of factors with a method for static factor models. Specifically, we deploy the IC_{p1} and IC_{p2} criteria of Bai and Ng (2002). For our simulation designs, both criteria select 22 factors in every Monte Carlo draw across all sample sizes. In fact, by plugging (C.3) into (3), we see that our DGP corresponds to a model with $k(d+1) = 22$ static factors.

E.2 Overidentified case with time-varying number of factors

In this section, we conduct Monte Carlo experiments with a model featuring a time-varying number of conditional factors.

E.2.1 Data generating process

The data generating process for the excess returns $y_{i,t}$ is (3)-(4) where the factor loadings $b_{i,t-1}$ are now generated by

$$b_{i,t-1} = D_{t-1}(b_{i,0} + b_{i,1}Z_{t-1}), \qquad D_{t-1} = \begin{bmatrix} I_{k_0} & 0 \\ 0 & \{\phi(Z_{t-1})\}_+ I_{k-k_0} \end{bmatrix}, \tag{C.8}$$

where $\phi(\cdot)_+ = \max\{0, \phi(\cdot)\}$ is the positive part of the scalar-valued function $\phi(\cdot)$, $k_0 < k$, and $b_{i,0}$, $b_{i,1}$, Z_{t-1} are generated in the same way as in Subsection 7. As a result, the number of non-zero factor loadings in $b_{i,t-1}$ is now time-varying:

$$k_t = \begin{cases} k, & \text{when} \quad \phi(Z_{t-1}) > 0 \\ k_0, & \text{otherwise} \end{cases}$$

The vector of instrumental variables $w_{i,t}$ is as in (C.4) with

$$A_{t-1} = \begin{bmatrix} \{(1 + \|Z_{t-1}\|^2)D_{t-1}^2 + D_{t-1}\mathbf{1}_k\mathbf{1_k}'D_{t-1}\}^{-1}_{k_t,k_t} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{(k-k_t)\times(k-k_t)} \end{bmatrix} C^{-1}, \tag{C.9}$$

where $\{\cdot\}_{k_t,k_t}$ denotes the upper-left (k_t, k_t) block of a matrix. The factors are i.i.d. conditionally heteroschedastic given Z_{t-1}, with $f_t \sim N(0, D_{t-1}^2)$ and $\nu_t = D_{t-1}(\nu_0 + \nu_1 Z_{t-1})$. Hence, the conditional variances of $k - k_0$ factors are phased out by function $\phi(Z_{t-1})_+$. The conditions in Assumption 2.TV and Assumption 4.TV are met (see Appendix D.2), with

$$E(\xi_t|\mathcal{F}_{t-1}) = \begin{pmatrix} S_{k_0} & \vdots & S_{k-k_0}\phi(Z_{t-1})_+ \\ & 0 & \end{pmatrix} (\nu_0 + \nu_1 Z_{t-1}), \tag{C.10}$$

$$V(\xi_t|\mathcal{F}_{t-1}) = \begin{pmatrix} S_{k_0}S'_{k_0} + \phi(Z_{t-1})_+^2 S_{k-k_0}S'_{k-k_0} & 0 \\ 0 & 0 \end{pmatrix}, \tag{C.11}$$

where we partition $S = \begin{bmatrix} S_{k_0} \vdots S_{k-k_0} \end{bmatrix}$ and S_{k_0} is the left $(k + k_R) \times k_0$ block.

Based on the same setting $k = 2$, $k_R = 1$, $k_W = 3$, $d = 10$ and the same values for ν_0, ν_1, R, Q, and C as DGP 5 in the time-invariant case, we set up two DGPs for the time-varying case:

DGP 6 :

$$\phi(Z_{t-1}) = \mathbf{1}\{\frac{1}{3} \leq \frac{t}{T} \leq \frac{2}{3}\}, \quad k_0 = 1,$$

where we include the indicator variable $\mathbf{1}\{\frac{1}{3} \leq \frac{t}{T} \leq \frac{2}{3}\}$ as a component of Z_{t-1}.

DGP 7 :

$$\phi(Z_{t-1}) = \mathbf{1}\{Z_{1,t-1} \geq 0\}, \quad k_0 = 1,$$

where $Z_{1,t-1} = \mu + \varphi(Z_{1,t-1} - \mu) + e_t$ with $e_t \sim i.i.N.(0, 1)$.

In DGP 6, the switch in the number of factors between 1 and 2 is deterministic, while in DGP 7 it is driven by the value of $Z_{1,t}$. Hence, in DGP 6 the econometrician is assumed to know the dates of structural breaks in the number of factors. In DGP 7, this assumption is relaxed.

E.2.2 Simulation Results

Figure 3 and 4 display histograms of $\hat{k_t}$ across dates $t = 1, ..., T$ and 100 Monte Carlo replications for different sample sizes under DGP6, distinguishing dates with $k_t = 1$ and $k_t = 2$, respectively. For all sample sizes (except the smallest one), the mode of the histogram is the true number of factors. The histograms become more peaked as the sample sizes increase.

E.3 Factor normalization in the MC designs

E.3.1 Time-invariant number of factors

Let us check the validity of Assumption 4 when the instruments are defined as in (C.4) and (C.6). We have:

$$\begin{aligned}\Gamma_t &= E_t^c[w_{i,t-1}b'_{i,t-1}] \\ &= E_t^c\left(\begin{bmatrix} SA_{t-1}(Cb_{i,t-1}+u_{i,t-1}) \\ Qu^W_{i,t-1}\end{bmatrix} b'_{i,t-1}\right) \\ &= \begin{bmatrix} SA_{t-1}CE_t^c[b_{i,t-1}b'_{i,t-1}] \\ \mathbf{0}\end{bmatrix}.\end{aligned} \tag{E.1}$$

Moreover:

$$\begin{aligned} E_t^c[b_{i,t-1}b'_{i,t-1}] &= E_t^c[b_{i,0}b'_{i,0}] + E_t^c[b_{i,1}Z_{t-1}Z'_{t-1}b'_{i,1}] \\ &= I_k + \mathbf{1_k}\mathbf{1_k}' + E_t^c[b_{i,1}Z_{t-1}Z'_{t-1}b'_{i,1}].\end{aligned}$$

Let us now compute the (m,l) element of matrix $E_t^c[b_{i,1}Z_{t-1}Z'_{t-1}b'_{i,1}]$. We denote by $b'_{i,1,k}$ the kth row of matrix $b_{i,1}$, then:

$$\begin{aligned}(E_t^c[b_{i,1}Z_{t-1}Z'_{t-1}b'_{i,1}])_{ml} =& E_t^c[b'_{i,1,m}Z_{t-1}Z'_{t-1}b_{i,1,l}] = Tr(Z_{t-1}Z'_{t-1}E_t^c[b_{i,1,l}b'_{i,1,m}]) \\ =& \begin{cases} Tr(Z_{t-1}Z'_{t-1}) = \|Z_{t-1}\|^2, \text{ if } \ m=l \\ 0, \text{ otherwise}\end{cases}\end{aligned}$$

Thus, we get:

$$\Gamma_t = \begin{bmatrix} SA_{t-1}C\{(1+\|Z_{t-1}\|^2)I_k + \mathbf{1_k}\mathbf{1_k}'\} \\ \mathbf{0}\end{bmatrix}.$$

Then, with the choice (C.6) for A_{t-1} and using $V(g_t|\mathcal{F}_{t-1}) = I_k$, we get

$$\Gamma_t = \begin{bmatrix} S \\ \mathbf{0}\end{bmatrix}, \qquad V(\xi_t|\mathcal{F}_{t-1}) = \Gamma_t V(g_t|\mathcal{F}_{t-1})\Gamma'_t = \begin{bmatrix} SS' & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}.$$

Since $S'S = I_k$, the eigenvectors of matrix $V(\xi_t|\mathcal{F}_{t-1})$ associated withe the non-zero eigenvalues are the columns of matrix Γ_t, i.e. $J_t = \Gamma_t$, and Assumption 4 is met.

Finally, we note that matrix A_{t-1} can be written as:

$$\begin{aligned} A_{t-1} =& [(1+\|Z_{t-1}\|^2)I_k + \mathbf{1_k}\mathbf{1_k}']^{-1}C^{-1} = [\lambda_{1,t}P_k + \lambda_{2,t}M_k]^{-1}C^{-1} \\ =& [\frac{1}{\lambda_{1,t}}P_k + \frac{1}{\lambda_{2,t}}M_k]C^{-1},\end{aligned}$$

where $\lambda_{1,t} = 1+k+\|Z_{t-1}\|^2$, $\lambda_{2,t} = 1+\|Z_{t-1}\|^2$, $P_k = \frac{1}{k}\mathbf{1_k}\mathbf{1_k}'$, $M_k = I_k - P_k$.

E.3.2 Time-varying number of factors

Let us now consider the case with time-varying number of factors k_t according to the loadings in (C.8), and check the validity of Assumption 4** when the instruments are defined as in (C.4) and (C.9). Matrix Γ_t is as in (E.1) with

$$E_t^c[b_{i,t-1}b'_{i,t-1}] = (1 + \|Z_{t-1}\|^2)D_{t-1}^2 + D_{t-1}\mathbf{1_k}\mathbf{1_k}' D_{t-1}.$$

With the definition of A_{t-1} in (C.9) we get:

$$\Gamma_t = \begin{bmatrix} S_{k_t} & 0_{(k+k_R)\times(k-k_t)} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} S_{k_0} & S_{k-k_0}1_{k_t=k} \\ 0 & 0 \end{bmatrix}. \tag{E.2}$$

Using that $V(g_t|\mathcal{F}_{t-1}) = D_{t-1}^2$ we get $V(\xi_t|\mathcal{F}_{t-1})$ as in (C.11). The non-zero eigenvalues of matrix $V(\xi_t|\mathcal{F}_{t-1})$ are 1 with multiplicity k_t and $\phi(Z_{t-1})_+$ with multiplicity $k - k_t$. Then:

$$J_t = \begin{bmatrix} S_{k_t} \\ 0 \end{bmatrix}.$$

Hence, Assumption 4** holds. Finally, the expression of $E(\xi_t|\mathcal{F}_{t-1})$ in (C.10) follows from $E(\xi_t|\mathcal{F}_{t-1}) = \Gamma_t \nu_t$ and (E.2).

F Tables and Figures

		$\hat{g}_t$					$\hat{\nu}_t$					$\hat{f}_t$				
n	T	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
100	50	0.2021	0.2488	0.2872	0.3389	0.4722	0.0832	0.1646	0.2384	0.3194	0.4284	0.2468	0.3129	0.3833	0.4591	0.6072
	100	0.2187	0.2556	0.2897	0.3390	0.4406	0.0652	0.1194	0.1751	0.2345	0.3130	0.2459	0.2982	0.3463	0.4063	0.5229
	500	0.2286	0.2581	0.2887	0.3443	0.4681	0.0255	0.0538	0.0781	0.1030	0.1399	0.2357	0.2655	0.3022	0.3583	0.4727
500	50	0.0896	0.1108	0.1294	0.1559	0.2013	0.0852	0.1751	0.2480	0.3238	0.4542	0.1440	0.2137	0.2806	0.3542	0.4894
	100	0.0970	0.1124	0.1298	0.1537	0.1992	0.0555	0.1217	0.1686	0.2218	0.3025	0.1303	0.1743	0.2167	0.2674	0.3475
	500	0.1022	0.1152	0.1307	0.1579	0.2087	0.0235	0.0546	0.0775	0.1021	0.1361	0.1128	0.1327	0.1562	0.1849	0.2443
1000	50	0.0637	0.0788	0.0926	0.1117	0.1520	0.0864	0.1786	0.2422	0.3277	0.4400	0.1184	0.2015	0.2615	0.3473	0.4604
	100	0.0688	0.0798	0.0918	0.1093	0.1423	0.0620	0.1204	0.1740	0.2312	0.3177	0.1051	0.1526	0.1995	0.2514	0.3395
	500	0.0725	0.0819	0.0920	0.1108	0.1487	0.0283	0.0558	0.0790	0.1041	0.1420	0.0852	0.1054	0.1238	0.1501	0.1931
5000	50	0.0286	0.0352	0.0417	0.0508	0.0668	0.0817	0.1686	0.2464	0.3285	0.4438	0.0899	0.1745	0.2490	0.3279	0.4517
	100	0.0312	0.0369	0.0423	0.0506	0.0668	0.0603	0.1195	0.1743	0.2237	0.3099	0.0720	0.1261	0.1764	0.2292	0.3093
	500	0.0324	0.0362	0.0403	0.0469	0.0632	0.0253	0.0534	0.0776	0.1038	0.1374	0.0452	0.0676	0.0890	0.1134	0.1476

Table 1: Percentiles of RMSE (C.7) of estimators $\hat{g}_t, \hat{\nu}_t, \hat{f}_t$ in DGP1 (Exactly Identified Case)

		$\hat{g}_t$					$\hat{\nu}_t$					$\hat{f}_t$				
n	T	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
100	50	0.2369	0.2834	0.3196	0.3652	0.4326	0.0581	0.1152	0.1675	0.2215	0.3106	0.2658	0.3182	0.3661	0.4213	0.5200
	100	0.2533	0.2890	0.3217	0.3629	0.4390	0.0350	0.0849	0.1177	0.1571	0.2147	0.2676	0.3077	0.3457	0.3896	0.4702
	500	0.2692	0.2959	0.3202	0.3550	0.4283	0.0171	0.0380	0.0563	0.0719	0.0985	0.2726	0.3001	0.3243	0.3632	0.4388
500	50	0.1089	0.1275	0.1438	0.1618	0.1952	0.0524	0.1194	0.1694	0.2214	0.3194	0.1393	0.1808	0.2226	0.2766	0.3621
	100	0.1131	0.1280	0.1427	0.1595	0.1927	0.0406	0.0870	0.1204	0.1626	0.2234	0.1319	0.1590	0.1883	0.2242	0.2733
	500	0.1207	0.1324	0.1428	0.1579	0.1862	0.0200	0.0398	0.0567	0.0726	0.0989	0.1267	0.1410	0.1532	0.1727	0.2096
1000	50	0.0775	0.0906	0.1029	0.1166	0.1407	0.0634	0.1239	0.1746	0.2366	0.3189	0.1129	0.1600	0.2058	0.2555	0.3544
	100	0.0803	0.0918	0.1019	0.1130	0.1367	0.0431	0.0915	0.1276	0.1654	0.2214	0.1003	0.1329	0.1632	0.2011	0.2508
	500	0.0853	0.0934	0.1014	0.1132	0.1355	0.0176	0.0381	0.0549	0.0691	0.0966	0.0914	0.1049	0.1170	0.1331	0.1606
5000	50	0.0351	0.0404	0.0454	0.0518	0.0634	0.0578	0.1201	0.1681	0.2250	0.3169	0.0694	0.1288	0.1757	0.2325	0.3194
	100	0.0362	0.0410	0.0450	0.0512	0.0618	0.0436	0.0881	0.1248	0.1634	0.2164	0.0611	0.0989	0.1337	0.1697	0.2214
	500	0.0380	0.0417	0.0452	0.0502	0.0607	0.0178	0.0389	0.0560	0.0732	0.0989	0.0459	0.0580	0.0727	0.0883	0.1144

Table 2: Percentiles of RMSE (C.7) of estimators $\hat{g}_t, \hat{\nu}_t, \hat{f}_t$ in DGP2 (Exactly Identified Case)

				$\hat{g}_t$					$\hat{\nu}_t$					$\hat{f}_t$		
n	T	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
100	50	0.3351	0.4039	0.4624	0.5301	0.6459	0.5110	0.5890	0.6442	0.7107	0.9440	0.6458	0.7373	0.8092	0.8949	1.0828
	100	0.3639	0.4197	0.4615	0.5122	0.5981	0.3834	0.4416	0.4817	0.5262	0.5898	0.5353	0.6149	0.6715	0.7310	0.8311
	500	0.4027	0.4369	0.4666	0.4998	0.5680	0.1857	0.2108	0.2282	0.2447	0.2757	0.4388	0.4798	0.5160	0.5604	0.6418
500	50	0.1526	0.1828	0.2055	0.2308	0.2884	0.5120	0.5913	0.6479	0.7149	1.0539	0.5503	0.6239	0.6863	0.7493	1.0761
	100	0.1663	0.1883	0.2079	0.2310	0.2669	0.3812	0.4415	0.4834	0.5251	0.5911	0.4269	0.4858	0.5261	0.5710	0.6377
	500	0.1812	0.1975	0.2122	0.2272	0.2542	0.1838	0.2124	0.2294	0.2476	0.2761	0.2577	0.2888	0.3119	0.3394	0.3822
1000	50	0.1074	0.1280	0.1463	0.1655	0.1993	0.5137	0.5948	0.6532	0.7235	0.9825	0.5293	0.6113	0.6681	0.7402	0.9943
	100	0.1165	0.1341	0.1486	0.1640	0.1891	0.3819	0.4460	0.4863	0.5288	0.5945	0.4029	0.4628	0.5092	0.5531	0.6217
	500	0.1277	0.1392	0.1486	0.1598	0.1799	0.1884	0.2141	0.2318	0.2501	0.2728	0.2261	0.2535	0.2761	0.2979	0.3353
5000	50	0.0479	0.0573	0.0646	0.0727	0.0902	0.5189	0.5936	0.6506	0.7213	1.0074	0.5198	0.5953	0.6514	0.7259	1.0153
	100	0.0514	0.0598	0.0661	0.0735	0.0850	0.3844	0.4427	0.4819	0.5290	0.5924	0.3928	0.4478	0.4863	0.5350	0.5929
	500	0.0572	0.0622	0.0669	0.0712	0.0804	0.1866	0.2082	0.2272	0.2456	0.2716	0.1913	0.2185	0.2355	0.2545	0.2865

Table 3: Percentiles of RMSE (C.7) of estimators $\hat{g}_t, \hat{\nu}_t, \hat{f}_t$ in DGP3 (Exactly Identified Case)

				$\hat{g}_t$					$\hat{\nu}_t$					$\hat{f}_t$		
n	T	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
100	50	0.2346	0.2744	0.3056	0.3377	0.4008	0.3769	0.4637	0.5273	0.6086	0.7453	0.4630	0.5516	0.6200	0.6887	0.8130
	100	0.2532	0.2847	0.3079	0.3358	0.3795	0.2765	0.3362	0.3847	0.4387	0.5340	0.3940	0.4484	0.4934	0.5445	0.6386
	500	0.2767	0.2942	0.3087	0.3235	0.3491	0.1262	0.1554	0.1758	0.1983	0.2312	0.3055	0.3337	0.3539	0.3784	0.4194
500	50	0.1037	0.1223	0.1362	0.1517	0.1752	0.3899	0.4700	0.5393	0.6158	0.7584	0.4144	0.4906	0.5549	0.6366	0.7641
	100	0.1138	0.1275	0.1376	0.1490	0.1664	0.2776	0.3406	0.3852	0.4355	0.5348	0.3085	0.3649	0.4090	0.4564	0.5551
	500	0.1241	0.1319	0.1381	0.1454	0.1552	0.1276	0.1544	0.1728	0.1960	0.2328	0.1826	0.2047	0.2222	0.2409	0.2768
1000	50	0.0751	0.0871	0.0971	0.1075	0.1256	0.3677	0.4693	0.5413	0.6190	0.7594	0.3820	0.4779	0.5462	0.6287	0.7713
	100	0.0803	0.0902	0.0974	0.1050	0.1199	0.2744	0.3413	0.3863	0.4400	0.5358	0.2857	0.3534	0.3992	0.4502	0.5440
	500	0.0879	0.0938	0.0982	0.1029	0.1103	0.1285	0.1545	0.1758	0.1983	0.2313	0.1582	0.1813	0.2016	0.2221	0.2583
5000	50	0.0329	0.0384	0.0426	0.0476	0.0561	0.3849	0.4747	0.5409	0.6172	0.7577	0.3854	0.4783	0.5413	0.6185	0.7567
	100	0.0361	0.0403	0.0434	0.0470	0.0528	0.2745	0.3428	0.3891	0.4467	0.5455	0.2778	0.3460	0.3904	0.4490	0.5433
	500	0.0391	0.0418	0.0439	0.0458	0.0489	0.1286	0.1551	0.1760	0.1967	0.2292	0.1345	0.1614	0.1818	0.1997	0.2330

Table 4: Percentiles of RMSE (C.7) of estimators $\hat{g}_t, \hat{\nu}_t, \hat{f}_t$ in DGP4 (Exactly Identified Case)

n	T	$\hat{k}=1$	$\hat{k}=2$	$\hat{k}=3$	$\hat{k}=4$	$\hat{k}=5$	$\hat{k}=6$
1000	250	24%	76%	0%	0%	0%	0%
	500	5%	95%	0%	0%	0%	0%
	1000	5%	95%	0%	0%	0%	0%
5000	250	0%	100%	0%	0%	0%	0%
	500	0%	100%	0%	0%	0%	0%
	1000	0%	100%	0%	0%	0%	0%
10000	250	0%	100%	0%	0%	0%	0%
	500	0%	100%	0%	0%	0%	0%
	1000	0%	100%	0%	0%	0%	0%

Table 5: Percentages of selected $\hat{k}$ in the overidentified case with constant number of factors $k = 2$. DGP5 is defined in Subsection 6.1.1. Results are computed with 100 Monte Carlo replications.

Figure 1: Diagram of a single hidden layer, feed-forward neural network

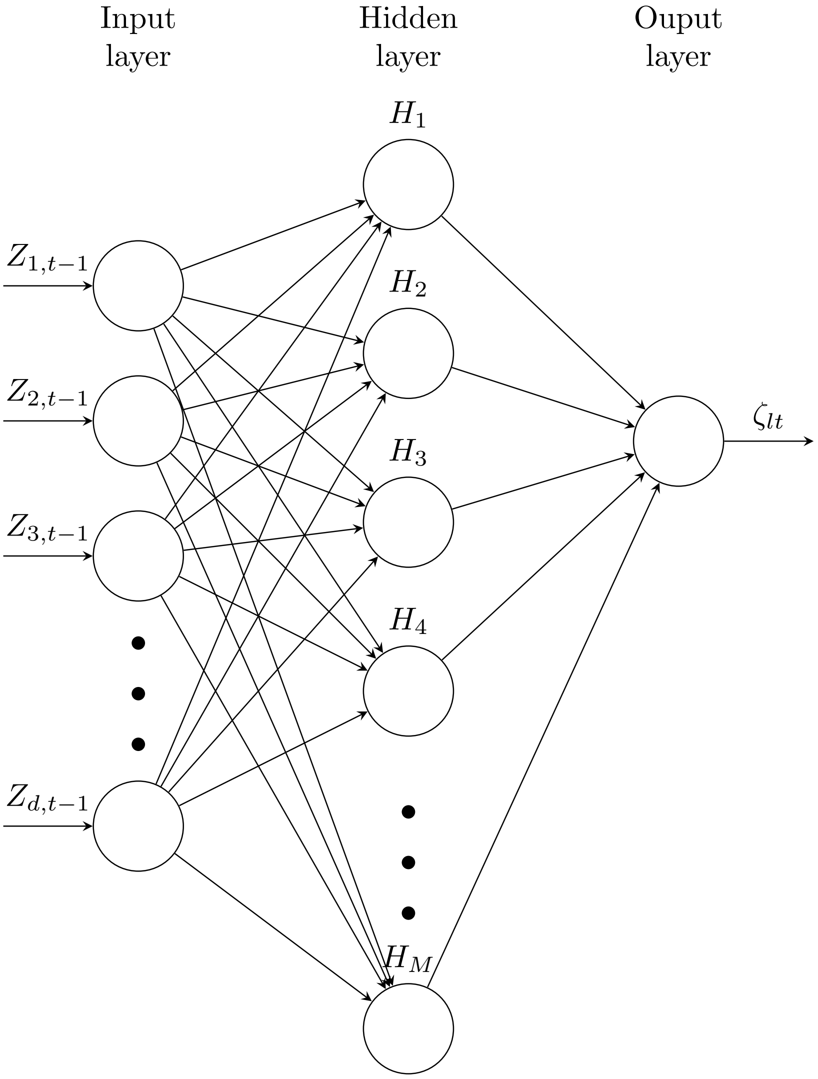
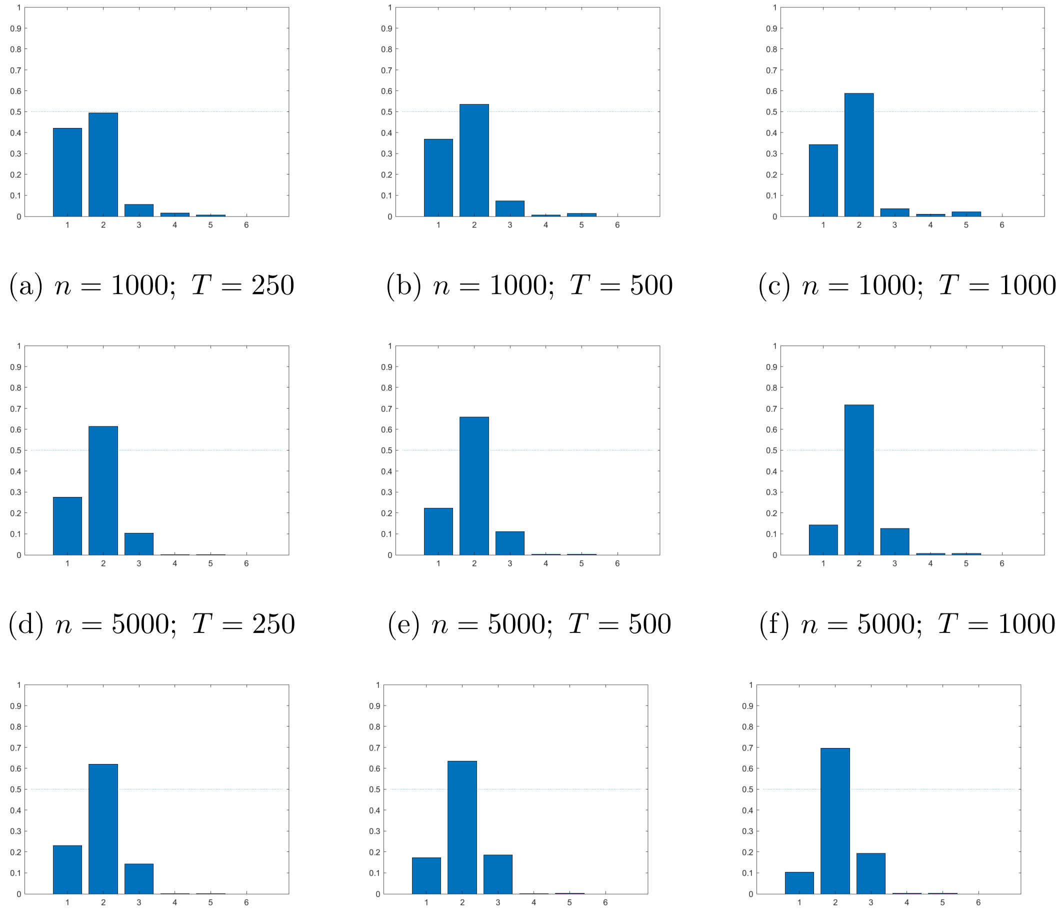


Figure 2: Histograms of $\hat{k}_t$, with time-invariant $k = 2$ (DGP5)



(a) $n = 1000;\ T = 250$

(b) $n = 1000;\ T = 500$

(c) $n = 1000;\ T = 1000$

(d) $n = 5000;\ T = 250$

(e) $n = 5000;\ T = 500$

(f) $n = 5000;\ T = 1000$

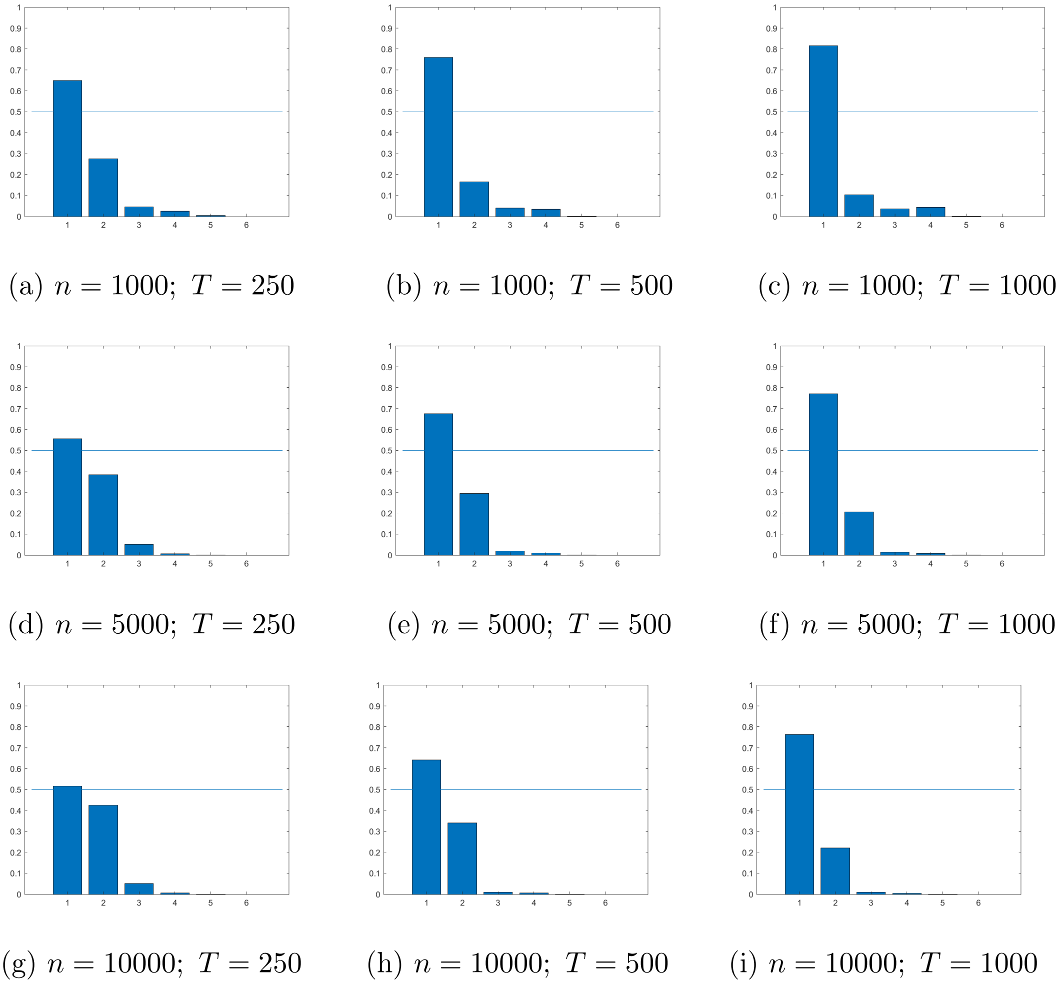
(g) $n = 10000;\ T = 250$

(h) $n = 10000;\ T = 500$

(i) $n = 10000;\ T = 1000$

The nine panels display the histograms of selected $\hat{k}_t$ across time t and Monte Carlo replications for different sample size settings. The number of true, redundant and useless factors are: $k = 2$, $k_R = 1$ and $k_W = 3$. Data are generated according to DGP5 defined in Subsection 6.1.1.

Figure 3: Histograms of $\hat{k}_t$, when $k_t = 1$ (DGP6)



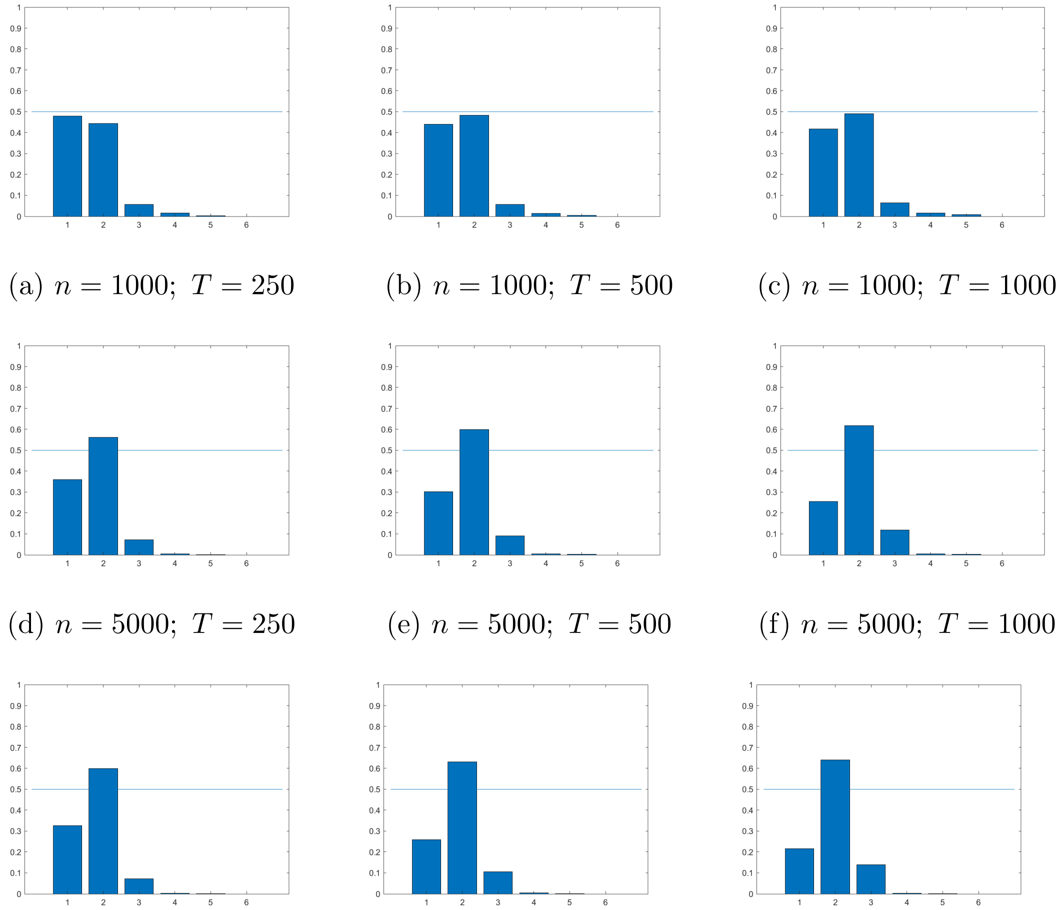
(a) $n = 1000;\ T = 250$ (b) $n = 1000;\ T = 500$ (c) $n = 1000;\ T = 1000$

(d) $n = 5000;\ T = 250$ (e) $n = 5000;\ T = 500$ (f) $n = 5000;\ T = 1000$

(g) $n = 10000;\ T = 250$ (h) $n = 10000;\ T = 500$ (i) $n = 10000;\ T = 1000$

The nine panels display the histograms of selected $\hat{k}_t$ across dates t with $k_t = 1$ and Monte Carlo replications for different sample size settings. The number of true factors is time-varying: $k_t = 2$ when $\frac{t}{T} \in [\frac{1}{3}, \frac{2}{3}]$ and $k_t = 1$ otherwise. Data are generated according to DGP6 defined in Subsection 6.2.1.

Figure 4: Histograms of $\hat{k}_t$, when $k_t = 2$ (DGP6)



(a) $n = 1000;\ T = 250$

(b) $n = 1000;\ T = 500$

(c) $n = 1000;\ T = 1000$

(d) $n = 5000;\ T = 250$

(e) $n = 5000;\ T = 500$

(f) $n = 5000;\ T = 1000$

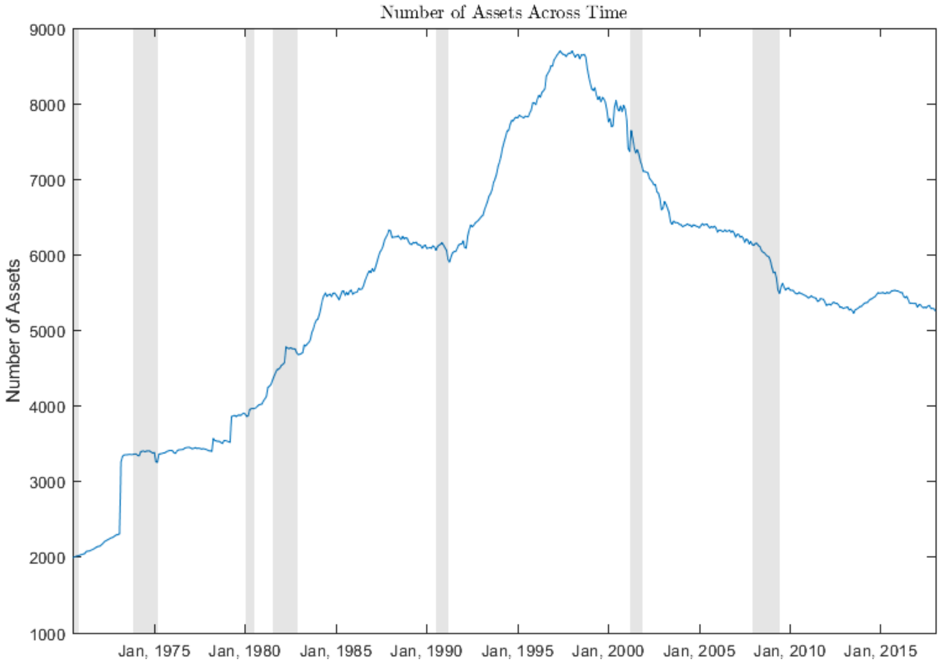
(g) $n = 10000;\ T = 250$

(h) $n = 10000;\ T = 500$

(i) $n = 10000;\ T = 1000$

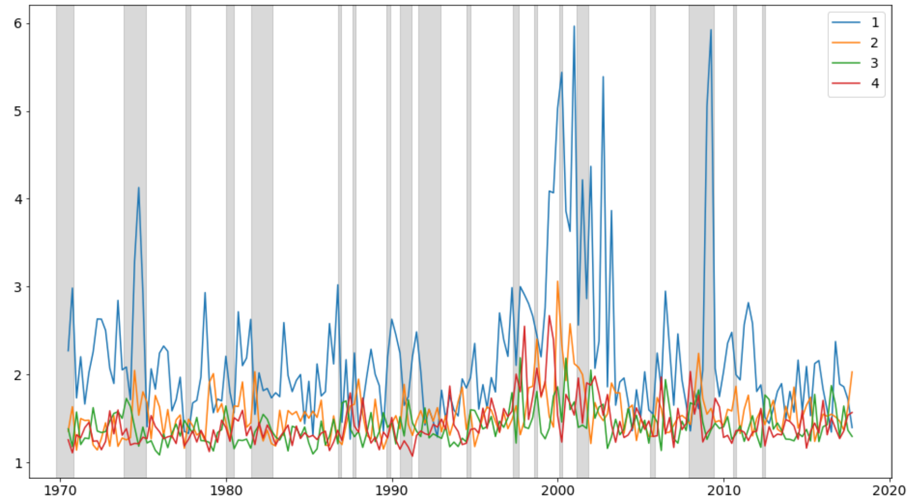
The nine panels display the histograms of selected $\hat{k}_t$ dates t with $k_t = 2$ and Monte Carlo replications for different sample size settings. The number of true factors is time-varying: $k_t = 2$ when $\frac{t}{T} \in [\frac{1}{3}, \frac{2}{3}]$ and $k_t = 1$ otherwise. Data are generated according to DGP6 defined in Subsection 6.2.1.

Figure 5: Cross-sectional sample size over time



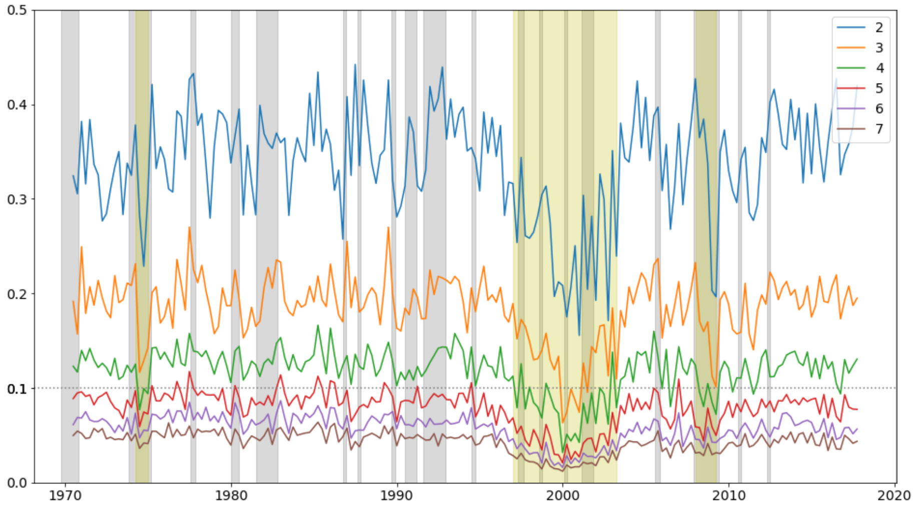
This figure displays the number of observations n_t available to compute the cross-sectional average $\hat{\xi}_t = \frac{1}{n_t}\sum_{i=1}^{n_t} w_{i,t-1}y_{i,t}$ at each month t in our sample. The vertical shaded bars denote recessions according to NBER.

Figure 6: Time series of eigenvalue ratios averaged by quarter



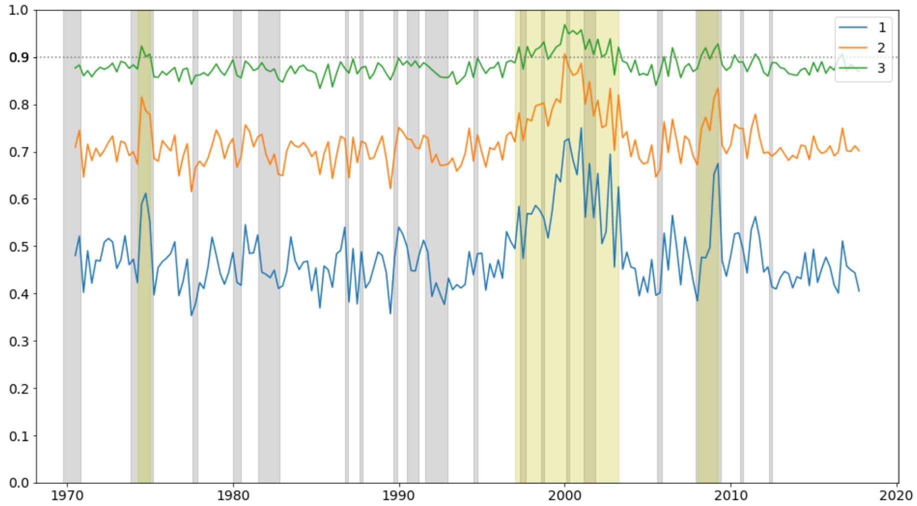
This figure displays the time series of eigenvalue ratios averaged by quarter, namely $\hat{\rho}_{r,\tau} = \frac{1}{3}\sum_{t\in\tau}\frac{\delta_r[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]}{\delta_{r+1}[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]}$, where τ denotes a quarter, for $r = 1, 2, 3, 4$. Grey vertical bars represent economic crises as from NBER and financial crises as in Zaffaroni (2019).

Figure 7: Time series of IEP ratios averaged by quarter



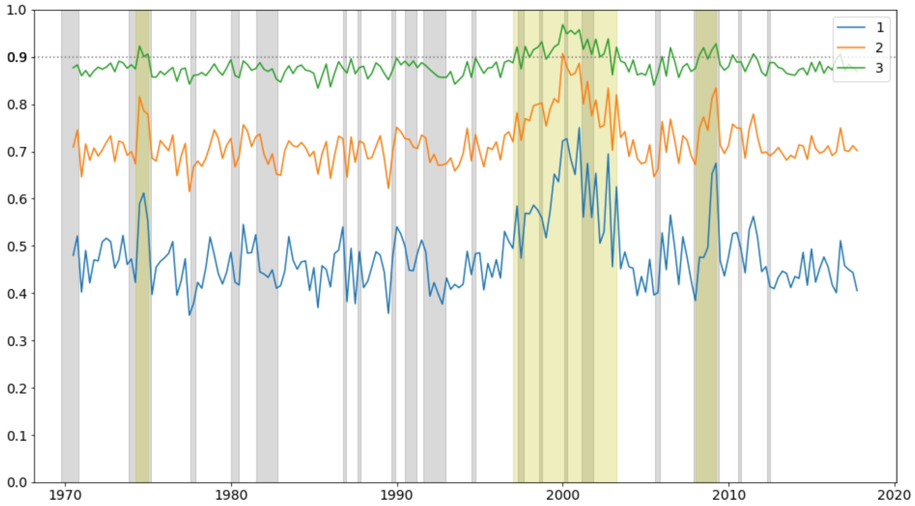
This figure displays the time series of incremental explanatory power (IEP) ratios averaged by quarter, namely $\tilde{\rho}^{I}_{r,\tau} = \frac{1}{3}\sum_{t\in\tau}\frac{\delta_r[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]}{\sum_{j=1}^{r}\delta_j[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]}$, where τ denotes a quarter, for $r = 2, 3, ..., 7$. Grey vertical bars represent economic crises as from NBER and financial crises as in Zaffaroni (2019).

Figure 8: Time series of AEP ratios averaged by quarter



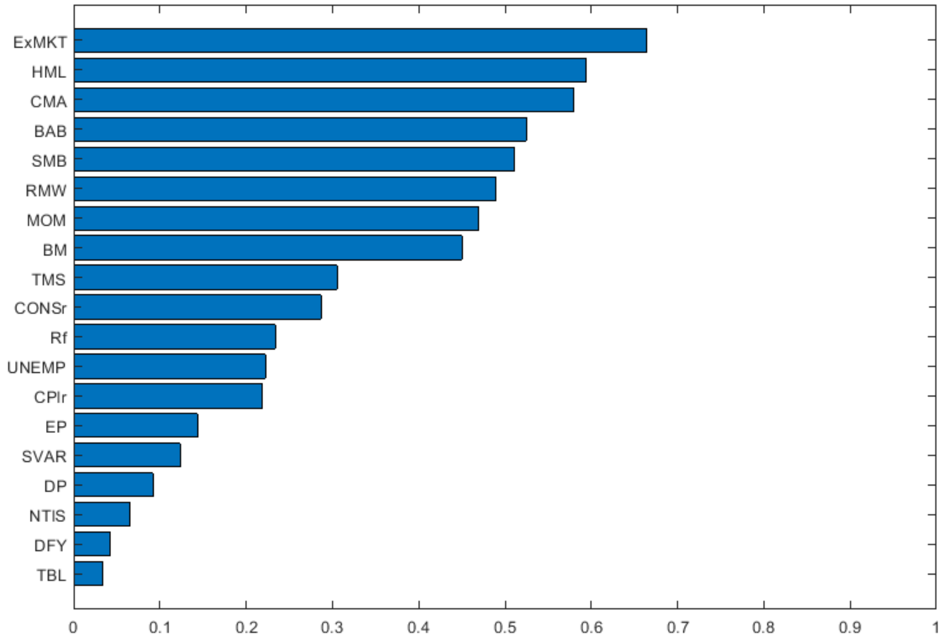
This figure displays the time series of accumulative explanatory power (AEP) ratios averaged by quarter, namely $\tilde{\rho}^{A}_{r,\tau} = \frac{1}{3}\sum_{t\in\tau}\frac{\sum_{j=1}^{r}\delta_j[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]}{\sum_{j=1}^{k_{\max}}\delta_j[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]}$ with $k_{\max}=4$, where τ denotes a quarter, for $r=1,2,3$. Grey vertical bars represent economic crises as from NBER and financial crises as in Zaffaroni (2019).

Figure 9: Time series of AEP ratios for squared eigenvalues, averaged by quarter



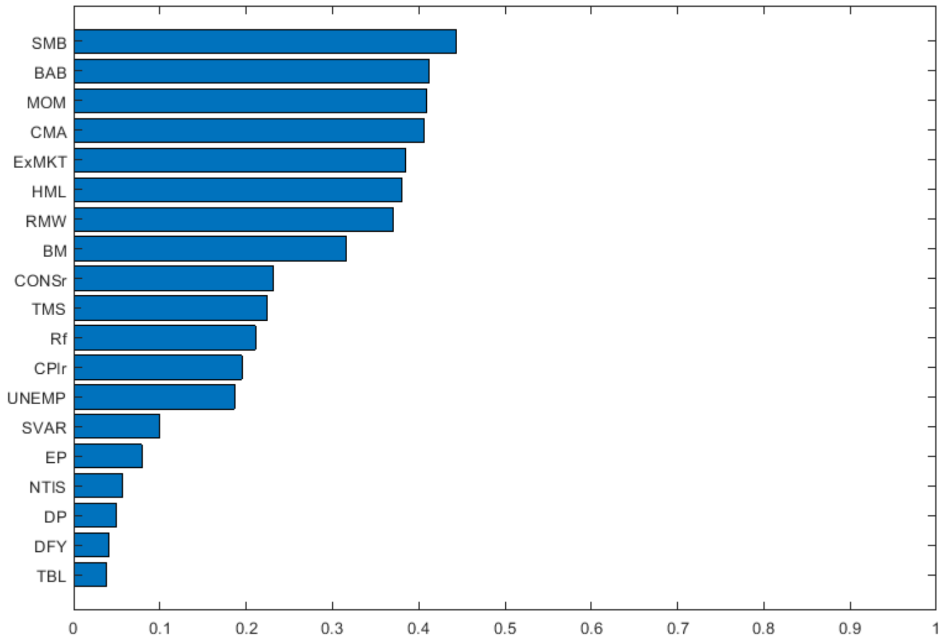
This figure displays the time series of accumulative explanatory power (AEP) ratios for squared eigenvalues, averaged by quarter, namely $\tilde{\rho}^{A,Z}_{r,\tau} = \frac{1}{3}\sum_{t\in\tau}\frac{\sum_{j=1}^{r}\delta_j[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]^2}{\sum_{j=1}^{k_{\max}}\delta_j[\hat{V}(\hat{\xi}_t|\mathcal{F}_{t-1})]^2}$ with $k_{\max} = 4$, where τ denotes a quarter, for $r = 1, 2, 3$. Grey vertical bars represent economic crises as from NBER and financial crises as in Zaffaroni (2019).

Figure 10: In-sample averages of conditional canonical correlation between $\hat{f}_{1,t}$ and one observable variable



This bar chart ranks the time averages of conditional canonical correlation between the first factor $\hat{f}_{1,t}$ and one observable variable.

Figure 11: In-sample averages of conditional canonical correlation between $\hat{f}_{2,t}$ and one observable variable



This bar chart ranks the time averages of conditional canonical correlation between the first factor $\hat{f}_{2,t}$ and one observable variable.