Journal of Econometrics 186 (2015) 294–316

Contents lists available at ScienceDirect

Journal of Econometrics

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

The three-pass regression filter: A new approach to forecasting using many predictors[☆]

Bryan Kelly[a,*], Seth Pruitt[b,c]

[a] University of Chicago, Booth School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637, United States

[b] Board of Governors of the Federal Reserve System, Washington DC, United States

[c] Arizona State University, W.P. Carey School of Business, Tempe, AZ 852877, United States

ARTICLE INFO

Article history:
Available online 17 March 2015

JEL classification:
C22
C23
C53

Keywords:
Forecast
Factor model
Principal components
Constrained least squares
Partial least squares

ABSTRACT

We forecast a single time series using many predictor variables with a new estimator called the three-pass regression filter (3PRF). It is calculated in closed form and conveniently represented as a set of ordinary least squares regressions. 3PRF forecasts are consistent for the infeasible best forecast when both the time dimension and cross section dimension become large. This requires specifying only the number of relevant factors driving the forecast target, regardless of the total number of common factors driving the cross section of predictors. The 3PRF is a constrained least squares estimator and reduces to partial least squares as a special case. Simulation evidence confirms the 3PRF's forecasting performance relative to alternatives. We explore two empirical applications: Forecasting macroeconomic aggregates with a large panel of economic indices, and forecasting stock market returns with price–dividend ratios of stock portfolios.

1. Introduction

A common interest among economists and policymakers is harnessing vast predictive information to forecast important economic aggregates like national product or stock market value. However, it can be difficult to use this wealth of information in practice. If the predictors number near or more than the number of observations, the standard ordinary least squares (OLS) forecaster is known to be poorly behaved or nonexistent.[1]

How then does one effectively use vast predictive information? A solution well known in the economics literature views the data as generated from a model in which latent factors drive the systematic variation of both the forecast target, $\boldsymbol{y}$, and the matrix of predictors, $\boldsymbol{X}$. In this setting, the best prediction of $\boldsymbol{y}$ is infeasible since the factors are unobserved. As a result, a factor estimation step is required. The literature's benchmark method extracts factors that are significant drivers of variation in $\boldsymbol{X}$ and then uses these to forecast $\boldsymbol{y}$.

Our procedure springs from the idea that the factors that are *relevant* to $\boldsymbol{y}$ may be a strict subset of all the factors driving $\boldsymbol{X}$. Our method, called the three-pass regression filter (3PRF), selectively identifies only the subset of factors that influence the forecast target while discarding factors that are irrelevant for the target but that may be pervasive among predictors. The 3PRF has the advantage of being expressed in closed form and virtually instantaneous to compute.

This paper makes four main contributions. The first is to develop asymptotic theory for the 3PRF. We begin by proving that the estimator converges in probability to the infeasible best forecast in the (simultaneous) limit as cross section size N and time series dimension T become large. This is true even when variation in predictors is dominated by target-irrelevant factors. We then derive the limiting distributions for the estimated forecasts and predictive coefficients, and provide consistent estimators of asymptotic covariance matrices that can be used to perform inference. The second contribution of the paper is to verify the

☆ We thank Andrew Ang, Gray Calhoun, John Cochrane, Drew Creal, Rob Engle, Jim Hamilton, Jim Nason, Serena Ng, Jim Stock, Eric Swanson, Matt Taddy and Mark Watson and participants at University of Chicago (Booth), NBER Summer Institute EFWW 2011, Federal Reserve Board, Econometric Society N.A. Summer Meeting 2011, SoFiE Annual Meeting 2011, NBER-NSF Time Series 2011 Conference and SITE 2011 for helpful comments. We especially thank the co-editor and two anonymous referees for comments which improved the paper. We thank Mark Watson and Domenico Giannone for kindly sharing data and code.

* Corresponding author. Tel.: +1 773 702 8359.
E-mail addresses: bryan.kelly@chicagobooth.edu (B. Kelly), seth.pruitt@asu.edu (S. Pruitt).

1 See Huber (1973) on the asymptotic difficulties of least squares when the number of regressors is large relative to the number of data points.

http://dx.doi.org/10.1016/j.jeconom.2015.02.011
0304-4076/

finite sample accuracy of our asymptotic theory through Monte Carlo simulations.

We also show that the method of partial least squares (PLS) is a special case of the 3PRF. Like partial least squares, the 3PRF can use the forecast target to discipline its dimension reduction. This emphasizes the covariance between predictors and target in the factor estimation step. But unlike PLS, the 3PRF also allows the econometrician to select additional disciplining variables, or factor proxies, on the basis of economic theory. Furthermore, because it is a special case of our methodology, the asymptotic theory we develop for the 3PRF applies directly to partial least squares. Recently Groen and Kapetanios (2009) showed the consistency of PLS under sequential N, T limits, while our approach proves consistency in the less restrictive simultaneous N, T limit. Those authors do not derive limiting distributions as we do here and so, to the best of our knowledge, our joint N and T asymptotics are new results to the PLS literature.

In our third contribution, we compare the 3PRF to other methods in order to illustrate the source of its improvement in forecasting performance. The economics literature has relied mainly on principal component regression (PCR) for forecasting problems involving many predictors, exemplified by Stock and Watson (1998, 2002a,b, 2006, 2012), Forni and Reichlin (1996, 1998), Bai and Ng (2002, 2006, 2008), Bai (2003) and Boivin and Ng (2006), among others.[2] Like the 3PRF, PCR can be calculated instantaneously for virtually any N and T. Stock and Watson's key insight is to condense information from the large cross section into a small number of predictive indices *before* estimating a linear forecast. PCR condenses the cross section according to *covariance within the predictors*. This identifies the factors driving the panel of predictors, some of which may be irrelevant for the dynamics of the forecast target, and uses those factors to forecast.

In contrast, the 3PRF condenses the cross section according to *covariance with the forecast target*. PCR must estimate *all* common factors among predictors to achieve consistency, including those that are irrelevant for forecasting. The 3PRF need only estimate the relevant factors, which are always less than or equal to the total number of factors required by PCR. While this difference is innocuous in large samples, it can be a crucial consideration in small samples.

We are not the first to investigate potential improvements upon PCR factor-based forecasts. Doz et al. (2012) propose quasi-maximum likelihood factor estimation as an alternative to PCR. Bai and Ng (2008) propose statistical thresholding rules that drop variables found to contain irrelevant information, building on the insights in Boivin and Ng (2006). In a similar vein, De Mol et al. (2008) propose Bayesian shrinkage methods. Thresholding and shrinkage methods are especially useful when relevant information is non-pervasive and confined to a subset of predictors. This does not solve the problem of pervasive irrelevant information among predictors. Our approach explicitly allows for both relevant and irrelevant pervasive factors.[3]

The final contribution of the paper is to provide empirical support for the 3PRF's strong forecasting performance in simulations and two separate empirical applications. We compare 3PRF to PCR, thresholding methods of Bai and Ng (2008), shrinkage methods of De Mol et al. (2008), and the factor analytic approach of Doz et al. (2012). Simulations show that the 3PRF often outperforms alternatives across a variety of factor model specifications. In empirical applications, we find that the 3PRF is a successful predictor of macroeconomic aggregates and equity market returns, and typically outperforms alternative methods.

The paper is structured as follows. Section 2 defines the 3PRF and proves its asymptotic properties. Section 3 reinterprets the 3PRF as a constrained least squares solution, then compares and contrasts it with partial least squares. Section 4 explores the finite sample performance of the 3PRF and other methods in Monte Carlo experiments. Section 5 reports empirical results for 3PRF and other methods' forecasts in asset pricing and macroeconomic applications. All proofs and supporting details are placed in the Appendix.

2. The three-pass regression filter

2.1. The estimator

There are several equivalent approaches to formulating our procedure, each emphasizing a related interpretation of the estimator. We begin with what we believe to be the most intuitive formulation of the filter, which is the sequence of OLS regressions that gives the estimator its name.

First we establish the environment wherein we use the 3PRF. There is a *target* variable which we wish to forecast. There exist many *predictors* which may contain information useful for predicting the target variable. The number of predictors N may be large and number near or more than the available time series observations T, which makes OLS problematic. Therefore we look to reduce the dimension of predictive information, and to do so we assume the data can be described by an approximate factor model. In order to make forecasts, the 3PRF uses *proxies*: These are variables, driven by the factors (and as we emphasize below, driven by *target-relevant* factors in particular), which we show are always available from the target and predictors themselves, but may alternatively be supplied to the econometrician on the basis of economic theory. The target is a linear function of a subset of the latent factors plus some unforecastable noise. The optimal forecast therefore comes from a regression on the true underlying relevant factors. However, since these factors are unobservable, we call this the *infeasible best forecast*.

We write $\boldsymbol{y}$ for the $T \times 1$ vector of the target variable time series from $2, 3, \ldots, T+1$.[4] Let $\boldsymbol{X}$ be the $T \times N$ matrix of predictors, $\boldsymbol{X} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_T')' = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ that have been standardized to have unit time series variance. Note that we are using two different typefaces to denote the N-dimensional cross section of predictors observed at time t ($\boldsymbol{x}_t$), and the T-dimensional time series of the ith predictor ($\mathbf{x}_i$). This is to distinguish the time series of predictors from the cross section of predictors in Table 1. We denote the $T \times L$ matrix of proxies as $\boldsymbol{Z}$, which stacks period-by-period proxy data as $\boldsymbol{Z} = (\boldsymbol{z}_1', \boldsymbol{z}_2', \ldots, \boldsymbol{z}_T')'$. We make no assumption on the relationship between N and T but assume $L \ll \min(N, T)$ in the spirit of dimension reduction. We provide additional details regarding the data generating processes for $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ in Assumption 1.

With this notation in mind, the 3PRF's regression-based construction is defined in Table 1. The first pass runs N separate *time series* regressions, one for each predictor. In these first pass regressions, the predictor is the dependent variable, the proxies are the regressors, and the estimated coefficients describe the sensitivity of the predictor to factors represented by the proxies. As we show later, proxies need not represent specific factors and

[2] The model investigated by Forni et al. (2000, 2004, 2005) concentrates on a frequency domain approach.

[3] We also demonstrate that the performance of 3PRF is robust to cases where relevant information is non-pervasive—that is, when only a subset of predictors have non-zero loadings on the relevant factors.

[4] Nothing prevents us from generalizing this to consider direct forecasts of y_{t+h} for $h \in \{1, 2, \ldots\}$—the theory is identical. For exposition's sake we deal only with y_{t+1}, knowing that $t+1$ could instead be $t+h$ but everything that follows would still hold.

Table 1
The three-pass regression filter.

Pass	Description
1.	Run time series regression of $\mathbf{x}_i$ on $\boldsymbol{Z}$ for $i = 1, \ldots, N$, $x_{i,t} = \phi_{0,i} + \boldsymbol{z}'\boldsymbol{\phi}_i + \epsilon_{it}$, retain slope estimate $\hat{\boldsymbol{\phi}}_i$.
2.	Run cross section regression of $\boldsymbol{x}_t$ on $\hat{\boldsymbol{\phi}}_i$ for $t = 1, \ldots, T$, $x_{i,t} = \phi_{0,t} + \hat{\boldsymbol{\phi}}'\boldsymbol{F}_t + \varepsilon_{it}$, retain slope estimate $\hat{\boldsymbol{F}}_t$.
3.	Run time series regression of y_{t+1} on predictive factors $\hat{\boldsymbol{F}}_t$, $y_{t+1} = \beta_0 + \hat{\boldsymbol{F}}'\boldsymbol{\beta} + \eta_{t+1}$, delivers forecast $\hat{y}_{t+1}$.

Notes: All regressions use OLS.

may be measured with noise. The important requirement is that their common components span the space of the target-relevant factors.

The second pass uses the estimated first-pass coefficients in T separate *cross section* regressions. In these second pass regressions, the predictors are again the dependent variable while the first-pass coefficients $\hat{\boldsymbol{\phi}}_i$ are the regressors. Fluctuations in the latent factors cause the cross section of predictors to fan out and compress over time. First-stage coefficient estimates map the cross-sectional distribution of predictors to the latent factors. Second-stage cross section regressions use this map to back out estimates of the factors at each point in time.[5]

We then carry forward the estimated second-pass predictive factors $\hat{\boldsymbol{F}}_t$ to the third pass. This is a single *time series* forecasting regression of the target variable y_{t+1} on the second-pass estimated predictive factors $\hat{\boldsymbol{F}}_t$. The third-pass fitted value $\hat{\beta}_0+\hat{\boldsymbol{F}}_t'\hat{\boldsymbol{\beta}}$ is the 3PRF time t forecast. Because the first-stage regression takes an errors-in-variables form, second-stage regressions produce an estimate for a unique but unknown rotation of the latent factors. Since the relevant factor space is spanned by $\hat{\boldsymbol{F}}_t$, the third-stage regression delivers consistent forecasts.

An alternative representation for the 3PRF is the one-step closed form:

$$\hat{\boldsymbol{y}} = \boldsymbol{\iota}_T\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\boldsymbol{W}_{XZ}\left(\boldsymbol{W}_{XZ}'\boldsymbol{S}_{XX}\boldsymbol{W}_{XZ}\right)^{-1}\boldsymbol{W}_{XZ}'\boldsymbol{s}_{Xy} \tag{1}$$

where $\boldsymbol{J}_T \equiv \boldsymbol{I}_T - \frac{1}{T}\boldsymbol{\iota}_T\boldsymbol{\iota}_T'$ for $\boldsymbol{I}_T$ the T-dimensional identity matrix and $\boldsymbol{\iota}_T$ the T-vector of ones ($\boldsymbol{J}_N$ is analogous), $\bar{y} = \boldsymbol{\iota}_T'\boldsymbol{y}/T$, $\boldsymbol{W}_{XZ} \equiv \boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}$, $\boldsymbol{S}_{XX} \equiv \boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}$ and $\boldsymbol{s}_{Xy} \equiv \boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}$. $\boldsymbol{J}$ matrices enter because each regression pass is run with a constant.

The closed form is central to the theoretical development that follows. Nonetheless, the regression-based procedure in Table 1 remains useful for two reasons. First, in practice (particularly with many predictors) one often faces unbalanced panels and missing data. The 3PRF as described in Table 1 easily handles these difficulties. Second, it is useful for developing intuition behind the procedure and for understanding its relation to partial least squares.

We can rewrite the forecast as

$$\hat{\boldsymbol{y}} = \boldsymbol{\iota}_T\bar{y} + \hat{\boldsymbol{F}}\hat{\boldsymbol{\beta}}$$

$$\hat{\boldsymbol{F}}' = \boldsymbol{S}_{ZZ}\left(\boldsymbol{W}_{XZ}'\boldsymbol{S}_{XZ}\right)^{-1}\boldsymbol{W}_{XZ}'\boldsymbol{X}', \tag{2}$$

$$\hat{\boldsymbol{\beta}} = \boldsymbol{S}_{ZZ}\boldsymbol{W}_{XZ}\boldsymbol{S}_{XZ}\left(\boldsymbol{W}_{XZ}'\boldsymbol{S}_{XX}\boldsymbol{W}_{XZ}\right)^{-1}\boldsymbol{W}_{XZ}'\boldsymbol{s}_{Xy} \tag{3}$$

where $\boldsymbol{S}_{XZ} \equiv \boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}$. Here we interpret $\hat{\boldsymbol{F}}$ as our predictive factor and $\hat{\boldsymbol{\beta}}$ the predictive coefficient on that factor. Since we have used the N predictors to construct a L-dimensional predictive factor, the 3PRF reduces the dimension of the forecasting problem.

Alternatively, we can rewrite the forecast

$$\hat{\boldsymbol{y}} = \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\hat{\boldsymbol{\alpha}}$$

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{W}_{XZ}\left(\boldsymbol{W}_{XZ}'\boldsymbol{S}_{XX}\boldsymbol{W}_{XZ}'\right)^{-1}\boldsymbol{W}_{XZ}'\boldsymbol{s}_{Xy} \tag{4}$$

interpreting $\hat{\boldsymbol{\alpha}}$ as the predictive coefficient on individual predictors. The regular OLS estimate of the projection coefficient $\boldsymbol{\alpha}$ is $(\boldsymbol{S}_{XX})^{-1}\boldsymbol{s}_{Xy}$. This representation suggests that our approach can be interpreted as a constrained version of least squares (Theorem 8 shows this formally below). We further discuss the properties of these estimators in Sections 2.3 and 2.4 after presenting our assumptions in the next subsection.

2.2. Assumptions

We next detail the modeling assumptions that provide a groundwork for developing asymptotic properties of the 3PRF.

Assumption 1 (*Factor Structure*)**.** The data are generated by the following:

$$\begin{aligned} &\boldsymbol{x}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t \quad && y_{t+1} = \beta_0 + \boldsymbol{\beta}'\boldsymbol{F}_t + \eta_{t+1} \quad && \boldsymbol{z}_t = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{F}_t + \boldsymbol{\omega}_t \\ &\boldsymbol{X} = \boldsymbol{\iota}\boldsymbol{\phi}_0' + \boldsymbol{F}\boldsymbol{\Phi}' + \boldsymbol{\varepsilon} && \boldsymbol{y} = \boldsymbol{\iota}\beta_0 + \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\eta} && \boldsymbol{Z} = \boldsymbol{\iota}\boldsymbol{\lambda}_0' + \boldsymbol{F}\boldsymbol{\Lambda}' + \boldsymbol{\omega} \end{aligned}$$

where $\boldsymbol{F}_t = (\boldsymbol{f}_t', \boldsymbol{g}_t')'$, $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_f, \boldsymbol{\Phi}_g)$, $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_f, \boldsymbol{\Lambda}_g)$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_f', \boldsymbol{0}')'$ with $|\boldsymbol{\beta}_f| > \boldsymbol{0}$. $K_f > 0$ is the dimension of vector $\boldsymbol{f}_t$, $K_g \geq 0$ is the dimension of vector $\boldsymbol{g}_t$, L is the dimension of vector $\boldsymbol{z}_t (0 < L < \min(N, T))$, and $K = K_f + K_g$.

Assumption 1 defines the factor structure. The target's factor loadings ($\boldsymbol{\beta} = (\boldsymbol{\beta}_f', \boldsymbol{0}')'$) allow the target to depend on a strict subset of the factors driving the predictors. We refer to this subset as the *relevant* factors, which are denoted $\boldsymbol{f}_t$. In contrast, *irrelevant* factors, $\boldsymbol{g}_t$, do not influence the forecast target but may drive the cross section of predictive information $\boldsymbol{x}_t$. The proxies $\boldsymbol{z}_t$ are driven by factors and proxy noise.

Assumption 2 (*Factors, Loadings and Residuals*)**.** Let $M < \infty$. For any i, s, t

1. $\mathbb{E}\|\boldsymbol{F}_t\|^4 < M$, $T^{-1}\sum_{s=1}^{T}\boldsymbol{F}_s \xrightarrow[T\to\infty]{p} \boldsymbol{\mu}$ and $T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F} \xrightarrow[T\to\infty]{p} \boldsymbol{\Delta}_F$
2. $\mathbb{E}\|\boldsymbol{\phi}_i\|^4 \leq M$, $N^{-1}\sum_{j=1}^{N}\boldsymbol{\phi}_j \xrightarrow[T\to\infty]{p} \bar{\boldsymbol{\phi}}$, $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi} \xrightarrow[N\to\infty]{p} \mathcal{P}$ and $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\phi}_0 \xrightarrow[N\to\infty]{p} \boldsymbol{P}_1$[6]
3. $\mathbb{E}(\varepsilon_{it}) = 0$, $\mathbb{E}|\varepsilon_{it}|^8 \leq M$
4. $\mathbb{E}(\boldsymbol{\omega}_t) = \boldsymbol{0}$, $\mathbb{E}\|\boldsymbol{\omega}_t\|^4 \leq M$, $T^{-1/2}\sum_{s=1}^{T}\boldsymbol{\omega}_s = \boldsymbol{O}_p(1)$ and $T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\omega} \xrightarrow[N\to\infty]{p} \boldsymbol{\Delta}_\omega$
5. $\mathbb{E}_t(\eta_{t+1}) = \mathbb{E}(\eta_{t+1}|y_t, F_t, y_{t-1}, F_{t-1}, \ldots) = 0$, $\mathbb{E}(\eta_{t+1}^4) \leq M$, and η_{t+1} is independent of $\phi_i(m)$ and $\varepsilon_{i,t}$.

Since η_{t+1} is a martingale difference sequence with respect to all information known at time t, $\beta_0 + \boldsymbol{\beta}_f'\boldsymbol{f}_t$ gives the best time t forecast. But it is infeasible since the relevant factors $\boldsymbol{f}_t$ are unobserved.

We require factors and loadings to be cross-sectionally regular in that they have well-behaved covariance matrices for large T and N, respectively. Assumption 2.4 does not exist in the work of Stock and Watson or Bai and Ng, and is required because the 3PRF uses proxies to extract factors. We bound the moments of proxy noise $\boldsymbol{\omega}_t$ in the same manner as the bounds on factor moments.

[5] If coefficients were observable, this mapping would be straightforward since factors could be directly estimated each period with cross section regressions of predictors on the loadings. While the loadings in our framework are unobservable, the same intuition for recovering the factor space applies to our cross section regressions. The difference is that we use estimated loadings as stand-ins for the unobservable true loadings.

[6] $\|\boldsymbol{\phi}_i\| \leq M$ can replace $\mathbb{E}\|\boldsymbol{\phi}_i\|^4 \leq M$ if $\boldsymbol{\phi}_i$ is non-stochastic.

Assumption 3 (*Dependence*). Let $x(m)$ denote the mth element of $\boldsymbol{x}$. For $M < \infty$ and any i, j, t, s, m_1, m_2

1. $\mathbb{E}(\varepsilon_{it}\varepsilon_{js}) = \sigma_{ij,ts}$, $|\sigma_{ij,ts}| \leq \bar{\sigma}_{ij}$ and $|\sigma_{ij,ts}| \leq \tau_{ts}$, and
   (a) $N^{-1}\sum_{i,j=1}^{N}\bar{\sigma}_{ij} \leq M$
   (b) $T^{-1}\sum_{t,s=1}^{T}\tau_{ts} \leq M$
   (c) $N^{-1}\sum_{i,s}|\sigma_{ii,ts}| \leq M$
   (d) $N^{-1}T^{-1}\sum_{i,j,t,s}|\sigma_{ij,ts}| \leq M$
2. $\mathbb{E}\left|N^{-1/2}T^{-1/2}\sum_{s=1}^{T}\sum_{i=1}^{N}\left[\varepsilon_{is}\varepsilon_{it} - \mathbb{E}\left(\varepsilon_{is}\varepsilon_{it}\right)\right]\right|^2 \leq M$
3. $\mathbb{E}\left|T^{-1/2}\sum_{t=1}^{T}F_t(m_1)\omega_t(m_2)\right|^2 \leq M$
4. $\mathbb{E}\left|T^{-1/2}\sum_{t=1}^{T}\omega_t(m_1)\varepsilon_{it}\right|^2 \leq M.$

Assumption 4 (*Central Limit Theorems*). For any i, t

1. $N^{-1/2}\sum_{i=1}^{N}\boldsymbol{\phi}_i\varepsilon_{it} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Gamma}_{\Phi\varepsilon})$, where $\boldsymbol{\Gamma}_{\Phi\varepsilon} = \text{plim}_{N\to\infty}N^{-1}\sum_{i,j=1}^{N}\mathbb{E}\left[\boldsymbol{\phi}_i\boldsymbol{\phi}_j'\varepsilon_{it}\varepsilon_{jt}\right]$
2. $T^{-1/2}\sum_{t=1}^{T}\boldsymbol{F}_t\eta_{t+1} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\Gamma}_{F\eta}\right)$, where $\boldsymbol{\Gamma}_{F\eta} = \text{plim}_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[\eta_{t+1}^2\boldsymbol{F}_t\boldsymbol{F}_t'\right] > 0$
3. $T^{-1/2}\sum_{t=1}^{T}\boldsymbol{F}_t\varepsilon_{it} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\Gamma}_{F\varepsilon,i}\right)$, where $\boldsymbol{\Gamma}_{F\varepsilon,i} = \text{plim}_{T\to\infty}T^{-1}\sum_{t,s=1}^{T}\mathbb{E}\left[\boldsymbol{F}_t\boldsymbol{F}_s'\varepsilon_{it}\varepsilon_{is}\right] > 0.$

Assumption 3 allows the factor structure to be approximate in the sense that some cross section correlation among ε_{it} is permitted, following Chamberlain and Rothschild (1983). Similarly, we allow for serial dependence among ε_{it} (including GARCH) as in Stock and Watson (2002a). In addition, we allow some proxy noise dependence with factors and idiosyncratic shocks. Assumption 4 requires that central limit theorems apply, and is satisfied when various mixing conditions hold among factors, loadings and shocks.

Assumption 5 (*Normalization*). $\mathcal{P} = \boldsymbol{I}, \boldsymbol{P}_1 = \boldsymbol{0}$ and $\boldsymbol{\Delta}_F$ is diagonal, positive definite, and each diagonal element is unique.

Assumption 5 recognizes that there exists an inherent unidentification between the factors and factor loadings.[7] It therefore selects a normalization in which the covariance of predictor loadings is the identity matrix, and in which factors are orthogonal to one another. As with principal components, the particular normalization is unimportant. We ultimately estimate a vector space spanned by the factors, and this space does not depend upon the choice of normalization.

Assumption 6 (*Relevant Proxies*). $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_f, \boldsymbol{0}]$ and $\boldsymbol{\Lambda}_f$ is nonsingular.

Assumption 6 states that proxies (i) have zero loading on irrelevant factors, (ii) have linearly independent loadings on the relevant factors, and (iii) number equal to the number of relevant factors. Combined with the normalization assumption, this says that the common component of proxies spans the relevant factor space, and that none of the proxy variation is due to irrelevant factors.

Note that Assumptions 2.4, 3.3 3.4 and 6 are the only conditions involving the proxy variables. We prove in Theorem 7 that automatic proxies, which are generally constructable using $\boldsymbol{X}$ and $\boldsymbol{y}$, are guaranteed to satisfy these proxy assumptions.

With these assumptions in place, we derive the asymptotic properties of the three-pass regression filter. Our proofs build upon the seminal theory of Stock and Watson (2002a), Bai (2003) and Bai and Ng (2002, 2006). Portions of auxiliary lemmas in the appendix draw directly from convergence results proved in these previous papers. Theorems reported in the main text are our central new results. In order to keep our theoretical development self-contained, we catalogue all theoretical results in the Appendix.

2.3. Consistency

Theorem 1. *Let Assumptions 1–6 hold. The three-pass regression filter forecast is consistent for the infeasible best forecast,* $\hat{y}_{t+1} \xrightarrow[T,N\to\infty]{p} \beta_0 + \boldsymbol{F}_t'\boldsymbol{\beta}$.

Theorem 1 says that the 3PRF is consistent so that for large N and T the difference between this feasible forecast and the infeasible best vanishes. This and our other asymptotic results are based on simultaneous N and T limits. As discussed by Bai (2003), the existence of a simultaneous limit implies the existence of coinciding sequential and pathwise limits, but the converse is not true. We refer readers to that paper for a more detailed comparison of these three types of joint limits.

The Appendix also establishes probability limits of first pass time series regression coefficients $\hat{\boldsymbol{\phi}}_i$, second pass cross section factor estimates $\hat{\boldsymbol{F}}_t$, and third stage predictive coefficients $\hat{\boldsymbol{\beta}}$. While primarily serving as intermediate inputs to the proof of Theorem 1, in certain applications these probability limits are useful in their own right. We refer interested readers to Lemmas 3 and 4 in the Appendix.

The estimated loadings on individual predictors, $\hat{\boldsymbol{\alpha}}$, play an important role in the interpretation of the 3PRF. The next theorem provides the probability limit for the loading on each predictor i.

Theorem 2. *Let $\hat{\alpha}_i$ denote the ith element of $\hat{\boldsymbol{\alpha}}$, and let Assumptions 1–6 hold. Then for any i,*

$$N\hat{\alpha}_i \xrightarrow[T,N\to\infty]{p} \left(\boldsymbol{\phi}_i - \bar{\boldsymbol{\phi}}\right)'\boldsymbol{\beta}.$$

The coefficient $\boldsymbol{\alpha}$ maps underlying factors to the forecast target via the observable predictors. As a result the probability limit of $\hat{\boldsymbol{\alpha}}$ is a product of the loadings of $\boldsymbol{X}$ and $\boldsymbol{y}$ on the relevant factors $\boldsymbol{f}$. This arises from the interpretation of $\hat{\boldsymbol{\alpha}}$ as a constrained least squares coefficient estimate, which we elaborate on in the next section. Note that $\hat{\boldsymbol{\alpha}}$ is multiplied by N in order to derive its limit. This is because the dimension of $\hat{\boldsymbol{\alpha}}$ grows with the number of predictors. As N grows, the predictive information in $\boldsymbol{f}$ is spread across a larger number of predictors so each predictor's contribution approaches zero. Standardizing by N is necessary to identify the non-degenerate limit.

What distinguishes these results from previous work using PCR is the fact that the 3PRF uses only as many predictive factors as the number of factors relevant to y_{t+1}. In contrast, the PCR forecast is asymptotically efficient when there are as many predictive factors as the total number of factors driving $\boldsymbol{x}_t$ (Stock and Watson, 2002a). This distinction is especially important when the number of relevant factors is strictly less than the number of total factors in the predictor data and the target-relevant principal components are dominated by other components in $\boldsymbol{x}_t$. In particular, if the factors driving the target are weak in

---

[7] Stock and Watson (2002a) summarize this point (we have replaced their symbols with our notation):

> [B]ecause $\boldsymbol{\Phi}\boldsymbol{F}_t = \boldsymbol{\Phi}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{F}_t$ for any nonsingular matrix $\boldsymbol{R}$, a normalization is required to uniquely define the factors. Said differently, the model with factor loadings $\boldsymbol{\Phi}\boldsymbol{R}$ and factors $\boldsymbol{R}^{-1}\boldsymbol{F}_t$ is observationally equivalent to the model with factor loadings $\boldsymbol{\Phi}$ and factors $\boldsymbol{F}_t$. Assumption 5 restricts $\boldsymbol{R}$ to be orthonormal and... restricts $\boldsymbol{R}$ to be a diagonal matrix with diagonal elements of ± 1.

We further discuss our normalization assumption in Appendix A.7. There we prove that a necessary condition for convergence to the infeasible best forecast is that the number of relevant proxies equals the number of relevant factors.

the sense that they contribute a only small fraction of the total variability in the predictors, then principal components may have difficulty identifying them. Said another way, there is no sense in which the method of principal components is assured to *first* extract predictive factors that are relevant to y_{t+1}. This point has in part motivated recent econometric work on thresholding (Bai and Ng, 2008) and shrinking (Stock and Watson, 2012) principal components for the purposes of forecasting.

On the other hand, the 3PRF identifies exactly those relevant factors in its second pass factor estimation. This step extracts *leading* indicators—estimated factors that are specifically valuable for forecasting a given target. To illustrate how this works, consider the special case in which there is only one relevant factor, and the sole proxy is the target variable y_{t+1} itself. We refer to this case as the *target-proxy three-pass regression filter*. The following corollary is immediate from Theorem 1.

Corollary 1. *Let Assumptions 1–5 hold with the exception of Assumptions 2.4, 3.3 and 3.4. Additionally, assume that there is only one relevant factor. Then the target-proxy three-pass regression filter forecaster is consistent for the infeasible best forecast.*

Corollary 1 holds regardless of the number of irrelevant factors driving $\boldsymbol{X}$ and regardless of where the relevant factor stands in the principal component ordering for $\boldsymbol{X}$. Compare this to PCR, whose first predictive factor is ensured to be the one that explains most of the covariance among $\boldsymbol{x}_t$, regardless of that factor's relationship to y_{t+1}. Only if the relevant factor happens to also drive most of the variation within the predictors does the first component achieve the infeasible best. It is in this sense that the forecast performance of the 3PRF is robust to the presence of irrelevant factors.

2.4. Asymptotic distributions

Not only is the 3PRF consistent for the infeasible best forecast, each forecast has a normal asymptotic distribution. We first derive the asymptotic distribution for $\hat{\boldsymbol{\alpha}}$ since this is useful for establishing the asymptotic distribution of forecasts.

Theorem 3. *Under Assumptions 1–6, as $N, T \to \infty$ we have*

$$\frac{\sqrt{T}N\left(\hat{\alpha}_i - \tilde{\alpha}_i\right)}{A_i} \xrightarrow{d} \mathcal{N}(0, 1)$$

where A_i^2 is the ith diagonal element of $\widehat{Avar}(\hat{\boldsymbol{\alpha}}) = \boldsymbol{\Omega}_\alpha\big(\frac{1}{T}\sum_t \hat{\eta}_{t+1}^2(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'\big)\boldsymbol{\Omega}_\alpha'$, $\hat{\eta}_{t+1}$ is the estimated 3PRF forecast error, $\tilde{\alpha}_i \equiv \boldsymbol{S}_i\boldsymbol{G}_\alpha\boldsymbol{\beta}$, where $\boldsymbol{S}_i$ is selects the ith element of vector $\boldsymbol{G}_\alpha\boldsymbol{\beta}$ and

$$\boldsymbol{G}_\alpha = \boldsymbol{J}_N\left(T^{-1}\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)\left(T^{-3}N^{-2}\boldsymbol{W}_{XZ}'\boldsymbol{S}_{XX}\boldsymbol{W}_{XZ}\right)^{-1}\left(N^{-1}T^{-2}\boldsymbol{W}_{XZ}'\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{F}\right),$$

and

$$\boldsymbol{\Omega}_\alpha = \boldsymbol{J}_N\left(\frac{1}{T}\boldsymbol{S}_{XZ}\right)\left(\frac{1}{T^3N^2}\boldsymbol{W}_{XZ}'\boldsymbol{S}_{XX}\boldsymbol{W}_{XZ}\right)^{-1}\left(\frac{1}{TN}\boldsymbol{W}_{XZ}'\right).$$

While Theorem 2 demonstrates that $\hat{\boldsymbol{\alpha}}$ may be used to measure the relative forecast contribution of each predictor, Theorem 3 offers a distribution theory, including feasible t-statistics, for inference. The $\boldsymbol{G}_\alpha$ matrix appears here because the factors are only identified up to an orthonormal rotation.

From here, we derive the asymptotic distribution of the 3PRF forecasts.

Theorem 4. *Under Assumptions 1–6, as $N, T \to \infty$ we have*

$$\frac{\sqrt{T}\left(\hat{y}_{t+1} - \mathbb{E}_t y_{t+1}\right)}{Q_t} \xrightarrow{d} \mathcal{N}(0, 1)$$

where $\mathbb{E}_t y_{t+1} = \beta_0 + \boldsymbol{\beta}'\boldsymbol{F}_t$ and Q_t^2 is the tth diagonal element of $\frac{1}{N^2}\boldsymbol{J}_T\boldsymbol{X}\widehat{Avar}(\hat{\boldsymbol{\alpha}})\boldsymbol{X}'\boldsymbol{J}_T$.

This result shows that besides being consistent for the infeasible best forecast $\mathbb{E}_t(y_{t+1}) \equiv \beta_0 + \boldsymbol{\beta}'\boldsymbol{F}_t$, the 3PRF forecast is asymptotically normal and provides a standard error estimator for constructing forecast confidence intervals. A subtle but interesting feature of this result is that we only need the asymptotic variance of individual predictor loadings $\widehat{Avar}(\hat{\boldsymbol{\alpha}})$ for the prediction intervals. This differs from the confidence intervals of PCR forecasts in Bai and Ng (2006), which require an estimate of the asymptotic variance for the predictive factor loadings (the analogue of our $\widehat{Avar}(\hat{\boldsymbol{\beta}})$ below) as well as an estimate for the asymptotic variance of the fitted latent factors, $\widehat{Avar}(\hat{\boldsymbol{F}})$. Unlike PCR, our framework allows us to represent loadings on individual predictors in a convenient algebraic form, $\hat{\boldsymbol{\alpha}}$. Inspection of $\hat{\boldsymbol{\alpha}}$ reveals why variability in both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{F}}$ is captured by $\widehat{Avar}(\hat{\boldsymbol{\alpha}})$.

Next, we provide the asymptotic distribution of predictive loadings on the latent factors and a consistent estimator of their asymptotic covariance matrix.

Theorem 5. *Under Assumptions 1–6, as $N, T \to \infty$ we have*

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{G}_\beta\boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\beta\right)$$

where $\boldsymbol{\Sigma}_\beta = \boldsymbol{\Sigma}_z^{-1}\boldsymbol{\Gamma}_{F\eta}\boldsymbol{\Sigma}_z^{-1}$ and $\boldsymbol{\Sigma}_z = \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega$. Furthermore,

$$\widehat{Avar}(\hat{\boldsymbol{\beta}}) = \left(T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1}T^{-1}\sum_t \hat{\eta}_{t+1}^2(\hat{\boldsymbol{F}}_t - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{F}}_t - \hat{\boldsymbol{\mu}})' \times \left(T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1}$$

is a consistent estimator of $\boldsymbol{\Sigma}_\beta$. $\boldsymbol{G}_\beta$ is defined in the Appendix.

We also derive the asymptotic distribution of the estimated relevant latent factor rotation.

Theorem 6. *Under Assumptions 1–6, as $N, T \to \infty$ we have for every t*

(i) *if $\sqrt{N}/T \to 0$, then*

$$\sqrt{N}\left[\hat{\boldsymbol{F}}_t - (\boldsymbol{H}_0 + \boldsymbol{H}\boldsymbol{F}_t)\right] \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_F\right)$$

(ii) *if $\liminf \sqrt{N}/T \geq \tau \geq 0$, then*

$$T\left[\hat{\boldsymbol{F}}_t - (\boldsymbol{H}_0 + \boldsymbol{H}\boldsymbol{F}_t)\right] = \boldsymbol{O}_p(1)$$

where $\boldsymbol{\Sigma}_F = \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega\right)\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F^2\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Gamma}_{\Phi\varepsilon}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F^2\boldsymbol{\Lambda}')^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega\right)$. $\boldsymbol{H}_0$ and $\boldsymbol{H}$ are defined in the Appendix.

The matrices $\boldsymbol{G}_\beta$ and $\boldsymbol{H}$ are present since we are in effect estimating a vector space. Quoting Bai and Ng (2006), Theorems 5 and 6 in fact "pertain to the difference between $[\hat{\boldsymbol{F}}_t/\hat{\boldsymbol{\beta}}]$ and the space spanned by $[\boldsymbol{F}_t/\boldsymbol{\beta}]$". Note that we do not provide an estimator the asymptotic variance of $\hat{\boldsymbol{F}}$. While under some circumstances such an estimator is available, this is not generally the case. In particular, when there exist irrelevant factors driving the predictors, the 3PRF only estimates the relevant factor subspace. This complicates the construction of a consistent estimator of $Avar(\hat{\boldsymbol{F}})$. Estimators for the asymptotic variance of $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{y}_{t+1}$ do not confront this difficulty for reasons discussed following Theorem 4.

2.5. Proxy selection

The formulation of the filter, and its success in forecasting even when principal components that dominate cross section variation are irrelevant to the forecast target, relies on the existence of proxies that depend only on target-relevant factors. This begs

Table 2
Automatic proxy-selection algorithm.

0.	Initialize $\boldsymbol{r}_0 = \boldsymbol{y}$.
	For $k = 1, \ldots, L$:
1.	Define the kth automatic proxy to be $\boldsymbol{r}_{k-1}$. Stop if $k = L$; otherwise proceed.
2.	Compute the 3PRF for target $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ using statistical proxies 1 through k. Denote the resulting forecast $\hat{\boldsymbol{y}}_k$.
3.	Calculate $\boldsymbol{r}_k = \boldsymbol{y} - \hat{\boldsymbol{y}}_k$, advance k, and go to step 1.

the question: Need we make an *a priori* assumption about the availability of such proxies? The answer is no—there always exist readily available proxies that satisfy the relevance criterion of Assumption 6. They are obtained from an *automatic proxy selection algorithm* which constructs proxies that depend *only* upon relevant factors. For now we treat the true number of relevant factors as known, and return to a discussion of statistical criteria for selecting the appropriate number of 3PRF factors in Section 4.2.

2.5.1. Automatic proxies

By definition, the target variable depends only on the relevant factors and therefore satisfies Assumptions 2.4, 3.3, 3.4 and 6 when there is one relevant factor ($K_f = 1$). This logic is exploited to prove Corollary 1. If $K_f > 1$, the target-proxy 3PRF does not extract enough factors to asymptotically attain the infeasible best.[8] In this case we can improve upon the target-proxy 3PRF by selecting additional proxies that depend only on relevant factors. We obtain the second proxy by noting that residuals from target-proxy 3PRF forecasts also satisfy Assumption 6 since they have non-zero loading on relevant factors (which follows from the insufficiency of the target-only proxy), have zero loading on irrelevant factors (by definition), and are linearly independent of the first proxy. From here, proxy construction proceeds iteratively: Use the residual from the target-proxy 3PRF as the second proxy, use the residual from this two-proxy 3PRF as the third proxy, etc. The details of the automatic proxy-selection algorithm are given in Table 2. When this algorithm is iterated to construct L predictive factors, we call the forecaster the *L-automatic-proxy* 3PRF.

In order to map the automatic proxy selection approach into the consistency and asymptotic normality results presented above, it is necessary to show that the proxies produced by the algorithm satisfy Assumptions 2.4, 3.3, 3.4 and 6. This is established by the following result.

Theorem 7. *Let Assumptions 1–5 hold with the exception of Assumptions 2.4, 3.3 and 3.4. Then the L-automatic-proxy three pass regression filter forecaster of* $\boldsymbol{y}$ *automatically satisfies Assumptions 2.4, 3.3, 3.4 and 6 when* $L = K_f$. *As a result, the L-automatic-proxy is consistent and asymptotically normal according to Theorems 1 and 4.*

Theorem 7 states that the 3PRF is generally available since the conditions of Theorems 1 and 4 can be satisfied by the construction of automatic proxies. Clearly, then, the only variables absolutely required to implement the filter are $\boldsymbol{y}$ and $\boldsymbol{X}$.

2.5.2. Theory proxies

The use of automatic proxies in the three-pass filter disciplines dimension reduction of the predictors by emphasizing the covariance between predictors and target in the factor estimation step. The filter may instead be employed using alternative disciplining variables (factor proxies) which may be distinct from the target and chosen on the basis of economic theory or by statistical arguments. Consider a situation in which K_f is one, so that the target and proxy are given by $y_{t+1} = \beta_0 + \beta f_t + \eta_{t+1}$ and $z_t = \lambda_0 + \Lambda f_t + \omega_t$. Also suppose that the population R^2 of the proxy equation is substantially higher than the population R^2 of the target equation.

The forecasts from using either z_t or the target as proxy are asymptotically identical. However, in finite samples, forecasts can be improved by proxying with z_t due to its higher signal-to-noise ratio.[9] To illustrate this point, in Section 5 we consider a macroeconomic application of theory proxies. We find that improved out-of-sample forecasts of inflation come by imposing a dynamic quantity theory of inflation. These forecasts have an attractive feature that they can accurately be described as embodying an economic narrative – that output and money growth fuel price inflation – that could serve to make the forecasts more appealing to policy-makers or institutional investors.

3. Related procedures

Comparing our procedure to other methods develops intuition for why the 3PRF produces powerful forecasts. Adding to our earlier comparisons with PCR, this section evaluates the link between the 3PRF and constrained least squares and partial least squares. Importantly, we show that the 3PRF is the constrained least squares estimate of the projection of $\boldsymbol{y}$ onto $\boldsymbol{X}$. The constraint we impose embodies the assumption that proxies span the relevant factor space.[10] It happens that partial least squares emerges as a special case of the 3PRF using automatic proxies.

3.1. Constrained least squares

Section 2.1 demonstrates that the forecast $\hat{y}_{t+1}$ may be represented not only in terms of factor loadings ($\hat{\boldsymbol{\beta}}$), but equivalently in terms of loadings on individual predictors ($\hat{\boldsymbol{\alpha}}$). The ith element of coefficient vector $\hat{\boldsymbol{\alpha}}$ provides a direct statistical description for the forecast contribution of predictor $\mathbf{x}_i$ when it is combined with the remaining $N - 1$ predictors. In fact, $\hat{\boldsymbol{\alpha}}$ is an N-dimensional projection coefficient, and is available when N is near or even greater than T. This object allows us to address questions that would typically be answered by the multiple regression coefficient in settings where OLS is unsatisfactory. As discussed by Cochrane (2011) in his presidential address to the American Finance Association:

> [W]e have to move past treating extra variables one or two at a time, and understand which of these variables are really important. Alas, huge multiple regression is impossible. So the challenge is, how to answer the great multiple-regression question, without actually running huge multiple regressions?

8 While we may always recast the system in terms of a single relevant factor $\boldsymbol{\beta}_f'\boldsymbol{f}_t$ and rotate the remaining factors to be orthogonal to it, this does not generally alleviate the requirement for as many proxies as relevant factors. As we demonstrate in Appendix A.7, this is because rotating the factors necessarily implies a rotation of factor loadings. Taking both rotations into account recovers the original requirement for as many relevant proxies as relevant factors.

9 On the other hand, if theory-motivated proxies are weakly correlated with the true relevant factors, then the 3PRF will break down and fail to identify a meaningful forecasting relationship. This point is raised by Kleibergen and Zhan (2013) in the context of Fama and MacBeth (1973) two-pass regression.

10 Of course, this span can be measured with error in the sense formalized by our assumptions regarding the proxy noise ω.

The 3PRF estimator $\hat{\boldsymbol{\alpha}}$ provides an answer. It is a projection coefficient relating y_{t+1} to $\boldsymbol{x}_t$ under the constraint that irrelevant factors do not influence forecasts. That is, the 3PRF forecaster may be derived as the solution to a constrained least squares problem, as we demonstrate in the following proposition.

Theorem 8. *The three-pass regression filter's implied N-dimensional predictive coefficient, $\hat{\boldsymbol{\alpha}}$, is the solution to*

$$\textit{arg}\min_{\alpha_0,\boldsymbol{\alpha}} \|\boldsymbol{y} - \alpha_0 - \boldsymbol{X}\boldsymbol{\alpha}\|$$

$$\textit{subject to} \quad (\boldsymbol{I} - \boldsymbol{W}_{XZ}(\boldsymbol{S}_{XZ}'\boldsymbol{W}_{XZ})^{-1}\boldsymbol{W}_{XZ})\boldsymbol{\alpha} = \boldsymbol{0}. \tag{5}$$

This solution is closely tied to the original motivation for dimension reduction: The unconstrained least squares forecaster is poorly behaved when N is large relative to T. The 3PRF's answer is to impose the constraint in Eq. (5), which exploits the proxies and has an intuitive interpretation. Premultiplying both sides of the equation by $\boldsymbol{J}_T\boldsymbol{X}$, we can rewrite the constraint as $(\boldsymbol{J}_T\boldsymbol{X} - \boldsymbol{J}_T\hat{\boldsymbol{F}}\hat{\boldsymbol{\Phi}}')\boldsymbol{\alpha} = \boldsymbol{0}$. For large N and T,

$$\boldsymbol{J}_T\boldsymbol{X} - \boldsymbol{J}_T\hat{\boldsymbol{F}}\hat{\boldsymbol{\Phi}}' \approx \boldsymbol{\varepsilon} + (\boldsymbol{F} - \boldsymbol{\mu})(\boldsymbol{I} - \boldsymbol{S}_{K_f})\boldsymbol{\Phi}'$$

which follows from Lemma 6 in the Appendix. Because the covariance between $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ is zero by the assumptions of the model,[11] the constraint simply imposes that the product of $\boldsymbol{\alpha}$ and the target-irrelevant common component of $\boldsymbol{X}$ is equal to zero. This is because the matrix $\boldsymbol{I} - \boldsymbol{S}_{K_f}$ selects only the terms in the total common component $\boldsymbol{F}\boldsymbol{\Phi}'$ that are associated with irrelevant factors. This constraint is important because it ensures that factors irrelevant to $\boldsymbol{y}$ drop out of the 3PRF forecast. It also ensures that $\hat{\boldsymbol{\alpha}}$ is consistent for the factor model's population projection coefficient of y_{t+1} on $\boldsymbol{x}_t$.

3.2. Partial least squares

The method of partial least squares, or PLS (Wold (1975), described in Appendix A.10), is a special case of the three-pass regression filter. In particular, partial least squares forecasts are identical to those from the 3PRF when (i) the predictors are demeaned and variance-standardized in a preliminary step, (ii) the first two regression passes are run without constant terms and (iii) proxies are automatically selected. As an illustration, consider the case where a single predictive index is constructed from the partial least squares algorithm. Assume, for the time being, that each predictor has been previously standardized to have mean zero and variance one. Following the construction of the PLS forecast given in Appendix A.10 we have

1. Set $\hat{\phi}_i = x_i'y$, and $\hat{\boldsymbol{\Phi}} = (\hat{\phi}_1, \ldots, \hat{\phi}_N)'$.
2. Set $\hat{u}_t = \boldsymbol{x}_t'\hat{\Phi}$, and $\hat{\boldsymbol{u}} = (\hat{u}_1, \ldots, \hat{u}_T)'$.
3. Run a predictive regression of $\boldsymbol{y}$ on $\hat{\boldsymbol{u}}$.

Constructing the forecast in this manner may be represented as a one-step estimator

$$\hat{\boldsymbol{y}}^{\text{PLS}} = \boldsymbol{X}\boldsymbol{X}'\boldsymbol{y}(\boldsymbol{y}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{y})^{-1}\boldsymbol{y}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{y}$$

which upon inspection is identical to the 1-automatic-proxy 3PRF forecast when constants are omitted from the first and second passes. Repeating the comparison of 3PRF and PLS when constructing additional predictive factors under conditions (i)–(iii) shows that this equivalence holds more generally.

How do the methodological differences between the auto-proxy 3PRF and PLS embodied by conditions (i)–(iii) affect forecast performance? First, since both methods (like PCR as well) lack scale-invariance, they each work with variance-standardized predictors. For PLS, the demeaning of predictors and omission of a constant in first pass regressions offset each other and produce no net difference versus the auto-proxy 3PRF. The primary difference therefore lies in the estimation of a constant in the second stage cross section regression of the auto-proxy 3PRF. A simple example in the context of the underlying factor model assumptions of this paper helps identify when estimating a constant in cross section regressions is useful. Consider the special case of Assumption 1 in which $K_f = 1$ and $K_g = 1$, the predictors and factors have mean zero, and the relevant factor's loadings are known. In this case, $x_{it} = \phi_{i1}f_t + \phi_{i2}g_t + \varepsilon_{it}$, and the second stage population regression of x_{it} on ϕ_{i1} when including a constant yields a slope estimate of $\lim_{N\to\infty}\hat{f}_t = f_t + g_t\frac{Cov(\phi_{i1},\phi_{i2})}{Var(\phi_{i1})}$, which reduces to f_t by Assumptions 2.2 and 5. The slope estimate omitting the constant is $\lim_{N\to\infty}\hat{f}_t = f_t + g_t\frac{\mathbb{E}[\phi_{i1}\phi_{i2}]}{\mathbb{E}[\phi_{i1}^2]}$. This is an error-ridden version of the true target-relevant factor, and thus can produce inferior forecasts.

Because PLS is a special case of our methodology, the asymptotic theory we have developed for the 3PRF applies directly to PLS estimates. Our results therefore provide a means of conducting inference when applying PLS. Groen and Kapetanios (2009) proved the consistency of PLS using sequential N, T limits and weak factor assumptions, but did not derive limiting distributions. To the best of our knowledge, our simultaneous N and T asymptotics are new results for the PLS literature.

4. Simulation evidence

4.1. Comparison against alternatives

We conduct Monte Carlo experiments to examine the finite sample accuracy of 3PRF forecasts.[12] Our simulations focus on out-of-sample forecast performance and compare this against five alternative procedures. The first alternative is PCR using the first five principal components (PCR5 henceforth), as advocated by Stock and Watson (2002a, 2012). The second and third are least-angle regression and LASSO versions of the "targeted predictors" approach proposed by Bai and Ng (2008). Here, the L_1 tuning parameter is adjusted to select a group of 30 targeted predictors, from which five principal components are then extracted and used for forecasting. We call these procedures PCLAR and PCLAS, respectively. The fourth alternative follows the Bayesian shrinkage approach proposed by De Mol et al. (2008). Shrinkage motivates LAR/LASSO wherein the L_1 tuning parameter is adjusted to select a group of 10 predictors.[13] We call this procedure 10LAR. Finally, we consider the quasi-maximum likelihood factor analysis approach of extracting five factors. We call this procedure FA. We compare each of those multivariate forecasts to forecasts from single predictive index constructed from the target-proxy 3PRF (denoted 3PRF1).

Our simulations use a range of specifications to examine how performance of the estimators is affected by various data features that may complicate factor extraction and forecasting. These include serial correlation in common factors and serial or cross-sectional correlation in idiosyncratic shocks. We also explore how the strength of the factor structure affects performance.

11 This follows from Theorem 2, which shows that $\hat{\boldsymbol{\alpha}}$ converges to $\boldsymbol{J}_N\boldsymbol{\Phi}\boldsymbol{\beta}$.

12 In the Appendix, we report Monte Carlo simulations that evaluate whether the asymptotic distribution theory developed in Section 2 is a good approximation of the finite sample distribution of 3PRF estimates.

13 De Mol et al. (2008) empirical exercise found that roughly ten predictors gave them the best forecast performance, and so we use that specification. Similarly, Bai and Ng (2008) also consider a LAR/LASSO procedure to select a group of five predictors.

Table 3
Simulated out-of-sample forecast performance.

ρ_f	ρ_g	a	d	$N = T = 100$						$N = T = 200$					
				3PRF1	PCR5	PCLAR	PCLAS	10LAR	FA	3PRF1	PCR5	PCLAR	PCLAS	10LAR	FA
				Panel A: Normal Factors											
0	0	0	0	28.4	33.5	26.0	24.6	12.6	**34.5**	38.2	42.0	24.9	34.4	25.6	**43.1**
0.3	0.9	0.3	0	26.3	**31.6**	24.1	22.7	11.0	31.2	37.4	41.9	34.9	34.4	25.3	**42.2**
0.3	0.9	0.3	1	**25.2**	22.9	22.0	20.3	12.6	23.6	36.8	40.0	34.1	33.7	26.1	**40.5**
0.3	0.9	0.9	0	**18.1**	−4.5	13.5	4.3	15.1	−4.7	**31.9**	−0.3	30.0	27.8	27.1	−0.9
0.3	0.9	0.9	1	**18.0**	−5.0	12.9	3.7	15.5	−4.2	**31.5**	−0.9	29.9	27.9	28.2	−1.0
0.9	0.3	0.3	0	33.0	36.0	30.9	29.3	20.6	**39.9**	41.0	44.8	38.5	38.0	29.1	**46.5**
0.9	0.3	0.3	1	30.8	26.7	28.8	26.4	21.2	**33.1**	40.3	42.5	37.6	37.4	29.9	**45.1**
0.9	0.3	0.9	0	**31.7**	20.7	29.8	23.5	26.0	21.7	**37.6**	23.3	37.0	35.3	33.1	23.1
0.9	0.3	0.9	1	**30.7**	19.4	29.3	22.6	26.9	20.6	36.3	20.7	**37.5**	34.4	33.3	21.2
				Panel B: Moderately Weak Factors											
0	0	0	0	21.3	23.5	16.4	15.2	2.0	**25.8**	34.5	37.6	28.1	27.8	17.4	**39.2**
0.3	0.9	0.3	0	**20.1**	20.0	14.4	13.6	1.0	18.7	33.5	36.8	27.8	27.7	17.3	**37.7**
0.3	0.9	0.3	1	**17.7**	5.8	11.9	10.0	2.4	6.4	**32.1**	30.4	26.8	26.7	17.6	31.9
0.3	0.9	0.9	0	**9.6**	−6.7	1.1	−2.7	5.2	−8.3	**24.4**	−1.8	20.2	17.6	19.6	−3.0
0.3	0.9	0.9	1	**9.4**	−7.1	0.7	−3.9	6.4	−7.9	**23.6**	−2.3	19.7	16.6	20.8	−2.6
0.9	0.3	0.3	0	27.5	24.6	22.0	20.8	11.7	**31.7**	37.5	39.6	32.1	31.9	21.2	**43.0**
0.9	0.3	0.3	1	**23.7**	11.5	18.3	17.1	11.7	18.3	35.6	33.3	30.9	30.3	21.3	**38.5**
0.9	0.3	0.9	0	**27.8**	18.9	25.4	20.9	21.8	20.3	**33.0**	17.2	32.0	30.8	28.6	17.0
0.9	0.3	0.9	1	**26.7**	17.8	24.8	19.6	21.9	18.8	**31.3**	15.5	31.8	29.9	28.1	17.6
				Pa C: Weak Factors											
0	0	0	0	8.4	4.4	2.0	1.9	−14.8	4.4	24.6	24.2	14.1	14.0	2.8	**27.4**
0.3	0.9	0.3	0	7.7	2.1	1.2	1.2	−14.2	−0.5	**23.6**	19.9	14.0	13.9	3.5	20.1
0.3	0.9	0.3	1	3.3	−1.2	−1.2	−0.9	−12.0	−3.5	**19.7**	4.1	12.1	12.1	4.0	4.3
0.3	0.9	0.9	0	−0.9	−7.1	−0.8	−0.8	−11.4	−10.1	**10.7**	−2.0	2.7	1.6	4.7	−2.7
0.3	0.9	0.9	1	−1.5	−7.4	−7.7	−8.0	−10.6	−9.7	**10.1**	−2.4	2.5	0.6	5.9	−3.0
0.9	0.3	0.3	0	**17.6**	8.8	9.5	8.9	−3.1	15.7	29.0	25.0	18.8	18.7	7.5	**32.4**
0.9	0.3	0.3	1	**11.6**	2.6	5.9	5.3	−3.0	4.9	**24.6**	10.0	17.0	16.6	8.1	15.3
0.9	0.3	0.9	0	**24.0**	17.7	21.9	18.2	15.7	18.2	**26.5**	13.3	26.4	25.3	22.5	13.4
0.9	0.3	0.9	1	**22.2**	16.8	21.4	16.6	15.9	17.4	24.9	11.7	**25.3**	23.7	21.5	13.6
				Panel D: Moderately Weak and Non-pervasive Factors											
0	0	0	0	3.6	4.0	3.8	2.9	−9.2	**6.0**	18.0	26.3	20.8	20.6	12.3	**30.5**
0.3	0.9	0.3	0	**3.0**	0.2	2.4	1.9	−8.9	−1.0	18.0	22.5	20.0	20.0	12.6	**24.8**
0.3	0.9	0.3	1	**0.5**	−4.0	−2.1	−2.1	−6.9	−6.1	15.6	4.4	**17.6**	17.2	14.3	6.7
0.3	0.9	0.9	0	−1.9	−8.1	−6.8	−6.8	−3.6	−10.8	9.7	−3.1	8.0	5.5	**16.2**	−3.6
0.3	0.9	0.9	1	−2.3	−8.2	−8.1	−7.6	−0.1	−10.3	9.2	−3.3	6.3	3.6	**18.0**	−3.7
0.9	0.3	0.3	0	13.0	7.2	10.3	9.2	0.4	**14.5**	23.4	27.3	23.7	23.6	15.7	**34.9**
0.9	0.3	0.3	1	**8.4**	−1.3	5.2	3.7	2.6	1.7	**20.6**	11.9	20.5	20.4	17.2	18.7
0.9	0.3	0.9	0	**23.2**	13.7	21.1	15.1	17.1	15.1	25.7	8.4	**27.1**	25.4	25.7	10.5
0.9	0.3	0.9	1	**21.7**	12.9	20.0	14.0	18.0	13.2	24.0	7.9	**25.7**	23.5	24.7	9.6

Notes: Out-of-sample percentage R^2 from recursive out-of-sample forecasts begun at the middle of the time series. Infeasible best is 50%. Serial correlation in factors is governed by ρ_f and ρ_g, while a and d govern serial and cross section correlation in predictor idiosyncrasies. Factor strength marked by the median percentage of predictor variation explained by factors: 30% for normal factors, 20% for moderately weak factors and 10% for weak. For simulations labeled "Non-pervasive Factors", half of the predictors have a loading of zero on the relevant factor, otherwise all predictors have non-zero loadings on all factors. We **bold** the best median performer for each specification when it outperforms the historical mean. The procedures are described in the text.

By factor strength, we mean the proportion of variation among predictors that is due to the common factors. Lastly, we consider how the pervasiveness of the factor structure impacts estimator performance, where we define "pervasiveness" as the fraction of predictors with non-zero loadings on common factors.

Table 3 reports the out-of-sample forecasting performance across estimators using simulated data. We simulate data according to Assumption 1 using one relevant factor and four irrelevant factors in all cases. We use data sets of dimension $N, T = 100$ or $N, T = 200$. For each parameter configuration, we conduct 5000 simulations and report the median out-of-sample forecast percentage R^2 for each method.[14]

The strength of the factor structure may be "normal" (Panel A), in which the predictors have a median R^2 of 30% on the factors, "moderately weak" (Panel B) with R^2 of 20%, or "weak" (Panel C) with R^2 of 10%. The normal structure is roughly in line with the degree of common variation documented in Stock and Watson's (2002b) analysis of macroeconomic data, while the weak structure is motivated by Groen and Kapetanios (2009) and Onatski (2012). Because the factor loadings are drawn at random in each simulation, there is variation across predictors in the fraction of their variance explained by the factors. In Panels A–C we simulate a pervasive relevant factor, meaning that all predictors have a non-zero loading on it. In Panel D, we report results when the relevant factor is non-pervasive by imposing that half of the predictors have a loading of zero on the relevant factor, which should give an advantage to variable-subset forecasters like PCLAR, PCLAS and 10LAR.

Our experimental design builds from Stock and Watson (2002a). We simulate factors as $f_t = \rho_f f_{t-1} + u_{f,t}$ and $\boldsymbol{g}_t = \rho_g \boldsymbol{g}_{t-1} + \boldsymbol{u}_{g,t}$ for $u_{f,t} \sim IIN(0, 1)$ and $\boldsymbol{u}_{g,t} \sim IIN(0, \boldsymbol{\Sigma}_g)$, with $u_{f,t}$ and $\boldsymbol{u}_{g,t}$ uncorrelated and $K_f = 1, K_g = 4$ so that $K = 5$. Parameters of the diagonal matrix $\boldsymbol{\Sigma}_g$ are chosen so that irrelevant factors are dominant, in the sense that they have variances 1.25, 1.75, 2.25 and 2.75 times larger than the relevant factor. The parameters ρ_f and ρ_g govern serial correlation among factors and take values of

14 The R^2 measure is related to the relative mean squared error ($RMSE$) statistic according to $R^2 = 1 - RMSE$. It summarizes the forecast performance of each estimator relative to a naive forecast based on the target's historical mean.

0, 0.3, or 0.9.[15] We set $y_{t+1} = f_t + \sigma_y \eta_{t+1}$ for $\eta_{t+1} \sim IIN(0, 1)$ and adjust σ_y to ensure that the infeasible best forecast has an R^2 of 50%. Idiosyncratic errors are modeled as $\varepsilon_{i,t} = a\varepsilon_{i,t-1} + \tilde{\varepsilon}_{i,t}$, where a governs their serial correlation and takes values of 0, 0.3 or 0.9. Cross-sectional correlation among idiosyncrasies is specified via $\tilde{\varepsilon}_{i,t} = (1 + d^2)\nu_{i,t} + d\nu_{i-1,t} + d\nu_{i+1,t}$ where $\nu_{i,t}$ is standard normal and the cross-correlation parameter d takes values of 0 or 1. The factor loadings for each predictor are drawn as standard normals, allowing cross section dispersion in the proportion of predictor variation explained by the factors.

The Monte Carlo results suggest that the single-factor 3PRF performs well under a variety of circumstances, often outperforming the alternative multi-factor methods considered. There are a number of instances in which PCR and FA outperform the 3PRF, but only slightly (in larger samples, when the factor structure is strong, when factors quickly mean revert, and when there is little serial or cross-sectional correlation among predictor idiosyncrasies). On the other hand, the 3PRF often outperforms the alternatives by a wide margin, for example in a weak factor structure when the sample size is small. Furthermore, in all comparisons, 3PRF forecasts are based on a single estimated factor, while alternatives use 5 or 10 factors in the forecasting equation.

The presence of irrelevant factors causes more difficulty for PCR and FA than for 3PRF. This is particularly evident when the sample is small, when there is strong serial correlation in factors, and when serial or cross section dependence among residuals is strong. In these circumstances, 3PRF generally continues to forecast successfully, while PCR or FA can fail to detect any out-of-sample predictability. These effects are exacerbated in the weak factor scenario, even absent dependence in the idiosyncratic shocks.

While no single forecasting method dominates across all data generating processes, our central conclusion from the Monte Carlo is that the 3PRF demonstrates competitive out-of-sample forecasting performance in finite samples under a wide range of specifications.[16]

4.2. Information criteria

An information criterion (IC) may be used to select the appropriate number of 3PRF factors. The IC approach to factor selection is especially relevant for the automatic proxy version of the 3PRF, which is a statistically-motivated estimation procedure and is appropriately subject to an IC. In contrast, the number of theoretically-motivated proxies may best be determined *a priori* by an underlying economic theory.

Kramer and Sugiyama (2011) provide an unbiased degrees of freedom calculation and associated information criterion for PLS forecasting problems. We adopt their approach to the automatic proxy 3PRF setting. As these authors highlight, the degrees of freedom calculation is complicated by PLS's reliance on the forecast target in the factor extraction stage. This renders PLS a non-linear method and implies that the degrees of freedom for a K-factor PLS forecast are generally greater than those for an OLS forecast based on K exogenous regressors.[17]

We calculate degrees of freedom via the Krylov representation method of Kramer and Sugiyama (2011), then use this to compute the Bayesian Information Criterion (BIC). Details of this approach are given in Appendix A.9.

To study the BIC accuracy in our setting we simulate data according to the same data generating processes used in Table 3. Results are reported in Table A.2 and show that the information criterion is typically accurate in selecting the correct number of 3PRF factors. For example, when $N = T = 200$ and the true number of factors is equal to one, the average number of factors selected across simulations equals 1.0 in 15 of 27 specifications, and is between 1.0 and 1.3 in 21 of 27. The BIC tends to overestimate the number of factors in smaller samples and when the irrelevant factors and residuals exhibit strong serial correlation. But even when too many 3PRF factors are selected, the method achieves powerful out-of-sample forecasting performance and is typically close to the R^2 achieved by the one-factor 3PRF. Further details are discussed in Appendix A.9.

5. Empirical evidence

Here we report the results of two separate empirical investigations. In the first empirical investigation, we forecast macroeconomic aggregates using a well-known panel of quarterly macroeconomic variables. In the second, we use a factor model to relate individual assets' price–dividend ratios to market returns. We use the automatic-proxy 3PRF and compare its performance to the forecast accuracy of alternative procedures. We consider the same alternative procedures used for the simulation study: PCR (with number of factors selected via Bai and Ng's (2002) information criterion), PCLAR and PCLAS following Bai and Ng (2008), 10LAR following De Mol et al. (2008), and FA following Doz et al. (2012). Tests for statistical significance are provided by Diebold and Mariano (1995) and West (1996) because the different forecast models do not necessarily encompass one another. We end by considering examples of theory-proxies in our macroeconomic application.

5.1. Forecasting macroeconomic aggregates

We examine the forecastability of macroeconomic aggregates based on a large number of potential predictor variables. To maintain comparability to the literature, we take as our predictors a set of 108 macroeconomic variables compiled by Stock and Watson (2012) updated through the end of 2009.[18] Any variable that we eventually target is removed from the set of predictors. We focus attention on pseudo out-of-sample forecasting exercises described in detail in Appendix A.11 and following Bai and Ng (2008) and Stock and Watson (2012). We focus attention on macroeconomic aggregates that receive considerably attention in the literature and policy-making circles. For 3PRF, PCR and FA we consider single factor implementations because ours and Bai and Ng's (2002) information criteria consistently choose a single factor across forecast targets and training samples.

Table 4 presents our recursive out-of-sample forecasting results. In these macroeconomic data we see a great deal similarity in different procedures' out-of-sample forecast performance. Even ordinary PCR1 does very well here, often beating more sophisticated procedures involving LAR.[19] The closest competitor is 10LAR

15 For each persistence parameter, the parameters Σ_g are adjusted so that the relative factor volatilities maintain the values specified above.

16 In supplementary Monte Carlo analyses, we find that out-of-sample forecasting performance for 3PRF and PCR can be improved when the relevant factor is persistent ($\rho_f = 0.9$) by including lags of the target and lags of the prediction indices.

17 The degrees of freedom calculation for the PCR forecasting problem is the same as that for OLS with exogenous regressors. The forecasting IC differs from that studied in Bai and Ng (2002), who develop an IC for selecting the appropriate number of factors to explain variation among the cross section of predictors.

18 Variants of this data set have been used by Bai and Ng (2008), Ludvigson and Ng (2009), and others.

19 These macroeconomic data correspond well to Stock and Watson's (2002b) findings for factor strength. The first five PCs explains an average of 29.5% of the predictors' variation.

Table 4
Out-of-sample macroeconomic forecasting.

	3PRF1	PCR1	PCLAR	PCLAS	10LAR	FA1
GDP	30.12	35.18[a]	29.70	29.51	26.38	20.11
Consumption	23.20[a,*]	7.06	9.12	7.32	−14.85	2.72
Investment	38.88[a]	37.37	36.81	36.30	24.01	34.35
Exports	16.75[a]	13.25	−11.58	−9.42	−61.36	16.44
Imports	37.18[a]	36.50	18.46	16.93	22.77	36.58
Industrial Production	16.56[a,*]	8.92	5.67	5.71	11.04	12.04
Capacity Utilization	54.32	54.79	53.77	54.85	64.69[a,*]	55.58
Total Hours	53.81[a]	50.47	48.58	47.39	39.56	42.53
Total Employment	48.84[a]	47.27	38.14	37.16	18.91	41.73
Average Hours	20.12[a]	10.12	18.52	13.89	17.55	15.84
Housing Starts	26.97	−0.14	31.54	29.66	46.89[a,*]	0.13
GDP Inflation	0.64	2.05	−0.94	1.38	−5.89	2.80[a]
PCE Inflation	−1.29	−3.73	10.60	9.82	12.22[a]	−2.50

Notes: Quarterly data from Stock and Watson (2012) for the sample 1959:I-2009:IV. Out-of-sample R^2 of one quarter ahead forecasts, in percentage. Recursive procedure starts out-of-sample forecasts halfway through the sample in 1985. 3PRF1 uses a single automatic proxy. PCR1 uses the first PC to forecast. PCLAR and PCLAS use LAR and LASSO, respectively, to select a predictor subset from which principal components are extracted and used to forecast following Bai and Ng (2008). 10LAR uses LAR to forecast using the best five predictors following De Mol et al. (2008).

[a] Denotes the best-performing procedure.

[*] Denotes statistical significance at the 10% level of the best-performing procedure relative to the second-best using the Diebold and Mariano (1995) and West (1996) statistic.

which provides significantly superior forecasting performance in two cases. The 3PRF is a powerful forecaster across forecast targets, and dominates alternatives in most cases. 3PRF provides the best forecasting performance in eight of the thirteen series, and for two of these (consumption and industrial production) its outperformance is statistically significant.

We find that different methods produce similar forecasts. The average time series correlation between 3PRF and PCR forecasts for the targets in Table 4 is 83%. As noted in De Mol et al. (2008), this is a reassuring indication that high-dimension methods are capturing genuine data features.

5.2. Forecasting market returns

Asset return forecastability has been extensively examined in the asset pricing literature.[20] Identifying return predictability is of interest to academic researchers because it measures the extent to which risk premia fluctuate over time, and identifying the sources of risk premia guides development of asset pricing theory.

The present value relationship between prices, discount rates and future cash flows has proved a valuable lens for understanding price changes. It reveals that price changes are wholly driven by fluctuations in investors' expectations of future returns and cash flow growth (Campbell and Shiller (1988) and Vuolteenaho (2002)). Building from the present value identity, Kelly and Pruitt (2013) map the cross section of price–dividend ratios into the approximate latent factor model of Assumption 1, and argue that this set of predictors should possess forecasting power for log returns on the aggregate market.

We estimate the extent of market return predictability using 25 log price–dividend ratios of portfolios sorted by market equity and book-to-market ratio. The data is annual over the post-war period 1945–2010 (following Fama and French (1992), see Appendix for details of our data construction). We assume that the predictors take the form $pd_{i,t} = \phi_{i,0} + \boldsymbol{\phi}_i' \boldsymbol{F}_t + \varepsilon_{i,t}$, while the target takes the form $r_{t+1} = \beta_0^r + \boldsymbol{F}_t' \boldsymbol{\beta}^r + \eta_{t+1}^r$.

Our out-of-sample analysis here is recursive, as in the case of the previous macroeconomic application, which is common to this literature and described in detail in Appendix A.11.[21] To maintain comparability to our previous macroeconomic results, we begin out-of-sample forecasts in 1985.

We consider the performance of 3PRF, PCR and FA with one or two factors. We also use ours or Bai and Ng's (2002) information criteria to estimate the number of factors present in the cross section of value ratios and report those results, as well as the (average) number of factors chosen across all periods of the out-of-sample procedure, for both PCR and FA. Finally, we report the 10LAR procedure of De Mol et al. (2008).[22]

Table 5 reports market return forecasts and shows that the 3PRF achieves strong out-of-sample performance. The BIC picks one or two factors in most samples, with an average of 1.4. The 3PRF—*IC* finds an out-of-sample R^2 of 31.1%, just below the 3PRF2 R^2 of 36.3%. This performance is significantly higher than what is achieved using PCR, 10LAR or FA.[23] In fact, using just the first two PCs results in negative out-of-sample performance, and it requires four or five PCs to extract relevant predictive information and obtain a 27% out-of-sample R^2. Among dimension reduction techniques, the 3PRF demonstrates the strongest out-of-sample predictive power for aggregate returns.

5.3. Examples of theory proxies for macroeconomic forecasts

Economic implementation of dimension reduction techniques often defy interpretation. They are an amalgamation of different predictors which represent many different economic forces.[24] As discussed in Section 2.5.2, the theory-proxy 3PRF provides an applied researcher the opportunity capture some economic interpretability within the dimension reduction procedure. We now provide an example of this approach in the context of inflation forecasting.

Consider the problem of forecasting GDP inflation. Table 4 shows that it is difficult for the previously-considered estimators

20 Seminal studies include Rozeff (1984), Campbell and Shiller (1988), Fama and French (1988), Stambaugh (1986), Cochrane (1992) and Hodrick (1992).

21 See Goyal and Welch (2008).

22 We do not report the PCLAR or PCLAS procedures since our cross-sectional dimension is 25 and so those procedures coincide with the PCR5 procedure which we implicitly consider in PC—*IC*.

23 As has been well-documented in the literature, these financial data have a strong factor structure. The first five PCs explain an average of 95.8% of the price–dividend ratios' variation.

24 Stock and Watson (2002b) address this issue by regressing the predictor series back onto the factors and grouping predictors into highly correlated groups. Variables in each group are then interpreted as representing a specific economic force.

Table 5
Out-of-sample market return forecasts.

	3PRF1	3PRF2	3PRF−*IC*	PC1	PC2	PC−*IC*	10LAR	FA1	FA2	FA−*IC*
Return R^2	27.63	36.34[a,*]	31.15	−10.45	−8.89	27.00	13.28	−10.08	−11.76	21.68
# of factors			1.36			4.58				4.58

Notes: R^2 in percentage. Annual data 1945–2010, from CRSP. Twenty-five size/book-to-market sorted portfolios of dividend-paying stocks. One year ahead, recursive out-of-sample forecasts of the aggregate market returns begin in 1985. PC−*L* denotes the forecast using *L* principal components. 3PRF*L* denotes the *L*-automatic-proxy 3PRF forecast. PC-*IC* and FA−*IC* uses the number of PCs are chosen by Bai and Ng's (2002) IC_p2. 3PRF-*IC* use the BIC provided earlier in the paper. The "# of factors" displays the average number of factors chosen across the training samples using the BICs.
[a] Denotes the best-performing procedure of any group (3PRF, PCR, 10LAR or FA).
[*] Denotes statistical significance at the 10% level of the best-performing procedure relative to the best member of the remaining groups, using the Diebold and Mariano (1995) and West (1996) statistic.

to achieve significant out-of-sample predictability. Can imposing an economic theory help? To answer this, we consider a dynamic version of the quantity theory of money, building upon Fama (1981, 1982), that links next future inflation π_{t+1} to the current growth in real output ($f_{q,t}$) and money supply ($f_{m,t}$)[25]

$$\pi_{t+1} = a_1 f_{q,t} + a_2 f_{m,t} + \text{error}_{t+1}.$$

Under this model, forecasts of future inflation may be obtained from observed output growth(q_t) and observed money growth (m_t). But these quantities may themselves be subject to measurement noise

$$q_t = b_0 + b_1 f_{q,t} + \omega_{y,t}, \qquad m_t = c_0 + c_1 f_{m,t} + \omega_{m,t}.$$

How do avoid an errors-in-variables problem? The error-ridden observables may be used as 3PRF theory-proxies for extracting inflation-relevant information from the cross section of macroeconomic predictors $\boldsymbol{x}$ at time t under the assumption that

$$\boldsymbol{x}_t = \boldsymbol{\Phi}(f_{q,t}, f_{m,t})' + \boldsymbol{\omega}_t.$$

Table 6 reports the results of using theoretically-motivated variables output growth and money growth to directly forecast GDP inflation. Such direct forecasts obtain a 1.3% out-of-sample R^2, just a bit less than the 2.1% obtained by PCR or 2.8% obtained by FA in Table 4. But when we instead use output growth and money growth as theory-proxies to extract predictive factors from the cross section of macroeconomic predictors, we obtain superior out-of-sample performance with an R^2 of 7.6%. This improvement is significant according to the Diebold–Mariano–West statistic at the 10% level.

We are essentially using the cross section of predictors to "clean" the noise in the theory-proxies. The resulting forecasts are simple to explain to a policy maker – we forecast tomorrow's inflation using today's output and money growth – because basic macroeconomic theory says that those variables determine inflation. We have only used the three-pass regression filter to better triangulate the latent output and money growth factors driving the predictable part of future inflation.

Table 6
Out-of-Sample GDP inflation forecasts using theory proxies.

Theory proxies	Theory direct	Theory 3PRF
GDP Growth, M1 Growth	1.28	7.61[*]

Notes: Last two columns report R^2, in percentage. Cross sections and out-of-sample periods identical to Table 4 for GDP Inflation.
[*] Denotes statistical significance at the 10% level relative to the best performing forecast in Table 4 using the Diebold and Mariano (1995) and West (1996) statistic.

6. Conclusion

This paper has introduced a new econometric technique called the three-pass regression filter which is effective for forecasting in a many-predictor environment. The key feature of the 3PRF is its ability to selectively identify the subset of factors that is useful for forecasting a given target variable while discarding factors that are irrelevant for the target but that may be pervasive among predictors. We prove that 3PRF forecasts converge in probability to the infeasible best forecast as N and T simultaneously become large. We also derive the limiting distributions of forecasts and estimated predictive coefficients. We compare our method to principal components regressions following Stock and Watson (2002a) as well as newer forecasting techniques found in Bai and Ng (2008), De Mol et al. (2008) and Doz et al. (2012). The 3PRF demonstrates strong forecasting performance, and is often superior to alternatives, across a variety of simulation specifications and in empirical applications using macroeconomic and financial data.

[25] This would be justified by an assumption that prices adjust slowly to underlying inflationary pressures, perhaps due to a Calvo-style pricing friction.

Appendix

A.1. Overview

This Appendix includes proofs of theorems in the main text. It is designed to be self-contained so that the reader need not reference results in appendices of other papers or translate those results to the current setting. We include auxiliary lemmas upon which our primary theorems are based. All of our theorems are novel results. They build upon certain parts of our auxiliary convergence lemmas that have been proved in the literature on large sample properties of principal components. The purpose of this overview is to outline where our lemmas draw on earlier literature and where we contribute new results. We also point to analogies between our asymptotic theory and that in the PCR literature.

Of Lemma 1, parts 8, 9, 13 and 15 are new whereas the others essentially appear in Stock and Watson (2002a). Lemma 2 collects matrices whose elements are sums addressed in Lemma 1. Lemmas 3–5 and Theorems 1–6 are new in that they (i) rely on the fact that the 3PRF may be written in closed form, (ii) use 3PRF's ability to estimate only the relevant subset of factors, and (iii) rely on factor proxies, which do not apply to PCR. Lemmas 3–5 follow analogous results for PCR in Bai and Ng (2006) and Stock and Watson (2002a) for part of their development. Theorem 2 is new in providing the limit of the resulting projection coefficient on the ith predictor, although one might be able to adapt results in Bai and Ng (2006) to obtain a similar result. Lemma 6 follows Bai and Ng (2006) except that the limit of our estimated "idiosyncrasies" can include a factor structure (induced by the existence of irrelevant factors). Lemmas 7 and 8 are similar to those in Bai and Ng (2006). Theorem 3 is new as it provides the asymptotic distribution of the projection coefficient on the ith predictor. Theorems 4 and 5 are similar to Bai and Ng's (2006) main result except that we do not require a relative rate condition on N and T. Lemma 9 and Theorem 6 are similar to Bai and Ng (2006), though we cannot provide a consistent estimator of the asymptotic variance of $\hat{\boldsymbol{F}}_t$ due to the fact that our estimated predictor idiosyncrasies can include

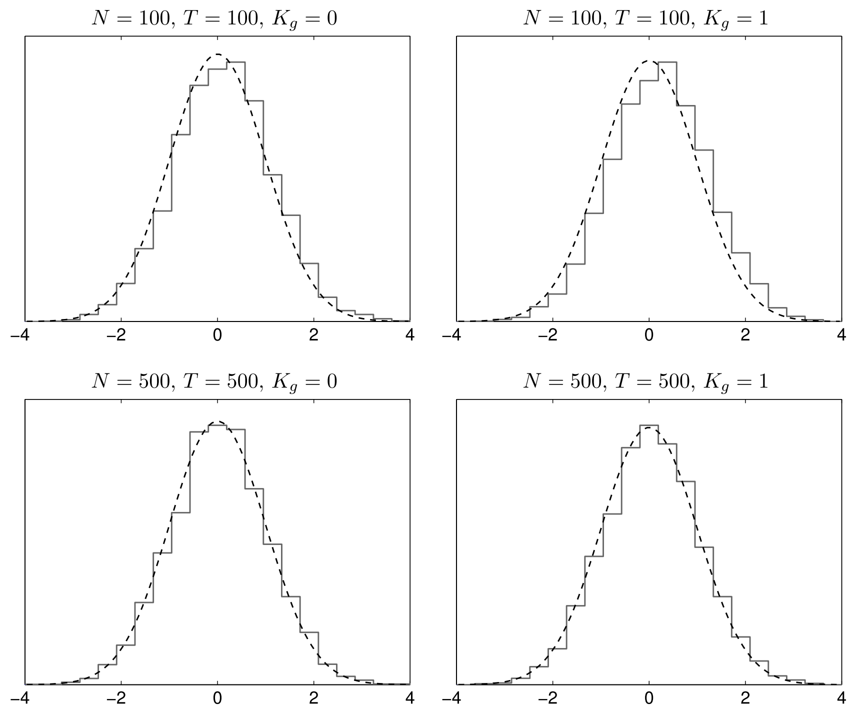

Fig. A.1. Simulated distribution, $\hat{y}_{t+1}$.

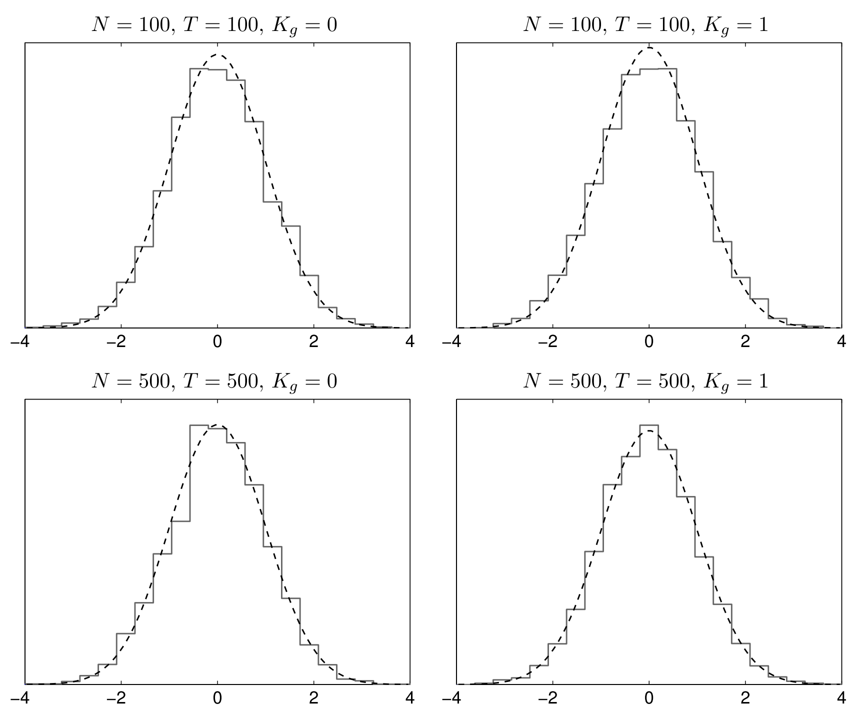

Fig. A.2. Simulated distribution, $\hat{\alpha}$.

a factor structure. Theorem 7 is a new but basic induction result, and Theorem 8 is new.

Finally, Appendix A.7 is provided to establish the necessity of the relevant proxy assumption, particularly that we need as many relevant proxies as there are relevant factors. We show that it is *not* generally possible to achieve consistent forecasts using a *single* 3PRF predictive index when there are multiple relevant factors. In fact, this only obtains in a knife-edge case wherein the relevant factors' time series variances and relevant loadings' cross-sectional variances are all equal. Absent this condition, consistency requires as many relevant proxies as there are relevant factors.

A.2. Assumptions

We restate the assumptions here so that the online appendix is self-contained and can be read without referring to assumptions in the main text.

Assumption 1 (*Factor Structure*)**.** The data are generated by the following:

$$\boldsymbol{x}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t \qquad y_{t+1} = \beta_0 + \boldsymbol{\beta}'\boldsymbol{F}_t + \eta_{t+1} \qquad \boldsymbol{z}_t = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{F}_t + \boldsymbol{\omega}_t$$
$$\boldsymbol{X} = \boldsymbol{\iota}\boldsymbol{\phi}_0' + \boldsymbol{F}\boldsymbol{\Phi}' + \boldsymbol{\varepsilon} \qquad \boldsymbol{y} = \boldsymbol{\iota}\beta_0 + \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\eta} \qquad \boldsymbol{Z} = \boldsymbol{\iota}\boldsymbol{\lambda}_0' + \boldsymbol{F}\boldsymbol{\Lambda}' + \boldsymbol{\omega}$$

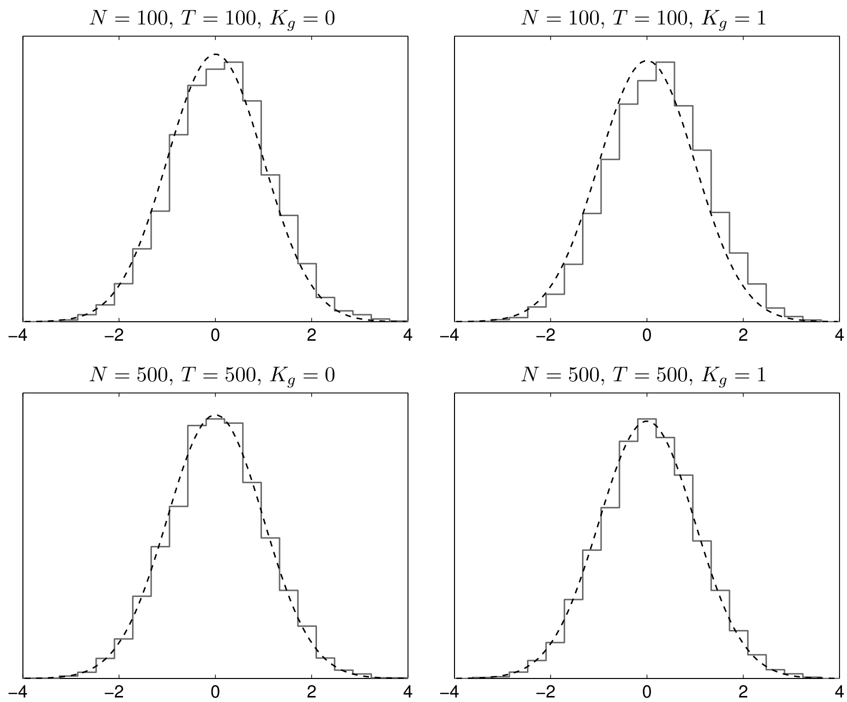

Fig. A.3. Simulated distribution, $\hat{\boldsymbol{\beta}}$.

where $h > 0$, $\boldsymbol{F}_t = (\boldsymbol{f}_t', \boldsymbol{g}_t')'$, $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_f, \boldsymbol{\Phi}_g)$, $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_f, \boldsymbol{\Lambda}_g)$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_f', \boldsymbol{0}')'$ with $|\boldsymbol{\beta}_f| > \boldsymbol{0}$. $K_f > 0$ is the dimension of vector $\boldsymbol{f}_t$, $K_g \geq 0$ is the dimension of vector $\boldsymbol{g}_t$, $L > 0$ is the dimension of vector $\boldsymbol{z}_t(0 < L < \min(N, T))$, and $K = K_f + K_g$.

Assumption 2 (*Factors, Loadings and Residuals*)**.** Let $M < \infty$. For any i, s, t

1. $\mathbb{E}\|\boldsymbol{F}_t\|^4 < M$, $T^{-1}\sum_{s=1}^T \boldsymbol{F}_s \xrightarrow[T\to\infty]{p} \boldsymbol{\mu}$ and $T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F} \xrightarrow[T\to\infty]{p} \boldsymbol{\Delta}_F$
2. $\mathbb{E}\|\boldsymbol{\phi}_i\|^4 \leq M$, $N^{-1}\sum_{j=1}^N \boldsymbol{\phi}_j \xrightarrow[T\to\infty]{p} \bar{\boldsymbol{\phi}}$, $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi} \xrightarrow[N\to\infty]{p} \mathcal{P}$ and $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\phi}_0 \xrightarrow[N\to\infty]{p} \boldsymbol{P}_1$[26]
3. $\mathbb{E}(\varepsilon_{it}) = 0$, $\mathbb{E}|\varepsilon_{it}|^8 \leq M$
4. $\mathbb{E}(\boldsymbol{\omega}_t) = \boldsymbol{0}$, $\mathbb{E}\|\boldsymbol{\omega}_t\|^4 \leq M$, $T^{-1/2}\sum_{s=1}^T \boldsymbol{\omega}_s = \boldsymbol{O}_p(1)$ and $T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\omega} \xrightarrow[N\to\infty]{p} \boldsymbol{\Delta}_\omega$
5. $\mathbb{E}_t(\eta_{t+1}) = \mathbb{E}(\eta_{t+1}|y_t, F_t, y_{t-1}, F_{t-1}, \ldots) = 0$, $\mathbb{E}(\eta_{t+1}^4) \leq M$, and η_{t+1} is independent of $\phi_i(m)$ and $\varepsilon_{i,t}$.

Assumption 3 (*Dependence*)**.** Let $x(m)$ denote the mth element of $\boldsymbol{x}$. For $M < \infty$ and any i, j, t, s, m_1, m_2

1. $\mathbb{E}(\varepsilon_{it}\varepsilon_{js}) = \sigma_{ij,ts}$, $|\sigma_{ij,ts}| \leq \bar{\sigma}_{ij}$ and $|\sigma_{ij,ts}| \leq \tau_{ts}$, and

$$\text{(a) } N^{-1}\sum_{i,j=1}^N \bar{\sigma}_{ij} \leq M \qquad \text{(c) } N^{-1}\sum_{i,s}|\sigma_{ii,ts}| \leq M$$
$$\text{(b) } T^{-1}\sum_{t,s=1}^T \tau_{ts} \leq M \qquad \text{(d) } N^{-1}T^{-1}\sum_{i,j,t,s}|\sigma_{ij,ts}| \leq M.$$

2. $\mathbb{E}\left|N^{-1/2}T^{-1/2}\sum_{s=1}^T\sum_{i=1}^N\left[\varepsilon_{is}\varepsilon_{it} - \mathbb{E}(\varepsilon_{is}\varepsilon_{it})\right]\right|^2 \leq M$
3. $\mathbb{E}\left|T^{-1/2}\sum_{t=1}^T F_t(m_1)\omega_t(m_2)\right|^2 \leq M$
4. $\mathbb{E}\left|T^{-1/2}\sum_{t=1}^T \omega_t(m_1)\varepsilon_{it}\right|^2 \leq M$.

Assumption 4 (*Central Limit Theorems*)**.** For any i, t

1. $N^{-1/2}\sum_{i=1}^N \boldsymbol{\phi}_i\varepsilon_{it} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Gamma}_{\Phi\varepsilon})$, where $\boldsymbol{\Gamma}_{\Phi\varepsilon} = \text{plim}_{N\to\infty}N^{-1}\sum_{i,j=1}^N \mathbb{E}\left[\boldsymbol{\phi}_i\boldsymbol{\phi}_j'\varepsilon_{it}\varepsilon_{jt}\right]$
2. $T^{-1/2}\sum_{t=1}^T \boldsymbol{F}_t\eta_{t+1} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\Gamma}_{F\eta}\right)$, where $\boldsymbol{\Gamma}_{F\eta} = \text{plim}_{T\to\infty}T^{-1}\sum_{t=1}^T \mathbb{E}\left[\eta_{t+1}^2\boldsymbol{F}_t\boldsymbol{F}_t'\right] > 0$
3. $T^{-1/2}\sum_{t=1}^T \boldsymbol{F}_t\varepsilon_{it} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\Gamma}_{F\varepsilon,i}\right)$, where $\boldsymbol{\Gamma}_{F\varepsilon,i} = \text{plim}_{T\to\infty}T^{-1}\sum_{t,s=1}^T \mathbb{E}\left[\boldsymbol{F}_t\boldsymbol{F}_s'\varepsilon_{it}\varepsilon_{is}\right] > 0$.

Assumption 5 (*Normalization*)**.** $\mathcal{P} = \boldsymbol{I}$, $\boldsymbol{P}_1 = \boldsymbol{0}$ and $\boldsymbol{\Delta}_F$ is diagonal, positive definite, and each diagonal element is unique.

Assumption 6 (*Relevant Proxies*)**.** $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_f \quad \boldsymbol{0}]$ and $\boldsymbol{\Lambda}_f$ is nonsingular.

A.3. Auxiliary lemmas

The following lemma collects basic results for various sums of products of the random variables appearing in our factor system. It repeatedly uses Cauchy–Schwarz and follows arguments appearing in Bai and Ng (2002), Stock and Watson (2002a) and Bai (2003).

Lemma 1. *Let Assumptions 1–4 hold. Then for all s, t, i, m, m_1, m_2*

1. $\mathbb{E}\left|(NT)^{-1/2}\sum_{i,s}F_s(m)\left[\varepsilon_{is}\varepsilon_{it} - \sigma_{ii,st}\right]\right|^2 \leq M$
2. $\mathbb{E}\left|(NT)^{-1/2}\sum_{i,s}\omega_s(m)\left[\varepsilon_{is}\varepsilon_{it} - \sigma_{ii,st}\right]\right|^2 \leq M$
3. $N^{-1/2}T^{-1/2}\sum_{i,t}\varepsilon_{it} = O_p(1)$,
   $N^{-1/2}\sum_i \varepsilon_{it} = O_p(1)$,
   $T^{-1/2}\sum_t \varepsilon_{it} = O_p(1)$
4. $T^{-1/2}\sum_t \eta_{t+1} = O_p(1)$,
5. $T^{-1/2}\sum_t \varepsilon_{it}\eta_{t+1} = O_p(1)$
6. $N^{-1/2}T^{-1/2}\sum_{i,t}\varepsilon_{it}\eta_{t+1} = O_p(1)$
7. $N^{-1}T^{-1/2}\sum_{i,t}\phi_i(m_1)\varepsilon_{it}F_t(m_2) = O_p(1)$
8. $N^{-1}T^{-1/2}\sum_{i,t}\phi_i(m_1)\varepsilon_{it}\omega_t(m_2) = O_p(1)$

26 $\|\boldsymbol{\phi}_i\| \leq M$ can replace $\mathbb{E}\|\boldsymbol{\phi}_i\|^4 \leq M$ if $\boldsymbol{\phi}_i$ is non-stochastic.

9. $N^{-1/2}T^{-1/2}\sum_{i,t}\phi_i(m)\varepsilon_{it}\eta_{t+1} = O_p(1)$
10. $N^{-1}T^{-1/2}\sum_{i,s}\varepsilon_{is}\varepsilon_{it} = O_p(\delta_{NT}^{-1})$
11. $N^{-1}T^{-3/2}\sum_{i,s,t}\varepsilon_{is}\varepsilon_{it}\eta_{t+1} = O_p(\delta_{NT}^{-1})$
12. $N^{-1}T^{-1/2}\sum_{i,s}F_s(m)\varepsilon_{is}\varepsilon_{it} = O_p(\delta_{NT}^{-1})$
13. $N^{-1}T^{-1/2}\sum_{i,s}\omega_s(m)\varepsilon_{is}\varepsilon_{it} = O_p(\delta_{NT}^{-1})$
14. $N^{-1}T^{-1}\sum_{i,s,t}F_s(m)\varepsilon_{is}\varepsilon_{it}\eta_{t+1} = O_p(1)$
15. $N^{-1}T^{-1}\sum_{i,s,t}\omega_s(m)\varepsilon_{is}\varepsilon_{it}\eta_{t+1} = O_p(1)$

The stochastic order is understood to hold as $N, T \to \infty$ and $\delta_{NT} \equiv \min(\sqrt{N}, \sqrt{T})$.

Proof. Item 1: Note that

$$\mathbb{E}\left|(NT)^{-1/2}\sum_{i,s}F_s(m)\left[\varepsilon_{is}\varepsilon_{it}-\sigma_{ii,st}\right]\right|^2$$
$$=\mathbb{E}\left[(NT)^{-1}\sum_{i,j,s,u}F_s(m)F_u(m)\left[\varepsilon_{is}\varepsilon_{it}-\sigma_{ii,st}\right]\left[\varepsilon_{ju}\varepsilon_{jt}-\sigma_{jj,ut}\right]\right]$$
$$\le \max_{s,u}\mathbb{E}|F_s(m)F_u(m)|\mathbb{E}$$
$$\times\left[(NT)^{-1}\sum_{i,j,s,u}\left[\varepsilon_{is}\varepsilon_{it}-\sigma_{ii,st}\right]\left[\varepsilon_{ju}\varepsilon_{jt}-\sigma_{jj,ut}\right]\right]$$
$$\le \max_{s,u}\mathbb{E}|F_s(m)|\mathbb{E}|F_u(m)|\mathbb{E}\left|(NT)^{-1/2}\sum_{i,s}\left[\varepsilon_{is}\varepsilon_{it}-\sigma_{ii,st}\right]\right|^2<\infty$$

by Assumptions 2.1 and 3.2. The same argument applies to Item 2 using Assumptions 2.4, 3.1

Item 3: The first part follows from $\mathbb{E}\left|N^{-1/2}T^{-1/2}\sum_{i,t}\varepsilon_{it}\right|^2 = N^{-1}T^{-1}\sum_{i,j,t,s}\sigma_{ij,ts} \le N^{-1}T^{-1}\sum_{i,j,t,s}|\sigma_{ij,ts}| \le M$ by Assumption 3.1. The second and third parts of Item 3 follow similar rationale.

Item 4 follows from $\mathbb{E}\left|T^{-1/2}\sum_t\eta_{t+1}\right|^2 = T^{-1}\sum_t\mathbb{E}[\eta_{t+1}^2] = O_p(1)$ by Assumption 2.5.

Item 5: Note that $\mathbb{E}\left|T^{-1/2}\sum_t\varepsilon_{it}\eta_{t+1}\right|^2 = T^{-1}\sum_t\sigma_{ii,tt}\mathbb{E}[\eta_{t+1}^2] \le T^{-1}\sum_t\mathbb{E}[\eta_{t+1}^2]\bar{\sigma}_{ii} = O_p(1)$ by Assumptions 2.5 and 3.1.

Item 6: Note that $\mathbb{E}\left|N^{-1/2}T^{-1/2}\sum_{i,t}\varepsilon_{it}\eta_{t+1}\right|^2 = N^{-1}T^{-1}\sum_{i,j,t}\sigma_{ij,tt}\mathbb{E}[\eta_{t+1}^2] \le T^{-1}\sum_t\mathbb{E}[\eta_{t+1}^2]N^{-1}\sum_{i,j}\bar{\sigma}_{ij} = O_p(1)$ by Assumptions 2.5 and 3.1.

Item 7 is bounded by $\left(N^{-1}\sum_i\phi_i(m_1)^2\right)^{1/2}(N^{-1}\sum_i[T^{-1/2}\sum_t\varepsilon_{it}F_t(m_2)]^2)^{1/2} = O_p(1)$ by Assumptions 2.2 and 4.3. Item 8 follows the same rationale using Assumptions 2.2 and 3.4.

Item 9: Note that $\mathbb{E}\left|N^{-1/2}T^{-1/2}\sum_{i,t}\phi_i(m)\varepsilon_{it}\eta_{t+1}\right|^2 = N^{-1}T^{-1}\sum_{i,j,t}\mathbb{E}\left[\phi_i(m)\phi_j(m)\varepsilon_{it}\varepsilon_{jt}\eta_{t+1}^2\right]$ since $\mathbb{E}\left[\eta_{t+1}\eta_{s+1}\right] = 0$ for $t \ne s$, which is in turn equal to $T^{-1}\sum_t\mathbb{E}\left[\eta_{t+1}^2\right]\mathbb{E}\left[\left(N^{-1/2}\sum_i\phi_i(m)\varepsilon_{it}\right)^2\right]$, by Assumption 2.5. That this expression is $O_p(1)$ follows from Assumptions 2.5 and 4.1.

Item 10: $N^{-1}T^{-1/2}\sum_{i,s}[\varepsilon_{is}\varepsilon_{it}-\sigma_{ii,st}] + T^{-1/2}N^{-1}\sum_{i,s}\sigma_{ii,st} = O_p(N^{-1/2}) + O_p(T^{-1/2})$ by Assumptions 3. 2 and 3.1.

Item 11: By Item 10 and Assumption 2.5,
$N^{-1}T^{-3/2}\sum_{i,s,t}\varepsilon_{is}\varepsilon_{it}\eta_{t+1} \le \left(T^{-1}\sum_t\eta_{t+1}^2\right)^{1/2}(T^{-1}\sum_t[N^{-1}T^{-1/2}\sum_{i,s}\varepsilon_{is}\varepsilon_{it}]^2)^{1/2} = O_p(\delta_{NT}^{-1})$.

Item 12: First, we have
$N^{-1}T^{-1/2}\sum_{i,s}F_s(m)\varepsilon_{is}\varepsilon_{it} = N^{-1/2}(N^{-1/2}T^{-1/2}\sum_{i,s}F_s(m)[\varepsilon_{is}\varepsilon_{it}-\sigma_{ii,st}]) + T^{-1/2}\left(N^{-1}\sum_{i,s}F_s(m)\sigma_{ii,st}\right)$. By Lemma Item 1 the first term is $O_p(N^{-1/2})$. Because $\mathbb{E}\left|N^{-1}\sum_{i,s}F_s(m)\sigma_{ii,st}\right| \le N^{-1}\max_s\mathbb{E}|F_s(m)|\sum_{i,s}|\sigma_{ii,st}| = O_p(1)$ by Assumption 3.1, the second term is $O_p(T^{-1/2})$. The same argument applies to Item 13 using Item 2.

Item 14: By Assumption 4.3 and Item 5,
$N^{-1}T^{-1}\sum_{i,s,t}F_s(m)\varepsilon_{is}\varepsilon_{it}\eta_{t+1} \le (N^{-1}\sum_i[T^{-1/2}\sum_t\varepsilon_{it}\eta_{t+1}]^2)^{1/2}\left(N^{-1}\sum_i\left[T^{-1/2}\sum_sF_s(m)\varepsilon_{is}\right]^2\right)^{1/2} = O_p(1)$. The same argument applies to Item 15 using Assumption 3.4 and Item 5. □

The following result builds on the previous lemma. It identifies finite-dimensional matrices that appear in the expression for the 3PRF, and then looks to find the stochastic order of any generic element of the matrix.

Lemma 2. *Let Assumptions 1–4 hold. Then*

1. $T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\omega} = \boldsymbol{O}_p(1)$
2. $T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta} = \boldsymbol{O}_p(1)$
3. $T^{-1/2}\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} = \boldsymbol{O}_p(1)$
4. $N^{-1/2}\boldsymbol{\varepsilon}_t'\boldsymbol{J}_N\boldsymbol{\Phi} = \boldsymbol{O}_p(1)$
5. $N^{-1}T^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p(\delta_{NT}^{-1})$
6. $N^{-1}T^{-1/2}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\omega} = \boldsymbol{O}_p(1)$
7. $N^{-1/2}T^{-1/2}\boldsymbol{\Phi}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} = \boldsymbol{O}_p(1)$
8. $N^{-1}T^{-3/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p(\delta_{NT}^{-1})$
9. $N^{-1}T^{-3/2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p(\delta_{NT}^{-1})$
10. $N^{-1}T^{-3/2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\omega} = \boldsymbol{O}_p(\delta_{NT}^{-1})$
11. $N^{-1}T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t = \boldsymbol{O}_p(\delta_{NT}^{-1})$
12. $N^{-1}T^{-1/2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t = \boldsymbol{O}_p(\delta_{NT}^{-1})$
13. $N^{-1}T^{-3/2}\boldsymbol{\eta}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p(\delta_{NT}^{-1})$
14. $N^{-1}T^{-3/2}\boldsymbol{\eta}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\omega} = \boldsymbol{O}_p(\delta_{NT}^{-1})$.

The stochastic order is understood to hold as $N, T \to \infty$, stochastic orders of matrices are understood to apply to each entry, and $\delta_{NT} \equiv \min(\sqrt{N}, \sqrt{T})$.

Proof. Item 1: $T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\omega} = T^{-1/2}\sum_t\boldsymbol{F}_t\boldsymbol{\omega}_t' - (T^{-1}\sum_t\boldsymbol{F}_t)(T^{-1/2}\sum_t\boldsymbol{\omega}_t') = \boldsymbol{O}_p(1)$ by Assumptions 2.1, 2.4 and 3.3.

Item 2: $T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta} = T^{-1/2}\sum_t\boldsymbol{F}_t\eta_{t+1} - (T^{-1}\sum_t\boldsymbol{F}_t)(T^{-1/2}\sum_t\eta_{t+1}) = \boldsymbol{O}_p(1)$ by Lemma 1.4 and Assumptions 2.1 and 4.2.

Item 3: Follows directly from Lemma 1.5, 1.6 and Assumption 2.3.

Item 4 has mth element $N^{-1/2}\sum_i\varepsilon_{it}\phi_i(m) - (N^{-1/2}\sum_i\varepsilon_{it})(N^{-1}\sum_i\phi_i(m)) = O_p(1)$ by Assumption Assumption 2.2, 2.3, 4.1 and Lemma 1.3.

Item 5 is a $K \times K$ matrix with generic (m_1, m_2) element[27]

$$N^{-1}T^{-1}\sum_{i,t}\phi_i(m_1)F_t(m_2)\varepsilon_{it} - N^{-2}T^{-1}\sum_{i,j,t}\phi_i(m_1)F_t(m_2)\varepsilon_{jt}$$
$$-N^{-1}T^{-2}\sum_{j,s,t}F_s(m_2)\phi_j(m_1)\varepsilon_{jt} + N^{-2}T^{-2}\sum_{i,j,s,t}F_s(m_2)\phi_i(m_1)\varepsilon_{jt}$$
$$= 5.\mathrm{I} - 5.\mathrm{II} - 5.\mathrm{III} + 5.\mathrm{IV}.$$

$5.\mathrm{I} = O_p\left(T^{-1/2}\right)$ by Lemma 1.7.

$5.\mathrm{II} = O_p(T^{-1/2})$ since $N^{-1}\sum_i\phi_i(m_1) = O_p(1)$ by Assumption 2.2 and $N^{-1}\sum_j\left(T^{-1/2}\sum_tF_t(m_2)\varepsilon_{jt}\right) = O_p(1)$ by Assumption 4.3.

$5.\mathrm{III} = O_p(N^{-1/2})$ since $T^{-1}\sum_sF_s(m_2) = O_p(1)$ by Assumption 2.1 and $T^{-1}\sum_t\left(N^{-1/2}\sum_j\phi_j(m_1)\varepsilon_{jt}\right) = O_p(1)$ by Assumption 4.1. For the following items in this lemma's proof we use the argument here and in Item 5.II without further elaboration except to change the referenced assumption or lemma items.

$5.\mathrm{IV} = O_p\left(T^{-1/2}N^{-1/2}\right)$ by Assumptions 2.1, 2.2 and Lemma 1.3.

Summing these terms, Item 5 is $\boldsymbol{O}_p(\delta_{NT}^{-1})$.

[27] The web appendix rearranges this and following items to cleanly show the terms.

Item 6 is a $K \times L$ matrix with generic (m_1, m_2) element

$$N^{-1}T^{-1/2}\sum_{i,t}\phi_i(m_1)\omega_t(m_2)\varepsilon_{it} - N^{-2}T^{-1/2}\sum_{i,j,t}\phi_i(m_1)\omega_t(m_2)\varepsilon_{jt}$$
$$-N^{-1}T^{-3/2}\sum_{j,s,t}\omega_s(m_2)\phi_j(m_1)\varepsilon_{jt}$$
$$+N^{-2}T^{-3/2}\sum_{i,j,s,t}\omega_s(m_2)\phi_i(m_1)\varepsilon_{jt} = 6.\text{I} - 6.\text{II} - 6.\text{III} + 6.\text{IV}.$$

6.I $= O_p(1)$ by Lemma 1.8.
6.II $= O_p(1)$ by Assumption 2.2, 3.4.
6.III $= O_p(N^{-1/2})$ by Assumption 2.4, 4.1.
6.IV $= O_p\left(T^{-1/2}N^{-1/2}\right)$ by Assumption 2.2, 2.4 and Lemma 1.3.
Summing these terms, Item 6 is $\boldsymbol{O}_p(1)$.

Item 7 has generic mth element

$$N^{-1/2}T^{-1/2}\sum_{i,t}\phi_i(m)\varepsilon_{it}\eta_{t+1} - N^{-1/2}T^{-3/2}\sum_{i,s,t}\phi_i(m)\varepsilon_{it}\eta_{s+h}$$
$$-N^{-3/2}T^{-1/2}\sum_{i,j,t}\phi_i(m)\varepsilon_{jt}\eta_{t+1} + N^{-3/2}T^{-3/2}\sum_{i,j,s,t}\phi_i(m)\varepsilon_{jt}\eta_{s+h}$$
$$= 7.\text{I} - 7.\text{II} - 7.\text{III} + 7.\text{IV}.$$

7.I $= O_p(1)$ by Lemma 1.9.
7.II $= O_p(1)$ by Assumption 4.1 and Lemma 1.4.
7.III $= O_p(1)$ by Assumption 2.2 and Lemma 1.6.
7.IV $= O_p(T^{-1/2})$ by Assumption 2.2 and Lemmas 1.3 and 1.4.
Summing these terms, Item 7 is $\boldsymbol{O}_p(1)$.

Item 8 is $K \times K$ with generic (m_1, m_2) element

$$N^{-1}T^{-3/2}\sum_{i,s,t}F_s(m_1)\varepsilon_{is}\varepsilon_{it}F_t(m_2) - N^{-1}T^{-5/2}\sum_{i,s,t,u}F_s(m_1)\varepsilon_{is}\varepsilon_{it}F_u(m_2)$$
$$-N^{-1}T^{-5/2}\sum_{i,s,t,u}F_s(m_1)\varepsilon_{it}\varepsilon_{iu}F_u(m_2)$$
$$+N^{-1}T^{-7/2}\sum_{i,s,t,u,v}F_s(m_1)\varepsilon_{it}\varepsilon_{iu}F_v(m_2)$$
$$+N^{-2}T^{-3/2}\sum_{i,j,s,t}F_s(m_1)\varepsilon_{is}\varepsilon_{jt}F_t(m_2)$$
$$+N^{-2}T^{-5/2}\sum_{i,j,s,t,u}F_s(m_1)\varepsilon_{is}\varepsilon_{jt}F_u(m_2)$$
$$+N^{-2}T^{-5/2}\sum_{i,j,s,t,u}F_s(m_1)\varepsilon_{it}\varepsilon_{ju}F_u(m_2)$$
$$-N^{-2}T^{-7/2}\sum_{i,j,s,t,u,v}F_s(m_1)\varepsilon_{it}\varepsilon_{ju}F_v(m_2) = 8.\text{I} - \cdots - 8.\text{VIII}.$$

8.I $= T^{-1/2}(N^{-1}\sum_{i,s,t}\left(T^{-1/2}\sum_s F_s(m_1)\varepsilon_{is}\right)\left(T^{-1/2}\sum_t F_t(m_2)\right.$ $\left.\varepsilon_{it}\right)) = O_p(T^{-1/2})$ by Assumption 4.3.

8.II $= O_p(\delta_{NT}^{-1})$ by Assumption 2.1 and Lemma 1.12. Item 8.III is identical.

8.IV $= O_p(\delta_{NT}^{-1})$ by Assumption 2.1 and Lemma 1.10.

8.V $= O_p(T^{-1/2})$ by Assumption 4.3.

8.VI $= O_p(N^{-1/2}T^{-1/2})$ by Assumptions 2.1, 4.3 and Lemma 1.3. Item 8.VII is identical.

8.VIII $= O_p(N^{-1}T^{-1/2})$ by Assumption 2.1 and Lemma 1.3.

Summing these terms, we have Item 8 is $\boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$.

Items 9 and 10 follow the same argument as Item 8 but replace where appropriate $w_s(m)$ for $F_s(m)$, Lemma 1.13 for 1.12 and Assumption 3.4 for 4.3. Substituting this way implies Items 9, 10 are no larger than $\boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$.

Item 11 has generic mth element given by

$$N^{-1}T^{-1/2}\sum_{i,s}F_s(m)\varepsilon_{is}\varepsilon_{it} - N^{-2}T^{-1/2}\sum_{i,j,s}F_s(m)\varepsilon_{is}\varepsilon_{jt}$$
$$-N^{-1}T^{-3/2}\sum_{i,s,u}F_s(m)\varepsilon_{iu}\varepsilon_{it} + N^{-2}T^{-3/2}\sum_{i,j,s,u}F_s(m)\varepsilon_{iu}\varepsilon_{jt}$$
$$= 11.\text{I} - 11.\text{II} - 11.\text{III} + 11.\text{IV}.$$

11.I $= O_p(\delta_{NT}^{-1})$ by Lemma 1.12.
11.I $= O_p(N^{-1/2})$ by Assumption 4.3 and Lemma 1.3.
11.III $= O_p(\delta_{NT}^{-1})$ by Assumption 2.1 and Lemma 1.10.
11.IV $= O_p(N^{-1})$ by Assumption 2.1 and Lemma 1.3.
Summing these terms, we have Item 11 is $\boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$.

Item 12 follows nearly the same argument as Item 11, but replaces $w_s(m)$ for $F_s(m)$ and Assumptions 3.4 for 4.3. Substituting this way implies that Item 12 is $\boldsymbol{O}_p(\delta_{NT}^{-1})$.

Item 13 has mth element

$$N^{-1}T^{-3/2}\sum_{i,s,t}F_s(m)\varepsilon_{is}\varepsilon_{it}\eta_{t+1} - N^{-1}T^{-3/2}\sum_{i,s,t,u}F_s(m)\varepsilon_{is}\varepsilon_{it}\eta_{u+h}$$
$$\times N^{-1}T^{-5/2}\sum_{i,s,t,u}F_s(m)\varepsilon_{it}\varepsilon_{iu}\eta_{u+h}$$
$$+N^{-1}T^{-7/2}\sum_{i,s,t,u,v}F_s(m)\varepsilon_{it}\varepsilon_{iu}\eta_{v+h}$$
$$-N^{-2}T^{-3/2}\sum_{i,j,s,t}F_s(m)\varepsilon_{is}\varepsilon_{jt}\eta_{t+1}$$
$$+N^{-2}T^{-5/2}\sum_{i,j,s,t,u}F_s(m)\varepsilon_{is}\varepsilon_{jt}\eta_{u+h}$$
$$+N^{-2}T^{-5/2}\sum_{i,j,s,t,u}F_s(m)\varepsilon_{it}\varepsilon_{ju}\eta_{u+h}$$
$$-N^{-2}T^{-7/2}\sum_{i,j,s,t,u,v}F_s(m)\varepsilon_{it}\varepsilon_{ju}\eta_{v+h} = 13.\text{I} - \cdots - 13.\text{VIII}.$$

13.I $= O_p(T^{-1/2})$ by Lemma 1.14.
13.II $= O_p(T^{-1/2}\delta_{NT}^{-1})$ by Lemmas 1.12 and 1.4.
13.III $= O_p(\delta_{NT}^{-1})$ by Assumption 2.1 and Lemma 1.11.
13.IV $= O_p(T^{-1/2}\delta_{NT}^{-1})$ by Assumption 2.1 and Lemmas Lemma 1.3 and 1.4.
13.V $= O_p(N^{-1/2}T^{-1/2})$ by Assumption 4.3 and Lemma 1.6.
13.VI $= O_p(N^{-1/2}T^{-1})$ by Assumption 4.3 and Lemmas 1.3 and 1.4.
13.VII $= O_p(N^{-1}T^{-1/2})$ by Assumption 2.1 and Lemmas 1.3 and 1.6.
13.VIII $= O_p(N^{-1}T^{-1/2})$ by Assumption 2.1 and Lemmas 1.3 and 1.4.
Summing these terms, Item 13 is $\boldsymbol{O}_p(\delta_{NT}^{-1})$.

Item 14 follows the same argument as Item 13 replacing Lemma 1.15 for 1.14, Lemma 1.13 for 1.12 and Assumption 3.4 for 4.3. Substituting this way implies that Item 14 is $\boldsymbol{O}_p(\delta_{NT}^{-1})$. □

A.4. Probability limits and forecast consistency

This lemma finds the probability limit for our factor estimator $\hat{\boldsymbol{F}}$. It expands out this expression to find terms involving $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}$ that can then be expressed using Assumption 1 as matrices appearing in Lemma 2.

Lemma 3. *Let Assumptions 1–4 hold. Then the probability limits of* $\hat{\boldsymbol{\Phi}}$ *and* $\hat{\boldsymbol{F}}_t$ *are*

$$\hat{\boldsymbol{\Phi}} \xrightarrow[T\to\infty]{p} \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Phi}'$$

and

$$\hat{\boldsymbol{F}}_t \xrightarrow[T,N\to\infty]{p} \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega\right)\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{P}_1 + \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{F}_t\right).$$

Proof. From Eq. (2), for any t the second stage 3PRF regression coefficient is

$$\begin{aligned}\hat{\boldsymbol{F}}_t &= T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(N^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{x}_t\\ &= \hat{\boldsymbol{F}}_A\hat{\boldsymbol{F}}_B^{-1}\hat{\boldsymbol{F}}_{C,t}.\end{aligned}$$

We handle each of these three terms individually.

$$\begin{aligned}\hat{\boldsymbol{F}}_A &= T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\\ &= \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\omega}\right)\\ &\quad + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}' + T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\omega}\\ &\xrightarrow[T,N\to\infty]{p} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega.\end{aligned}$$

$$\begin{aligned}\hat{\boldsymbol{F}}_B &= N^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\\ &= \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}'\\ &\quad + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\omega}\right)\\ &\quad + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}T^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}'\\ &\quad + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}T^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}\boldsymbol{J}_T\boldsymbol{\omega}\right)\\ &\quad + \boldsymbol{\Lambda}\left(N^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}'\\ &\quad + \boldsymbol{\Lambda}\left(N^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\omega}\right)\\ &\quad + \boldsymbol{\Lambda}\left(N^{-1}T^{-2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}'\\ &\quad + \boldsymbol{\Lambda}\left(N^{-1}T^{-2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\omega}\right)\\ &\quad + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}'\\ &\quad + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\omega}\right)\\ &\quad + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}T^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}'\\ &\quad + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}T^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}\boldsymbol{J}_T\boldsymbol{\omega}\right)\\ &\quad + \left(N^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}'\\ &\quad + \left(N^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\omega}\right)\\ &\quad + \left(N^{-1}T^{-2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\Lambda}'\\ &\quad + \left(N^{-1}T^{-2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}\boldsymbol{J}_T\boldsymbol{\omega}\right)\\ &\xrightarrow[T,N\to\infty]{p} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'.\end{aligned}$$

$$\begin{aligned}\hat{\boldsymbol{F}}_{C,t} &= N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{x}_t \qquad\qquad (A.1)\\ &= \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\phi}_0\right) + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi}\right)\boldsymbol{F}_t\\ &\quad + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right)\\ &\quad + \boldsymbol{\Lambda}\left(N^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\phi}_0\right) + \boldsymbol{\Lambda}\left(N^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\Phi}\right)\boldsymbol{F}_t\\ &\quad + \boldsymbol{\Lambda}\left(N^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right)\\ &\quad + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\phi}_0\right) + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi}\right)\boldsymbol{F}_t\\ &\quad + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right)\\ &\quad + \left(N^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\phi}_0\right) + \left(N^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\Phi}\right)\boldsymbol{F}_t\\ &\quad + \left(N^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right)\\ &\xrightarrow[T,N\to\infty]{p} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{P}_1 + \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{F}_t.\end{aligned}$$

Each convergence result follows from Lemma 2 and Assumptions 1–4. The final result is obtained via the continuous mapping theorem. The result for $\hat{\boldsymbol{\Phi}}$ proceeds similarly, using the result above for $\hat{\boldsymbol{F}}_A$ and the fact that $\text{plim}_{N,T\to\infty}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X} = \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Phi}'$ using Lemma 2. □

This lemma finds the probability limit for our factor estimator $\hat{\boldsymbol{\beta}}$. It expands out this expression to find terms involving $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y}$ that can then be expressed using Assumption 1 as matrices appearing in Lemma 2.

Lemma 4. *Let Assumptions 1–4 hold. Then the probability limit of estimated third stage predictive coefficients $\hat{\boldsymbol{\beta}}$ is*

$$\begin{aligned}\hat{\boldsymbol{\beta}} &\xrightarrow[T,N\to\infty]{p} \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\\ &\quad\times\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\beta}. \qquad (A.2)\end{aligned}$$

Proof. From Eq. (3), the third stage 3PRF regression coefficient is

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left(T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}N^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\\ &\quad\times\left(N^{-2}T^{-3}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}N^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}\\ &= \hat{\boldsymbol{\beta}}_1^{-1}\hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\beta}}_3^{-1}\hat{\boldsymbol{\beta}}_4.\end{aligned}$$

We handle each of these three terms individually. Note that $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{F}}_A$ and $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{F}}_B$ and these probability limits are established in Lemma 3. The expressions for $\hat{\boldsymbol{\beta}}_3$ and $\hat{\boldsymbol{\beta}}_4$ are more tedious and require an additional lemma (that holds given Assumptions 1–4) which we place in the web appendix. Therefore we have that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_3 &= N^{-2}T^{-3}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\\ &\xrightarrow[T,N\to\infty]{p} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\end{aligned}$$

and

$$\begin{aligned}\hat{\boldsymbol{\beta}}_4 &= N^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}\\ &\xrightarrow[T,N\to\infty]{p} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\beta}.\end{aligned}$$

Each convergence result follows from Lemma 2 and Assumptions 1–4. The final result is obtained via the continuous mapping theorem. □

This lemma finds the probability limit for our factor estimator $\hat{\boldsymbol{y}}$, but is immediate from the two preceding proofs.

Lemma 5. *Let Assumptions 1–3 hold. Then the three pass regression filter forecast satisfies*

$$\begin{aligned}\hat{y}_{t+1} &\xrightarrow[T,N\to\infty]{p} \beta_0 + \boldsymbol{\mu}'\boldsymbol{\beta} + (\boldsymbol{F}_t - \boldsymbol{\mu})'\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\\ &\quad\times\left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right]^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\beta}. \qquad (A.3)\end{aligned}$$

Proof. Immediate from Eq. (1) and Lemmas 3 and 4. □

This theorem uses the probability limit just found for $\hat{\boldsymbol{y}}$ and adds the assumption that the proxies are relevant. This allows certain off-diagonal matrices to go to zero, ensuring consistency.

Theorem 1. *Let Assumptions 1–6 hold. The three-pass regression filter forecast is consistent for the infeasible best forecast, $\hat{y}_{t+1} \xrightarrow[T,N\to\infty]{p} \beta_0 + \boldsymbol{F}_t'\boldsymbol{\beta}$.*

Proof. Given Assumptions 1–3, Lemma 5 holds and we can therefore manipulate (A.3). Partition $\mathcal{P}$ and $\boldsymbol{\Delta}_F$ as

$$\mathcal{P} = \begin{bmatrix}\mathcal{P}_1 & \mathcal{P}_{12}\\ \mathcal{P}_{12}' & \mathcal{P}_2\end{bmatrix}, \qquad \boldsymbol{\Delta}_F = \begin{bmatrix}\boldsymbol{\Delta}_{F,1} & \boldsymbol{\Delta}_{F,12}\\ \boldsymbol{\Delta}_{F,12}' & \boldsymbol{\Delta}_{F,2}\end{bmatrix}$$

such that the block dimensions of $\mathcal{P}$ and $\boldsymbol{\Delta}_F$ coincide. By Assumption 5, the off-diagonal blocks, $\mathcal{P}_{12}$ and $\boldsymbol{\Delta}_{F,12}$, are zero. As a result, the first diagonal block of the term $\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F$ in Eq. (A.3) is $\boldsymbol{\Delta}_{F,1}\mathcal{P}_1\boldsymbol{\Delta}_{F,1}\mathcal{P}_1\boldsymbol{\Delta}_{F,1}$. By Assumption 6, pre- and post-multiplying by $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_f, \boldsymbol{0}]$ reduces the term in square brackets to $\boldsymbol{\Lambda}_f\boldsymbol{\Delta}_{F,1}\mathcal{P}_1\boldsymbol{\Delta}_{F,1}\mathcal{P}_1\boldsymbol{\Delta}_{F,1}\boldsymbol{\Lambda}_f$. Similarly, $\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' = \left[\boldsymbol{\Lambda}_f\mathcal{P}_1\boldsymbol{\Delta}_{F,1}, \boldsymbol{0}\right]'$ and

$\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F = \left[\boldsymbol{\Lambda}_f\boldsymbol{\Delta}_{F,1}\mathcal{P}_1\boldsymbol{\Delta}_{F,1}, \mathbf{0}\right]$. By Assumption 6, $\boldsymbol{\Lambda}_f$ is invertible and therefore the expression for $\hat{y}_{t+1}$ reduces to $\beta_0 + \boldsymbol{F}_t'\boldsymbol{\beta}$.[28] □

Corollary 1. *Let Assumptions 1–5 hold. Additionally, assume that there is only one relevant factor. Then the target-proxy three pass regression filter forecaster is consistent for the infeasible best forecast.*

Proof. It follows directly from previous result by noting that the loading of $\boldsymbol{y}$ on $\boldsymbol{F}$ is $\boldsymbol{\beta} = (\beta_1, \mathbf{0}')'$ with $\beta_1 \neq 0$. Therefore the target satisfies the condition of Assumption 6. □

We consider a generic element of the projection coefficient $\boldsymbol{\alpha}$ and obtain its probability limit, which boils down to performing matrix algebra.

Theorem 2. *Let $\hat{\alpha}_i$ denote the ith element of $\hat{\boldsymbol{\alpha}}$, and let Assumptions 1–6 hold. Then for any i,*

$$N\hat{\alpha}_i \xrightarrow[T,N\to\infty]{p} \left(\boldsymbol{\phi}_i - \bar{\boldsymbol{\phi}}\right)'\boldsymbol{\beta}.$$

Proof. Rewrite $\hat{\alpha}_i = \boldsymbol{S}_i\hat{\boldsymbol{\alpha}}$, where $\boldsymbol{S}_i$ is the $(1 \times N)$ selector vector with ith element equal to one and remaining elements zero. Expanding the expression for $\hat{\boldsymbol{\alpha}}$ in Eq. (4), the first term in $\boldsymbol{S}_i\hat{\boldsymbol{\alpha}}$ is the $(1 \times K)$ matrix $\boldsymbol{S}_i\boldsymbol{J}_N\boldsymbol{\Phi}$, which has probability limit $\left(\boldsymbol{\phi}_i - \bar{\boldsymbol{\phi}}\right)$ as $N, T \to \infty$. It then follows directly from previous results that

$$N\hat{\alpha}_i \xrightarrow[T,N\to\infty]{p} \left(\boldsymbol{\phi}_i - \bar{\boldsymbol{\phi}}\right)'\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\beta}.$$

Under Assumptions 5 and 6, this reduces to $\left(\boldsymbol{\phi}_i - \bar{\boldsymbol{\phi}}\right)'\boldsymbol{\beta}$. □

The following lemma finds the probability limit of the predictors' "residuals" that are unexplained by the factor estimator $\hat{\boldsymbol{F}}$ in the limit. Notice that $\hat{\boldsymbol{\varepsilon}}$ is consistent for the true idiosyncratic errors (for which cross-sectional dependence is limited by Assumption 3) and a linear combination of the irrelevant factors $\boldsymbol{g}$ which can be pervasive across predictors. This fact complicates the construction of a consistent estimator for the asymptotic variance of $\hat{\boldsymbol{F}}_t$.

Lemma 6. *Define $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{X} - \boldsymbol{\iota}\hat{\boldsymbol{\phi}}_0 - \hat{\boldsymbol{F}}\hat{\boldsymbol{\Phi}}'$, where $\hat{\boldsymbol{\phi}}_0 = T^{-1}\sum_t \boldsymbol{x}_t - \hat{\boldsymbol{\Phi}}(T^{-1}\sum_t \hat{\boldsymbol{F}}_t)$. Under Assumptions 1–6, $\hat{\boldsymbol{F}}\hat{\boldsymbol{\Phi}}' \xrightarrow[T,N\to\infty]{p} \boldsymbol{f}\boldsymbol{\Phi}_f'$ and $\hat{\boldsymbol{\varepsilon}} \xrightarrow[T,N\to\infty]{p} \boldsymbol{\varepsilon} + \boldsymbol{g}\boldsymbol{\Phi}_g'$.*

Proof. Let $\boldsymbol{S}_k$ be a $K \times K$ selector matrix that has ones in the first K_f main diagonal positions and zeros elsewhere. The fact that

$$\hat{\boldsymbol{F}}\hat{\boldsymbol{\Phi}}' \xrightarrow[T,N\to\infty]{p} \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{P}_1 + \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{F}'\right)'\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Phi}'$$

follows directly from Lemma 3. By Assumptions 5 and 6, this reduces to $\boldsymbol{F}\boldsymbol{S}_k\boldsymbol{\Phi}' = \boldsymbol{f}\boldsymbol{\Phi}_f'$, which also implies the stated probability limit of $\hat{\boldsymbol{\varepsilon}}$. □

The following lemma establishes the asymptotic independence of $\hat{\boldsymbol{F}}_t$ and η_{t+1}, which is used to find the asymptotic distribution of $\hat{\boldsymbol{\alpha}}$.

Lemma 7. *Under Assumptions 1–4, $\text{plim}_{N,T\to\infty} T^{-1}\sum_t \hat{\boldsymbol{F}}_t\eta_{t+1} = 0$ for all h.*

Proof. It suffices to show that $\text{plim}_{N,T\to\infty} T^{-1}\sum_t \hat{\boldsymbol{F}}_{C,t}\eta_{t+1} = 0$ for all h, and to do so we examine each term in Eq. (A.1). The four terms involving $\boldsymbol{\phi}_0$ becomes $o_p(1)$ since each is $O_p(1)$ by Lemma 2, since they do not possess t subscripts, and since $T^{-1}\sum_t \eta_{t+1} = o_p(1)$. By similar rationale, the four terms that are post-multiplied by $\boldsymbol{F}_t$ are $o_p(1)$ since $T^{-1}\sum_t \boldsymbol{F}_t\eta_{t+1} = o_p(1)$ by Assumption 4.3. Two of the remaining terms depend on the expression $T^{-1}\sum_t \left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right)\eta_{t+1}$, which is $o_p(1)$ because

$$\left|T^{-1}N^{-1}\sum_{i,t}\boldsymbol{\phi}_i\varepsilon_{it}\eta_{t+1}\right| \leq N^{-1/2}\left\{T^{-1}\sum_t\left(N^{-1/2}\sum_i\boldsymbol{\phi}_i\varepsilon_{it}\right)^2\right\}^{1/2} \times \left(T^{-1}\sum_t \eta_{t+1}^2\right)^{1/2} = o_p(1).$$

The last two remaining terms depend on $T^{-1}\sum_t(N^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t)\eta_{t+1}$, which is $o_p(1)$ following the same argument used to prove Lemma 2.14. □

A.5. Asymptotic distributions

Lemma 8. *Under Assumptions 1–4, as $N, T \to \infty$ we have*

$$N^{-1}T^{-3/2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{\eta} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Delta}_{\boldsymbol{F}}\mathcal{P}\boldsymbol{\Gamma}_{F\eta}\mathcal{P}\boldsymbol{\Delta}_{\boldsymbol{F}}\boldsymbol{\Lambda}'\right).$$

Proof.

$$\begin{aligned}
N^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{\eta} &= N^{-1}T^{-2}\boldsymbol{\Lambda}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta} \\
&\quad + N^{-1}T^{-2}\boldsymbol{\Lambda}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} \\
&\quad + N^{-1}T^{-2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta} + N^{-1}T^{-2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} \\
&\quad + N^{-1}T^{-2}\boldsymbol{\Lambda}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\Phi}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta} \\
&\quad + N^{-1}T^{-2}\boldsymbol{\Lambda}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} + N^{-1}T^{-2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\Phi}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta} \\
&\quad + N^{-1}T^{-2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} \\
&= \boldsymbol{O}_p(T^{-1/2}) + \boldsymbol{O}_p(T^{-1/2}N^{-1/2}) + \boldsymbol{O}_p(T^{-1}) \\
&\quad + \boldsymbol{O}_p(N^{-1/2}T^{-1}) + \boldsymbol{O}_p(T^{-1/2}\delta_{NT}^{-1}) \\
&\quad + \boldsymbol{O}_p(T^{-1/2}\delta_{NT}^{-1}) + \boldsymbol{O}_p(T^{-1}) + \boldsymbol{O}_p(T^{-1/2}\delta_{NT}^{-1}).
\end{aligned}$$

As $N, T \to \infty$, the first term is dominant and the stated asymptotic distribution obtains by Lemma 2 and Assumption 4.2. □

Theorem 3. *Under Assumptions 1–6, as $N, T \to \infty$ we have*

$$\frac{\sqrt{T}N\left(\hat{\alpha}_i - \tilde{\alpha}_i\right)}{A_i} \xrightarrow{d} \mathcal{N}(0, 1)$$

where A_i^2 is the ith diagonal element of $\widehat{Avar}(\hat{\boldsymbol{\alpha}}) = \boldsymbol{\Omega}_\alpha(\frac{1}{T}\sum_t \hat{\eta}_{t+1}^2(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})')\boldsymbol{\Omega}_\alpha'$, $\hat{\eta}_{t+1}$ is the estimated 3PRF forecast error and

$$\boldsymbol{\Omega}_\alpha = \boldsymbol{J}_N\left(\frac{1}{T}\boldsymbol{S}_{XZ}\right)\left(\frac{1}{T^3N^2}\boldsymbol{W}_{XZ}'\boldsymbol{S}_{XX}\boldsymbol{W}_{XZ}\right)^{-1}\left(\frac{1}{TN}\boldsymbol{W}_{XZ}'\right).$$

Proof. Given the definition of $\tilde{\alpha}_i$, note that

$$N\hat{\alpha}_i - N\tilde{\alpha}_i \stackrel{d}{=} \boldsymbol{S}_iT^{-1}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(T^{-3}N^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1} \times T^{-2}N^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{\eta}.$$

The asymptotic distribution and consistent variance estimator follow directly from Lemma 8 and previously derived limits, Assumptions 5 and 6, and noting that $\hat{\eta}_{t+1} = \eta_{t+1} + o_p(1)$ by Theorem 1. □

[28] This proof shows that Assumption 5 is stronger than is necessary. All we require is that $\mathcal{P}$ and $\boldsymbol{\Delta}_F$ are block diagonal.

Theorem 4. *Under Assumptions 1–6, as $N, T \to \infty$ we have*

$$\frac{\sqrt{T}\left(\hat{y}_{t+1}-\tilde{y}_{t+1}\right)}{Q_t} \xrightarrow{d} \mathcal{N}(0,1)$$

where $\tilde{y}_{t+1} = \bar{y} + \boldsymbol{x}_t'\boldsymbol{G}_\alpha\boldsymbol{\beta}$ and Q_t^2 is the tth diagonal element of $\frac{1}{N^2}\boldsymbol{J}_T\boldsymbol{X}\widehat{Avar}(\hat{\boldsymbol{\alpha}})\boldsymbol{X}'\boldsymbol{J}_T$.

Proof. The result follows directly from Theorems 2 and 3. Note that the theorem may be restated replacing $\tilde{y}_{t+1}$ with $\mathbb{E}_t y_{t+1}$ since the argument leading up to Theorem 1 implies that $\sqrt{T}\tilde{y}_{t+1} \xrightarrow[T,N\to\infty]{p} \mathbb{E}_t y_{t+1}$. By Slutsky's theorem convergence in distribution follows, yielding the theorem statement in the paper's text. □

Theorem 5. *Under Assumptions 1–6, as $N, T \to \infty$ we have*

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{G}_\beta\boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_\beta\right)$$

where $\boldsymbol{\Sigma}_\beta = \boldsymbol{\Sigma}_z^{-1}\boldsymbol{\Gamma}_{F\eta}\boldsymbol{\Sigma}_z^{-1}$ and $\boldsymbol{\Sigma}_z = \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega$. Furthermore,

$$\widehat{Avar}(\hat{\boldsymbol{\beta}}) = \left(T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1} T^{-1}\sum_t \hat{\eta}_{t+1}^2(\hat{\boldsymbol{F}}_t-\hat{\boldsymbol{\mu}})(\hat{\boldsymbol{F}}_t-\hat{\boldsymbol{\mu}})'\left(T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1}$$

is a consistent estimator of $\boldsymbol{\Sigma}_\beta$.

Proof. Define $\boldsymbol{G}_\beta = \hat{\boldsymbol{\beta}}_1^{-1}\hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\beta}}_3^{-1}\left(N^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{F}\right)$. The asymptotic distribution follows directly from Lemma 8 noting that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{G}_\beta\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_1^{-1}\hat{\boldsymbol{\beta}}_2\hat{\boldsymbol{\beta}}_3^{-1}\left(N^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{\eta}\right).$$

The asymptotic covariance matrix (before employing Assumptions 5 and 6) is $\boldsymbol{\Sigma}_\beta = \boldsymbol{\Psi}_\beta\boldsymbol{\Gamma}_{F\eta}\boldsymbol{\Psi}_\beta'$, where $\boldsymbol{\Psi}_\beta = \boldsymbol{\Sigma}_z^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}$. This expression follows from Lemma 8 and the probability limits derived in the proof of Lemma 4. Assumptions 5 and 6 together with the derivation in the proof of Theorem 1 reduces $\boldsymbol{\Sigma}_\beta$ to the stated form.

To show consistency of $\widehat{Avar}(\hat{\boldsymbol{\beta}})$, note that $\sqrt{T}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{G}_\beta\boldsymbol{\beta}\right) = \left(T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1}T^{-1/2}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\boldsymbol{\eta}$, which implies that the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is equal to the probability limit of

$$\left(T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1}T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\boldsymbol{\eta}\boldsymbol{\eta}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\left(T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1}. \tag{A.4}$$

Assumption 2.5 and Lemma 7 imply that $\text{plim}_{T,N\to\infty}T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\boldsymbol{\eta}\boldsymbol{\eta}'\boldsymbol{J}_T\hat{\boldsymbol{F}} = \text{plim}_{T,N\to\infty}T^{-1}\sum_t\eta_{t+1}^2(\hat{\boldsymbol{F}}_t-\hat{\boldsymbol{\mu}})(\hat{\boldsymbol{F}}_t-\hat{\boldsymbol{\mu}})'$. By Theorem 1, $\eta_{t+1} = \hat{\eta}_{t+1} + o_p(1)$, which implies that $\widehat{Avar}(\hat{\boldsymbol{\beta}})$ and (A.4) share the same probability limit, therefore $\widehat{Avar}(\hat{\boldsymbol{\beta}})$ is a consistent estimator of $\boldsymbol{\Sigma}_\beta$. □

Lemma 9. *Under Assumptions 1–4, as $N, T \to \infty$ we have*

(i) *if $\sqrt{N}/T \to 0$, then for every t*

$$N^{-1/2}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Gamma}_{\Phi\varepsilon}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)$$

(ii) *if $\liminf \sqrt{N}/T \ge \tau \ge 0$, then*

$$N^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t = \boldsymbol{O}_p(1).$$

Proof. From Lemma 2 we have

$$\begin{aligned}
N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t &= \hat{\boldsymbol{F}}_{3,t} - N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\left(\boldsymbol{\phi}_0 + \boldsymbol{\Phi}\boldsymbol{F}_t\right)\\
&= \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right) + \boldsymbol{\Lambda}\left(N^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right)\\
&\quad + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right) + \left(N^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right)\\
&= \boldsymbol{O}_p(N^{-1/2}) + \boldsymbol{O}_p(\delta_{NT}^{-1}T^{-1/2}) + \boldsymbol{O}_p(N^{-1/2}T^{-1/2})\\
&\quad + \boldsymbol{O}_p(\delta_{NT}^{-1}T^{-1/2}).
\end{aligned}$$

When $\sqrt{N}/T \to 0$, the first term determines the limiting distribution, in which case result (i) obtains by Assumption 4.1.

When $\liminf\sqrt{N}/T \ge \tau > 0$, we have $T\left(N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right) = \boldsymbol{O}_p(1)$ since $\liminf T/\sqrt{N} \le 1/\tau < \infty$. □

Define

$$\begin{aligned}
\boldsymbol{H}_0 &= \hat{\boldsymbol{F}}_A\hat{\boldsymbol{F}}_B^{-1}N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\phi_0 \quad \text{and}\\
\boldsymbol{H} &= \hat{\boldsymbol{F}}_A\hat{\boldsymbol{F}}_B^{-1}N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{\Phi}.
\end{aligned} \tag{A.5}$$

Theorem 6. *Under Assumptions 1–6, as $N, T \to \infty$ we have for every t*

(i) *if $\sqrt{N}/T \to 0$, then*

$$\sqrt{N}\left[\hat{\boldsymbol{F}}_t - (\boldsymbol{H}_0 + \boldsymbol{H}\boldsymbol{F}_t)\right] \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_F\right)$$

(ii) *if $\liminf\sqrt{N}/T \ge \tau \ge 0$, then*

$$T\left[\hat{\boldsymbol{F}}_t - (\boldsymbol{H}_0 + \boldsymbol{H}\boldsymbol{F}_t)\right] = \boldsymbol{O}_p(1)$$

where $\boldsymbol{\Sigma}_F = \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega\right)\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F^2\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Gamma}_{\Phi\varepsilon}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F^2\boldsymbol{\Lambda}')^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega\right)$.

Proof. The result follows directly from Lemma 9, noting that $\hat{\boldsymbol{F}}_t - (\boldsymbol{H}_0 + \boldsymbol{H}\boldsymbol{F}_t) = \hat{\boldsymbol{F}}_A\hat{\boldsymbol{F}}_B^{-1}N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t$. The asymptotic covariance matrix $\boldsymbol{\Sigma}_F$ is found from Lemma 9, the probability limits derived in the proof of Lemma 3, and by Assumption 5 (which sets $\mathcal{P} = \boldsymbol{I}$). □

A.6. Automatic proxy selection

Theorem 7. *Let Assumptions 1–5 hold with the exception of Assumption 2.4, 3.3, 3.4. Then the L-automatic-proxy three pass regression filter forecaster of $\boldsymbol{y}$ automatically satisfies Assumption 2.4, 3.3, 3.4, 6 when $L = K_f$. As a result, the L-automatic-proxy is consistent and asymptotically normal according to Theorems 1 and 4.*

Proof. We begin by showing that Assumption 6 is generally satisfied. If $K_f = 1$, Assumption 6 is satisfied by using $\boldsymbol{y}$ as the proxy (see Corollary 1). For $K_f > 1$, we proceed by induction to show that the automatic proxy selection algorithm constructs a set of proxies that satisfies Assumption 6. In particular, we wish to show that the automatically-selected proxies have a loading matrix on relevant factors ($\boldsymbol{\Lambda}_f$) that is full rank, and that their loadings on irrelevant factors are zero. We use superscript (k) to denote the use of k automatic proxies.

Denote the 1-automatic-proxy 3PRF forecast by $\hat{\boldsymbol{y}}^{(1)}$. We have from Eq. (1) that

$$\begin{aligned}
\boldsymbol{r}^{(1)} &= \boldsymbol{y} - \hat{\boldsymbol{y}}^{(1)} = \boldsymbol{\eta} + \boldsymbol{F}\boldsymbol{\beta} - \hat{\boldsymbol{F}}^{(1)}\hat{\boldsymbol{\beta}}^{(1)}\\
&= \boldsymbol{F}\left(\boldsymbol{\beta} - \boldsymbol{\Phi}'\boldsymbol{\Omega}^{(1)}\boldsymbol{F}\boldsymbol{\beta}\right) + \boldsymbol{\eta} + \boldsymbol{\varepsilon}\boldsymbol{\Omega}^{(1)}\boldsymbol{\eta},
\end{aligned}$$

where $\boldsymbol{\Omega}^{(1)} = \boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T$. For $\boldsymbol{r}^{(1)}$, $\boldsymbol{\Omega}^{(1)}$ is constructed based on $\boldsymbol{Z} = \boldsymbol{y}$. Recalling that $\boldsymbol{\beta} =$

$(\beta_f', \mathbf{0}')'$, it follows that $\boldsymbol{y}$ has zero covariance with irrelevant factors, so $\hat{\boldsymbol{y}}^{(1)}$ also has zero covariance with irrelevant factors and therefore $\boldsymbol{r}^{(1)}$ has population loadings of zero on irrelevant factors. To see this, note that irrelevant factors can be represented as $\boldsymbol{F}[\mathbf{0}, \boldsymbol{I}]'$, where the zero matrix is $K_g \times K_f$ and the identity matrix is dimension K_g. This, together with Assumption 2.5, 4.3, implies that the cross product matrix $[\mathbf{0}, \boldsymbol{I}]\boldsymbol{F}'\boldsymbol{r}^{(1)}$ is zero in expectation.

The induction step proceeds as follows. By hypothesis, suppose we have $k < K_f$ automatically-selected proxies with factor loadings $[\boldsymbol{\Lambda}_{f,k}, \mathbf{0}]$, where $\boldsymbol{\Lambda}_{f,k}$ is $k \times K_f$ and full row rank. The residual from the k-automatic-proxy 3PRF forecast is $\boldsymbol{r}^{(k)} = \boldsymbol{y} - \hat{\boldsymbol{y}}^{(k)}$, which has zero population covariance with irrelevant factors by the same argument given in the $k = 1$ case. It is left to show that the $\boldsymbol{r}^{(k)}$'s loading on relevant factors is linearly independent of the rows of $\boldsymbol{\Lambda}_{f,k}$. To this end, note that these relevant-factor loadings take the form $\beta_f - \boldsymbol{\Phi}_f'\boldsymbol{\Omega}^{(k)}\boldsymbol{f}\beta_f$, where $\boldsymbol{f} = \boldsymbol{F}\boldsymbol{S}_{K_f}$ and $\boldsymbol{S}_{K_f} = [\boldsymbol{I}, \mathbf{0}]'$ is the matrix that selects the first K_f columns of the matrix that it multiplies (the form of this loading matrix follows again from $\beta = [\beta_f', \mathbf{0}']'$). Also note that as part of the induction hypothesis, $\boldsymbol{\Omega}^{(k)}$ is constructed based on $\boldsymbol{Z} = (\boldsymbol{r}^{(1)}, \ldots, \boldsymbol{r}^{(k-1)})$.

Next, project $\boldsymbol{r}^{(k)}$'s relevant-factor loadings onto the column space of $\boldsymbol{\Lambda}_{f,k}'$. The residual's loading vector is linearly independent of $\boldsymbol{\Lambda}_{f,k}'$ if the difference between it and its projection on $\boldsymbol{\Lambda}_{f,k}'$ is non-zero. Calculating this difference, we find $(\boldsymbol{I} - \boldsymbol{\Lambda}_{f,k}'(\boldsymbol{\Lambda}_{f,k}\boldsymbol{\Lambda}_{f,k}')^{-1}\boldsymbol{\Lambda}_{f,k})\left(\boldsymbol{I} - \boldsymbol{\Phi}_f'\boldsymbol{\Omega}^{(k)}\boldsymbol{f}\right)\beta_f$. Because $\left(\boldsymbol{I} - \boldsymbol{\Phi}_f'\boldsymbol{\Omega}^{(k)}\boldsymbol{f}\right) \neq \mathbf{0}$ with probability one, this difference is zero only when $\boldsymbol{\Lambda}_{f,k}'(\boldsymbol{\Lambda}_{f,k}\boldsymbol{\Lambda}_{f,k}')^{-1}\boldsymbol{\Lambda}_{f,k} = \boldsymbol{I}$. But the induction hypothesis ensures that this is not the case so long as $k < K_f$. Therefore the difference between the $\boldsymbol{r}^{(k)}$'s loading vector and its projection onto the column space of $\boldsymbol{\Lambda}_{f,k}'$ is nonzero, thus its loading vector is linearly independent of the rows of $\boldsymbol{\Lambda}_{f,k}$. Therefore we have constructed proxies that satisfy Assumption 6.

We next show that the L-automatic-proxy 3PRF satisfies Assumptions 2.4, 3.3 and 3.4 when the remaining parts of Assumptions 1–6 hold. Each automatic proxy is a forecast error, $z_t = y_{t+1} - \hat{y}_{t+1}$, where the forecast $\hat{y}_{t+1}$ is a linear combination of predictors. By similar limiting arguments as those leading up to Theorem 1, this linear combination can be generically expressed as $N^{-1}\boldsymbol{a}'\boldsymbol{x}_t$, where $\boldsymbol{a} = \boldsymbol{O}_p(1)$. We can rewrite an automatic proxy z_t (suppressing constants) as $z_t = \boldsymbol{b}'\boldsymbol{f}_t + \omega_t$ with $\omega_t = \eta_{t+1} + N^{-1}\boldsymbol{a}'\boldsymbol{\varepsilon}_t$.

By Assumption 2.5, the η_{t+1} and $\boldsymbol{\varepsilon}_t$ components of ω_t are independent and can be handled separately. By Assumption 2.5, 4.2, the η_{t+1} component directly satisfies Assumption 2.4, 3.3, 3.4.

Assumption 2.4 also requires $\mathbb{E}(\frac{1}{N}\sum_j a_j\varepsilon_{jt}) = 0$, and $\mathbb{E}\|\frac{1}{N}\sum_j a_j\varepsilon_{jt}\|^4 \leq M$, $\frac{1}{\sqrt{TN}}\sum_{j,t} a_j\varepsilon_{jt} = O_p(1)$, which are satisfied by Assumptions 2.3 and 3.2. Assumption 3.3 requires $\mathbb{E}\left|\frac{1}{\sqrt{TN}}\sum_{j,t} a_j\varepsilon_{jt}F_t(m)\right|^2 \leq M$, which is satisfied by Assumption 4.3. Assumption 3.4 requires $\mathbb{E}\left|\frac{1}{\sqrt{TN}}\sum_{j,t} b_j\varepsilon_{jt}\varepsilon_{it}\right|^2 \leq M$, which is satisfied by Assumption 3.2.

Together these results imply that the L-automatic-proxy satisfies the conditions of Theorems 1 and 4 when $L = K_f$. □

The following proposition simply shows that the 3PRF is the constrained least squares estimator of a projection of $\boldsymbol{y}$ onto $\boldsymbol{X}$. The body of the text interprets this constraint as the assumption that the relevant factor space is spanned by one's choice of proxy variables.

Theorem 8. *The three-pass regression filter's implied N-dimensional predictive coefficient, $\hat{\alpha}$, is the solution to*

$$\arg\min_{\alpha_0, \alpha} \|\boldsymbol{y} - \alpha_0 - \boldsymbol{X}\alpha\|$$

$$\text{subject to} \quad (\boldsymbol{I} - \boldsymbol{W}_{XZ}(\boldsymbol{S}_{XZ}'\boldsymbol{W}_{XZ})^{-1}\boldsymbol{W}_{XZ})\alpha = \mathbf{0}.$$

Proof. By the Frisch–Waugh–Lovell Theorem, the value of α that solves this problem is the same as the value that solves the least squares problem for $\|\boldsymbol{J}_T\boldsymbol{y} - \boldsymbol{J}_T\boldsymbol{X}\alpha\|$. From Amemiya (1985, Section 1.4), the estimate of α that minimizes the sum of squared residuals $(\boldsymbol{J}_T\boldsymbol{y} - \boldsymbol{J}_T\boldsymbol{X}\alpha)'(\boldsymbol{J}_T\boldsymbol{y} - \boldsymbol{J}_T\boldsymbol{X}\alpha)$ subject to the constraint $\boldsymbol{Q}'\alpha = \boldsymbol{c}$ is found by

$$\boldsymbol{R}(\boldsymbol{R}'\boldsymbol{S}_{XX}\boldsymbol{R})^{-1}\boldsymbol{R}'\boldsymbol{X}'\boldsymbol{s}_{Xy} + [\boldsymbol{I} - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{S}_{XX}\boldsymbol{R})^{-1}\boldsymbol{R}'\boldsymbol{S}_{XX}]\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{Q})^{-1}\boldsymbol{c}$$

for $\boldsymbol{R}$ such that $\boldsymbol{R}'\boldsymbol{Q} = \mathbf{0}$ and $[\boldsymbol{Q} \quad \boldsymbol{R}]$ is nonsingular. In our case,

$$\boldsymbol{c} = \mathbf{0} \quad \text{and} \quad \boldsymbol{Q} = (\boldsymbol{I} - \boldsymbol{W}_{XZ}(\boldsymbol{S}_{XZ}'\boldsymbol{W}_{XZ})^{-1}\boldsymbol{W}_{XZ}),$$

hence we can let $\boldsymbol{R} = \boldsymbol{W}_{XZ}$ and the result follows. □

A.7. Relevant proxies and relevant factors

This section explores whether, given our normalization assumptions, it is possible in general to reformulate the multi-factor system as a one-factor system, and achieve consistent forecasts with the 3PRF using a single automatically selected proxy (that is, the target-proxy 3PRF). The answer is that this is not generally possible. We demonstrate this both algebraically and in simulations. The summary of this section is:

I. There is a knife-edge case (which is ruled out by Assumption 5) in which the target-proxy 3PRF is always consistent regardless of K_f.
II. In the more general case (consistent with Assumption 5) the target-proxy 3PRF is inconsistent for $K_f > 1$ but the K_f-automatic-proxy 3PRF is consistent.

To demonstrate points 1 and 2, we begin from our normalization assumptions and show that three necessary conditions for consistency must hold for any rotation of the factor model. Second, we show that in the knife-edge case the target-proxy 3PRF is consistent (ruled out in our main development by assumption) but that the general case consistency continues to require as many proxies as there are relevant factors. This remains true when the multi-factor model is reformulated in terms of a single factor. Third, we provide simulation evidence that supports these conclusions.

Heuristically speaking, the main intuition of this section is the following: The 3PRF's consistency requires that the first-pass and second-pass regressions be consistent, which in turn requires that they have no omitted variable bias. For the first pass regression this is satisfied by the assumption that the factors are orthogonal. For the second pass regression, since it is on the *loadings* this is satisfied only once all the relevant factors have been spanned since we only require that relevant factors' loadings and irrelevant factors' loadings are orthogonal (a simple normalization assumption) and not that each relevant factor's loading is orthogonal to every other (an assumption that is stronger than mere normalization).

A.7.1. Our original representation

Our analysis centers on the probability limit given in Lemma 5. For simplicity, we assume in this appendix that y, $\boldsymbol{x}$, $\boldsymbol{F}$ and $\boldsymbol{\phi}$ are mean zero, $K_f = dim(\boldsymbol{f}) > 1$, suppress time subscripts, and assume

$$\mathbb{E}(\boldsymbol{F}\boldsymbol{F}') = \boldsymbol{\Delta}_F = \begin{bmatrix} \boldsymbol{\Delta}_f & \boldsymbol{\Delta}_{fg} \\ \boldsymbol{\Delta}_{fg}' & \boldsymbol{\Delta}_g \end{bmatrix}, \qquad \mathbb{E}(\boldsymbol{f}\boldsymbol{\varepsilon}') = \mathbf{0}, \qquad \mathbb{E}(\boldsymbol{g}\boldsymbol{\varepsilon}') = \mathbf{0}.$$

The points we make in this simpler case transfer directly to the model described in the main text. The probability limit of $\hat{y}$ may therefore be rewritten as

$$\hat{y} \xrightarrow[T,N\to\infty]{p} \boldsymbol{F}'\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right]^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\beta. \tag{A.6}$$

By inspection, consistency requires three conditions to ensure that the coefficient vector post-multiplying $\boldsymbol{F}'$ in (A.6) reduces to $(\beta_f', \mathbf{0})'$. These conditions are:

1. $\mathbf{\Lambda} = \begin{bmatrix}\mathbf{\Lambda}_f & \mathbf{0}\end{bmatrix}$ (Relevant proxies)
2. $\mathbf{\Delta}_{fg} = \mathbf{0}$ (Relevant factors orthogonal to irrelevant factors)
3. $\mathcal{P}_{fg} = \mathbf{0}$ (Relevant factors loadings orthogonal to irrelevant factors loadings).

To see that these are necessary, first note that condition 1 implies that $\mathcal{P}\mathbf{\Delta}_F\mathbf{\Lambda}'$ reduces to

$$\begin{bmatrix}\mathcal{P}_f\mathbf{\Delta}_f\mathbf{\Lambda}_f' + \mathcal{P}_{fg}\mathbf{\Delta}_{fg}'\mathbf{\Lambda}_f' \\ \mathcal{P}_{fg}'\mathbf{\Delta}_f\mathbf{\Lambda}_f' + \mathcal{P}_g\mathbf{\Delta}_{fg}\mathbf{\Lambda}_f'\end{bmatrix}. \tag{A.7}$$

Since the same matrix $(\left[\mathbf{\Lambda}\mathbf{\Delta}_F\mathcal{P}\mathbf{\Delta}_F\mathcal{P}\mathbf{\Delta}_F\mathbf{\Lambda}'\right]^{-1}\mathbf{\Lambda}\mathbf{\Delta}_F\mathcal{P}\mathbf{\Delta}_F\boldsymbol{\beta})$ post-multiplies both of these rows, we can here determine the necessity of conditions 2 and 3. The bottom row of (A.7) must be $\mathbf{0}$ for the irrelevant factors to drop out. Conditions 2 and 3 achieve this while avoiding degeneracy of the underlying factors and factor loadings.

Given necessary conditions 1–3, we have that $\hat{\boldsymbol{y}}$ is reduced to

$$\boldsymbol{f}'\mathcal{P}_f\mathbf{\Delta}_f\mathbf{\Lambda}_f'\left[\mathbf{\Lambda}_f\mathbf{\Delta}_f\mathcal{P}_f\mathbf{\Delta}_f\mathcal{P}_f\mathbf{\Delta}_f\mathbf{\Lambda}_f'\right]^{-1}\mathbf{\Lambda}_f\mathbf{\Delta}_f\mathcal{P}_f\mathbf{\Delta}_f\boldsymbol{\beta}_f. \tag{A.8}$$

Consistency requires that (A.8) reduces to $\boldsymbol{f}'\boldsymbol{\beta}_f$. We are now in a position to identify the knife-edge and general cases. The knife-edge case occurs when $\mathcal{P}_f\mathbf{\Delta}_f = \sigma\boldsymbol{I}$ and $\mathbf{\Lambda}_f = \boldsymbol{\beta}_f$, for positive scalar σ. In this case (A.8) becomes

$$\sigma\boldsymbol{\beta}_f\left[\sigma^2\boldsymbol{\beta}_f'\mathbf{\Delta}_f\boldsymbol{\beta}_f\right]^{-1}\sigma\boldsymbol{\beta}_f'\mathbf{\Delta}_f\boldsymbol{\beta}_f = \boldsymbol{\beta}_f.$$

The target-proxy 3PRF is consistent even though there are $K_f > 1$ relevant factors in the original system.

In the general case, we only assume $\boldsymbol{P}_f$, $\mathbf{\Delta}_f$, $\mathbf{\Lambda}_f$ are invertible (so that $\boldsymbol{P}_f\mathbf{\Delta}_f$ need not be an equivariance matrix). In this case (A.8) reduces to $\boldsymbol{f}'\boldsymbol{\beta}_f$. The key condition here is the invertibility of these matrices, which requires using $K_f > 1$ relevant proxies (obtainable by the auto-proxy algorithm). This is the paper's main result.

Recalling the discussion in Stock and Watson (2002a) and Section 2.2, it is quite natural that the final condition required for consistency involves both the factor (time-series) variances and the (cross-sectional) variances of the factor loadings: This is the nature of identification in factor models. The general point is that requirements for identification and consistent estimation of factor models requires assumptions regarding both factors and loadings. By convention we assume that factors are orthogonal to one another. The loadings can then be rotated in relation to the factor space we have assumed, but not all rotations are observationally-equivalent once we have pinned down the factor space.

A.7.2. A One-factor representation of the multi-factor system

Let us rewrite the factor system by condensing multiple relevant factors into a single relevant factor:

$$h = \boldsymbol{\beta}_f'\boldsymbol{f}.$$

In addition, we can rotate the original factors so that the first factor h is orthogonal to all others. Let this rotation be achieved by some matrix $\boldsymbol{M}$ such that

$$\boldsymbol{m} = \boldsymbol{M}'\boldsymbol{f},$$

$$\mathbb{E}\left[\begin{pmatrix}h \\ \boldsymbol{m}\end{pmatrix}\begin{pmatrix}h & \boldsymbol{m}\end{pmatrix}\right] = \begin{pmatrix}\boldsymbol{\beta}_f' \\ \boldsymbol{M}'\end{pmatrix}\mathbf{\Delta}_f\begin{pmatrix}\boldsymbol{\beta}_f & \boldsymbol{M}\end{pmatrix} = \begin{bmatrix}\Delta_h & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}_m\end{bmatrix}. \tag{A.9}$$

The new formulation therefore satisfies

$$y = h + \eta$$
$$\boldsymbol{x} = \mathbf{\Psi}_h h + \mathbf{\Psi}_m\boldsymbol{m} + \mathbf{\Psi}_g\boldsymbol{g} + \boldsymbol{\varepsilon}$$
$$\mathbf{\Lambda} = \begin{bmatrix}1 & \mathbf{0}\end{bmatrix}.$$

Now h is the single relevant factor while $(\boldsymbol{m}', \boldsymbol{g}')'$ are the irrelevant factors. We have represented the system such that first two necessary conditions for consistency are satisfied. We now show that the third necessary condition will not be satisfied in general.

Let us write the loadings in this rotated system $(\mathbf{\Psi}_h, \mathbf{\Psi}_m, \mathbf{\Psi}_g)$ in terms of the loadings in the original system $(\mathbf{\Phi}_f, \mathbf{\Phi}_g)$. Because $\mathbb{E}(h\boldsymbol{m}')$, $\mathbb{E}(h\boldsymbol{g})$, $\mathbb{E}(\boldsymbol{m}\boldsymbol{g}')$ are all zero, we recover

$$\mathbb{E}\left((\boldsymbol{x} - \mathbf{\Psi}_h h)h\right) = 0 \Rightarrow \mathbf{\Psi}_h = \frac{1}{\boldsymbol{\beta}_f'\mathbf{\Delta}_f\boldsymbol{\beta}_f}\mathbf{\Phi}_f\mathbf{\Delta}_f\boldsymbol{\beta}_f$$
$$\mathbb{E}\left((\boldsymbol{x} - \mathbf{\Psi}_m\boldsymbol{m})\boldsymbol{m}'\right) = \mathbf{0} \Rightarrow \mathbf{\Psi}_m = \mathbf{\Phi}_f\mathbf{\Delta}_f\boldsymbol{M}\left(\boldsymbol{M}'\mathbf{\Delta}_f\boldsymbol{M}\right)^{-1}$$
$$\mathbb{E}\left((\boldsymbol{x} - \mathbf{\Psi}_g\boldsymbol{g})\boldsymbol{g}'\right) = \mathbf{0} \Rightarrow \mathbf{\Psi}_g = \mathbf{\Phi}_g.$$

The covariance matrix of loadings is therefore

$$N^{-1}\sum_{i=1}^N\begin{pmatrix}\psi_{h,i} \\ \boldsymbol{\psi}_{m,i} \\ \boldsymbol{\psi}_{g,i}\end{pmatrix}\begin{pmatrix}\psi_{h,i} & \boldsymbol{\psi}_{m,i}' & \boldsymbol{\psi}_{g,i}'\end{pmatrix}$$
$$= N^{-1}\sum_{i=1}^N\begin{bmatrix}\psi_{h,i}^2 & \psi_{h,i}\boldsymbol{\psi}_{m,i}' & \psi_{h,i}\boldsymbol{\psi}_{g,i}' \\ \psi_{h,i}\boldsymbol{\psi}_{m,i} & \boldsymbol{\psi}_{m,i}\boldsymbol{\psi}_{m,i}' & \boldsymbol{\psi}_{m,i}\boldsymbol{\psi}_{g,i}' \\ \psi_{h,i}\boldsymbol{\psi}_{g,i} & \boldsymbol{\psi}_{g,i}\boldsymbol{\psi}_{m,i}' & \boldsymbol{\psi}_{g,i}\boldsymbol{\psi}_{g,i}'\end{bmatrix}$$

and the third necessary condition is determined by whether or not the matrix

$$N^{-1}\sum_{i=1}^N\left[\psi_{h,i}\boldsymbol{\psi}_{m,i}' \quad \psi_{h,i}\boldsymbol{\psi}_{g,i}\right]$$

equals zero in the limit. The second element $\psi_{h,i}\boldsymbol{\psi}_{g,i}$ has a zero limit whenever the original system satisfies its three necessary conditions. But the first element $\psi_{h,i}\boldsymbol{\psi}_{m,i}'$ has a limit determined by whether the knife-edge or the general case holds since

$$N^{-1}\sum_{i=1}^N\psi_{h,i}\boldsymbol{\psi}_{m,i}' \xrightarrow[N\to\infty]{p} \frac{1}{\boldsymbol{\beta}_f'\mathbf{\Delta}_f\boldsymbol{\beta}_f}\boldsymbol{\beta}_f'\mathbf{\Delta}_f\mathcal{P}_f\mathbf{\Delta}_f\boldsymbol{M}\left(\boldsymbol{M}'\mathbf{\Delta}_f\boldsymbol{M}\right)^{-1}.$$

The critical term in determining whether this expression reduces to zero is $\boldsymbol{\beta}_f'\mathbf{\Delta}_f\mathcal{P}_f\mathbf{\Delta}_f\boldsymbol{M}$. If the knife-edge condition holds, then we have $\boldsymbol{\beta}_f'\mathbf{\Delta}_f\mathcal{P}_f\mathbf{\Delta}_f\boldsymbol{M} = \sigma\boldsymbol{\beta}_f'\mathbf{\Delta}_f\boldsymbol{M} = \mathbf{0}$ in light of (A.9). However, in the general case, $\boldsymbol{\beta}_f'\mathbf{\Delta}_f\mathcal{P}_f\mathbf{\Delta}_f\boldsymbol{M} \neq \mathbf{0}$ even though (A.9) holds and the third necessary condition cannot generally be satisfied in this rewritten system.

A.7.3. Simulation study

We now run a Monte Carlo to demonstrate that, when there are multiple relevant factors, a target-proxy achieves the infeasible best only when the knife-edge case holds. Our simulation design uses the following:

$$\boldsymbol{y} = \boldsymbol{f}\boldsymbol{\iota} + \boldsymbol{\eta}, \qquad \boldsymbol{X} = \begin{bmatrix}\boldsymbol{f} & \boldsymbol{g}\end{bmatrix}\mathbf{\Phi}' + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\iota}$ is $K_f \times 1$ ones vector, $\boldsymbol{g}(T \times K_g)$, $\mathbf{\Phi}(N \times K_f + K_g)$, $\boldsymbol{\eta}(T \times 1)$, and $\boldsymbol{\varepsilon}(T \times N)$ are iid standard normal, and $\boldsymbol{f}(T \times K_f)$ is iid normal with standard deviation $\boldsymbol{\sigma}_f$.

The infeasible best forecast for this system is $\boldsymbol{f}\boldsymbol{\iota}$. We use six factors, three relevant and three irrelevant ($K_f = K_g = 3$) and consider different values for N, T and $\boldsymbol{\sigma}_f$. We consider $N = T = 200$ and $N = T = 2000$. We use an identity covariance matrix for factor loadings ($\mathcal{P} = \boldsymbol{I}$) and consider two values for $\boldsymbol{\sigma}_f$: a knife-edge (equivariant) case $[1 \quad 1 \quad 1]$and a more general (non-equivariant) case $\begin{bmatrix}0.5 & 1 & 2\end{bmatrix}$.

Table A.1 lends simulation support to our algebraic proof. We focus on in-sample results since out-of-sample results are qualitatively similar.

In the knife-edge case the target-proxy 3PRF appears consistent. For $N = T = 2000$ the correlation between the 3PRF forecast and the infeasible best forecast is 0.993, and their relative R^2 is 0.9901.

Table A.1
Simulation study.

# auto proxies:	In-Sample			Out-of-Sample		
	1	2	3	1	2	3
$N = T = 200$						
$\boldsymbol{\sigma}_f = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$						
$\frac{\hat{y}R^2}{\boldsymbol{f}\boldsymbol{\iota}R^2}$	0.9607			0.9316		
$\rho(\hat{y}, \boldsymbol{f}\boldsymbol{\iota})$	0.9678			0.9649		
$\boldsymbol{\sigma}_f = \begin{bmatrix} 0.5 & 1 & 2 \end{bmatrix}$						
$\frac{\hat{y}R^2}{\boldsymbol{f}\boldsymbol{\iota}R^2}$	0.8425	0.9736	0.9938	0.8307	0.9580	0.9735
$\rho(\hat{y}, \boldsymbol{f}\boldsymbol{\iota})$	0.9169	0.9806	0.9892	0.9136	0.9791	0.9884
$N = T = 2000$						
$\boldsymbol{\sigma}_f = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$						
$\frac{\hat{y}R^2}{\boldsymbol{f}\boldsymbol{\iota}R^2}$	0.9901			0.9850		
$\rho(\hat{y}, \boldsymbol{f}\boldsymbol{\iota})$	0.9930			0.9929		
$\boldsymbol{\sigma}_f = \begin{bmatrix} 0.5 & 1 & 2 \end{bmatrix}$						
$\frac{\hat{y}R^2}{\boldsymbol{f}\boldsymbol{\iota}R^2}$	0.8586	0.9762	0.9983	0.8575	0.9746	0.9962
$\rho(\hat{y}, \boldsymbol{f}\boldsymbol{\iota})$	0.9241	0.9877	0.9981	0.9238	0.9876	0.9981

Notes: $\frac{\hat{y}R^2}{\boldsymbol{f}\boldsymbol{\iota}R^2}$ denotes the average ratio of 3PRF R^2 to the infeasible best R^2. $\rho(\hat{y}, \boldsymbol{f}\boldsymbol{\iota})$ gives the average time series correlation between the 3PRF forecast and the infeasible best forecast.

For $N = T = 200$ these numbers are lower, but that is attributable to the smaller sample.

In the general case the target-proxy 3PRF appears inconsistent. The relative R^2 is 0.8425 for $N = T = 200$ and 0.8586 for $N = T = 2000$; the correlation is 0.9169 for $N = T = 200$ and 0.9241 for $N = T = 2000$. This agreement across the two sample sizes is strongly suggestive that the inconsistency is not a small sample issue, but rather holds in large N, T for which 2000 is a good approximation. Furthermore, the relative R^2 increases notably as we move to 2 auto-proxies: 0.9736 for $N = T = 200$ and 0.9762 for $N = T = 2000$. Once we have 3 auto-proxies (as our theorem states) the simulation evidence suggests that the 3PRF is consistent. The relative R^2 is 0.9938 for $N = T = 200$ and 0.9983 for $N = T = 2000$.

A.8. Accuracy of asymptotic theory in finite samples

Our first experiment evaluates the accuracy of finite sample approximations based on the asymptotic distributions we have derived. We examine the distributions of predictive coefficient estimates as well as the forecasts themselves. For each Monte Carlo draw, we first compute the estimates $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. Then we standardize each estimate in accordance with Theorems 3–5 by subtracting off the mean adjustment term and dividing by the respective asymptotic standard error estimate. According to the theory presented in Section 2, these standardized estimates should follow a standard normal distribution for large N and T.

For each estimator (corresponding to Figs. A.1–A.3) we plot the distribution of standardized estimates across simulations (solid line) versus the standard normal pdf (dashed line). The four panels of each figure correspond to $N = 100$, $T = 100$ and $N = 500$, $T = 500$ in the cases that (i) there is a single relevant factor and (ii) there is one relevant and one irrelevant factor. Factors, factor loadings and shocks are drawn from a standard normal distribution. The predictive loading on the relevant factor is set to one (that is, the infeasible best R^2 is set equal to 50%). We simulate 5000 samples for each set of parameter values.

These results show that the standard normal distribution successfully describes the finite sample behavior of these estimates, consistent with the results in Section 2. In all cases but one we fail to reject the standard normal null hypothesis for standardized estimates. The exception occurs for $\hat{\boldsymbol{\beta}}$ when $N = 100$ and $T = 100$, which demonstrates a minor small sample bias (Fig. A.3, upper right). This bias vanishes when the sample size increases (Fig. A.3, lower right). The simulated coverage rates of a 0.95 confidence interval for $\hat{y}_{t+1}$ are also well behaved. For $N = 100$ and $T = 100$ the simulated coverage is 0.95 when there is no irrelevant factor and 0.94 when an irrelevant factor exists. For $N = 500$ and $T = 500$ the simulated coverage is 0.947 and 0.949 and, respectively. Altogether, simulations provide evidence that the 3PRF accurately estimates the infeasible best forecasts and predictive coefficients, and that its theoretical asymptotic distributions accurately approximate the finite sample distributions for 3PRF estimates.

A.9. Information criterion Monte Carlo

In Section 4.2 we discuss an information criterion for selecting the number of predictive indices when using the auto-proxy 3PRF. Our degrees of freedom calculation uses the "Trace of the Krylov Representation" method of Kramer and Sugiyama (2011). In particular, when using m 3PRF automatic proxies, the degrees of freedom are given by

$$\widehat{DoF}(m) = 1 + \sum_{j=1}^{m} c_j \text{trace}(\boldsymbol{K}^j) - \sum_{l,j=1}^{m} \boldsymbol{t}_l' \boldsymbol{K}^j \boldsymbol{t}_l + (\boldsymbol{y} - \hat{\boldsymbol{y}}_m)' \sum_{j=1}^{m} \boldsymbol{K}^j \boldsymbol{v}_j + m$$

where $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}'$, c_j are elements of the vector $\boldsymbol{c} = \boldsymbol{B}^{-1}\boldsymbol{T}\boldsymbol{y}$, $\boldsymbol{B}$ is a Krylov basis decomposition, $\boldsymbol{T}$ is the matrix of PLS factor estimate vectors $\boldsymbol{t}_j$, and $\boldsymbol{v}_j$ are columns of the matrix $\boldsymbol{T}(\boldsymbol{B}^{-1})'$. The BIC is then calculated as $\sum_t (y_t - \hat{y}_{m,t})^2/T + \log(T)\hat{\sigma}^2\widehat{DoF}(m)/T$ where $\hat{\sigma} = \sqrt{\sum_t (y_t - \hat{y}_{m,t})^2/(T - DoF(m))}$. We refer readers to Kramer and Sugiyama (2011) for additional details.

Table A.2 studies the accuracy of the information criterion in the simulation specifications of Table 3, and compares how 3PRF1

Table A.2
Simulated out-of-sample forecast performance using information criterion.

ρ_f	ρ_g	a	d	$N = T = 100$			$N = T = 200$		
				3PRF1	3PRFIC	#IC	3PRF1	3PRFIC	#IC
				Panel A: Normal Factors					
0	0	0	0	**28.4**	16.8	1.1	**38.2**	34.4	1.0
0.3	0.9	0.3	0	**26.3**	16.8	2.0	**37.4**	34.8	1.0
0.3	0.9	0.3	0.9	**25.2**	17.8	1.4	**36.8**	35.3	1.0
0.3	0.9	0.9	0	18.1	**23.8**	2.4	31.9	**38.1**	2.0
0.3	0.9	0.9	0.9	18.0	**24.1**	2.5	31.5	**38.3**	2.0
0.9	0.3	0.3	0	**33.0**	20.9	1.1	**41.0**	34.3	1.0
0.9	0.3	0.3	0.9	**30.8**	21.8	1.3	**40.3**	36.0	1.0
0.9	0.3	0.9	0	**31.7**	27.1	1.4	**37.6**	37.1	1.2
0.9	0.3	0.9	0.9	**30.7**	27.9	1.6	36.3	**36.5**	1.3
				Panel B: Moderately Weak Factors					
0	0	0	0	**21.3**	7.2	1.1	**34.5**	30.4	1.0
0.3	0.9	0.3	0	**20.1**	8.0	1.1	**33.5**	30.6	1.0
0.3	0.9	0.3	0.9	**17.7**	8.0	1.4	**32.1**	30.6	1.0
0.3	0.9	0.9	0	9.6	**16.1**	2.8	24.4	**32.8**	2.2
0.3	0.9	0.9	0.9	9.4	**17.5**	2.9	23.6	**33.4**	2.4
0.9	0.3	0.3	0	**27.5**	10.9	1.1	**37.5**	29.9	1.0
0.9	0.3	0.3	0.9	**23.7**	11.4	1.3	**35.6**	31.2	1.0
0.9	0.3	0.9	0	**27.8**	23.5	1.4	**33.0**	32.6	1.2
0.9	0.3	0.9	0.9	**26.7**	23.7	1.6	31.3	**31.8**	1.3
				Panel C: Weak Factors					
0	0	0	0	**8.4**	−16.5	1.2	**24.6**	19.4	1.0
0.3	0.9	0.3	0	**7.7**	−11.9	1.3	**23.6**	19.8	1.0
0.3	0.9	0.3	0.9	**3.3**	−10.9	1.9	**19.7**	17.1	1.0
0.3	0.9	0.9	0	−0.9	**2.1**	3.1	10.7	**20.9**	2.8
0.3	0.9	0.9	0.9	−1.5	**3.5**	3.4	10.1	**21.5**	3.0
0.9	0.3	0.3	0	**17.6**	−5.8	1.1	**29.0**	20.2	1.0
0.9	0.3	0.3	0.9	**11.6**	−4.4	1.6	**24.6**	19.3	1.0
0.9	0.3	0.9	0	**24.0**	17.5	1.4	**26.5**	26.4	1.2
0.9	0.3	0.9	0.9	**22.2**	17.8	1.5	**24.9**	25.3	1.3

Notes: The table reports performance in terms of out-of-sample forecast percentage R^2 for the 3PRF based on the actual number of relevant factors (column 3PRF1) or the number of factors selected by an information criterion (column 3PRFIC). For the IC version, we report the average number of factors chosen by the criterion (column #IC). The data generating processes are described in Section 4 and Table 3.

forecasts compare to those using the number of 3PRF factors selected by the information criterion (denoted 3PRFIC). We report the out-of-sample forecast R^2 for 3PRF1 and 3PRFIC, as well as the average number of factors selected by the information criterion. In the $T, N = 100$ case, the performance of 3PRF suffers when the number of factors is chosen according to the information criterion. The largest setbacks occur when the irrelevant factors or idiosyncrasies display strong serial dependence. The average number of factors chosen ranges from 1.1 to 3.4

In larger sample ($T, N = 200$), 3PRF performance is much less affected by relying on the information criterion to select the number of factors. The drop in R^2 versus 3PRF tends to be small, and for most parameter configurations the average number of predictors chosen is 1.0.

The table also includes some pathological cases in which 3PRFIC outperforms 3PRF1. This occurs when both the irrelevant factors and the idiosyncrasies are strongly serially correlated, but the relevant factors are quickly mean reverting. In this case, the first 3PRF factor is corrupted by persistent irrelevant information, and additional 3PRF factors allow the procedure to pick up residual relevant information missed by the first factor.

A.10. Partial least squares

Like the three-pass regression filter and principal components, partial least squares (PLS) constructs forecasting indices as linear combinations of the underlying predictors. These predictive indices are referred to as "directions" in the language of PLS. The PLS forecast based on the first K PLS directions, $\hat{\boldsymbol{y}}^{(k)}$, is constructed according to the following algorithm (as stated in Hastie et al. (2009)):

1. Standardize each $\mathbf{x}_i$ to have mean zero and variance one by setting $\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i - \hat{\mathbb{E}}[\mathbf{x}_{it}]}{\hat{\sigma}(\mathbf{x}_{it})}$, $i = 1, \ldots, N$
2. Set $\hat{\boldsymbol{y}}^{(0)} = \bar{y}$, and $\mathbf{x}_i^{(0)} = \tilde{\mathbf{x}}_i$, $i = 1, \ldots, N$
3. For $k = 1, 2, \ldots, K$
   (a) $\boldsymbol{u}_k = \sum_{i=1}^N \hat{\phi}_{ki}\mathbf{x}_i^{(k-1)}$, where $\hat{\phi}_{ki} = \widehat{Cov}(\mathbf{x}_i^{(k-1)}, \boldsymbol{y})$
   (b) $\hat{\beta}_k = \widehat{Cov}(\boldsymbol{u}_k, \boldsymbol{y})/\widehat{Var}(\boldsymbol{u}_k)$
   (c) $\hat{\boldsymbol{y}}^{(k)} = \hat{\boldsymbol{y}}^{(k-1)} + \hat{\beta}_k\boldsymbol{u}_k$
   (d) Orthogonalize each $\mathbf{x}_i^{(k-1)}$ with respect to $\boldsymbol{u}_k$:
   $$\mathbf{x}_i^{(k)} = \mathbf{x}_i^{(k-1)} - \left(\widehat{Cov}(\boldsymbol{u}_k, \mathbf{x}_i^{(k-1)})/\widehat{Var}(\boldsymbol{u}_k)\right)\boldsymbol{u}_k,$$
   $$i = 1, 2, \ldots, N.$$

A.11. Empirical procedures

The recursive out-of-sample forecasting procedure for macroeconomic data following Bai and Ng (2008) and Stock and Watson (2012) is as follows. Before forecasting each target, we first transform the data by partialing the target and predictors with respect to a constant and four lags of the target. To construct a time $t+1$ out-of-sample forecast, consider the data known at time t: $\mathcal{Y}_t \equiv \{y_t, \boldsymbol{x}_t \boldsymbol{z}_t, y_{t-1}, \boldsymbol{x}_{t-1}, \boldsymbol{z}_{t-1}, \ldots\}$. Calculate either the 3PRF's three passes or PCR's eigenvalue decomposition on $\mathcal{Y}_t$. For the target-proxy 3PRF: the first pass regressions are of $x_{i,\tau-1}$ on y_τ and a constant for $\tau = 1, 2, \ldots, t$, separately run for each $i = 1, 2, \ldots, N$, yielding $\hat{\phi}_i$; the second pass regression is of $x_{i,\tau}$ on $\hat{\phi}_i$ and a constant for $i = 1, 2, \ldots, N$, separately run for each $\tau = 1, 2, \ldots, t$, yielding $\hat{f}_\tau$; the third pass regression is of y_τ on $\hat{f}_{\tau-1}$ and a constant for $\tau = 1, 2, \ldots, t$, yielding $\hat{\beta}_0, \hat{\beta}$. Then the out-of-sample forecast is constructed as $\hat{\beta}_0 + \hat{f}_t\hat{\beta}$.

For financial data, we do not partial the target or predictors as a first step. But the remaining steps of the recursive out-of-sample forecasting procedure are done, to ensure that a time t forecast (of the time $t+1$ realization) uses only information (and estimates) available at time t.

A.12. Portfolio data construction

We construct portfolio-level log price–dividend ratios from the CRSP monthly stock file using data on prices and returns with and without dividends. Twenty-five portfolios (five-by-five sorts) are formed on the basis of underlying firms' market equity and book-to-market ratio, mimicking the methodology of Fama and French (1992). Characteristics for year t are constructed as follows. Market equity is price multiplied by common shares outstanding at the end of December. Book-to-market is the ratio of book equity in year $t-1$ to market equity at the end of year t. Book equity is calculated from the Compustat file as book value of stockholders' equity plus balance sheet deferred taxes and investment tax credit (if available) minus book value of preferred stock. Book value of preferred stock is defined as either the redemption, liquidation or par value of preferred stock (in that order). When Compustat data is unavailable, we use Moody's book equity data (if available) from Davis et al. (2000). We focus on annual data to avoid seasonality in dividends, as is common in the literature. Unlike Fama and French, we rebalance the characteristic-based portfolios each month. Using portfolio returns with and without dividends, we calculate the log price–dividend ratio for these portfolios at the end of December the following year.

For a stock to be assigned to a portfolio at time t, we require that it is classified as common equity (CRSP share codes 10 and 11) traded on NYSE, AMEX or NASDAQ, and that its $t-1$ year-end market equity value is non-missing. When forming portfolios on the basis of book-to-market we require that a firm has positive book equity at $t-1$. Because we are working with log price–dividend ratios, a firm is included only if it paid a dividend at any time in the twelve months prior to t. We perform sorts simultaneously rather than sequentially to ensure uniformity in characteristics across portfolios in both dimensions. Stock sorts for characteristic-based portfolio assignments are performed using equally-spaced quantiles as breakpoints to avoid excessively lop-sided allocations of firms to portfolios. That is, for a K-bin sort, portfolio breakpoints are set equal to the $\{\frac{100}{K}, 2\frac{100}{K}, \ldots, (K-1)\frac{100}{K}\}$ quantiles of a given characteristic.

References

Bai, J., 2003. Inferential theory for factor models of large dimensions. Econometrica 71 (1), 135–171.
Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.
Bai, J., Ng, S., 2006. Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. Econometrica 74 (4), 1133–1150.
Bai, J., Ng, S., 2008. Forecasting economic time series using targeted predictors. J. Econometrics 146 (2), 304–317.
Boivin, J., Ng, S., 2006. Are more data always better for factor analysis? J. Econometrics 132 (1), 169–194.
Campbell, J., Shiller, R., 1988. The Dividend-price ratio and expectations of future dividends and discount factors. Rev. Financ. Stud. 1 (3), 195–228.
Chamberlain, G., Rothschild, M., 1983. Arbitrage, factor structure, and mean–variance analysis on large asset markets. Econometrica 51 (5), 1281–1304.
Cochrane, J., 1992. Explaining the variance of price-dividend ratios. Rev. Financ. Stud. 5 (2), 243–280.
Cochrane, J., 2011. Presidential address: Discount rates. J. Finance 66, 1047–1108.
Davis, J., Fama, E., French, K., 2000. Characteristics, covariances, and average returns: 1929 to 1997. J. Finance 55 (1), 389–406.
De Mol, C., Giannone, D., Reichlin, L., 2008. Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? J. Econometrics 146 (2), 318–328.
Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13, 253–263.
Doz, C., Giannone, D., Reichlin, L., 2012. A Quasi–maximum likelihood approach for large, approximate dynamic factor models. Rev. Econ. Stat. 94 (4), 1014–1024.
Fama, E., 1981. Stock returns, real activity, inflation, and money. Am. Econ. Rev. 71 (4), 545–565.
Fama, E., 1982. Inflation, output, and money. J. Bus. 55 (2), 201–231.
Fama, E., French, K., 1988. Permanent and temporary components of stock prices. J. Political Economy 96 (2), 246–273.
Fama, E., French, K., 1992. The cross-section of expected stock returns. J. Finance 47 (2), 427–465.
Fama, E.F., MacBeth, J.D., 1973. Risk, return, and equilibrium: Empirical tests. J. Political Economy 607–636.
Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2000. The generalized dynamic-factor model: Identification and estimation. Rev. Econ. Stat. 82 (4), 540–554.
Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2004. The generalized dynamic factor model consistency and rates. J. Econometrics 119 (2), 231–255.
Forni, M., Hallin, M., Lippi, M., Reichlin, L., 2005. The generalized dynamic factor model: One-sided estimation and forecasting. J. Amer. Statist. Assoc. 100, 830–840.
Forni, M., Reichlin, L., 1996. Dynamic common factors in large cross-sections. Empir. Econ. 21 (1), 27–42.
Forni, M., Reichlin, L., 1998. Let's get real: A factor analytical approach to disaggregated business cycle dynamics. Rev. Econom. Stud. 65 (3), 453–473.
Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21 (4), 1455–1508.
Groen, J.J. J., Kapetanios, G., 2009, Revisiting Useful Approaches to Data-Rich Macroeconomic Forecasting, Working Papers 327, Federal Reserve Bank of New York.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer.
Hodrick, R., 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. Rev. Financ. Stud. 5 (3), 357.
Huber, P.J., 1973. Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Stat. 1 (5), 799–821.
Kelly, B., Pruitt, S., 2013. Market expectations in the cross-section of present values. J. Finance 68 (5), 1721–1756.
Kleibergen, F., Zhan, Z., 2013, Unexplained Factors and Their Effects on Second Pass R-Squareds and t-Tests, Discussion paper, Brown University.
Kramer, N., Sugiyama, M., 2011. The degrees of freedom of partial least squares regression. J. Am. Stat. Assoc. 106 (494), 697–705.
Ludvigson, S.C., Ng, S., 2009. Macro factors in bond risk premia. Rev. Financ. Stud. 22 (12), 5027–5067.
Onatski, A., 2012. Asymptotics of the principal components estimator of large factor models with weakly influential factors. J. Econometrics 168 (2), 244–258.
Rozeff, M., 1984. Dividend yields are equity risk premiums. J. Portfolio Manag. 11 (1), 68–75.
Stambaugh, R., 1986, Bias in Regressions with Lagged Stochastic Regressors, Working Paper, University of Chicago.
Stock, J.H., Watson, M.W., 1998, Diffusion Indexes, NBER Working Paper.
Stock, J.H., Watson, M.W., 2002a. Forecasting using principal components from a large number of predictors. J. Amer. Statist. Assoc. 97 (460), 1167–1179.
Stock, J.H., Watson, M.W., 2002b. Macroeconomic forecasting using diffusion indexes. J. Bus. Econom. Statist. 20 (2), 147–162.
Stock, J.H., Watson, M.W., 2006. In: Elliott, G., Granger, C., Timmermann, A. (Eds.), Forecasting with Many Predictors. In: Handbook of Economic Forecasting, vol. 1. Elsevier, pp. 515–554.
Stock, J.H., Watson, M.W., 2012. Generalized shrinkage methods for forecasting using many predictors. J. Bus. Econom. Statist. 30 (4), 481–493.
Vuolteenaho, T., 2002. What drives firm-level stock returns?. J. Finance 57 (1), 233–264.
West, K.D., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.
Wold, H., 1975. Estimation of principal components and related models by iterative least squares. In: Krishnaiaah, P. (Ed.), Multivariate Analysis. Academic Press, New York, pp. 391–420.