CHICAGO BOOTH

The University of Chicago Booth School of Business

Working Paper No. 11-19
Fama-Miller Paper Series

The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors

Bryan Kelly
University of Chicago Booth School of Business

Seth Pruitt
Federal Reserve Board of Governors

The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors[*]

Bryan Kelly[†]
University of Chicago
Booth School of Business

Seth Pruitt[‡]
Federal Reserve
Board of Governors

First Draft: January 2011
This Draft: September 2011

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 converge to the infeasible best forecast when both the time dimension and cross section dimension become large. This requires only specifying the number of relevant factors driving the forecast target, regardless of the total number of common (including potentially irrelevant) factors driving the cross section of predictors. We derive inferential theory including limiting distributions for estimated relevant factors, predictive coefficients and forecasts, and provide consistent standard error estimators. We explore two empirical applications: Forecasting macroeconomic aggregates with a large panel of economic indices, and forecasting stock market aggregates with many individual assets' price-dividend ratios. These, combined with a range of Monte Carlo experiments, demonstrate the 3PRF's forecasting power.

Key words: forecast, many predictors, factor model, principal components, Kalman filter, constrained least squares, partial least squares

[*] We thank Andrew Ang, Gray Calhoun, John Cochrane, Drew Creal, Jim Hamilton, Serena Ng, Jim Stock and Mark Watson and participants at University of Chicago (Booth), NBER Summer Institute, Federal Reserve Board, Econometric Society N.A. Summer Meeting, SoFiE Annual Meeting, NBER-NSF Time Series Conference and SITE for helpful comments. We thank Mark Watson for kindly sharing data.

[†] Mail: 5807 S. Woodlawn Ave., Chicago, IL 60637. E-mail: bryan.kelly@chicagobooth.edu. Phone: (773) 702-8359.

[‡] Mail: Board of Governors of the Federal Reserve System, Washington, DC 20551. E-mail: seth.j.pruitt@frb.gov. Phone: (202) 452-2344.

This paper reflects the views and opinions of the authors and does not reflect the views or opinions of the Federal Reserve Board, the Federal Reserve System, or their respective staffs.

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.

In addition to proposing the estimator, 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 relevant factors, predictive coefficients, and forecasts, and provide consistent estimators of asymptotic covariance matrices that can be used to perform inference. Second, we compare the 3PRF to other methods in order to illustrate the source of its improvement in forecasting performance. We show that the 3PRF is the solution to a constrained least squares problem, that the 3PRF resembles a restricted Kalman filter, and that the method of partial least squares is a special case of the 3PRF. Throughout we develop numerous comparisons with principal components regression, which is the economics literature's benchmark method of forecasting using many predictors. The third contribution of the paper is to investigate the finite sample accuracy of our asymptotic theory through Monte Carlo simulations. We find that the 3PRF accurately estimates the (infeasible) best possible forecast in a variety of experimental designs and in small samples. The final contribution of

[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.

the paper is to provide empirical support for the 3PRF's strong forecasting performance in two separate applications. The first applies the procedure to a well-known macroeconomic data set in order to forecast key macroeconomic aggregates. The second application is an asset-pricing study linking the cross section of price-dividend ratios to market expectations.

1.1 Existing Procedures

The fact that the 3PRF is a constrained least squares forecast is closely tied to the original motivation for dimension reduction: Unconstrained least squares forecasts are poorly behaved when N is large relative to T. The 3PRF imposes an intuitive constraint which ensures that the factors irrelevant to $\boldsymbol{y}$ drop out of the 3PRF forecast.

There is a link between the 3PRF and the Kalman filter, which is the theoretically optimal state space method (see Maybeck (1979) and Hamilton (1994)) but sometimes carries a debilitating computational burden.[2] We show that the 3PRF acts like a restricted Kalman filter. The restrictions implicit in this interpretation of the 3PRF include using OLS in place of generalized least squares, using observable proxies in lieu of more computationally-expensive factor estimates, and ignoring the temporal pooling of past information. Importantly, the 3PRF retains the Kalman filter's cross-sectional combination of information via least squares.[3] Imposing these restrictions is valuable because they achieve an accurate and easily implemented estimator while sacrificing theoretical (though often infeasible) efficiency. As N and T grow, the cost of these restrictions vanishes since the 3PRF converges to the optimal forecast, while the computational benefit becomes increasingly valuable.

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. The important distinction from PLS is that the 3PRF also allows the econometrician to select alternative disciplining variables, or factor proxies, on the basis of economic theory. Furthermore, because it is a special case of our methodology, the asymp-

[2] The Kalman filter likelihood-based parameter estimates are not available in closed form and must be obtained via numerical optimization. Computational demands become substantial as the number of predictors grows. As Bai (2003) notes, "As N increases, the state space and the number of parameters to be estimated increase very quickly, rendering the estimation problem challenging, if not impossible." Accordingly, large N forecasting applications often avoid the Kalman filter due to the difficulty of parameter estimation. Jungbacker and Koopman (2008) show that, in some applications, the filtering algorithm may be rewritten to speed up these computations and restore feasibility of Kalman filter maximum likelihood estimation.

[3] Watson and Engle's (1983) EM algorithm approach to state space estimation illustrates how the cross sectional pooling of information in a Kalman may be obtained from GLS regressions. See Section 3 and Appendix A.8 for further detail.

totic theory we develop for the 3PRF applies directly to partial least squares. To the best of our knowledge, these joint N and T asymptotics are a new result to the PLS literature.

The economics literature relies on principal component regression (PCR) for many-predictor forecasting problems, exemplified by Stock and Watson (1989, 1998, 2002a,b, 2006, 2009), Forni and Reichlin (1996, 1998), Forni, Hallin, Lippi and Reichlin (2000, 2004, 2005), Bai and Ng (2002, 2006, 2008, 2009) and Bai (2003), among others. 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 factors *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.

The key difference between PCR and 3PRF is their method of dimension reduction. The 3PRF condenses the cross section according to *covariance with the forecast target.* To do so, the 3PRF first assesses how strongly each predictor is related to the relevant factors. This is achieved by calculating predictors' covariances with either theoretically motivated or automatically selected proxies for the relevant latent factors (as we prove, automatic proxies are always available by construction). Next, a linear combination of predictors is constructed to mimic each relevant factor. The weights of individual predictors in this linear combination are based on the strength of predictors' estimated covariance with the proxies. This step consistently estimates a rotation of the relevant factors that is in turn used to forecast the target. The ultimate prediction is a discerningly constructed linear combination of individual predictors with powerful forecast performance.

We are not the first to investigate potential improvements upon PCR forecasts. Bai and Ng (2008) tackle this issue with statistical thresholding rules that drop variables found to contain irrelevant information. Thresholding requires that irrelevant factors only affect a relatively small subset of predictors since dropping predictors works against the large N feature of PCR. Bai and Ng's (2009) statistically boosted PCR forecasts are closer in spirit to our paper. This approach recognizes that some principal components may not help in forecasting the target, then uses forecast error to guide its component selection. In a similar vein, Stock and Watson (2011) use shrinkage methods to downweight components that are unrelated to the target. Each of those papers first uses principal components to reduce the dimension of the predictors. Our approach differs in that we explicitly allow irrelevant information to be pervasive among all predictors within the basic model specification. Because of this, the notion of "relevance" is directly incorporated into the way we perform our dimension reduction. Furthermore, this allows us to directly derive the limiting distribution of our estimator despite the presence of irrelevant factors.

1.2 Empirical Results

In the first empirical investigation, we forecast macroeconomic aggregates using a well-known panel of quarterly macroeconomic variables that has been explored in PCR forecasting studies (see Stock and Watson (2002b, 2006, 2011) and Ludvigson and Ng (2009), among others). There we find that the 3PRF uncovers factors that significantly improve upon the performance of autoregressive forecasts for key macroeconomic variables. These results link our paper to macroeconomic theories built upon dynamic factor models including Geweke (1977), Sargent and Sims (1977), Stock and Watson (1989), Bernanke, Boivin and Eliasz (2005) and Aruoba, Diebold and Scotti (2009), among others. We consider forecasts that are purely statistical, similar to PCR, and derived from our automatic procedure described below. We also consider forecasts that improve upon purely statistical forecasts by exploiting the unique ability of the 3PRF to directly incorporate economic theory within its procedure.

A second and separate application analyzes asset prices. We use a factor model that ties individual assets' price-dividend ratios to aggregate stock market fluctuations in order to uncover investors' discount rates and dividend growth expectations. There we find an unprecedented level of predictability, even out-of-sample. These results highlight the link between this paper and the pricing theories of Sharpe (1964), Lintner (1965), Treynor (1961), Merton (1973) and Ross (1976), among others, each of which suggests that the cross section of individual asset prices contains information about a few common factors. This literature prompted the use of principal components in Connor and Korajczyk (1988, 1993), building upon Ross's arbitrage pricing theory and the approximate factor formulation of Chamberlain and Rothschild (1983). Kelly and Pruitt (2011) investigate asset pricing applications of the 3PRF in more detail.

Finally, we explore the finite sample properties of the 3PRF in Monte Carlo experiments, as well as analyze its performance relative to PCR and OLS under a variety of specifications.

1.3 Outline

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 state space methods and partial least squares. Section 4 explores the finite sample performance of the 3PRF, PCR and OLS in Monte Carlo experiments.

Section 5 reports empirical results for 3PRF 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. Here and throughout, matrices and vectors are shown in boldface.

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). 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 $h, h+1, \ldots, T+h$, where h is the forecast horizon. Let $\boldsymbol{X}$ be the $T \times N$ matrix of predictors, $\boldsymbol{X} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_T')' = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ that have been variance standardized. 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 i^{th} predictor ($\mathbf{x}_i$). 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. We provide additional details regarding the data generating processes for $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ in Assumption 1 below.

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

Table 1: The Three-Pass Regression Filter

Pass	Description
1.	Run time series regression of $\mathbf{x}_i$ on $\boldsymbol{Z}$ for $i = 1, \ldots, N$, $x_{i,t} = \tilde{\phi}_{0,i} + \boldsymbol{z}_t' \tilde{\boldsymbol{\phi}}_i + \tilde{\varepsilon}_{it}$, retain slope estimate $\hat{\boldsymbol{\phi}}_i$
2.	Run cross section regression of $\boldsymbol{x}_t$ on $\hat{\boldsymbol{\phi}}_i$ for $t = 1, \ldots, T$, $x_{i,t} = \ddot{\phi}_{0,t} + \hat{\boldsymbol{\phi}}_i' \ddot{\boldsymbol{F}}_t + \ddot{\varepsilon}_{it}$, retain slope estimate $\hat{\boldsymbol{F}}_t$
3.	Run time series regression of y_{t+h} on predictive factors $\hat{\boldsymbol{F}}_t$, $y_{t+h} = \breve{\beta}_0 + \hat{\boldsymbol{F}}_t' \breve{\boldsymbol{\beta}} + \breve{\eta}_{t+h}$, delivers forecast $\hat{y}_{t+h}$

Notes: All regressions use OLS.

by the proxies. As we show later, proxies need not represent specific factors and 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 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.[4]

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+h} 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.

The following proposition gives an alternative representation for the 3PRF. It shows that

[4] 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.

the estimator is available in a condensed, one-step closed form. We denote $\boldsymbol{J}_N \equiv \boldsymbol{I}_N - \frac{1}{N}\boldsymbol{\iota}_N\boldsymbol{\iota}_N'$ where $\boldsymbol{I}_N$ is the N-dimensional identity matrix and $\boldsymbol{\iota}_N$ is a N-vector of ones. These $\boldsymbol{J}$ matrices enter because each regression pass is run with a constant.

Proposition 1. *The three-pass regression filter forecast of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and proxies $\boldsymbol{Z}$ is*

$$\hat{\boldsymbol{y}} = \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}. \tag{1}$$

where $\bar{y}$ is the sample mean of $\boldsymbol{y}$. The second stage factor estimate used to construct this forecast is

$$\hat{\boldsymbol{F}}' = \boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'. \tag{2}$$

The third stage predictive coefficient estimate is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}. \tag{3}$$

The implied predictive coefficient on the cross section of predictors is

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}. \tag{4}$$

Proposition 1 provides a convenient closed form for 3PRF estimates that is useful in the theoretical development that follows. Nonetheless, the regression-based procedure in Table 1 remains useful for two reasons. First, it is useful for developing intuition behind the procedure and for understanding its relation to the Kalman filter and partial least squares. Second, in practice (particularly with many predictors) one often faces unbalanced panels and missing data. The 3PRF as described in Table 2 easily handles these difficulties.[5] In addition to the formula for the vector of forecasts $\hat{\boldsymbol{y}}$, Proposition 1 also provides formulas for estimates of the underlying factors, $\hat{\boldsymbol{F}}$, the predictive coefficients for the factors, $\hat{\boldsymbol{\beta}}$, and the vector of estimated predictive loadings on each individual predictor, $\hat{\boldsymbol{\alpha}}$. Equation (4) shows that forecasts may be equivalently written as $\hat{\boldsymbol{y}} = \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\hat{\boldsymbol{\alpha}}$, interpreting $\hat{\boldsymbol{\alpha}}$ as the predictive coefficient for individual predictors. We further discuss the properties of these estimators in detail below.

[5] In contrast, PCR requires more involved EM techniques when data are missing, as Stock and Watson (2002b) explain.

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:*

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

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* $\boldsymbol{g}_t$, $L > 0$ *is the dimension of vector* $\boldsymbol{z}_t$, *and* $K = K_f + K_g$.

Assumption 1 gives the factor structure that allows us to reduce the dimension of predictor information. The structure of the target's factor loadings ($\boldsymbol{\beta} = (\boldsymbol{\beta}_f', \boldsymbol{0}')'$) allows 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 the factors as well as proxy noise. Since η_{t+h} is a martingale difference sequence with respect to all information known at time t (see Assumption 2.5 below), $\beta_0 + \boldsymbol{\beta}_f'\boldsymbol{f}_t$ gives the best time t forecast. But it is infeasible since the relevant factors $\boldsymbol{f}_t$ are unobserved.

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. $\mathbb{E}\|\boldsymbol{\phi}_i\|^4 \leq M$, $N^{-1}\sum_{j=1}^N \boldsymbol{\phi}_j \xrightarrow[T\to\infty]{p} \bar{\boldsymbol{\phi}}$, $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi} \xrightarrow[N\to\infty]{p} \boldsymbol{\mathcal{P}}$ *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. $\mathbb{E}_t(\eta_{t+h}) = \mathbb{E}(\eta_{t+h}|y_t, F_t, y_{t-1}, F_{t-1}, ...) = 0$, $\mathbb{E}(\eta_{t+h}^2) = \delta_\eta < \infty$ *for any* $h > 0$, *and* η_t *is independent of* $\phi_i(m)$ *and* $\varepsilon_{i,s}$.

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

We require factors and loadings to be cross-sectionally regular insofar as they have well-behaved covariance matrices for large T and N, respectively, and these matrices are finite and nonsingular. Assumption 2.4 is the only assumption that materially differs from the work of Stock and Watson or Bai and Ng. This is because proxy noise, $\boldsymbol{\omega}_t$, does not play a role in principal components. We bound the moments of $\boldsymbol{\omega}_t$ in a manner analogous to the bounds on factor moments.

Assumption 3 (Dependence). *Let $x(m)$ denote the m^{th} 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. $N^{-1/2}\sum_{i=1}^{N} \boldsymbol{\phi}_i\varepsilon_{it} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\Gamma}_{\Phi\varepsilon}\right)$, *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. $T^{-1/2}\sum_{t=1}^{T} \boldsymbol{F}_t\eta_{t+h} \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\Gamma}_{F\eta}\right)$, *where* $\boldsymbol{\Gamma}_{F\eta} = \text{plim}_{T\to\infty} T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[\eta_{t+h}^2\boldsymbol{F}_t\boldsymbol{F}_t'\right] > 0$
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). *$\boldsymbol{\mathcal{P}} = \boldsymbol{I}$, $\boldsymbol{P}_1 = \boldsymbol{0}$ 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). *$\boldsymbol{\Lambda} = [\ \boldsymbol{\Lambda}_f \quad \mathbf{0}\ ]$ 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. We prove in Theorem 7 that automatic proxies satisfying Assumption 6 are generally available.

With these assumptions in place, we next derive the asymptotic properties of the three-pass filter.

2.3 Consistency

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

Theorem 1 says that the 3PRF is "asymptotically efficient" in the words of Stock and Watson (2002a): For large cross section and time series dimensions, 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.

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

> [B]ecause $\boldsymbol{\Phi F}_t = \boldsymbol{\Phi R R}^{-1}\boldsymbol{F}_t$ for any nonsingular matrix $\boldsymbol{R}$, a normalization is required to uniquely define the factors. Said differently, the model with factor loadings $\boldsymbol{\Phi R}$ 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 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 Appendix A.3.

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 i^{th} element of $\hat{\boldsymbol{\alpha}}$, and let Assumptions 1-6 hold. Then for any i,*

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

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+h}. 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 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+h}. This point has in part motivated recent econometric work on thresholding (Bai and Ng (2008)), boosting (Bai and Ng (2009)) and shrinking (Stock and Watson (2011)) principal components for the purposes of forecasting.

On the other hand, the 3PRF identifies exactly these relevant factors in its second pass factor estimation. This step effectively extracts *leading* indicators. 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+h} 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. 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 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 predictors' covariance, regardless of that factor's relationship to y_{t+h}. 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 PCR may be foiled by the presence irrelevant factors.

2.4 Asymptotic Distributions

Not only is the 3PRF consistent for the infeasible best forecast, each forecast has a normal asymptotic distribution.[8] 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 $N, T \rightarrow \infty$ we have*

$$\sqrt{T}N\left(\boldsymbol{S}_{N^*}\hat{\boldsymbol{\alpha}} - \boldsymbol{S}_{N^*}\boldsymbol{G}_{\alpha}\boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{S}_{N^*}\boldsymbol{\Sigma}_{\alpha}\boldsymbol{S}'_{N^*}\right)$$

where $\boldsymbol{\Sigma}_{\alpha} = \boldsymbol{J}_N\boldsymbol{\Phi}\boldsymbol{\Delta}_F^{-1}\boldsymbol{\Gamma}_{F\eta}\boldsymbol{\Delta}_F^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N$. Furthermore,

$$\widehat{Avar}(\boldsymbol{S}_{N^*}\hat{\boldsymbol{\alpha}}) = \boldsymbol{\Omega}_{\alpha,N^*}\left(\frac{1}{T}\sum_t \hat{\eta}^2_{t+h}(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'\right)\boldsymbol{\Omega}'_{\alpha,N^*}$$

is a consistent estimator of $\boldsymbol{S}_{N^*}\boldsymbol{\Sigma}_{\alpha}\boldsymbol{S}'_{N^*}$, where

$$\boldsymbol{\Omega}_{\alpha,N^*} = \boldsymbol{S}_{N^*}\boldsymbol{J}_N\left(\frac{1}{T}\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)\left(\frac{1}{T^3N^2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\left(\frac{1}{TN}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\right)$$

and $\boldsymbol{G}_{\alpha}$ is defined in the appendix.

[8] Our asymptotic theory builds upon the seminal theory of Bai (2003) and Bai and Ng (2002, 2006).

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 $(N^* \times N)$ selector matrix $\boldsymbol{S}_{N^*}$ is present to ensure that the limit involves a finite-dimensional object. That is, each row of $\boldsymbol{S}_{N^*}$ has a single element equal to one and remaining elements zero, no two rows are identical, the highest column index for a non-zero element is $N^* << N$, and the positions of non-zero elements are fixed and independent of N.

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

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

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

where $\mathbb{E}_t y_{t+h} = \beta_0 + \boldsymbol{\beta}'\boldsymbol{F}_t$ and Q_t^2 is the t^{th} 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 $\mathbb{E}_t(y_{t+h}) \equiv \beta_0 + \boldsymbol{\beta}'\boldsymbol{F}_t$, 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 be 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 $N, T \to \infty$ and $T/N \to 0$ 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,

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

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

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

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

Theorem 5 is analogous to the first theorem of Bai and Ng (2006). Asymptotic normality of $\hat{\boldsymbol{\beta}}$ requires the additional condition that $T/N \to 0$. This is due to the fact that the relevant factors must be estimated. Theorem 6 is the 3PRF analogue to the first theorem of Bai (2003). 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 $\hat{y}_{t+h}$ do not confront this difficulty for the 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 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.

Table 2: Automatic Proxy-Selection Algorithm

0. Initialize $\boldsymbol{r}_0 = \boldsymbol{y}$.

For $k = 1, \ldots, L$:

1. Define the k^{th} 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.

2.5.1 Automatic Proxies

By definition the target variable depends only on the relevant factors and therefore satisfies Assumption 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 fails to asymptotically attain the infeasible best.[9] Hence in general we can improve upon the target-proxy 3PRF forecast 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 Assumption 6. This is established by the following result.

Theorem 7. *Let Assumptions 1-5 hold. Then the L-automatic-proxy three-pass regression filter forecaster of* $\boldsymbol{y}$ *satisfies Assumption 6 when* $L = K_f$.

[9] 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.

Theorem 7 states that as long as Assumptions 1-5 are satisfied, the 3PRF is generally available since the final condition of Theorems 1-4 (that is, Assumption 6) can be satisfied by construction. Moreover, the only variables necessary to implement the filter are $\boldsymbol{y}$ and $\boldsymbol{X}$ since the proxies are constructed by the algorithm.

2.5.2 Theory-Motivated 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.

As a statistical example, consider a situation in which K_f is one, so that the target and proxy are given by $y_{t+h} = \beta_0 + \beta f_t + \eta_{t+h}$ 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.

Next, consider the economic example of forecasting asset returns. It is well known that, especially over short horizons of one month or one year, the bulk of variation in asset returns comes in the form of an unpredictable shock, consistent with the theory of efficient markets (Fama (1965, 1970)). Efficient market theory implies that the remaining predictable part of return variation arises from persistent equilibrium fluctuations in risk compensation over time. Many asset pricing theories have been proposed that generate a small predictable return component of this form and link it to potentially observable "state variables" (Merton (1973) is a representative framework). An example of this framework might allow the predictable component of returns to be a linear function of a persistent market volatility process:

$$r_{t+1} = \beta_0 + \beta\sigma_t + \eta_{t+1}, \qquad \sigma_t = \bar{\sigma} + \rho\sigma_{t-1} + \xi_t, \qquad \hat{\sigma}_t = \sigma_t + \omega_t$$

where $\hat{\sigma}_t$ is an estimate of market volatility. Suppose further that conditional return volatility can be accurately measured, so that the R^2 of the $\hat{\sigma}_t$ equation is relatively large, and by an efficient markets argument, the R^2 of the r_{t+1} equation is small. Then, based on this theory, $\hat{\sigma}_t$ would be a superior proxy to the automatic proxy in finite samples.

An example of the usefulness of theory-based proxies is given in Section 5.1.1. We use the quantity theory of money to view inflation as driven by growth in real activity and the money supply. Using natural proxies for these factors, we find that the resulting out-of-sample forecasts of GDP inflation are more accurate than those obtained by either PCR or the automatic-proxy 3PRF.

3 Other 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, the Kalman filter, and partial least squares.

3.1 Constrained Least Squares

Proposition 1 demonstrates that in addition to representing the forecast $\hat{y}_{t+h}$ in terms of a dimension reduction ($\hat{\boldsymbol{F}}'_t\hat{\boldsymbol{\beta}}$), it may be equivalently represented in terms of individual predictors ($\boldsymbol{x}'_t\hat{\boldsymbol{\alpha}}$). The i^{th} 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?

The 3PRF estimator $\hat{\boldsymbol{\alpha}}$ provides an answer. It is a projection coefficient relating y_{t+h} 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.

Proposition 2. *The three-pass regression filter's implied N-dimensional predictive coeffi-*

cient, $\hat{\boldsymbol{\alpha}}$, is the solution to

$$\begin{aligned} & \mathit{arg} \min_{\alpha_0, \boldsymbol{\alpha}} ||\boldsymbol{y} - \alpha_0 - \boldsymbol{X}\boldsymbol{\alpha}|| \\ & \mathit{subject\ to} \quad (\boldsymbol{I} - \boldsymbol{J}_N \boldsymbol{X}' \boldsymbol{J}_T \boldsymbol{Z} (\boldsymbol{Z}' \boldsymbol{J}_T \boldsymbol{X} \boldsymbol{J}_N \boldsymbol{X}' \boldsymbol{J}_T \boldsymbol{Z})^{-1} \boldsymbol{Z}' \boldsymbol{J}_T \boldsymbol{X} \boldsymbol{J}_N) \boldsymbol{\alpha} = \boldsymbol{0}. \end{aligned} \tag{5}$$

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 equation (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,[10] 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+h} on $\boldsymbol{x}_t$.

3.2 Kalman Filter

The least squares prediction in the linear system of Assumption 1 is provided by the Kalman filter (Maybeck (1979)) and the system parameters are efficiently estimated by maximum likelihood (Hamilton (1994)). To simplify, assume that all variables are mean zero and suppose that $h = 1$. The prediction of the augmented state vector $\boldsymbol{\Pi}_t = (\boldsymbol{F}_t', \boldsymbol{F}_{t-1}')'$ can be written

$$\boldsymbol{\Pi}_{t|t} = \left(\boldsymbol{\Pi}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{\Upsilon}_{t|t-1}\right) + \boldsymbol{K}_t\boldsymbol{\Upsilon}_t \tag{6}$$

for $\boldsymbol{\Upsilon}_t$ the vector of variables observed by time t. $\boldsymbol{\Upsilon}_t$ includes variables *that are known at time* t, which includes the predictors, target and proxies. The predictors $\boldsymbol{x}_t$ depend on $\boldsymbol{F}_t$, the target y_t depends on $\boldsymbol{F}_{t-1}$ and the proxies may depend on either $\boldsymbol{F}_t$ or $\boldsymbol{F}_{t-1}$ according to their particular timing. The first term of (6) combines information both cross-sectionally and

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

temporally, while the second term combines information only cross-sectionally (see Appendix A.8 for more detail). The Kalman gain can be written

$$\boldsymbol{K}_t = \left(\boldsymbol{P}_{t|t-1}^{-1} + \boldsymbol{\Psi}'\boldsymbol{R}^{-1}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}\boldsymbol{R}^{-1}. \tag{7}$$

The matrix $\boldsymbol{\Psi}$ determines how observable variables are related to the latent factors, and $\boldsymbol{P}_{t|t-1}$ is the covariance matrix of the time t state vector conditional on time $t-1$ information. The vector $(\boldsymbol{0}', \boldsymbol{\beta}')$ is the row of $\boldsymbol{\Psi}$ corresponding to the target variable $\boldsymbol{y}$. Then the optimal linear predictor of y_{t+h} conditional on $\{\boldsymbol{\Upsilon}_t, \boldsymbol{\Upsilon}_{t-1}, \boldsymbol{\Upsilon}_{t-2}, \ldots\}$ is given by $(\boldsymbol{0}', \boldsymbol{\beta}')\boldsymbol{\Pi}_{t|t}$.

Consider what it means to ignore the components that temporally pool information. This affects the parts of (6) and (7) that are conditioned on past information by setting $\boldsymbol{\Pi}_{t|t-1}$ and $\boldsymbol{\Upsilon}_{t|t-1}$ to their unconditional mean of zero. Moreover, the idea that past information gives *no information* is expressed by an arbitrarily large $\boldsymbol{P}_{t|t-1}$, which implies that $\boldsymbol{P}_{t|t-1}^{-1}$ vanishes. Restricting the Kalman filter's information set to no longer temporally pool information delivers

$$\boldsymbol{\Pi}_{t|t} = \left(\boldsymbol{\Psi}'\boldsymbol{R}^{-1}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}'\boldsymbol{R}^{-1}\boldsymbol{\Upsilon}_t \tag{8}$$

$$y_{t+h|t} = \boldsymbol{\beta}'\boldsymbol{F}_{t|t}. \tag{9}$$

Equations (8) and (9) give the restricted Kalman filter's prediction of y_{t+h} conditional on $\boldsymbol{\Upsilon}_t$.

Watson and Engle's (1983) discussion of the EM algorithm demonstrates that the maximum likelihood parameter estimates of $\boldsymbol{\Psi}$ may be obtained from GLS regressions (see Appendix A.8 for more detail) in analogy to our first pass coefficient estimates.

The predictive factor $\hat{\boldsymbol{F}}_t$ from our second pass regression is comparable to $\boldsymbol{\Pi}_{t|t}$ in (6). By (8), the factor estimate $\boldsymbol{\Pi}_{t|t}$ is a GLS regression of $\boldsymbol{\Upsilon}_t$ on $\boldsymbol{\Psi}$ using the observation equations' error covariance matrix $\boldsymbol{R}$ as the weighting matrix. Finally, our third stage predictive regression is the precise analogue of equation (9). Therefore, the 3PRF can be viewed as similar to an implementation of a restricted Kalman that uses OLS in place of GLS, replaces unobserved factors with observable proxies, and shuts down the temporal pooling of past information. Crucially, the 3PRF retains the Kalman filter's cross-sectional signal extraction by least squares. In effect, large cross section asymptotics in the 3PRF substitute for the Kalman filter's time aggregation.

3.3 Partial Least Squares

The method of partial least squares, or PLS (Wold (1975), described in Appendix A.9), 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.9 we have

1. Set $\hat{\phi}_i = x_i'y$, and $\hat{\boldsymbol{\Phi}} = (\hat{\phi}_1, ..., \hat{\phi}_N)'$
2. Set $\hat{u}_t = \boldsymbol{x}_t'\hat{\Phi}$, and $\hat{\boldsymbol{u}} = (\hat{u}_1, ..., \hat{u}_T)'$
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 help 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 $\hat{f}_t = f_t + g_t\frac{Cov(\phi_{i1},\phi_{i2})}{Var(\phi_{i1})}$, which reduces to f_t

by Assumption 2.2. The slope estimate omitting the constant is $\hat{f}_t = f_t + g_t \frac{\mathbb{E}[\phi_{i1}\phi_{i2}]}{\mathbb{E}[\phi_{i1}^2]}$. This is an error-ridden version of the true target-relevant factor, and thus can produce inferior forecasts.

The most important distinction vis-à-vis PLS is the flexibility afforded by 3PRF. As discussed in Section 2.5, the three-pass filter allows the econometrician to select proxies for latent factors on the basis of economic theory, a feature which has no PLS analogue.

Because PLS is a special case of our methodology, the asymptotic theory we have developed for the 3PRF applies directly to PLS estimates under the minor modifications discussed above. These include treating basic model components (factors and predictors) as mean zero and omitting constants from first and second pass regressions (that is, replace each $\boldsymbol{J}$ matrix with the conformable identity matrix). Our results therefore provide a means of conducting inference when applying PLS. To the best of our knowledge, our joint N and T asymptotics are new results for the PLS literature.

4 Simulation Evidence

To assess the finite sample accuracy of the theoretical asymptotic results derived above we conduct a series of Monte Carlo experiments. First, we examine the accuracy of 3PRF forecasts relative to the infeasible best forecast. The data are generated according to Assumption 1 with K_f =1, 2 or 3 and K_g =0, 1 or 3. We use cross section and time dimensions between 25 and 500. Factors, loadings and shocks are drawn from a standard normal distribution. We begin by comparing the relative forecasting performance of PCR and the 3PRF. We report the ratio of the R^2 obtained by either method to the infeasible best R^2, which is set equal to 50%. For large N and T, Theorem 1 states that this ratio should be close to one for the 3PRF when we estimate the correct number of relevant factors. The rows of Table 3 use PCL to denote the forecast using L principal components and 3PRFL to denote the L-automatic-proxy 3PRF forecast. Across 1000 simulations, small sample estimates are in line with the consistency results proved above. The 3PRF R^2 is close to the infeasible best as signified by ratios near one in all cases. In contrast, the performance of principal components forecasts deteriorates substantially in the presence of irrelevant factors, as our earlier arguments would suggest.

Our next experiment evaluates predictive coefficient estimates in addition to the forecasts themselves. Our analysis here focuses on the accuracy of finite sample approximations based on the asymptotic distributions we have derived. For each Monte Carlo draw, we compute estimates for $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. Then we standardize each estimate in accordance with Theorems

Table 3: Simulated Forecast Performance, Relative to Infeasible Best

	$N = 25$	$N = 100$	$N = 100$	$N = 500$		$N = 25$	$N = 100$	$N = 100$	$N = 500$
	$T = 70$	$T = 200$	$T = 500$	$T = 100$		$T = 70$	$T = 200$	$T = 500$	$T = 100$
	2 Relevant, 0 Irrelevant					**3 Relevant, 3 Irrelevant**			
PC1	0.5891	0.6236	0.5962	0.6576	PC1	0.2425	0.1906	0.1404	0.2434
3PRF1	0.9255	0.9790	0.9805	1.0038	3PRF1	0.8227	0.9064	0.9232	0.9218
PC2	0.9275	0.9916	0.9854	1.0003	PC2	0.5618	0.4687	0.4046	0.5735
3PRF2	1.0379	1.0311	1.0018	1.1016	3PRF2	0.9802	1.0045	0.9933	1.0518
	1 Relevant, 1 Irrelevant								
PC1	0.6118	0.5843	0.5573	0.6731	PC3	0.7325	0.6993	0.6086	0.7633
3PRF1	0.9219	0.9815	0.9785	0.9961	3PRF3	1.0487	1.0330	1.0043	1.1215
PC2	0.9346	0.9848	0.9882	1.0070	PC6	0.9592	1.0038	0.9953	1.0285
3PRF2	1.0390	1.0234	1.0030	1.1002					

Notes: Forecast R^2 relative to infeasible best, median across 1000 simulations. PCL denotes the forecast using L principal components; 3PRFL denotes the L-automatic-proxy 3PRF forecast.

3, 4 and 5 by subtracting off the theoretical adjustment term and dividing by the respective asymptotic standard error estimates. 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 Figures 1-3) 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.

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 (Figure 3, upper right). This bias vanishes when the sample size increases (Figure 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 0.945 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, respectively. Altogether, simulations provide evidence

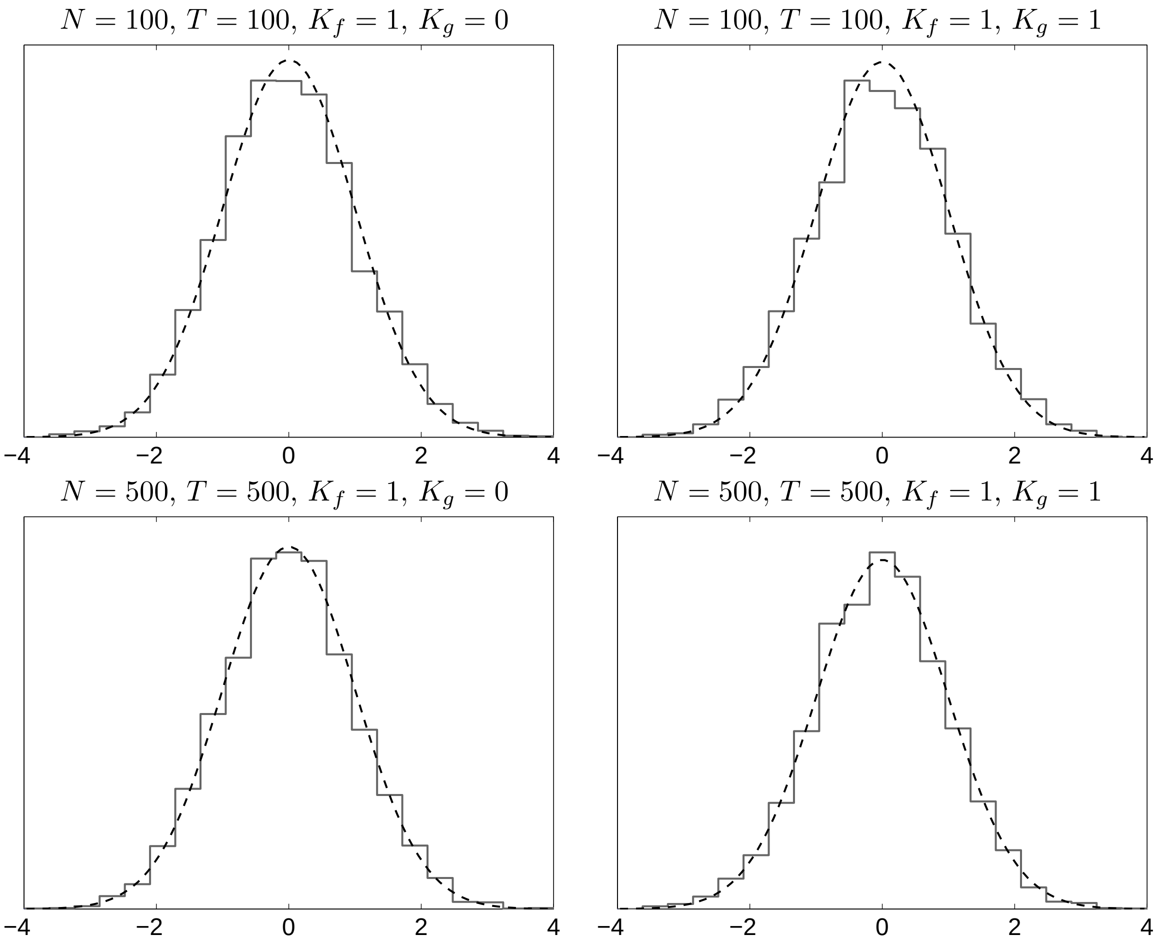

Figure 1: Simulated Distribution, $\hat{y}_{t+1}$

Notes: 5000 simulations.

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.

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 empirical investigation we use a factor model to relate individual assets' price-dividend ratios to market returns and aggregate dividend growth. We use the automatic-proxy 3PRF and compare its performance to the forecast accuracy of PCR and OLS. In the macroeconomic investigation we additionally consider the usefulness

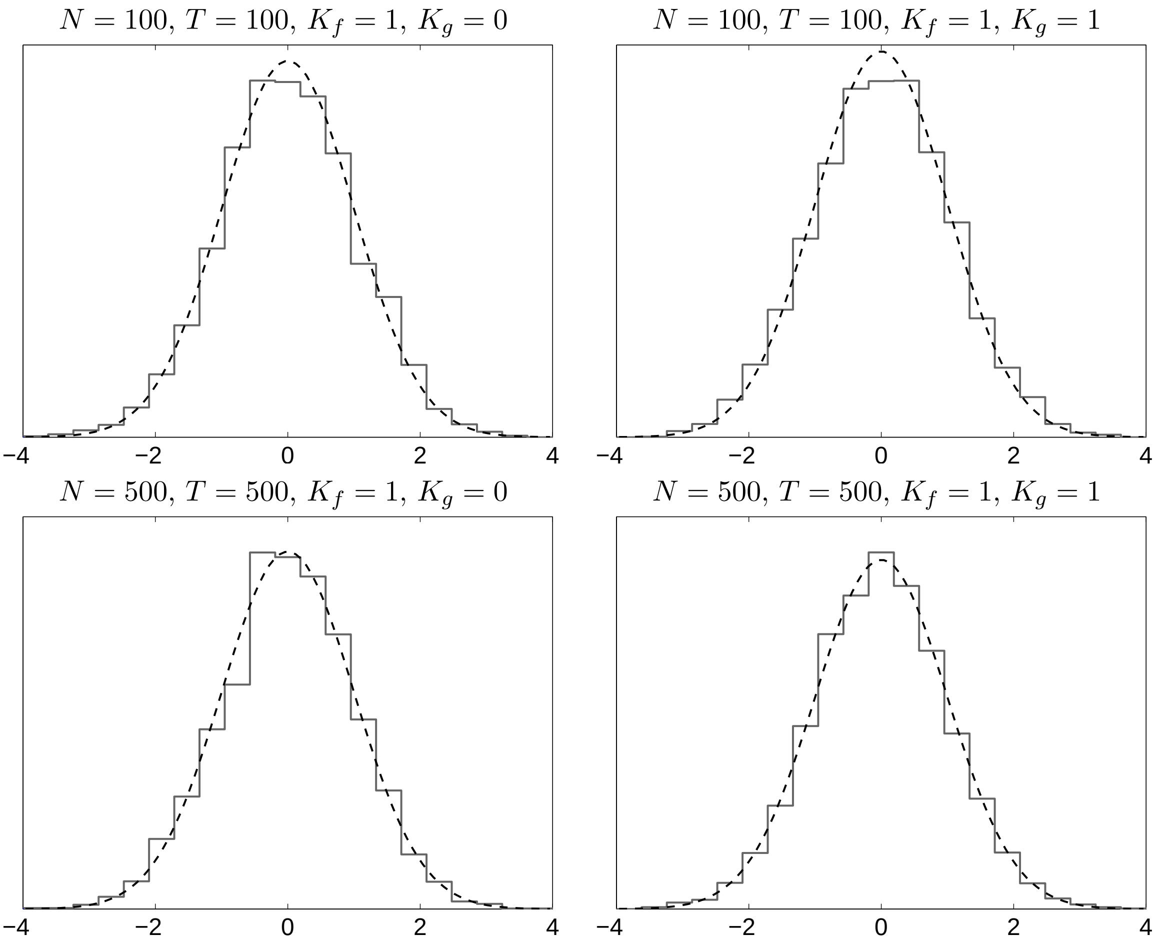

Figure 2: SIMULATED DISTRIBUTION, $\hat{\boldsymbol{\alpha}}$

Notes: 5000 simulations.

theory-motivated proxies in forecasting inflation.[11] Throughout this section, the in-sample R^2 we report is the standard, centered coefficient of determination. Out-of-sample R^2 is the ratio of explained out-of-sample variance to total out-of-sample variance around the training sample mean (see, for example, Goyal and Welch (2008)).

5.1 Macroeconomic Forecasting

We explore 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 and filtered by Stock and Watson (2011). Variants of this data set have been used by those authors, as well as by Bai and

[11] Kelly and Pruitt (2011) consider in detail the usefulness of using theory-motivated proxies to forecast market aggregates using price-dividend ratios, and so we omit that analysis from our second investigation.

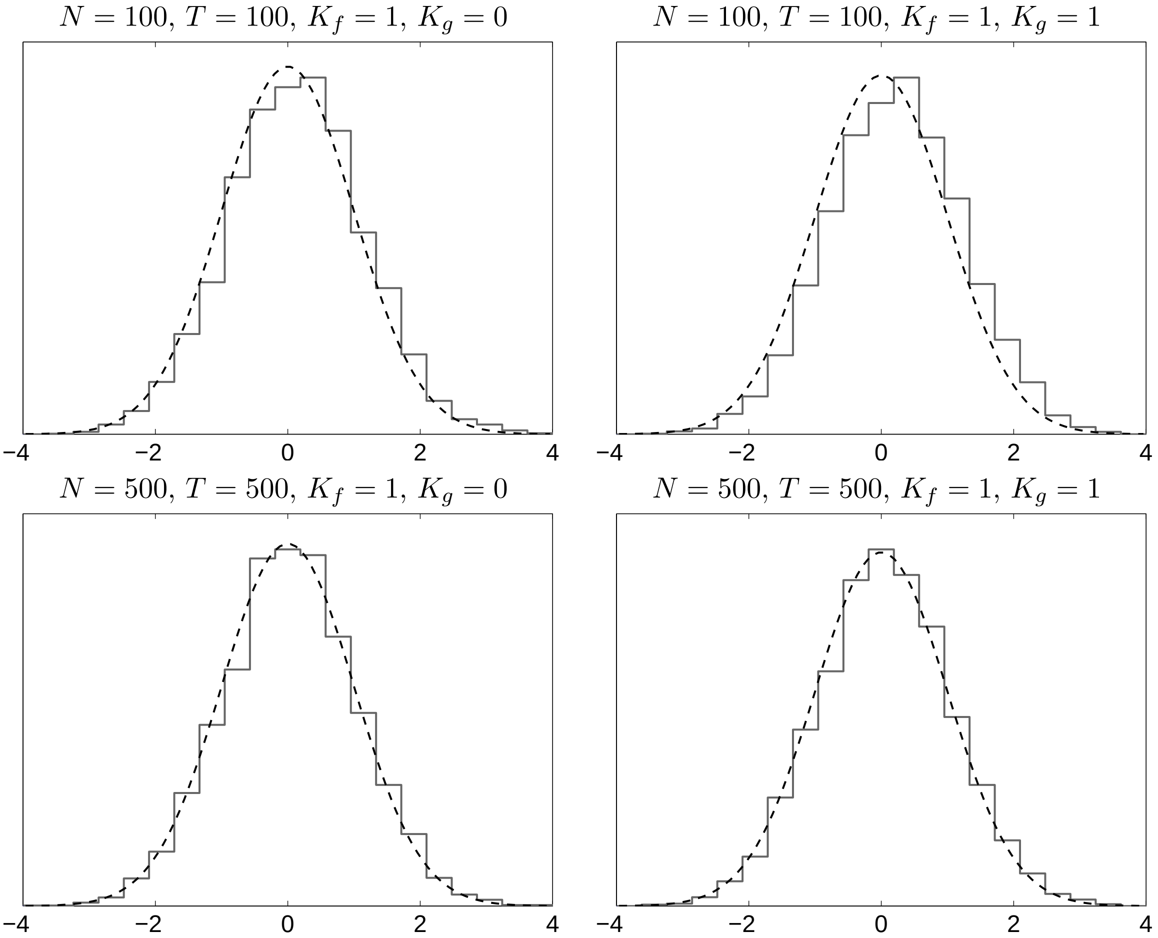

Figure 3: Simulated Distribution, $\hat{\boldsymbol{\beta}}$

Notes: 5000 simulations.

Ng (2008, 2009) and Ludvigson and Ng (2009). Any variable that we eventually target is removed from the set of predictors.[12]

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, as in the studies cited above. This generates the following variables:

$$\begin{aligned} \ddot{y}_{t+1} &= y_{t+1} - \hat{\mathbb{E}}(y_{t+1}|y_t, y_{t-1}, y_{t-2}, y_{t-3}) \\ \ddot{\boldsymbol{x}}_t &= \boldsymbol{x}_t - \hat{\mathbb{E}}(\boldsymbol{x}_t|y_t, y_{t-1}, y_{t-2}, y_{t-3}) \end{aligned}$$

[12] Forecasts from the three-pass regression filter, like principal components and partial least squares, depend on the scale of predictors, thus we standardize all predictors to have a variance of one. We also standardize proxies if they are not automatically selected. Predictors are not demeaned since constants are estimated within the algorithm.

Table 4: Out-of-Sample Macroeconomic Forecasting

	In-Sample			**Out-of-Sample**		
Target	OLS	PCR1	3PRF1	OLS	PCR1	3PRF1
Output	82.31	48.18	56.23	−28.48	45.48	48.43
Consumption	78.32	13.27	43.89	−32.15	9.61	32.22
Investment	77.99	45.57	49.30	−42.89	43.60	42.20
Exports	74.93	6.57	25.44	−53.02	5.07	13.96
Imports	79.08	23.70	38.22	−51.27	19.43	29.14
Industrial Production	61.24	4.84	23.13	−143.02	2.15	8.77
Capacity Utilization	90.76	66.19	69.82	38.14	62.72	62.68
Total Hours	86.32	62.80	67.16	8.74	60.01	61.43
Total Employment	73.51	29.94	40.86	−74.39	27.08	31.65
Average Hours	74.97	25.19	37.51	−63.87	21.70	28.80
Labor Productivity	71.72	0.55	33.39	−98.04	−2.85	12.63
Housing Starts	86.09	0.33	48.76	12.56	−4.45	26.37
GDP Inflation	74.13	4.37	18.68	−74.28	3.02	−0.01
PCE Inflation	75.64	1.73	24.30	−48.32	1.14	5.81

Notes: R^2 in percentage. Quarterly data from Stock and Watson (2011), 1960-2009. For each dependent variable, we work work with data "partialed" with respect to four lags of the variable: for dependent variable y_{t+1}, we forecast $\ddot{y}_{t+1} \equiv y_{t+1} - \hat{\mathbb{E}}(y_{t+1}|y_t, y_{t-1}, y_{t-2}, y_{t-3})$ using $\ddot{\boldsymbol{x}}_t \equiv \boldsymbol{x}_t - \hat{\mathbb{E}}(\boldsymbol{x}_t|y_t, y_{t-1}, y_{t-2}, y_{t-3})$ as predictors. The original predictors are 108 macroeconomic variables. Cross Validation drops observations $\{t-4, \ldots, t, \ldots, t+4\}$ from parameter estimation steps, where the t^{th} observation is $(\ddot{y}_{t+1}, \ddot{\boldsymbol{x}}_t')'$. Bai and Ng's (2002) IC_{p2} chooses one factor in the full sample.

where $\hat{\mathbb{E}}(\cdot|\Omega)$ denotes linear projection on Ω and a constant. These residuals become our target and predictors. Forecasting results therefore have the interpretation of performance above and beyond that provided by a simple AR(4) forecast.

In addition to applying the 3PRF in-sample, we also perform a cross-validation out-of-sample analysis to reduce any effect of small sample bias in our results. To construct each period-t forecast, we run the 3PRF omitting observations for nine quarters surrounding the forecasts. That is, the training sample drops data for periods $\{t-4, ..., t-1, t, t+1, ..., t+4\}$. The cross-validation forecast is then constructed using parameters estimated from this restricted data set: $\hat{y}_t^{CV} = \boldsymbol{x}_{t-1}'\hat{\boldsymbol{\alpha}}^{CV,t}$. This cross-validation approach, which is used in the studies cited above, is also applied for OLS and PCR forecasts in our comparisons.[13]

[13] We focus on cross validation here to remain consistent with prior literature. A second and more stark alternative is a pure out-of-sample analysis. For instance, our asset pricing results below consider a standard recursive out-of-sample estimation scheme which has been well-studied in the literature (see, for example, Clark and McCracken (2001)). Recursive out-of-sample forecast results corresponding to our macro variable forecasts corroborate our cross-validation results and are available upon request.

Table 4 presents our macroeconomic forecasting results. OLS displays the obvious signs of overfit: High in-sample predictability alongside extremely poor out-of-sample performance. In contrast, 3PRF and PCR dimension reduction techniques display notable protection from overfit.

As part of our comparison with PCR, we estimate the number of factors among the cross section of predictors following Bai and Ng (2002). For almost all targets, the information criterion (IC_{p2}) chooses one factor.

We focus our discussion on out-of-sample forecasts. There are some cases in which the 3PRF finds substantially more predictability than PCR. The 3PRF R^2 is more than three times stronger for consumption, exports, industrial production and PCE inflation. Moreover, the 3PRF provides strong positive predictive power for labor productivity and housing starts, while PCR fails to outperform the sample mean. In all but three cases the 3PRF improves over PCR in out-of-sample predictability. In two of these cases, investment and capacity utilization, the difference between PCR and the 3PRF is small. The third case is GDP inflation, where the first principal component uncovers 3.38% predictability and the 3PRF finds 0.37%. We explore inflation in further detail next.

5.1.1 Theory-Motivated Proxies

Section 2.5.2 discussed the potential usefulness of selecting proxies on the basis of economic theory rather than relying on the automatic selection algorithm. In this section we use theory-motivated proxies for the purpose of inflation forecasting. Perhaps the most basic theory of inflation comes from the quantity theory of money

$$\frac{\Delta(\text{Money supply}) \times \Delta(\text{Velocity of money})}{\Delta(\text{Real Product})} = \Delta(\text{Price level}).$$

This equation states that product inflation is a function of money supply growth, changes in the velocity of money and growth in real activity. Fama (1981) tests the quantity theory by regressing future inflation on growth in output and money supply. Here we are interested in producing inflation forecasts that are explicable in terms of this simple model while using our new forecasting method to exploit the wealth of available macroeconomic information. We proxy for changes in real activity using log output growth and proxy for changes in money supply using log growth in M1. As in Fama (1981), changes in velocity, which are inherently difficult to quantify, serve as the error term in the forecasting relationship. Timing is aligned so that proxies observed at time t are used to extract information from the predictors at time t for forecasting GDP inflation at time $t+1$.

Table 5: GDP Inflation Forecasts based on Theory-Motivated Proxies

Proxies	**In-Sample**	**Out-of-Sample**
Output, Money Growth	7.03	4.47
Output Growth	4.71	4.04
Money Growth	1.63	−1.26

Notes: R^2 in percentage. Quarterly data from Stock and Watson (2011), 1960-2009. Forecasted variable is GDP Inflation. See notes of Table 4 for description of "partialing" and cross-validation out-of-sample procedure.

Table 5 contains the forecasting results using these theory-motivated proxies. Real output is particularly useful for extracting relevant information from the predictive cross section. Money growth by itself does little, but is incrementally useful when combined with output.

Recall from Table 4 that the first principal component is strongly related to activity measures like output, total employment and capacity utilization. The output-proxied 3PRF builds this relationship directly into the forecast and allows us to attribute 4% to 5% of inflation's variation to a real activity factor, more than the first principal component obtained. It also achieves a higher out-of-sample forecast R^2 than achieved by the target-proxy 3PRF, and outperforms a direct regression of GDP inflation on lagged output growth and growth in M1.[14]

5.2 Forecasting Market Returns and Dividend Growth

Asset return forecastability has been extensively examined in the asset pricing literature.[15] 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)). Building from the present value identity, Kelly and Pruitt (2011) map the cross section of price-dividend ratios into the approximate latent factor model of Assumption 1. The predictors are the cross section of log price-dividend ratios, $pd_{i,t} = \phi_{i,0} + \boldsymbol{\phi}_i' \boldsymbol{F}_t + \varepsilon_{i,t}$, and the targets are log returns and log dividend growth for the

[14] Direct regression results in an in-sample R^2 of 1.97% and out-of-sample R^2 of 0.4%.

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

aggregate market, $r_{t+1} = \beta_0^r + \boldsymbol{F}'_t\boldsymbol{\beta}^r + \eta_{t+1}^r$ and $\Delta d_{t+1} = \beta_0^d + \boldsymbol{F}'_t\boldsymbol{\beta}^d + \eta_{t+1}^d$. This structure motivates an exploration of the predictive content of portfolios' price ratios for market returns and dividend growth using dimension reduction techniques.

Our analysis here considers twenty-five price-dividend ratios of portfolios sorted by size and book-to-market characteristics over the post-war period 1946-2009 (following Fama and French (1993); see appendix for details of our data construction). Our out-of-sample analysis uses a recursive procedure common to this literature[16] and proceeds as follows. We split the 1946-2009 sample at 1980, using the first 35 observations as a training sample and the last 29 observations as the out-of-sample horizon. Beginning with $t = 35$, we estimate first-stage factor loadings using observations $\{1, ..., t\}$. Then, for each period $\tau \in \{1, ..., t\}$, we estimate the time τ value of our predictor variable using the cross section of valuation ratios at τ and first-stage coefficients (which are based on data from $\{1, ..., t\}$). We then estimate the predictive coefficient in a third-stage forecasting regression of realized returns (or dividend growth) for periods $\{2, ..., t\}$ on our predictor from $\{1, ..., t-1\}$. Finally, our out-of-sample forecast of the $t + 1$ return is the product of the third-stage predictive coefficient and the time t second-stage factor estimate. At time $t + 1$, we construct our forecast of the return at $t + 2$ by repeating the entire three stage procedure using data from $\{1, ..., t + 1\}$. This process is iterated forward each year until the entire time series has been exhausted.

Additionally we consider the results of out-of-sample cross-validation as in our macro forecasts above. Our emphasis is on factor model parsimony, hence we focus on one-factor and two-factor implementations of PCR and the 3PRF. We also use Bai and Ng's (2002) IC_{p2} 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 PCs chosen across all periods of the out-of-sample procedure.

For both return and dividend growth forecasts in Table 6 we find that the 3PRF provides strong in-sample and out-of-sample forecasts using one and two factors. With one or two factors, the 3PRF demonstrates substantially better overall performance than PCR. Seven PCR predictive factors are selected based on the IC_{p2} criterion in-sample. PCR forecasts with a high number of factors as selected by IC_{p2} become more competitive with two-factor 3PRF forecasts. Overall, the 3PRF demonstrates the ability to extract leading indicators with strong predictive power.

[16] See Goyal and Welch (2008).

Table 6: Out-of-Sample Market Return and Dividend Growth Forecasting

	Returns			Dividend Growth		
	IS	OOS-R	OOS-CV	IS	OOS-R	OOS-CV
OLS	70.77	−314.83	−22.49	72.40	−111.38	−31.18
PC1	6.20	−6.35	−6.48	0.09	−7.73	−11.80
3PRF1	15.01	17.91	2.85	26.93	12.84	−0.05
PC2	9.30	−3.05	−6.12	0.11	−10.34	−17.85
3PRF2	31.29	18.44	7.98	43.42	34.34	17.41
PC-*IC*	36.12	22.23	7.65	41.02	25.79	13.88
memo:	7.00	5.46	6.67	7.00	5.46	6.67

Notes: R^2 in percent. Annual data 1946–2009, from CRSP. Twenty-five size/book-to-market sorted portfolios of dividend-paying stocks. Out-of-sample recursive (OOS-R) forecasts begin in 1983. IS denotes In-sample forecasts. Out-of-sample cross-validation (OOS-CV) forecasts leave out 3 observations from the estimation subsample for each period. All forecasts are one year ahead: OLS denotes the forecast from simple linear projection of the target on the twenty-five predictors; PCL denotes the forecast using L principal components; 3PRFL denotes the L-automatic-proxy 3PRF forecast. PC-IC denotes the number of PCs are chosen by Bai and Ng's (2002) IC_p2 for each sample; *memo* displays the average number of factors chosen across the subsamples entering into OOS results.

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 influences the forecast target 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, which condenses cross section information according to covariance within the predictors. The 3PRF, on the other hand, condenses cross section information according to covariance between predictors and the forecast target. The intuition behind the 3PRF is further illustrated through comparisons with constrained least squares, the Kalman filter, and partial least squares. Finally, we demonstrate its efficacy for forecasting in Monte Carlo experiments and empirical applications from the macroeconomics and finance literature.

References

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press.

Anderson, T. W. (2003): *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, third edn.

Aruoba, S. B., F. X. Diebold, and C. Scotti (2009): "Real-Time Measurement of Business Conditions," *Journal of Business & Economic Statistics*, 27(4), 417–427.

Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71(1), 135–171.

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

——— (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74(4), 1133–1150.

——— (2008): "Forecasting economic time series using targeted predictors," *Journal of Econometrics*, 146(2), 304–317.

——— (2009): "Boosting diffusion indices," *Journal of Applied Econometrics*, 24(4), 607–629.

Bernanke, B., J. Boivin, and P. S. Eliasz (2005): "Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach," *Quarterly Journal of Economics*, 120(1), 387–422.

Campbell, J., and R. Shiller (1988): "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1(3), 195–228.

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51(5), 1281–304.

Cochrane, J. (1992): "Explaining the Variance of Price-Dividend Ratios," *Review of Financial Studies*, 5(2), 243–80.

——— (2011): "Presidential Address: Discount Rates," *Journal of Finance*, 66, 1047–1108.

Connor, G., and R. A. Korajczyk (1988): "Risk and return in an equilibrium APT : Application of a new test methodology," *Journal of Financial Economics*, 21(2), 255–289.

Connor, G., and R. A. Korajczyk (1993): "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance*, 48(4), 1263–91.

Davis, J., E. Fama, and K. French (2000): "Characteristics, covariances, and average returns: 1929 to 1997," *Journal of Finance*, 55(1), 389–406.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

Fama, E. (1965): "The behavior of stock-market prices," *Journal of Business*, 38(1), 34–105.

——— (1970): "Efficient capital markets: A review of theory and empirical work," *Journal of Finance*, 25(2), 383–417.

——— (1981): "Stock Returns, Real Activity, Inflation, and Money," *American Economic Review*, 71(4), 545–565.

Fama, E., and K. French (1988): "Permanent and Temporary Components of Stock Prices," *Journal of Political Economy*, 96(2), 246–73.

——— (1993): "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33(1), 3–56.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000): "The generalized dynamic-factor model: Identification and estimation," *Review of Economics and Statistics*, 82(4), 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2004): "The generalized dynamic factor model consistency and rates," *Journal of Econometrics*, 119(2), 231–255.

——— (2005): "The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting," *Journal of the American Statistical Association*, 100, 830–840.

Forni, M., and L. Reichlin (1996): "Dynamic common factors in large cross-sections," *Empirical Economics*, 21(1), 27–42.

Forni, M., and L. Reichlin (1998): "Let's Get Real: A Factor Analytical Approach to Disaggregated Business Cycle Dynamics," *Review of Economic Studies*, 65(3), 453–73.

Geweke, J. F. (1977): "The Dynamic Factor Analysis of Economic Time Series," in *Latent Variables in Socio-Economic Models*, chap. 19. Amsterdam: North-Holland.

Goyal, A., and I. Welch (2008): "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 21(4), 1455–1508.

Hamilton, J. D. (1994): *Time Series Analysis*. Princeton University Press.

Hastie, T., R. Tibshirani, and J. Friedman (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," *Review of Financial Studies*, 5(3), 357.

Huber, P. J. (1973): "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *Annals of Statistics*, 1(5), 799–821.

Jungbacker, B., and S. Koopman (2008): "Likelihood-based Analysis for Dynamic Factor Models," *Tinbergen Institute Discussion Paper.*

Kelly, B. T., and S. J. Pruitt (2011): "Market Expectations in the Cross Section of Present Values," Working paper, Chicago Booth.

Lintner, J. (1965): "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, 47(1), 13–37.

Ludvigson, S. C., and S. Ng (2009): "Macro Factors in Bond Risk Premia," *Review of Financial Studies*, 22(12), 5027–5067.

Maybeck, P. S. (1979): *Stochastic Models, Estimation, and Control, Vol. 1*, vol. 141. Academic Press; Mathematics in Science and Engineering.

Merton, R. (1973): "An intertemporal capital asset pricing model," *Econometrica*, pp. 867–887.

Ross, S. (1976): "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, 13(3), 341–360.

Rozeff, M. (1984): "Dividend yields are equity risk premiums," *Journal of Portfolio Management*, 11(1), 68–75.

Sargent, T. J., and C. A. Sims (1977): "Business cycle modeling without pretending to have too much a priori economic theory," Working Papers 55, Federal Reserve Bank of Minneapolis.

Sharpe, W. (1964): "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *Journal of Finance*, 19(3), 425–442.

Simon, D. (2006): *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches.* Wiley-Interscience.

Stambaugh, R. (1986): "Bias in Regressions with Lagged Stochastic Regressors," *Working Paper, University of Chicago.*

Stock, J. H., and M. W. Watson (1989): "New Indexes of Coincident and Leading Economic Indicators," in *NBER Macroeconomics Annual 1989, Volume 4*, pp. 351–409. National Bureau of Economic Research, Inc.

——— (1998): "Diffusion indexes," *NBER Working Paper.*

——— (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97(460), 1167–1179.

——— (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20(2), 147–62.

——— (2006): "Forecasting with Many Predictors," in *Handbook of Economic Forecasting*, ed. by G. Elliott, C. Granger, and A. Timmermann, vol. 1, chap. 10, pp. 515–554. Elsevier.

——— (2011): "Generalized Shrinkage Methods for Forecasting Using Many Predictors," Working papers, Princeton University.

Treynor, J. (1961): "Toward a Theory of Market Value of Risky Assets," *Unpublished manuscript.*

Watson, M. W., and R. F. Engle (1983): "Alternative Algorithms for the Estimation of Dynamic Factor, MIMIC, and Varying Coefficient Regression Models," *Journal of Econometrics*, 23, 385–400.

Wold, H. (1975): "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*, ed. by P. Krishnaiaah, pp. 391–420. New York: Academic Press.

A Appendix

A.1 Assumptions

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

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

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* $\boldsymbol{g}_t$, $L > 0$ *is the dimension of vector* $\boldsymbol{z}_t$, *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. $\mathbb{E}\|\boldsymbol{\phi}_i\|^4 \leq M$, $N^{-1}\sum_{j=1}^{N}\boldsymbol{\phi}_j \xrightarrow[T\to\infty]{p} \bar{\boldsymbol{\phi}}$, $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Phi} \xrightarrow[N\to\infty]{p} \boldsymbol{\mathcal{P}}$ *and* $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\phi}_0 \xrightarrow[N\to\infty]{p} \boldsymbol{P}_1$[17]
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. $\mathbb{E}_t(\eta_{t+h}) = \mathbb{E}(\eta_{t+h}|y_t, F_t, y_{t-1}, F_{t-1}, ...) = 0$, $\mathbb{E}(\eta_{t+h}^2) = \delta_\eta < \infty$ *for any* $h > 0$, *and* η_t *is independent of* $\phi_i(m)$ *and* $\varepsilon_{i,s}$.

Assumption 3 (Dependence)**.** *Let* $x(m)$ *denote the* m^{th} *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. $T^{-1/2}\sum_{t=1}^{T}\boldsymbol{F}_t\eta_{t+h} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Gamma}_{F\eta})$, *where* $\boldsymbol{\Gamma}_{F\eta} = \text{plim}_{T\to\infty}T^{-1}\sum_{t=1}^{T}\mathbb{E}\left[\eta_{t+h}^2\boldsymbol{F}_t\boldsymbol{F}_t'\right] > 0$
3. $T^{-1/2}\sum_{t=1}^{T}\boldsymbol{F}_t\varepsilon_{it} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Gamma}_{F\varepsilon,i})$, *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)**.** $\boldsymbol{\mathcal{P}} = \boldsymbol{I}$, $\boldsymbol{P}_1 = \boldsymbol{0}$ *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.*

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

A.2 Auxiliary Lemmas

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. $T^{-1/2}\sum_t \eta_{t+h} = O_p(1)$,
5. $T^{-1/2}\sum_t \varepsilon_{it}\eta_{t+h} = O_p(1)$
6. $N^{-1/2}T^{-1/2}\sum_{i,t}\varepsilon_{it}\eta_{t+h} = O_p(1)$
7. $N^{-1}T^{-1/2}\sum_{i,t}\phi_i(m_1)\varepsilon_{it}F_t(m_2) = O_p(1)$
8. $N^{-1}T^{-1/2}\sum_{i,t}\phi_i(m_1)\varepsilon_{it}\omega_t(m_2) = O_p(1)$
9. $N^{-1/2}T^{-1}\sum_{i,t}\phi_i(m)\varepsilon_{it}\eta_{t+h} = O_p(1)$
10. $N^{-1}T^{-1/2}\sum_{i,s}\varepsilon_{is}\varepsilon_{it} = O_p(\delta_{NT}^{-1})$
11. $N^{-1}T^{-3/2}\sum_{i,s,t}\varepsilon_{is}\varepsilon_{it}\eta_{t+h} = O_p(\delta_{NT}^{-1})$
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. $N^{-1}T^{-1}\sum_{i,s,t}F_s(m)\varepsilon_{is}\varepsilon_{it}\eta_{t+h} = O_p(1)$
15. $N^{-1}T^{-1}\sum_{i,s,t}\omega_s(m)\varepsilon_{is}\varepsilon_{it}\eta_{t+h} = O_p(1)$

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

Proof: Item 1: Note that

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

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

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

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

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

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

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

Item 9 is bounded by $\left(T^{-1}\sum_t \eta_{t+h}^2\right)^{1/2}\left(T^{-1}\sum_t\left[N^{-1/2}\sum_i \phi_i(m)\varepsilon_{it}\right]^2\right)^{1/2} = O_p(1)$ by 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 Assumption 3.2 and 3.1.

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

Item 12: First, we have
$N^{-1}T^{-1/2}\sum_{i,s}F_s(m)\varepsilon_{is}\varepsilon_{it} = N^{-1/2}\left(N^{-1/2}T^{-1/2}\sum_{i,s}F_s(m)[\varepsilon_{is}\varepsilon_{it}-\sigma_{ii,st}]\right)+T^{-1/2}\left(N^{-1}\sum_{i,s}F_s(m)\sigma_{ii,st}\right)$.
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,
$N^{-1}T^{-1}\sum_{i,s,t}F_s(m)\varepsilon_{is}\varepsilon_{it}\eta_{t+h} \le \left(N^{-1}\sum_i\left[T^{-1/2}\sum_t\varepsilon_{it}\eta_{t+h}\right]^2\right)^{1/2}\left(N^{-1}\sum_i\left[T^{-1/2}\sum_s F_s(m)\varepsilon_{is}\right]^2\right)^{1/2} = O_p(1)$. The same argument applies to Item 15 using Assumption 3.4 and Item 5.

QED

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

1. $T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\omega} = \boldsymbol{O}_p\left(1\right)$
2. $T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta} = \boldsymbol{O}_p\left(1\right)$
3. $T^{-1/2}\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} = \boldsymbol{O}_p\left(1\right)$
4. $N^{-1/2}\boldsymbol{\varepsilon}_t'\boldsymbol{J}_N\boldsymbol{\Phi} = \boldsymbol{O}_p\left(1\right)$
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. $N^{-1/2}T^{-1}\boldsymbol{\Phi}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} = \boldsymbol{O}_p\left(1\right)$
8. $N^{-1}T^{-3/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$
9. $N^{-1}T^{-3/2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$
10. $N^{-1}T^{-3/2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\omega} = \boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$
11. $N^{-1}T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t = \boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$
12. $N^{-1}T^{-1/2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}_t = \boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$
13. $N^{-1}T^{-3/2}\boldsymbol{\eta}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$
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: $T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta} = T^{-1/2}\sum_t \boldsymbol{F}_t\eta_{t+h} - (T^{-1}\sum_t \boldsymbol{F}_t)(T^{-1/2}\sum_t \eta_{t+h}) = \boldsymbol{O}_p(1)$ by Lemma 1.4 and Assumptions 2.1 and 4.2.

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

Item 4 has m^{th} 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 Assumptions 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[18]

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

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

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

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

5.III $= O_p(N^{-1/2})$ since $T^{-1}\sum_s F_s(m_2) = O_p(1)$ 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.IV $= O_p\left(T^{-1/2}N^{-1/2}\right)$ by Assumption 2.1, 2.2 and Lemma 1.3.

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

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

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

6.I $= O_p(1)$ by Lemma 1.8.

6.II $= O_p(1)$ by Assumptions 2.2 and 3.4.

6.III $= O_p(N^{-1/2})$ by Assumptions 2.4 and 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 m^{th} element

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

7.I $= O_p(1)$ by Lemma 1.9.

7.II $= O_p(T^{-1/2})$ by Assumption 4.1 and Lemma 1.4.

7.III $= O_p(T^{-1/2})$ by Assumption 2.2 and Lemma 1.6.

7.IV $= O_p(T^{-1})$ 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

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

8.I $= T^{-1/2}\left(N^{-1}\sum_{i,s,t}\left(T^{-1/2}\sum_s F_s(m_1)\varepsilon_{is}\right)\left(T^{-1/2}\sum_t F_t(m_2)\varepsilon_{it}\right)\right) = O_p(T^{-1/2})$ 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 and 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 and 10 are no larger than $\boldsymbol{O}_p\left(\delta_{NT}^{-1}\right)$.

Item 11 has generic m^{th} element given by

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

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 Assumption 3.4 for 4.3. Substituting this way implies that Item 12 is $\boldsymbol{O}_p(\delta_{NT}^{-1})$.

Item 13 has m^{th} element

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

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 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 $\boldsymbol{O}_p(\delta_{NT}^{-1})$.

QED

A.3 Estimators

Proposition 1. *The three pass regression filter forecast of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and proxies $\boldsymbol{Z}$ is*

$$\hat{\boldsymbol{y}} = \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}. \tag{A1}$$

where $\bar{y}$ is the sample mean of $\boldsymbol{y}$. The second stage factor estimate used to construct this forecast is

$$\hat{\boldsymbol{F}}' = \boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'. \tag{A2}$$

The third stage predictive coefficient estimate is

$$\hat{\boldsymbol{\beta}} \;=\; \left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}. \tag{A3}$$

The implied predictive coefficient on the cross section of predictors is

$$\hat{\boldsymbol{\alpha}} \;=\; \boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}. \tag{A4}$$

Proof: The first stage regression is

$$\boldsymbol{X} = \boldsymbol{\iota}\tilde{\boldsymbol{\Phi}}_0 + \boldsymbol{Z}\tilde{\boldsymbol{\Phi}}' + \tilde{\boldsymbol{\epsilon}}$$

and the first stage coefficient estimate of $\tilde{\boldsymbol{\Phi}}'$ is

$$\hat{\boldsymbol{\Phi}}' = \left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}.$$

The second stage regression is

$$\boldsymbol{X} = \boldsymbol{\iota}\ddot{\boldsymbol{\phi}}_{0,t} + \ddot{\boldsymbol{F}}\hat{\boldsymbol{\Phi}}' + \ddot{\boldsymbol{\epsilon}}$$

and the second stage coefficient estimate of $\ddot{\boldsymbol{F}}'$ is

$$\begin{aligned}
\hat{\boldsymbol{F}}' &= \left(\hat{\boldsymbol{\Phi}}'\boldsymbol{J}_N\hat{\boldsymbol{\Phi}}\right)^{-1}\hat{\boldsymbol{\Phi}}'\boldsymbol{J}_N\boldsymbol{X}' \\
&= \left\{\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\right\}^{-1}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}' \\
&= \boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'.
\end{aligned}$$

The third stage regression is

$$\boldsymbol{y} = \boldsymbol{\iota}\breve{\beta}_0 + \hat{\boldsymbol{F}}\breve{\boldsymbol{\beta}} + \breve{\boldsymbol{\eta}}$$

and the third stage coefficient estimate of $\breve{\boldsymbol{\beta}}$ is

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\boldsymbol{y}' \\
&= \left\{\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right\}^{-1} \\
&\quad \times\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y} \\
&= \left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}
\end{aligned}$$

with intercept estimate $\boldsymbol{\iota}\hat{\beta}_0 = T^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\left(\boldsymbol{y} - \hat{\boldsymbol{F}}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\iota}\bar{y} - T^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\hat{\boldsymbol{F}}\hat{\boldsymbol{\beta}}$. The corresponding Y forecast

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\hat{\boldsymbol{F}}\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z} \\
&\quad \times\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y} \\
&= \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}
\end{aligned}$$

which may be rewritten $\hat{\boldsymbol{y}} = \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\hat{\boldsymbol{\alpha}}$.

QED

Proposition 2. *The three-pass regression filter's implied N-dimensional predictive coefficient, $\hat{\boldsymbol{\alpha}}$, is the solution to*

$$\begin{aligned}
&arg \min_{\alpha_0,\boldsymbol{\alpha}} ||\boldsymbol{y} - \alpha_0 - \boldsymbol{X}\boldsymbol{\alpha}|| \\
&subject\ to \quad (\boldsymbol{I} - \boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N)\boldsymbol{\alpha} = \boldsymbol{0}.
\end{aligned}$$

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

$$\boldsymbol{R}(\boldsymbol{R}'\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{R})^{-1}\boldsymbol{R}'\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y} + [\boldsymbol{I} - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{R})^{-1}\boldsymbol{R}'\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}]\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{Q})^{-1}\boldsymbol{c}$$

for $\boldsymbol{R}$ such that $\boldsymbol{R}'\boldsymbol{Q} = \boldsymbol{0}$ and $[\begin{array}{cc} \boldsymbol{Q} & \boldsymbol{R} \end{array}]$ is nonsingular. In our case,

$$\boldsymbol{c} = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{Q} = (\boldsymbol{I} - \boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N),$$

hence we can let $\boldsymbol{R} = \boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}$ and the result follows.

QED

A.4 Probability Limits and Forecast Consistency

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

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

Proof: From Proposition 1, 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.

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

$$
\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}' + \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}' + \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}' + \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}' + \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}' + \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}' + \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}' + \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}' + \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\boldsymbol{\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 \text{(A5)}\\
&= \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\phi}_{\mathbf{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 + \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}_{\mathbf{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 + \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}_{\mathbf{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 + \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}_{\mathbf{0}}\right) + \left(N^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{J}_N\boldsymbol{\Phi}\right)\boldsymbol{F}_t + \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\boldsymbol{\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.

QED

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

$$
\hat{\boldsymbol{\beta}} \xrightarrow[T,N\to\infty]{p} \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Delta}_\omega\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\beta}. \qquad \text{(A6)}
$$

Proof: From Proposition 1, 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

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

and

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

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

QED

Lemma 5. *Let Assumptions 1, 2 and 3 hold. Then the three pass regression filter forecast satisfies*

$$\hat{y}_{t+h} \xrightarrow[T,N\to\infty]{p} \beta_0 + \boldsymbol{\mu}'\boldsymbol{\beta} + (\boldsymbol{F}_t - \boldsymbol{\mu})'\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' \left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right]^{-1} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\beta}. \tag{A7}$$

Proof: Immediate from Proposition 1 and Lemmas 3 and 4.

QED

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

Proof: Given Assumptions 1, 2 and 3, Lemma 5 holds and we can therefore manipulate (A7). Partition $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\Delta}_F$ as

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

such that the block dimensions of $\boldsymbol{\mathcal{P}}$ and $\boldsymbol{\Delta}_F$ coincide. By Assumption 5, the off-diagonal blocks, $\boldsymbol{\mathcal{P}}_{12}$ and $\boldsymbol{\Delta}_{F,12}$, are zero. As a result, the first diagonal block of the term $\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F$ in Equation A7 is $\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{P}}_1\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{P}}_1\boldsymbol{\Delta}_{F,1}$. By Assumption 6, pre- and post-multiplying by $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_f, \boldsymbol{0}]$ reduces the term in square brackets to $\boldsymbol{\Lambda}_f\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{P}}_1\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{P}}_1\boldsymbol{\Delta}_{F,1}\boldsymbol{\Lambda}_f$. Similarly, $\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' = [\boldsymbol{\Lambda}_f\boldsymbol{\mathcal{P}}_1\boldsymbol{\Delta}_{F,1}, \boldsymbol{0}]'$ and $\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F = [\boldsymbol{\Lambda}_f\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{P}}_1\boldsymbol{\Delta}_{F,1}, \boldsymbol{0}]$. By Assumption 6, $\boldsymbol{\Lambda}_f$ is invertible and therefore the expression for $\hat{y}_{t+h}$ reduces to $\beta_0 + \boldsymbol{F}_t'\boldsymbol{\beta}$.[19]

QED

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 $\boldsymbol{\beta} = (\beta_1, \boldsymbol{0}')'$ with $\beta_1 \neq 0$. Therefore the target satisfies the condition of Assumption 6.

QED

Theorem 2. *Let $\hat{\alpha}_i$ denote the i^{th} 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 i^{th} element equal to one and remaining elements zero. Expanding the expression for $\hat{\boldsymbol{\alpha}}$ in Proposition 1, 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

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

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

QED

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

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

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

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 $\hat{\boldsymbol{\varepsilon}}$.

QED

It will be useful for subsequent results to establish the asymptotic independence of $\hat{\boldsymbol{F}}_t$ and η_{t+h}.

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

Proof: It suffices to show that $\text{plim}_{N,T\to\infty} T^{-1}\sum_t \hat{\boldsymbol{F}}_{C,t}\eta_{t+h} = 0$ for all h, and to do so we examine each term in Equation A5. 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 $T^{-1}\sum_t \eta_{t+h} = o_p(1)$. By similar rationale, the four terms that are post-multiplied by $\boldsymbol{F}_t$ are $o_p(1)$ since $T^{-1}\sum_t \boldsymbol{F}_t\eta_{t+h} = o_p(1)$ by Assumption 4.3. Two of the remaining terms depend on the expression $T^{-1}\sum_t \left(N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\varepsilon}_t\right)\eta_{t+h}$, which is $o_p(1)$ because

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

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

QED

A.5 Asymptotic Distributions

Lemma 8. *Under Assumptions 1-4, as* $N, T \to \infty$ *and* $T/N \to 0$ *we have*

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

Proof:

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

As $N, T \to \infty$ and $T/N \to 0$, the first term is dominant and the stated asymptotic distribution obtains by Lemma 2 and Assumption 4.2.

QED

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

$$\sqrt{T}N\left(\boldsymbol{S}_{N^*}\hat{\boldsymbol{\alpha}} - \boldsymbol{S}_{N^*}\boldsymbol{G}_\alpha\boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{S}_{N^*}\boldsymbol{\Sigma}_\alpha\boldsymbol{S}'_{N^*}\right)$$

where $\boldsymbol{\Sigma}_\alpha = \boldsymbol{J}_N\boldsymbol{\Phi}\boldsymbol{\Delta}_F^{-1}\boldsymbol{\Gamma}_{F\eta}\boldsymbol{\Delta}_F^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N$. *Furthermore,*

$$\widehat{Avar}(\boldsymbol{S}_{N^*}\hat{\boldsymbol{\alpha}}) = \boldsymbol{\Omega}_{\alpha,N^*}\left(\frac{1}{T}\sum_t \hat{\eta}^2_{t+h}(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'\right)\boldsymbol{\Omega}'_{\alpha,N^*}$$

is a consistent estimator of $\boldsymbol{S}_{N^*}\boldsymbol{\Sigma}_\alpha\boldsymbol{S}'_{N^*}$, *where*

$$\boldsymbol{\Omega}_{\alpha,N^*} = \boldsymbol{S}_{N^*}\boldsymbol{J}_N\left(\frac{1}{T}\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)\left(\frac{1}{T^3N^2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\left(\frac{1}{TN}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\right).$$

Proof: Define

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

Also define $\boldsymbol{S}_{N^*}$ to be a $(N^* \times N)$ selector matrix. That is, each row of $\boldsymbol{S}_{N^*}$ has a single element equal to one and remaining elements zero, no two rows are identical, the highest column index for a non-zero element is $N^* << N$, and the positions of non-zero elements are fixed and independent of N.

The first term in $\boldsymbol{S}_{N^*}\hat{\boldsymbol{\alpha}}$ is the $(N^* \times K)$ matrix $\boldsymbol{S}_{N^*}\boldsymbol{J}_N\boldsymbol{\Phi}$, which has probability limit $\left(\boldsymbol{S}_{N^*}\boldsymbol{\phi}_i - \boldsymbol{S}_{N^*}\iota\bar{\boldsymbol{\phi}}\right)$ as $N, T \to \infty$. The asymptotic distribution and consistent variance estimator follow directly from Lemma 8 and previously derived limits, Assumptions 5 and 6, noting that $\hat{\eta}_{t+h} = \eta_{t+h} + o_p(1)$ by Theorem 1, and noting that

$$N\boldsymbol{S}_{N^*}\hat{\boldsymbol{\alpha}} - N\boldsymbol{S}_{N^*}\boldsymbol{G}_\alpha\boldsymbol{\beta} \overset{d}{=} \boldsymbol{S}_{N^*}T^{-1}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(T^{-3}N^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}T^{-2}N^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{\eta}.$$

QED

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

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

where $\tilde{y}_{t+1} = \bar{y} + \boldsymbol{x}'_t\boldsymbol{G}_\alpha\boldsymbol{\beta}$ *and* Q_t^2 *is the* t^{th} *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.

QED

Theorem 5. *Under Assumptions 1-6, as $N, T \to \infty$ and $T/N \to 0$ 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,*

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

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 $\boldsymbol{\Psi}_\beta = \boldsymbol{\Sigma}_z^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_{\boldsymbol{F}}\boldsymbol{\mathcal{P}}$. 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

$$\left(T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1}T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\boldsymbol{\eta}\boldsymbol{\eta}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\left(T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\hat{\boldsymbol{F}}\right)^{-1}. \tag{A8}$$

Assumption 2.5 and Lemma 7 imply that $\text{plim}_{T,N\to\infty}\, T^{-1}\hat{\boldsymbol{F}}'\boldsymbol{J}_T\boldsymbol{\eta}\boldsymbol{\eta}'\boldsymbol{J}_T\hat{\boldsymbol{F}} = \text{plim}_{T,N\to\infty}T^{-1}\sum_t \eta_{t+h}^2(\hat{\boldsymbol{F}}_t - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{F}}_t - \hat{\boldsymbol{\mu}})'$. By Theorem 1, $\eta_{t+h} = \hat{\eta}_{t+h} + o_p(1)$, which implies that $\widehat{Avar}(\hat{\boldsymbol{\beta}})$ and (A8) share the same probability limit, therefore $\widehat{Avar}(\hat{\boldsymbol{\beta}})$ is a consistent estimator of $\boldsymbol{\Sigma}_\beta$.

QED

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 $\liminf \sqrt{N}/T \geq \tau \geq 0$, 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

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

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

When $\liminf \sqrt{N}/T \geq \tau > 0$, 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 $\liminf T/\sqrt{N} \leq 1/\tau < \infty$.

QED

Define

$$\boldsymbol{H}_0 = \hat{\boldsymbol{F}}_A\hat{\boldsymbol{F}}_B^{-1}N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\phi_0 \quad \text{and} \quad \boldsymbol{H} = \hat{\boldsymbol{F}}_A\hat{\boldsymbol{F}}_B^{-1}N^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{\Phi}. \tag{A9}$$

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

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 $\boldsymbol{\mathcal{P}} = \boldsymbol{I}$).

QED

A.6 Automatic Proxy Selection

Theorem 7. *Let Assumptions 1, 2, 3, and 5 hold. Then the L-automatic-proxy three pass regression filter forecaster of* $\boldsymbol{y}$ *satisfies Assumption 6 when* $L = K_f$.

Proof: If $K_f = 1$, Assumption 6 is satisfied by using $\boldsymbol{y}$ has 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 Proposition 1 and Equation 1 that

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

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 $\boldsymbol{\beta} = (\boldsymbol{\beta}_f', \boldsymbol{0}')'$, 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 $\boldsymbol{F}[\boldsymbol{0}, \boldsymbol{I}]'$, where the zero matrix is $K_g \times K_f$ and the identity matrix is dimension K_g. This, together with Assumptions 2.5 and 4.3, implies that the cross product matrix $[\boldsymbol{0}, \boldsymbol{I}]\boldsymbol{F}'\boldsymbol{r}^{(1)}$ is zero in expectation.

The induction step proceeds as follows. By hypothesis, suppose we have $k < K_f$ automatically-selected proxies with factor loadings $[\boldsymbol{\Lambda}_{f,k}, \boldsymbol{0}]$, 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 $\boldsymbol{\beta}_f - \boldsymbol{\Phi}_f'\boldsymbol{\Omega}^{(k)}\boldsymbol{f}\boldsymbol{\beta}_f$, where $\boldsymbol{f} = \boldsymbol{F}\boldsymbol{S}_{K_f}$ and $\boldsymbol{S}_{K_f} = [\boldsymbol{I}, \boldsymbol{0}]'$ 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 $\boldsymbol{\beta} = [\boldsymbol{\beta}_f', \boldsymbol{0}']'$). Also note that as part of the induction hypothesis, $\boldsymbol{\Omega}^{(k)}$ is constructed based on $\boldsymbol{Z} = (\boldsymbol{r}^{(1)}, ..., \boldsymbol{r}^{(k-1)})$.

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 $(\boldsymbol{I} - \boldsymbol{\Lambda}_{f,k}'(\boldsymbol{\Lambda}_{f,k}\boldsymbol{\Lambda}_{f,k}')^{-1}\boldsymbol{\Lambda}_{f,k})\left(\boldsymbol{I} - \boldsymbol{\Phi}_f'\boldsymbol{\Omega}^{(k)}\boldsymbol{f}\right)\boldsymbol{\beta}_f$. Because $\left(\boldsymbol{I} - \boldsymbol{\Phi}_f'\boldsymbol{\Omega}^{(k)}\boldsymbol{f}\right) \neq \boldsymbol{0}$ 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.

QED

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.

A.7.1 Our Original Representation

Our analysis centers on the 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

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

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

$$\hat{y} \xrightarrow[T,N\to\infty]{p} \boldsymbol{F}'\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' \left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right]^{-1} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\beta}. \tag{A10}$$

By inspection, consistency requires three conditions to ensure that the coefficient vector post-multiplying $\boldsymbol{F}'$ in (A10) reduces to $(\boldsymbol{\beta}'_f, \mathbf{0})'$. These conditions are:

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

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

$$\begin{bmatrix} \boldsymbol{\mathcal{P}}_f\boldsymbol{\Delta}_f\boldsymbol{\Lambda}'_f + \boldsymbol{\mathcal{P}}_{fg}\boldsymbol{\Delta}'_{fg}\boldsymbol{\Lambda}'_f \\ \boldsymbol{\mathcal{P}}'_{fg}\boldsymbol{\Delta}_f\boldsymbol{\Lambda}'_f + \boldsymbol{\mathcal{P}}_g\boldsymbol{\Delta}_{fg}\boldsymbol{\Lambda}'_f \end{bmatrix}. \tag{A11}$$

Since the same matrix ($\left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right]^{-1} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{P}}\boldsymbol{\Delta}_F\boldsymbol{\beta}$) post-multiplies both of these rows, we can here determine the necessity of conditions 2 and 3. The bottom row of (A11) 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

$$\boldsymbol{f}'\boldsymbol{\mathcal{P}}_f\boldsymbol{\Delta}_f\boldsymbol{\Lambda}'_f \left[\boldsymbol{\Lambda}_f\boldsymbol{\Delta}_f\boldsymbol{\mathcal{P}}_f\boldsymbol{\Delta}_f\boldsymbol{\mathcal{P}}_f\boldsymbol{\Delta}_f\boldsymbol{\Lambda}'_f\right]^{-1} \boldsymbol{\Lambda}_f\boldsymbol{\Delta}_f\boldsymbol{\mathcal{P}}_f\boldsymbol{\Delta}_f\boldsymbol{\beta}_f. \tag{A12}$$

Consistency requires that (A12) 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 $\boldsymbol{\mathcal{P}}_f\boldsymbol{\Delta}_f = \sigma\boldsymbol{I}$ and $\boldsymbol{\Lambda}_f = \boldsymbol{\beta}_f$, for positive scalar σ. In this case (A12) becomes

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

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 $\boldsymbol{P}_f, \boldsymbol{\Delta}_f, \boldsymbol{\Lambda}_f$ are invertible (so that $\boldsymbol{P}_f\boldsymbol{\Delta}_f$ need not be an equivariance matrix). In this case (A12) 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've assumed, but not all rotations are observationally-equivalent once we've 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

$$\boldsymbol{m} = \boldsymbol{M}'\boldsymbol{f} \quad , \quad \mathbb{E}\left[\begin{pmatrix} h \\ \boldsymbol{m} \end{pmatrix}\begin{pmatrix} h & \boldsymbol{m} \end{pmatrix}\right] = \begin{pmatrix} \boldsymbol{\beta}_f' \\ \boldsymbol{M}' \end{pmatrix}\boldsymbol{\Delta}_f\begin{pmatrix} \boldsymbol{\beta}_f & \boldsymbol{M} \end{pmatrix} = \begin{bmatrix} \Delta_h & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Delta}_m \end{bmatrix}. \tag{A13}$$

The new formulation therefore satisfies

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

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

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

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

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

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

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

The critical term in determining whether this expression reduces to zero is $\boldsymbol{\beta}'_f\boldsymbol{\Delta}_f\boldsymbol{\mathcal{P}}_f\boldsymbol{\Delta}_f\boldsymbol{M}$. If the knife-edge condition holds, then we have $\boldsymbol{\beta}'_f\boldsymbol{\Delta}_f\boldsymbol{\mathcal{P}}_f\boldsymbol{\Delta}_f\boldsymbol{M}=\sigma\boldsymbol{\beta}'_f\boldsymbol{\Delta}_f\boldsymbol{M}=\boldsymbol{0}$ in light of (A13). However, in the general case, $\boldsymbol{\beta}'_f\boldsymbol{\Delta}_f\boldsymbol{\mathcal{P}}_f\boldsymbol{\Delta}_f\boldsymbol{M}\neq\boldsymbol{0}$ even though (A13) 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:

$$\boldsymbol{y}=\boldsymbol{f}\boldsymbol{\iota}+\boldsymbol{\eta},\qquad \boldsymbol{X}=\left[\begin{array}{cc}\boldsymbol{f} & \boldsymbol{g}\end{array}\right]\boldsymbol{\Phi}'+\boldsymbol{\varepsilon}$$

where $\boldsymbol{\iota}$ is $K_f\times 1$ ones vector, $\boldsymbol{g}$ $(T\times K_g)$, $\boldsymbol{\Phi}$ $(N\times K_f+K_g)$, $\boldsymbol{\eta}$ $(T\times 1)$, 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 $N=T=2,000$. We use an identity covariance matrix for factor loadings ($\boldsymbol{\mathcal{P}}=\boldsymbol{I}$) and consider two values for $\boldsymbol{\sigma}_f$: a knife-edge (equivariant) case $\left[\begin{array}{ccc}1 & 1 & 1\end{array}\right]$ and a more general (non-equivariant) case $\left[\begin{array}{ccc}0.5 & 1 & 2\end{array}\right]$.

Table A1 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 $N=T=2,000$ the correlation between the 3PRF forecast and the infeasible best forecast is 0.993, and their relative R^2 is 0.9901. 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 $N=T=2,000$; the correlation is 0.9169 for $N=T=200$ and 0.9241 for $N=T=2,000$. 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 2,000 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 $N=T=2,000$. 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 $N=T=2,000$.

Table A1: Simulation Study

	In-Sample			Out-of-Sample		
# auto proxies:	1	2	3	1	2	3
$N = T = 200$						
$\boldsymbol{\sigma}_f = \left[\ 1\quad 1\quad 1\ \right]$						
$\frac{\hat{y}R^2}{\boldsymbol{f\iota}R^2}$	0.9607			0.9316		
$\rho(\hat{y}, \boldsymbol{f\iota})$	0.9678			0.9649		
$\boldsymbol{\sigma}_f = \left[\ 0.5\quad 1\quad 2\ \right]$						
$\frac{\hat{y}R^2}{\boldsymbol{f\iota}R^2}$	0.8425	0.9736	0.9938	0.8307	0.9580	0.9735
$\rho(\hat{y}, \boldsymbol{f\iota})$	0.9169	0.9806	0.9892	0.9136	0.9791	0.9884
$N = T = 2,000$						
$\boldsymbol{\sigma}_f = \left[\ 1\quad 1\quad 1\ \right]$						
$\frac{\hat{y}R^2}{\boldsymbol{f\iota}R^2}$	0.9901			0.9850		
$\rho(\hat{y}, \boldsymbol{f\iota})$	0.9930			0.9929		
$\boldsymbol{\sigma}_f = \left[\ 0.5\quad 1\quad 2\ \right]$						
$\frac{\hat{y}R^2}{\boldsymbol{f\iota}R^2}$	0.8586	0.9762	0.9983	0.8575	0.9746	0.9962
$\rho(\hat{y}, \boldsymbol{f\iota})$	0.9241	0.9877	0.9981	0.9238	0.9876	0.9981

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

A.8 The Kalman Filter

This system is defined by the state space

$$\boldsymbol{\Pi}_t = \boldsymbol{M}_0 + \boldsymbol{M}\boldsymbol{\Pi}_{t-1} + \mathbf{error}_t^F, \qquad \mathbf{error}_t^F \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}) \tag{A14}$$

$$\boldsymbol{\Upsilon}_t = \boldsymbol{\Psi}_0 + \boldsymbol{\Psi}\boldsymbol{\Pi}_t + \mathbf{error}_t^{\Upsilon}, \qquad \mathbf{error}_t^{\Upsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}) \tag{A15}$$

$$\boldsymbol{\Pi}_t = \begin{pmatrix} \boldsymbol{F}_t \\ \boldsymbol{F}_{t-1} \end{pmatrix} \tag{A16}$$

$$\boldsymbol{\Upsilon}_t = \begin{pmatrix} \tilde{\boldsymbol{z}}_t \\ \boldsymbol{x}_t \end{pmatrix} \tag{A17}$$

$\boldsymbol{\Pi}_t$ is an augmented state vector containing both the current and lagged values of the $(K_f + K_g)$-dimensional factor vector $\boldsymbol{F}_t$. We assume that each element of the proxy vector depends only on the current or the lagged factor, not both. Given the system parameters $\{\boldsymbol{M}, \boldsymbol{M}_0, \boldsymbol{Q}, \boldsymbol{\Psi}, \boldsymbol{\Psi}_0, \boldsymbol{R}\}$, the Kalman filter provides the conditional expectation $\mathbb{E}(\boldsymbol{\Pi}_t | \boldsymbol{\Upsilon}_t, \boldsymbol{\Upsilon}_{t-1}, \ldots)$ if initialized at $\mathbb{E}(\boldsymbol{\Pi}_0)$: therefore it provides the least squares predictor (see Maybeck (1979)). The well-known filter equations (see Hamilton (1994)) are:

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{M}\boldsymbol{P}_{t-1|t-1}\boldsymbol{M}' + \boldsymbol{Q} \tag{A18}$$

$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1}\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1} \tag{A19}$$

$$\boldsymbol{K}_t = \boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1} \tag{A20}$$

$$\boldsymbol{\Pi}_{t|t-1} = \boldsymbol{M}_0 + \boldsymbol{M}\boldsymbol{\Pi}_{t-1|t-1} \tag{A21}$$

$$\boldsymbol{\Upsilon}_{t|t-1} = \boldsymbol{\Psi}_0 + \boldsymbol{\Psi}\boldsymbol{\Pi}_{t|t-1} \tag{A22}$$

$$\boldsymbol{\Pi}_{t|t} = \boldsymbol{\Pi}_{t|t-1} + \boldsymbol{K}_t\left(\boldsymbol{\Upsilon}_t - \boldsymbol{\Upsilon}_{t|t-1}\right) \tag{A23}$$

We first note that the matrix inversion lemma lets us rewrite (A19) as

$$\boldsymbol{P}_{t|t} = \left(\boldsymbol{P}_{t|t-1}^{-1} + \boldsymbol{\Psi}'\boldsymbol{R}^{-1}\boldsymbol{\Psi}\right)^{-1}$$

Then, following Simon (2006), (A20) can be rewritten in a form similar to (7) by seeing that

$$\begin{aligned}
\boldsymbol{K}_t &= \boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1} \\
&= \boldsymbol{P}_{t|t}\boldsymbol{P}_{t|t}^{-1}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1} \\
&= \boldsymbol{P}_{t|t}\left(\boldsymbol{P}_{t|t-1}^{-1} + \boldsymbol{\Psi}'\boldsymbol{R}^{-1}\boldsymbol{\Psi}\right)\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1} \\
&= \boldsymbol{P}_{t|t}\left(\boldsymbol{\Psi}' + \boldsymbol{\Psi}'\boldsymbol{R}^{-1}\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\right)\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1} \\
&= \boldsymbol{P}_{t|t}\boldsymbol{\Psi}'\left(\boldsymbol{I} + \boldsymbol{R}^{-1}\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\right)\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1} \\
&= \boldsymbol{P}_{t|t}\boldsymbol{\Psi}'\boldsymbol{R}^{-1}\left(\boldsymbol{R} + \boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\right)\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1} \\
&= \boldsymbol{P}_{t|t}\boldsymbol{\Psi}'\boldsymbol{R}^{-1},
\end{aligned}$$

where we have premultiplied by $\boldsymbol{I} = \boldsymbol{P}_{t|t}\boldsymbol{P}_{t|t}^{-1}$ in the second line, we have rewritten $\boldsymbol{P}_{t|t}^{-1}$ in the third line, we have distributed in lines four and five, we have rewritten $\boldsymbol{I}$ as $\boldsymbol{R}^{-1}\boldsymbol{R}$ and then distributed in the sixth line, and simplified in the final line.

Next, we look understand what is the maximum likelihood estimate (MLE) of the system parameters. According to Watson and Engle (1983) the parameters that maximize the likelihood can be found using the EM algorithm of Dempster, Laird, and Rubin (1977). To simplify, assume $\boldsymbol{\Psi}_0$ and $\boldsymbol{M}_0$ are zero. Hence, the MLE of $\boldsymbol{\Psi}$ satisfies the following

$$\widehat{vec(\boldsymbol{\Psi})} = \left(\hat{\boldsymbol{\Pi}}'\hat{\boldsymbol{\Pi}} \otimes \hat{\boldsymbol{R}}^{-1}\right)^{-1}\left(\hat{\boldsymbol{\Pi}}'\boldsymbol{\Upsilon} \otimes \hat{\boldsymbol{R}}^{-1}\right) \tag{A24}$$

for $\hat{\boldsymbol{\Pi}} = \left(\hat{\boldsymbol{\Pi}}_1, \ldots, \hat{\boldsymbol{\Pi}}_T\right)'$, $\boldsymbol{\Upsilon} = (\boldsymbol{\Upsilon}_1, \ldots, \boldsymbol{\Upsilon}_T)'$, and

$$\hat{\boldsymbol{R}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Upsilon}_t - \boldsymbol{\Psi}\hat{\boldsymbol{\Pi}}_t \tag{A25}$$

and $\hat{\boldsymbol{\Pi}}_t$ denotes the best possible estimate of the latent factors on the basis of the MLE of the system parameters (this is usually given by the smoothed estimates $\boldsymbol{\Pi}_{t|T}$ using the MLE parameters). Equations (A24) and (A25) make it clear that the MLE of $\boldsymbol{\Psi}$ is obtained by a GLS regression of the observable variables $\boldsymbol{\Upsilon}$ on the best estimate of the latent factors.

Finally, consider the optimal linear prediction of $\boldsymbol{\Upsilon}_{t+1}$ on the basis of $\{\boldsymbol{\Upsilon}_t, \boldsymbol{\Upsilon}_{t-1}, \ldots\}$, ignoring the Kalman filter's temporal pooling of information. We do this by considering $\boldsymbol{M}$, which we can partition into four square submatrices. $\boldsymbol{M}_{21} = \boldsymbol{I}$ and $\boldsymbol{M}_{22} = \boldsymbol{0}$. Ignoring the temporal pooling is equivalent to restricting $\boldsymbol{M}_{11}$ and $\boldsymbol{M}_{12}$ to be zero. Clearly, $\boldsymbol{M}_{11} = \boldsymbol{0}$ imposes that the latent factors are serially uncorrelated, a restriction that is almost always invalidated by the data. This restriction also forces the estimate of $\mathbf{\Pi}_t$ to be based solely on time t information from the N-dimensional cross section. As N gets large, time t information alone is sufficient to estimate the latent factor precisely. Therefore

$$\begin{aligned}\mathbf{\Upsilon}_{t+1|t} &= \mathbf{\Psi}\mathbf{\Pi}_{t+1|t} \\ &= \mathbf{\Psi}\boldsymbol{M}\mathbf{\Pi}_{t|t} \\ &= \mathbf{\Psi}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{I} & \mathbf{0} \end{bmatrix}\mathbf{\Pi}_{t|t} \qquad \text{(A26)}\end{aligned}$$

Recall that y is an element of $\mathbf{\Upsilon}$, and we can therefore let the first row of $\mathbf{\Psi}$ be $(\mathbf{0}', \boldsymbol{\beta}')$, which with (A26) says

$$y_{t+1|t} = \boldsymbol{\beta}'\boldsymbol{F}_{t|t}.$$

This coincides with the infeasible best forecast that we refer to throughout.

A.9 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, Tibshirani, and Friedman (2009)):

1. Standardize each $\mathbf{x}_i$ to have mean zero and variance one by setting $\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i - \hat{\mathbb{E}}[\mathbf{x}_{it}]}{\hat{\sigma}(\mathbf{x}_{it})}, \ i = 1, ..., N$
2. Set $\hat{\boldsymbol{y}}^{(0)} = \bar{y}$, and $\mathbf{x}_i^{(0)} = \tilde{\mathbf{x}}_i, \ i = 1, ..., N$
3. For $k = 1, 2, ..., K$
   (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 $\boldsymbol{u}_k$:

$$\mathbf{x}_i^{(k)} = \mathbf{x}_i^{(k-1)} - \left(\widehat{Cov}(\boldsymbol{u}_k, \mathbf{x}_i^{(k-1)})/\widehat{Var}(\boldsymbol{u}_k)\right)\boldsymbol{u}_k, \ i = 1, 2, ..., N.$$

A.10 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 (1993). 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 Moodys book equity data (if available) from Davis, Fama and French (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 $\{\frac{100}{K}, 2\frac{100}{K}, ..., (K-1)\frac{100}{K}\}$ quantiles of a given characteristic.