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

January 2011

Abstract

We propose a new approach to forecasting a single time series when the number of predictor variables is large. Our procedure can be conveniently represented and implemented as a series of ordinary least squares regressions, and can be interpreted as a solution to a restricted version of the Kalman filtering problem. Like principal components, it efficiently estimates underlying factors among the predictors. Unlike principal components, it can provide efficient forecasts even when the number of extracted factors is fewer than the total number of factors in the data. It also produces superior forecasts when the factors dominating the forecast target's variation contribute only weakly to variance among the predictors. The estimator performs exceedingly well in two empirical applications that exemplify the many predictor problem: Forecasting equity returns and macroeconomic aggregates.

[*] We thank Jim Hamilton 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 on the views or opinions of the Federal Reserve Board, the Federal Reserve System, or their respective staffs.

1 Introduction

There is widespread academic and financial interest in forecasting important macroeconomic quantities. For example, since Burn's and Mitchell's work a vast array of employment and production variables have been linked to the fluctuations of aggregate output and hours (see Geweke 1977, Sargent and Sims 1977, Stock and Watson 1989, Bernanke, Boivin and Eliasz 2005 and Aruoba, Diebold and Scotti 2009 among others). Similarly, since the introduction of today's benchmark pricing models the cross section of asset prices have been related to the values of market prices and dividends (see Sharpe 1964, Lintner 1965 and Treynor 1961 among others). These theories state that there are myriad variables economically linked to a target macroeconomic variable in question, which presents a problem: how does one effectively use the large cross section in a time series analysis? The standard solution is to view the data as coming from an approximate factor model where unobservables drive the systematic variation of both the target variable and the cross section of predictive information. Analysis then proceeds by reducing the dimension of the large cross section to its important predictive factors and relating the factors to the target. As the benchmark, a large literature uses the method of principal components for dimension reduction followed by least squares to estimate the factors' relationship to the target (see Forni and Reichlin 1998, Forni, Hallin, Reichlin, and Lippi 2004, Stock and Watson 1989, 2002a, 2002b, 2006, 2009, Bai and Ng 2002, 2008 and Ludvigson and Ng 2009 among others). We introduce an alternative that uses least squares for both steps. This alternative is called the Three-Pass Regression Filter and shows superior performance to the benchmark in many relevant contexts.

The Three-Pass Regression Filter (3PRF) is asymptotically efficient in Stock and Watson's (2002a) sense, meaning that the feasible forecasts converge in probability to the infeasible best forecasts constructed using the actual values of the latent factors. The 3PRF obtains asymptotic efficiency when one estimates every factor in the data, just as does the principal components (PC) forecaster (Stock and Watson 2002a).Furthermore, the 3PRF can obtain asymptotic efficiency when one estimates *fewer* than the total number of factors in the data. If the target is driven by *relevant* factors while the data additionally contain *irrelevant* factors, the 3PRF can be asymptotically efficient by estimating only the number of relevant factors. This is important because there is no assured connection between which factors are relevant to the target variable and which factors are responsible for the predictors' variation. Irrelevant factors can be either economically interesting or uninteresting, yet regardless of their interpretation economic theory often cannot rule out their existence. Thus, the 3PRF's performance in these cases is quite advantageous.

It is well-known that for a fixed cross-sectional dimension N the Kalman filter (KF) provides the least squares predictor of the linear system we have in mind (see Maybeck 1979 or Hamilton 1994). The KF pools both current information (cross-sectionally) and past information (temporally) and is most often implemented along with maximum likelihood parameter estimation. The resulting information extraction is based on the covariance between each cross-sectional unit and latent factors that drive the target variable. However, the maximum likelihood parameter estimates are not available in closed form and must be obtained via numerical optimization. When the number of predictors becomes large, the difficulty of satisfactorily analyzing the likelihood surface is apparent: 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 KF due to the difficulty of parameter estimation. We show how the 3PRF is a special restricted form of the KF that ignores the temporal pooling of past information, but retains the cross-sectional combination of information via least squares.[1] By putting restrictions on the conditioning set and eschewing the KF's use of GLS, the 3PRF is available in closed form via OLS. Hence, the 3PRF is instantaneously computable for virtually arbitrary time-series length T and cross section dimension N.[2]

We compare the 3PRF to the widely used PC benchmark.[3] We give the condition in which 3PRF outperforms PC, and argue that this condition is essentially always feasible. The key difference between PC and 3PRF is in their respective methods of *dimension reduction*. Stock and Watson's (2002a) key insight is to recognize that we require a dimension reduction method to condense information from the large cross section into a small number of highly informative predictors, because linear projections work poorly when the number of regressors is near the number of observations.[4] PC achieves the dimension reduction by considering the eigenvalue decomposition of the predictors' sample covariance matrix, and so predictive

[1] The 3PRF is a modification of Fama and MacBeth's (1973) well known procedure. Although our exposition starts with the KF because of logical flow, the 3PRF's intellectual debt to Fama-MacBeth is noteworthy. We discuss this below.

[2] For instance, if one has 500 predictors and 1000 time periods, calculation of the 3PRF takes less than one-tenth of a second on a standard desktop using Matlab 7.7.

[3] Another method is De Jong and Kiers's (1992) method of principal covariate regression. Heij, Groenen, and van Dijk's (2007) analyze this method and demonstrate promising performance. However, it involves a nonlinear optimization problem that they solve by iterative majorization, which hampers its widespread use.

[4] Heuristically, this poor performance flows from Huber's (1973) point about the asymptotic difficulties of OLS when the number of regressors is allowed to grow too fast. Of course, if $N > T$ LP is not defined at all, which goes to the point.

information is extracted according to *covariance within the cross section.* The 3PRF achieves the dimension reduction by remaining close to KF's least squares mechanics, and so predictive information is extracted according to *covariance with the factors driving the target variable.* This least squares intuition explains the 3PRF's superior forecast performance found in theory and supported by both monte carlo simulation and empirical application.

The 3PRF is based on three types of variables: a target variable, a cross section of predictors, and a set of proxies. Proxies are variables that are driven by the factors, and therefore any combination of the predictors or target can act as a proxy. The target itself is a very good proxy because by definition it is driven only by relevant factors. Finding other proxies can involve economic reasoning, as Kelly and Pruitt (2010) do when using the 3PRF to estimate expected stock market returns and dividend growth from the cross section of disaggregate price-dividend ratios. On the other hand, statistical methods can be used to find proxies: in fact, principal components of the predictors work as proxies, and this is what allows the 3PRF to work at least as well as PC. We discuss the issue of proxy choice more in-depth below.

The paper is structured as follows: Section 2 formally defines the 3PRF and presents proofs of its consistency.[5] Section 3 compares the 3PRF to the KF as well as PC. Section 4 covers some implementation issues. Section 5 compares the 3PRF to PC via simulation evidence, and Section 6 continues the comparison by presenting empirical results from forecasting market stock returns and important macroeconomic aggregates. We then conclude. All proofs and supporting details are placed in the Appendix.

[5] The present draft is incomplete: future drafts will show the asymptotic distribution of the estimator.

2 The Three-Pass Regression Filter

We describe the Three-Pass Regression Filter and present propositions as to its statistical properties. In the process, we compare the procedure to the principal components (PC) forecaster, a widely used benchmark. We then relate the procedure to the well-known Fama-MacBeth procedure. All lemmas and propositions are proved in Appendix A.

2.1 Procedure

The following is the environment wherein we use the Three-Pass Regression Filter (3PRF): There is a *target* variable which we wish to forecast. There exist many *predictors* which may possibly give predictive information about this target variable. The number of predictors N is large and close to- or greater than the number of available observations T, which makes naïve linear projection problematic.[6] Therefore we look to reduce the dimension of the predictive content, and to do so we suppose the data can be described by an approximate factor model. In order to make forecasts, the 3PRF uses *proxies* which are variables driven by the factors – these can be taken from amongst the set of predictors and target. The target is a linear function of relevant factors, plus some unforecastable noise. Therefore, the best possible forecast is a linear function of the unobservable factors – we call this the *infeasible best*. Following Stock and Watson (2002a), when feasible forecasts converge in probability to the infeasible best, we say that the forecaster is *asymptotically efficient*.

We write $\boldsymbol{y}$ for the $(T \times 1)$ vector of the target variable's time series from $2, 3, \ldots, T+1$. Let $\boldsymbol{X}$ be the $(T \times N)$ matrix of predictors, $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)' = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_N \end{pmatrix}$: note that we are using two different typefaces to denote the N-dimensional cross section of predictors $\boldsymbol{x}_t$ observed at time t, and the T-dimensional time series of the i^{th} predictor $\mathbf{x}_i$, respectively. We denote the $(T \times M)$ matrix of proxies as $\boldsymbol{Z}$ and we write $\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_T)'$. We make no assumption on the relationship between N and T. With this notation in mind, the 3PRF is defined in Table 1.

In the first pass of the 3PRF, we run N separate *time series* regressions. In these first stage regressions, the predictors are the left-hand-side variable while the proxies are the regressors. We carry the estimated first-stage coefficients on the proxies into the second stage. In the second pass of the 3PRF, we run T separate *cross section* regressions. In these second stage regressions, the predictors are (once again) the left-hand-side variable while

[6] We say "naïve linear projection" to mean projection of the target on all the predictors. We are interested with contexts in which this is well-known to perform poorly (when $N < T$ but N is close to T) or is not even defined ($N \geq T$).

Table 1: The Three-Pass Regression Filter

Pass	Description
1	Run time series regression of $\mathbf{x}_i$ on $\boldsymbol{Z}$ for $i = 1, \ldots, N$ $x_{i,t} = \phi_{0,i} + \boldsymbol{z}_t'\boldsymbol{\phi}_i + \varepsilon_{i,t}$
2	Run cross section regression of $\boldsymbol{x}_t$ on $\hat{\boldsymbol{\phi}}_i$ for $t = 1, \ldots, T$ $x_{i,t} = \varsigma_{0,t} + \hat{\boldsymbol{\phi}}_i'\boldsymbol{\varsigma}_t + \epsilon_{i,t}$
3	Run time series regression of $\boldsymbol{y}$ on the predictive factors $\hat{\boldsymbol{\varsigma}}_1, \ldots, \hat{\boldsymbol{\varsigma}}_T$ $y_{t+1} = \beta_0 + \hat{\boldsymbol{\varsigma}}_t'\boldsymbol{\beta} + \eta_{t+1}$

Notes: All regressions are via OLS.

the first-stage coefficients are the regressors. We carry the estimated second-stage predictive factors (what we call the estimated coefficients on the first-stage estimated coefficients) into the third stage. In the third pass of the 3PRF, we run one *time series* regression. In this third stage regression, the target variable is the left-hand-side variable while the second-stage predictive factors are the regressors. All regressions are via OLS and therefore the entire procedure is nearly instantaneous. We refer to the third-stage fitted value for y_{t+1} as the 3PRF's time t forecast.

The following lemma gives an alternative representation for the 3PRF that highlights its availability in closed form. We use the notation $\boldsymbol{J}_L \equiv \boldsymbol{I}_L - L^{-1}\boldsymbol{\iota}_L\boldsymbol{\iota}_L'$ where $\boldsymbol{I}_L$ is the L-dimensional identity matrix and $\boldsymbol{\iota}_L$ is a L-vector of ones. Then we have the following:

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

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

2.2 Properties

We will refer to the following assumptions:

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

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

where $\boldsymbol{F}_t = \begin{pmatrix} \boldsymbol{f}_t' & \boldsymbol{g}_t' \end{pmatrix}'$, $\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 \end{pmatrix}$, $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{\Lambda}_2 \end{pmatrix}$, *and* $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1' & \boldsymbol{0}' \end{pmatrix}'$.

$K_f > 0$ *is the dimension of the vector* $\boldsymbol{f}_t$*, and the column dimension of* $\boldsymbol{\Phi}_1, \boldsymbol{\Lambda}_1$ *and the row dimension of* $\boldsymbol{\beta}_1$.

$K_g \geq 0$ *is the dimension of the vector* $\boldsymbol{g}_t$*, and the column dimension of* $\boldsymbol{\Phi}_2, \boldsymbol{\Lambda}_2$ *– when* $K_g = 0$*,* $\boldsymbol{g}_t, \boldsymbol{\Phi}_2, \boldsymbol{\Lambda}_2$ *disappear.*

$M > 0$ *is the dimension of the vector* $\boldsymbol{z}_t$*, and the row dimension of* $\boldsymbol{\Lambda}$.

Let $K = K_f + K_g$.

The unforecastable shock is such that $\mathbb{E}_t(\eta_{t+1}) = 0, \mathbb{E}_t(\eta_{t+1}^2) = \delta_\eta < \infty$ *for all* t*. Hence, the "infeasible best" forecast of* y_{t+1} *conditional on time* t *information is given by* $\beta_0 + \boldsymbol{\beta}'\boldsymbol{F}_t = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{f}_t$.

Assumption 2 (Factors and Loadings)**.** *For* $\tilde{\boldsymbol{\phi}}_i \equiv (\phi_{i,0}, \boldsymbol{\phi}_i')'$ *and* $\bar{\phi} < \infty$

1. $T^{-1}\sum_{t=1}^T \boldsymbol{F}_t \xrightarrow[T\to\infty]{p} \boldsymbol{\mu}$ *and* $T^{-1}\sum_{t=1}^T \boldsymbol{F}_t\boldsymbol{J}_T\boldsymbol{F}_t' \xrightarrow[T\to\infty]{p} \boldsymbol{\Delta}_F$.
2. $|\tilde{\boldsymbol{\phi}}_i| \leq \bar{\phi}\ \forall i$.
3. $N^{-1}\sum_{i=1}^N \tilde{\boldsymbol{\phi}}_i\boldsymbol{J}_N\tilde{\boldsymbol{\phi}}_i' \xrightarrow[N\to\infty]{p} \begin{bmatrix} B_0 & \boldsymbol{B}_1' \\ \boldsymbol{B}_1 & \boldsymbol{\mathcal{B}} \end{bmatrix}$ *with* $\boldsymbol{\mathcal{B}}$ *nonsingular.*

Assumption 3 (Error Moments)**.** *There exists a constant* $A < \infty$ *such that*

1. $T^{-1}\sum_{t=1}^T \varepsilon_{i,t} \xrightarrow[T\to\infty]{p} \boldsymbol{0}\ \forall i$ *and* $N^{-1}\sum_{i=1}^N \varepsilon_{i,t} \xrightarrow[N\to\infty]{p} \boldsymbol{0}\ \forall t$
2. $T^{-1}\sum_{t=1}^T \boldsymbol{\omega}_t \xrightarrow[T\to\infty]{p} \boldsymbol{0}$ *and* $T^{-1}\sum_{t=1}^T \boldsymbol{\eta}_t \xrightarrow[T\to\infty]{p} \boldsymbol{0}$.
3. $T^{-1}\sum_{t=1}^T \boldsymbol{\varepsilon}_t\boldsymbol{\eta}_t' \xrightarrow[T\to\infty]{p} \boldsymbol{0}, T^{-1}\sum_{t=1}^T \boldsymbol{F}_t\boldsymbol{\eta}_t' \xrightarrow[T\to\infty]{p} \boldsymbol{0}$ *and* $T^{-1}\sum_{t=1}^T \boldsymbol{F}_t\boldsymbol{\omega}_t' \xrightarrow[T\to\infty]{p} \boldsymbol{0}$.
4. $T^{-1}\sum_{t=1}^T \varepsilon_{i,t}\omega_{k,t} \xrightarrow[T\to\infty]{p} \gamma(i,k),$ *and* $\lim_{N\to\infty}\sup_k \sum_{i=1}^N |\gamma(i,k)| \leq A$.
5. $T^{-1}\sum_{t=1}^T \varepsilon_{i,t}\varepsilon_{j,t} \xrightarrow[T\to\infty]{p} \delta(i,j) = \delta(j,i),$ *and* $\lim_{N\to\infty}\sup_j \sum_{i=1}^N |\delta(i,j)| \leq A$.
6. $N^{-1}\sum_{i=1}^N \varepsilon_{i,t}\varepsilon_{i,s} \xrightarrow[N\to\infty]{p} \kappa(t,s) = \kappa(s,t),$ *and* $\lim_{T\to\infty}\sup_s \sum_{t=1}^T |\kappa(s,t)| \leq A$.

Assumption 4 (Rank Condition)**.** *The matrix* $\boldsymbol{\Lambda}$ *is nonsingular.*

Assumption 5 (Alternative Rank Condition)**.** $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \mathbf{0})$ *and* $\boldsymbol{\Lambda}_1$ *is nonsingular.*

Assumption 6 (Normalization)**.** $\boldsymbol{\mathcal{B}} = \boldsymbol{I}$ *and* $\boldsymbol{\Delta}_F$ *is diagonal.*

Assumption 1 gives the factor structure that allows us to reduce the dimension of the information contained in the predictors. It states that the target depends only on *relevant* factors $\boldsymbol{f}_t$ while some *irrelevant* factors $\boldsymbol{g}_t$ may exist and drive predictors and proxies. When we state that the 3PRF is asymptotically efficient, we mean

$$\hat{\boldsymbol{y}} \xrightarrow[N,T\to\infty]{p} \boldsymbol{\iota}_T \beta_0 + \boldsymbol{F}\boldsymbol{\beta} \tag{2}$$

for $\hat{\boldsymbol{y}}$ defined in (A1). Note that Assumption 1 imposes no dynamic structure on the factor evolution, but it does impose that the "relationship between $x_{i,t}$ and $\boldsymbol{F}_t$ is still static," in the words of Bai (2003). Yet, Stock and Watson (2002a) note that if we assume a linear dynamic factor transition equation, then we can use $\boldsymbol{F}_t$ to capture all the lags of the dynamic factor that enter into $x_{i,t}$ (Stock and Watson 2002a, p.1168). In practice, the structure is not very restrictive because one can partial out the target's own lags to capture dynamic effects: Stock and Watson (2009) and the exercise of Section 6.2 are examples.

Assumption 2 allows the factor structure to be approximate with serially correlated idiosyncratic shocks, following Chamberlain and Rothschild (1983) and Stock and Watson (2002a). Factors can have any finite mean and variance. We require loadings to be cross-sectionally regular insofar as there is a well-defined "covariance" matrix of the loadings for large N and they are finite. Note that $B_0, \boldsymbol{B}_1$ can be zero or nonzero, but $\boldsymbol{\mathcal{B}}$ is nonsingular which rules out the case that all predictors' factor loadings are identical.

Assumption 3 follows Stock and Watson (2002) and Bai (2003) in modestly limiting the time- and cross-sectional dependency amongst the idiosyncratic shocks. This structure rules out stochastic trends, but allows for serial- and cross-sectional correlation in proxies and noise, as well as temporal and cross-sectional heteroskedasticity (such as GARCH). In addition, we structure the noise properties and allow for temporal heteroskedasticity (such as GARCH). We additionally allow for some dependence between the idiosyncratic shocks and proxy noise.

Our first result is Proposition 1. The conditions and statement resemble Stock and Watson's (2002a) Theorem 2 for the principal components (PC) forecaster.[7] Assumption 4 assumes that we know the *total* number of factors driving the data, the same assumption as Stock and Watson and Bai make. Those authors argue that the assumption is warranted

[7] Or the obvious corollary to Bai's (2003) Proposition 2.

because there exist consistent estimators of this number, and we agree that there are many cases where this assumption quite reasonable[8] Our first result is given by the following:

Proposition 1. *Let Assumptions 1, 2, 3 and 4 hold. The three pass regression filter forecaster of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and proxies $\boldsymbol{Z}$ is asymptotically efficient.*

Proposition 1 says that the 3PRF is asymptotically efficient when we span the complete factor space by our choice of proxies. We do not view this condition as any stronger than the condition that we know the total number of factors K, because we can find these proxies as combinations of the target or predictors themselves, and hence Proposition 1 as always practically available.

We now consider what the 3PRF delivers in cases where Assumption 4 does not hold. For instance: in out-of-sample forecasting applications it is typically desirable to be as parsimonious as possible with the number of predictive factors used; or, we may have reason to believe that many factors exist in the data, but only a small number drive predictable variation in the target.[9] Hence we consider the following conditions. Assumption 6 is a normalization like that imposed in Stock and Watson's (2002a) assumption F1 (p. 1168) for the PC forecaster: this simply provides an identification of each factor because there exists an inherent unidentification between the factors and factor loadings, and is more a definition than an imposition.[10] Assumption 5 imposes that our choice of proxies is judicious insofar as they are driven by only relevant factors. We then obtain our second result:

Proposition 2. *Let Assumptions 1, 2, 3, 5 and 6 hold. The three pass regression filter forecaster of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and proxies $\boldsymbol{Z}$ is asymptotically efficient.*

Let "the target-proxied 3PRF" denote the 3PRF using as its only proxy the target variable itself. Proposition 2 has an noteworthy implication:

[8] Bai and Ng (2002) and Onatski (2009) provide consistent estimators of the number of factors, both of which use principal components.

[9] For instance, in Kelly and Pruitt (2010a) we note that a log price-dividend ratio is approximately the sum of time t expectations of all future returns and dividend growth rates. It is a strong assumption to state that we know the number of factors that are used by the market in forming expectations for *every future horizon*. However, we might feel more comfortable with the assumption that a small number of factors drive a certain horizon's expectations.

[10] Using our notation, Stock and Watson say:

> [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 F1(a) restricts $\boldsymbol{R}$ to be orthonormal, and this together with Assumption F1(b) restricts $\boldsymbol{R}$ to be a diagonal matrix with diagonal elements of ±1.

Corollary 1. *Let Assumptions 1, 2, 3, 6 hold. Additionally, assume that $K_f = 1$. The target-proxied three pass regression filter forecaster of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and proxy $\boldsymbol{y}$ is asymptotically efficient, for any value of K_g.*

Corollary 1 is important because the target is always available to us and by definition satisfies Assumption 5. It also points to a contrast between the 3PRF and PC. The key feature of PC is that its factor estimates are driven completely by the covariation within the elements of $\boldsymbol{x}$. The first estimated factor (by definition) is the linear combination of $\boldsymbol{x}$ that explains most of the variation within the predictors; the second estimated factor (by definition) is the linear combination of $\boldsymbol{x}$ that is orthogonal to the first estimated factor and explains most of the variation within the predictors; the third, etc. But in general the factor(s) driving $\boldsymbol{y}$ may have *any relationship* to the linear combinations of $\boldsymbol{x}$ that explain most of the variation within $\boldsymbol{x}$. For example, suppose there is a single relevant factor and Assumption 6 holds. If the relevant factor happens to also drive most of the variation within the predictors, the PC's single factor forecast is asymptotically efficient. But if the relevant factor drives less of the predictors' variation than does their first principal component, PC cannot obtain the infeasible best using only one factor. Meanwhile, the target-proxied 3PRF is asymptotically efficient regardless of the relevant factor's role in driving variation amongst the predictors

The next proposition additionally imposes that the factors are all equally variable:

Proposition 3. *Let Assumptions 1, 2, 3, 5 and 6 hold. Further assume that the factors $\boldsymbol{f}_t$ have equal variance σ_f^2. Then the target-proxied three pass regression filter forecaster of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and proxy $\boldsymbol{y}$ is asymptotically efficient, for any number of factors K_f and K_g.*

We view Proposition 3 as remarkable. Using a single proxy for the 3PRF is like using one principal component for the PC in that a single predictive factor estimate results. How could the 3PRF's single predictive factor be asymptotically similar to the infeasible best that is constructed from multiple factors? We come at the answer from an indirect way. Recall that we are assured that both the 3PRF and PC are asymptotically efficient when they use K predictive factors. But whatever this number of predictive factors, both of these forecasts eventually boil down to a particular linear combination of the cross section of predictors. It turns out when the relevant factors have the same variance, the target-proxied 3PRF nails this linear combination in the limit.

Our final proposition gives the condition under which the 3PRF forecast is superior to the PC forecast, in the sense that the former's forecast error is weakly smaller in the limit.

We refer to this as being asymptotically more efficient:

Proposition 4. *Let Assumptions 1, 2, and 3 hold. Define the matrix*

$$\boldsymbol{\Omega} = \boldsymbol{\Delta}_F \boldsymbol{\Lambda}' \left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F^3\boldsymbol{\Lambda}'\right]^{-1} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F - \boldsymbol{S}' \left[\boldsymbol{S}\boldsymbol{\Delta}_F\boldsymbol{S}'\right]^{-1} \boldsymbol{S}$$

where $\boldsymbol{S}$ is a $M \times K$ $(M < K)$ matrix with ones on the main diagonal and zeros elsewhere. Then the three pass regression filter forecast of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and any set of M proxies $\boldsymbol{Z}$ is (weakly) asymptotically more efficient than the forecast using the first M principal components of $\boldsymbol{X}$ if and only if $\boldsymbol{\Omega}$ is positive (semi-)definite.

We find Proposition 4 to always be feasible because principal components themselves can always be used as proxies. Therefore, if a principal component estimates a relevant factor, we can use it as a proxy and the 3PRF achieves the same asymptotic performance as does PC by adding that principal component to the forecast. Once one reaches a principal component that is not relevant, the target can be used as a proxy. Of course, once we have spanned the entire proxy space both methods deliver the same performance.

2.3 Remarks

It is forecasting folk wisdom that parsimony is vital to out-of-sample forecast performance. This is why in many-predictor contexts naïve linear projection is avoided. The fact that regressors number near (or more than) observations means that the OLS estimates of the projection coefficients are poor if they even exist. We continue this discussion below, but it suffices to say here that this fact leads to Stock and Watson's (2002a) key insight: one requires a dimension reduction of the predictive information. They linearly project the target variable on the most important linear combinations of the predictors, where the *importance* is determined by an eigenvalue decomposition of the predictors' covariance matrix. Our difference from their technique is that we have measured *importance* by focusing on covariance between predictors and proxies that either act like the target or are exactly the target itself. This allows the 3PRF to capture predictive information in a more parsimonious manner than PC in most cases – this is the tenor of Propositions 2 and 3, and this property is concordant with forecasting folk wisdom.

At the very least, the 3PRF is assured to do as well or better than PC in cases that one can assure by wisely choosing proxies – this is the tenor of Propositions 1 and 4. Therefore, one interpretation of this theory is that the 3PRF does well when PC does well, and can do well when PC does not. We argue that this interpretation is broadly correct, and we find support by both simulation and empirical application in Sections 5 and 6.

Fama-MacBeth Fama and MacBeth (1973) propose a two-pass regression method to be used in risk-return tests. Following Fama's (1976) discussion, one first finds beta (covariance with the market return), squared beta, and residual volatility estimates for assets or portfolios. Second, one runs a cross-sectional regression at each time period. As Cochrane (2001) points out, these cross-sectional estimates are useful for testing purposes, especially when the initial time series estimations are run in the full sample. Nonetheless, Fama (1976) emphasizes the *predictive* nature of the original Fama-MacBeth (FM) procedure, looking for a stable relationship between hypothesized risk factors and realized returns.

The 3PRF has modified this original FM procedure. Whereas FM posited variables that represented predictive (risk) factors, the 3PRF allows the proxies to represent linearly independent combinations of the predictive factors. Whereas FM used the T cross-sectional regressions for the purposes of statistically-testing the estimated price of risk, the 3PRF uses these T cross-sectional regressions to project the value of these predictive factors in each time period. Finally, recognizing the errors-in-variables problem which was a main point of Fama and MacBeth's (1973) analysis, the 3PRF uses a third-stage regression to effectively rotate the second-stage predictive factors into the best possible forecast of the target variable.

While the intent of FM and the 3PRF differ, the latter's genesis lies in consideration of the former. As we move into further discussion of the 3PRF's features, we will move towards explicit consideration of established forecasting methods of which FM is not one. However, it would be amiss to neglect mention of the contribution that Fama and MacBeth (1973) makes on the current enterprise.

3 Least Squares Perspective

In order to better understand and gain intuition about the 3PRF and its performance, we show that it is a special restricted form of the Kalman filter (KF), which is known to be the least squares predictor of the approximate factor model for any fixed N. We then continue our comparison of the 3PRF to the principal components forecaster (PC).

To simplify the exposition, we suppose that $\beta_0, \boldsymbol{\phi}_0, \boldsymbol{\lambda}_0, \boldsymbol{\mu}$ are zero throughout this section – this is without loss of generality regarding the points we make. Further details supporting the statements made here are placed in the Appendix.

3.1 The 3PRF and the KF

We can express the 3PRF time t predictor of y_{t+1} as

$$\hat{y}_{t+1} = \bar{y} + \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\varsigma}}_t \tag{3}$$

$$= \bar{y} + \left(\begin{bmatrix} \hat{\boldsymbol{\varsigma}}_1 & \cdots & \hat{\boldsymbol{\varsigma}}_T \end{bmatrix} \begin{bmatrix} y_2 \\ \vdots \\ y_{T+1} \end{bmatrix} \right)' \times \left(\begin{bmatrix} \hat{\boldsymbol{\varsigma}}_1 & \cdots & \hat{\boldsymbol{\varsigma}}_T \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\varsigma}}_1' \\ \vdots \\ \hat{\boldsymbol{\varsigma}}_T' \end{bmatrix} \right)^{-1} \times \hat{\boldsymbol{\varsigma}}_t \tag{4}$$

where

$$\hat{\boldsymbol{\varsigma}}_t = \left(\hat{\boldsymbol{\Phi}}\hat{\boldsymbol{\Phi}}' \right)^{-1} \hat{\boldsymbol{\Phi}}\boldsymbol{x}_t \tag{5}$$

$$\hat{\boldsymbol{\Phi}} = \begin{pmatrix} \hat{\boldsymbol{\phi}}_1 & \cdots & \hat{\boldsymbol{\phi}}_N \end{pmatrix} \tag{6}$$

$$\hat{\boldsymbol{\phi}}_i = \left(\sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t' \right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{z}_t x_{i,t} \right). \tag{7}$$

The mathematical statements involved in (3)-(4) are straightforward. First, (3) states that the 3PRF's forecast is a *linear* predictor given by an estimated parameter vector $\hat{\boldsymbol{\beta}}$ times an estimated M-dimensional predictive factor $\hat{\boldsymbol{\varsigma}}_t$. Equations (5)-(7) show that $\hat{\boldsymbol{\varsigma}}_t$ is found as a regression of the predictors on the first-stage coefficients. The first two terms of (4) state that the parameter vector $\hat{\boldsymbol{\beta}}$ comes from OLS regression of y_{t+1} on $\hat{\boldsymbol{\varsigma}}_t$. The hallmark of the 3PRF's forecasts is exactly this second-stage predictive factor $\hat{\boldsymbol{\varsigma}}_t$ and its particular construction via (5)-(7).

The least squares predictor of this linear state space is provided by the Kalman filter (Maybeck 1979) and the system parameters are efficiently estimated by maximum likelihood (Hamilton 1994). The prediction of the augmented state vector $\boldsymbol{\Pi}_t = (\boldsymbol{F}_t', \boldsymbol{F}_{t-1}')'$ can be written

$$\boldsymbol{\Pi}_{t|t} = \left(\boldsymbol{\Pi}_{t|t-1} - \boldsymbol{K}_t \boldsymbol{\Upsilon}_{t|t-1} \right) + \boldsymbol{K}_t \boldsymbol{\Upsilon}_t. \tag{8}$$

for $\boldsymbol{\Upsilon}_t$ the vector of variables observed by time t. $\boldsymbol{\Upsilon}_t$ includes variables *that are known at time* t, which definitely include the predictors $\boldsymbol{x}_t$. To finish the vector we must define the "as observed at time t" vector of proxies $\tilde{\boldsymbol{z}}_t$: since some proxies that we've included in $\boldsymbol{z}_t$ are future realized variables, we must be careful to distinguish in order to make sure the information sets stay honest. We imagine the target variable at time t (y_t) is one of the elements of $\tilde{\boldsymbol{z}}_t$. The first term of (8) combines information both cross-sectionally and temporally, while the second term combines information only cross-sectionally. The Kalman gain can be written

$$\boldsymbol{K}_t = \left(\boldsymbol{P}_{t|t-1}^{-1} + \boldsymbol{\Psi}'\boldsymbol{R}^{-1}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}\boldsymbol{R}^{-1}. \tag{9}$$

$\boldsymbol{\Psi}$ is the matrix determining 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. Suppose that $\boldsymbol{\beta}'$ is the nonzero $K_f + K_g$ columns of the row of $\boldsymbol{\Psi}$ corresponding to the target variable y. Then the optimal linear predictor of y_{t+1} conditional on $\{\boldsymbol{\Upsilon}_t, \boldsymbol{\Upsilon}_{t-1}, \boldsymbol{\Upsilon}_{t-2}, \ldots\}$ is given by premultiplying $\boldsymbol{\beta}'$ to $\boldsymbol{\Pi}_{t|t}$.

Let us ignore the components that temporally pool information. This affects the parts of (8) and (9) that are conditioned on past information, setting $\boldsymbol{\Pi}_{t|t-1} = \boldsymbol{0}$ and $\boldsymbol{\Upsilon}_{t|t-1} = \boldsymbol{0}$, their unconditional means. 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. Therefore, constraining the KF's information set to no longer temporally pool information delivers

$$\boldsymbol{\Pi}_{t|t} = \left(\boldsymbol{\Psi}'\boldsymbol{R}^{-1}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}'\boldsymbol{R}^{-1}\boldsymbol{\Upsilon}_t \tag{10}$$

$$y_{t+1|t} = \boldsymbol{\beta}'\boldsymbol{F}_{t|t}. \tag{11}$$

Equations (10) and (11) give the restricted KF's prediction of y_{t+1} conditional on $\boldsymbol{\Upsilon}_t$.

Let us compare the restricted KF's (10)-(11) to the 3PRF's (3)-(7). To do so, we recall that Watson and Engle's (1983) discussion of the EM algorithm makes clear that the parameters' MLE can be obtained (upon convergence) from GLS regressions – more detail is in the Appendix.

The $\hat{\boldsymbol{\varsigma}}_t$ in (3) is comparable to the $\boldsymbol{\Pi}_{t|t}$ in (8). By (10), the factor estimate $\boldsymbol{\Pi}_{t|t}$ is a GLS regression of $\boldsymbol{\Upsilon}_t$ on $\boldsymbol{\Psi}$ using the observation equations' errors' covariance matrix $\boldsymbol{R}$ as the weighting matrix.[11] The maximum likelihood estimate of $\boldsymbol{\Psi}$ is the matrix of coefficients from a feasible GLS regression of the observable variables on true factor $\boldsymbol{F}$. Meanwhile, recall that $\hat{\boldsymbol{\varsigma}}_t$ is a OLS regression of the predictors $\boldsymbol{x}$ on $\hat{\boldsymbol{\Phi}}$.

From (6) and (7), we see that $\hat{\boldsymbol{\Phi}}$ is the matrix of coefficients from an OLS regression of the predictors on noisy linear combinations of the true factor $\boldsymbol{\Pi}$ as provided by the proxies

[11] At first glance a regression on coefficients may appear strange. Yet it is nothing more than the calculation to find a conditional expectation for normal random variables. Suppose $(\boldsymbol{A}', \boldsymbol{B}') \sim N(\boldsymbol{0}', \boldsymbol{\Sigma}')$ and suitably partition $\boldsymbol{\Sigma}$. If there exists nonzero covariance between $\boldsymbol{A}$ and $\boldsymbol{\Phi}$, then knowing one gives information about the other. Thus we can write $\mathbb{E}(\boldsymbol{A}|\boldsymbol{B})$ as $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{B}$: this translates information about $\boldsymbol{B}$ into a conditional expectation for $\boldsymbol{A}$. $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ gives us the optimal translation and operates like a regression on parameters of the overall covariance matrix.

$\boldsymbol{z}$. Next, since $\boldsymbol{\beta}'$ in (11) is a row in $\boldsymbol{\Psi}$, it is found as part of a feasible GLS regression. Meanwhile, $\hat{\boldsymbol{\beta}}$ in (3) comes from an OLS regression of the target variable on our second-stage predictive factors $\hat{\boldsymbol{\varsigma}}$.

Therefore, the 3PRF is an implementation of a restricted KF that eschews GLS for OLS. Crucially, the 3PRF retains the KF's cross-sectional signal extraction by least squares.

3.2 Restrictions and Time Effects

An important aspect of the 3PRF's performance comes from an estimated "time effect", the second-pass regression constant $\varsigma_{0,t}$. To see why, consider the following. First suppose we have proxied-for and spanned every factor in the data. In this case, the second stage constant converges in probability to zero for large N. This is because the only omitted stochastic terms are the idiosyncratic shocks and the proxy noise. Assumption 3 ensures that the former and the pair's product average to zero in the limit. Now consider what happens when we have proxied-for and spanned *only* relevant factors. In this case, the second stage constant will not generally be zero in the limit – it will be the sum of each irrelevant factor's time t realization multiplied by that irrelevant factor's average (across the cross section) loading. Hence, in the case that Proposition 1 holds, this second stage constant is unimportant and usually near zero anyway. In the case that Proposition 2 holds, this second stage constant is usually not zero, but contains the influence of irrelevant factors we'd like to ignore. Therefore, the estimated time effect contributes to the 3PRF's success, and allows the 3PRF to be somewhat robust to misspecification of the total factor space.

The standard KF is not usually written down with a time effect. Obviously, one can easily adapt the KF to estimate this constant for each t, but in practice this is rarely done since parameter estimation is difficult enough without the addition of T more parameters. Nonetheless, a time effect is valuable in providing robustness to misspecification, and we see that the 3PRF's second stage constant soaks up and discards the influence of irrelevant factors.[12]

The impact of the 3PRF's restrictions are ultimately judged by applied researchers along a number of dimensions. First: The restrictions allow the 3PRF to be computed nearly instantaneously and in closed form, without the need for time-consuming numerical methods that leave lingering doubts as the global nature of the optimum obtained. Whether or not

[12] As an example, in Kelly and Pruitt (2010) we discuss how an important low-frequency component of price-dividend ratios is driven by long-horizon expectations which appear quite persistent. Hence, when looking for short-horizon expectations from these value ratios, it is crucial to have a means of separating out these persistent but irrelevant factors. The second stage constant does just this.

this issue of speed and closed form is important is a judgment call for each application. Second: The proofs in Section 2 show that the information lost by forgoing temporal pooling is inconsequential, to first order, in the large N limit. Whether or not N is large enough to make the actual results similar to the limit theory is questionable: the simulation evidence to be presented below suggests N does not need to be too large for this constraint to seem loose. Third: The possibility of correlation in the cross-sectional regressions suggests efficiency gains from GLS. However, GLS can often be out-performed in application by OLS when the model is misspecified (see Thursby 1987) or the requisite weighting matrix is poorly estimated. Alternatively, viewing our problem instead as a forecast combination problem, a large body of literature surveyed by Timmermann (2006) suggests theoretically-optimal weighting schemes (similar to GLS logic) are out-performed by equal-weighting schemes (similar to OLS logic) in practice. The evidence presented below suggests that the efficiency lost from using OLS is insubstantial.[13]

3.3 PC

To gain more insight into the 3PRF and PC, we consider what each estimates as the i^{th} predictor's weight in the single predictive factor case. For simplicity, suppose we use the first principal component or the target-proxied 3PRF. One can show (Appendix C) that the i^{th} predictor receives weight $\hat{\alpha}_i^{PC}$ under PC and $\hat{\alpha}_i^{3PRF}$ under 3PRF:

$$\hat{\alpha}_i^{PC} = \boldsymbol{s}'\boldsymbol{l}_{1,i}e_1^{-1} \tag{12}$$

$$\hat{\alpha}_i^{3PRF} = s_i\left(\left[\boldsymbol{s}'\boldsymbol{\Sigma}_X\boldsymbol{s}\right]^{-1}\left[\boldsymbol{s}'\boldsymbol{s}\right]\right) \tag{13}$$

where s_i is the i^{th} element of $\boldsymbol{s}$, $\boldsymbol{s} = \boldsymbol{X}'\boldsymbol{y}$, and $\boldsymbol{l}_{1,i}$ is the i^{th} column of $\boldsymbol{L}_1 = \tilde{\boldsymbol{l}}_1\tilde{\boldsymbol{l}}_1'$ the outer product of the normalized eigenvector corresponding to the largest eigenvalue e_1 of $\boldsymbol{\Sigma}_X$. Note a few properties of these weights. One, PC and the 3PRF are similar in that for both of them the i^{th} predictor's weightalmost surely depends positively on the i^{th} predictor's with-target covariance (s_i). Two, whereas the PC allows i^{th} predictor's weight to depend on non-i^{th} predictor with-target covariances, in the 3PRF the i^{th} predictor's weight depends only on the i^{th} predictor's with-target covariance.

[13] As a side note, one can view the 3PRF as a procedure whose obsolescence will be brought on by advances in the available computing technology. The KF is difficult to use because numerical optimization of a high dimensional surface is problematic. Numerical optimization is problematic because we must use hill-climbers or genetic algorithms to obtain solutions in reasonable amounts of time. As soon as computing power allows for the feasibility of grid searches in these high dimensional cases, applied work should use the Kalman filter with parameters estimated by maximum likelihood. Until such a time, the 3PRF is quite useful.

To make sense of these properties, we consider the following question: *What is it about the many-predictor context that causes naïve linear projection to work poorly?* Recall, this is formed using OLS to estimate the necessary predictor weights:

$$\begin{aligned} \hat{\boldsymbol{\alpha}}^{LP} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{s} \\ \hat{y}_{t+1}^{LP} &= \boldsymbol{x}_t'\hat{\boldsymbol{\alpha}}^{LP} \end{aligned} \tag{14}$$

In the introduction, we suggested this linear projection's poor performance stems from N being close to T and the ensuing poor OLS estimates. But this is only a partial answer, because there is a case wherein the precision of (14) is unaffected by how close N is to T: in the razor's edge case that the predictors are orthogonal *in-sample* to one another.[14] We refer to this case as "razor's edge" because even when the predictors are orthogonal in population, it is a probability zero event that they are literally orthogonal *in-sample*. When the predictors are orthogonal *in-sample* it is easy to see that the i^{th} element of $\hat{\boldsymbol{\alpha}}$ estimated in (14) must be numerically identical to what one estimates in the bivariate regression

$$\begin{aligned} \hat{\alpha}_i &= \left(\sum_t x_{i,t}^2\right)^{-1}\left(\sum_t x_{i,t}y_{t+1}\right) \\ &= s_i\ddot{\sigma}_{i,i} \end{aligned} \tag{15}$$

We have written this out to make clear that in this regression there is only one regressor but T observations, so for large T this regression estimate becomes more precise. Therefore, in this razor's edge case the i^{th} predictor's weight is accurately estimated by OLS *and* the i^{th} predictor's weight depends only on the i^{th} predictor's with-target covariance.

We think that this suggests that *the many-predictor context causes naïve linear projection to work poorly due to the poor sample estimate of the information matrix.* Generally, when N is close to T, the sample estimate of the information matrix is very inaccurate. Only when the predictors are orthogonal *in-sample* is this sample estimate accurate, and when this is the case the matrix is diagonal. In other words, in the razor's edge case that LP is working well, it is weighting the i^{th} predictor using only s_i and no other with-target covariance. But with probability one the sample information matrix has nonzero off-diagonal entries, which means the LP's i^{th} predictor weight is a linear combination of *all* with-target covariances. We attribute the poor performance of naïve linear projection to exactly this poorly estimated linear combination. What the 3PRF does is restrict the i^{th} predictor's weight to depend only

[14] This statement might be obvious to the reader; nonetheless see Appendix C for exact details supporting this statement.

on s_i, which is like imposing zeros into the fore mentioned linear combination.

It is worth noting, what does PC do in the razor's edge case that predictors are orthogonal *in-sample*? It returns a forecast based *solely on the predictor with largest sample variance.*[15] That is, PC weights one predictor by a nonzero amount, and weights all other predictors by 0. Although all predictors help to forecast, the PC ignores every predictor except one.

Let us for a moment take a forecast combination perspective, where we have N forecasts given by $s_i x_{i,t}$. The 3PRF uses a forecast combination that is *equally-weighted*. As Timmermann (2006) notes, equal-weighted forecast combinations often show strong performance in practice even when they are theoretically dominated by more sophisticated weighting schemes. By choosing OLS for the second-stage, the target-proxied 3PRF suggests an equal-weighting forecast combination. Moreover, if we have standardized the predictors before forecasting (as PC suggests we do), LP itself *also* suggests an equal-weighting forecast combination and 3PRF and LP give the identical forecast. On the other hand, PC throws away useful information, which stems from the fact that its dimension reduction focuses on within-predictor covariance. This can lead to strong divergence between the 3PRF and PC in pathological cases. For instance, assume that there are no factors in the data, but the target variable is defined as a certain linear combination of the predictors plus unforecastable noise. The PC forecasts will show no predictability of the target on the basis of the factors, as it produces predictive factors by looking only at the within-predictor covariance. The 3PRF forecasts will pick up the predictability because it will identify this certain linear combination by look at with-target covariance.

Having analyzed a razor's edge case wherein predictors exhibit no factor structure, what happens in the case that the predictors have a strong factor structure? One can imagine the situation we have in mind by supposing that each predictor is basically one underlying variable plus a very small amount of noise. This implies that the predictors are nearly collinear.

In this case, the PC returns an eigenvector with essentially the same number in every element: therefore, the PC weight on the i^{th} predictor is virtually s_i scaled by a number that is the same across i (because in this case $s_i \approx s_j$ for all i, j). In other words, in the strong factor case PC suggests an equal-weighting of the N forecasts given by $s_i x_{i,t}$ – *exactly*

[15] Technically-speaking, the PC forecaster is used on predictors that have been standardized and therefore should all have the same variance of 1. Hence theoretically the within-predictor covariance matrix has a lone eigenvalue 1 with multiplicity N, and the PC forecast is not uniquely defined: any of the eigenvectors could be returned to find the first principal component, and each of these eigenvectors will be all 0s except for one element that is a 1. In practice, numerical rounding with lead the computer to uniquely solve the problem as one "standardized" predictor will usually have a largest variance due to rounding error. The point is: the eigenvector returned will be all 0s except for one element that is a 1.

what 3PRF does all along. Therefore in the strongest factor case the performance of PC and 3PRF will be almost indistinguishable.

In this strong factor case, naïve linear projection runs into significant problems. These flow from the poor sample estimate of the information matrix $\mathbf{\Sigma}_X^{-1}$, caused by the poor conditionedness of $\mathbf{\Sigma}_X$ stemming from $\boldsymbol{X}$'s near-collinearity. In a sense, PC gets around this problem by a change of basis, converting the inversion of $\mathbf{\Sigma}_X$ into an inversion of a scalar, which is just what 3PRF is doing.

4 Implementation

Small sample bias emerges from in-sample algorithms wherein the time $t \in [1, T]$ forecast of y_{t+1} nontrivially depends on y_{t+1} itself. The predictive OLS regression $y_{t+1}|x_t$ gives perhaps the simplest example: the coefficient estimate is $\sum_{\tau=1}^{T} x_\tau y_{\tau+1} / \sum_{\tau=1}^{T} x_\tau^2$ and when $T < \infty$ it is clear that y_{t+1} is a nontrivial part of the coefficient estimate, and therefore the prediction $x_t \times \sum_{\tau=1}^{T} x_\tau y_{\tau+1} / \sum_{\tau=1}^{T} x_\tau^2$. This phenomenon is well-known to bias up estimates of in-sample predictability.

We refer to our basic procedure as described in Table 1 as the *full information* version of the 3PRF. In this version, first-stage regressions use the full time series of data to estimate first stage coefficients. Second-stage predictive factors use only the predictors at time t and the first-stage coefficients. Finally, third-stage predictive regressions are run in-sample. In the full information version it is possible that we have small sample bias in our predictors since first-stage and third-stage coefficients are estimated on the full time series, and we assume that the target itself is a proxy.

As a first alternative to the full information procedure, we calculate a *no-peek* version of our filter that reduces the effect of small sample bias while sacrificing minimal precision in estimating first-stage coefficients. In each period t beginning with $t = 1$, we run first-stage time series regressions omitting observations for dates $\{t + 1, \ldots, t + h_{np}\}$. The number of observations dropped (h_{np}) can be chosen by the user according to the persistence of the target variable. We construct the date t observation for our predictor variable from the second-stage cross section regression of time t price-dividend ratios on first-stage estimates (which ignore observations $\{t+1, \ldots, t+h_{np}\}$). At $t+1$ we re-estimate first-stage coefficients dropping observations for dates $\{t+2, \ldots, t+h_{np}+1\}$, and calculate the date $t+1$ value of our forecasting variable from the second-stage regression, and so on. Once we have exhausted the time series, we run a single third-stage forecasting regression based on the no-peek predictor, which has been constructed to explicitly preclude any unduly favorable effect from small

Table 2: Implementation Summary

Version	**First Stage**	**Third Stage**
Full Information	Same for all t: estimated from full sample $\{1, ..., T\}$.	Same for all t: estimated from full sample $\{1, ..., T\}$.
no-peek	For each t: estimated from sample $\{1, ..., t, t + h_{np}, ..., T\}$.	Same for all t: estimated from full sample $\{1, ..., T\}$.
Recursive Out-of-sample	For each $t \geq t_R$: estimated from sample $\{1, ..., t\}$.	For each $t \geq t_R$: estimated from sample $\{1, ..., t\}$.

Notes: Summary of three-pass regression filter implementation schemes. In all cases, the second stage regresses the time t cross section of predictors on the first-stage coefficients for time t, and the time $t+1$ predictor is formed by multiplying the time t second-stage predictive factor by the time t third-stage coefficients.

samples. It is important to note that no-peek estimates sacrifice observations, which can non-negligibly increase sampling error. Ultimately, we cannot distinguish if differences in forecasting results between full information and no-peek estimates are due to changes in look-ahead bias or sampling variation.[16] Nonetheless, the third-stage coefficient nontrivially depends on each target realization, so small sample bias may still be present.

A second and more stark alternative to the full information filter is a pure out-of-sample analysis. The procedure we use is a standard recursive out-of-sample estimation scheme which has been well-studied in the literature and affords us "encompassing" tests for the statistical significance of out-of-sample performance (see, for example, Clark and McCracken 2001 and Goyal and Welch 2008). Beginning with a user-determined initial training sample endpoint $t = t_R$, 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 predictors at τ and first-stage coefficients (which are based on data $\{1, ..., t\}$). We then estimate the third-stage coefficient 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 result. This process is iterated forward each year until the entire time series has been exhausted.

For the reader's reference, we summarize the key characteristics of these three procedures

[16] More precisely, the no-peek procedure has two effects on the filter's estimates. It eliminates the effect of small-sample bias by omitting observations that are eventually forecasted from preliminary parameter estimation. This decreases the absolute correlation between our predictor variable and forecasting target, giving a cleaner measurement of forecasting power. This effect on the R^2 is useful since it clarifies assessments of our predictor's power. A costly side-effect is that the sample is shortened, thereby weakening the precision of our estimates and obscuring correlation between the predictor and target. This tends to also decrease the R^2, which is harmful to our ability to detect the predictive power. Unfortunately, both of these effects ultimately decreases the no-peek R^2 relative to the full information case, making it difficult to ascertain whether the effect is due to small sample bias or increased estimation noise.

in Table 2. The differences lie in the sample that is used to estimate the first- and third-stage coefficients. In the full information procedure, these first-stage coefficients are the same for every time t; in the no-peek and out-of-sample procedures, these first-stage coefficients are different for each t due to the varying estimation samples. For every procedure, the second stage is the same: A time t cross section regression of predictors on the first-stage coefficients. In all cases, the eventual predictor is formed the same way: the second-stage predictive factor for time t is multiplied by the third-stage coefficient. In the full information and no-peek procedures, these third-stage coefficients are the same for every time t; in the out-of-sample procedure these third-stage coefficients are different for each t.

For statistical inference, since the 3PRF resembles the Kalman filter and OLS is a special case of maximum likelihood, there is no reason for the asymptotic distribution to fail to be well-approximated by bootstrapping. For our in-sample procedures, we use the circular block bootstrap to ensure the time-dependency of the data is considered when bootstrapping. We have experimented with different block lengths and found little difference in our asset-pricing exercise results.

Proxies can be the target, a predictor, or any combination of these variables. The main idea of choosing a proxy is to identify the relevant factors which are driving predictable variation in the target. Therefore the target is an excellent proxy. Relevant principal components work well. Future drafts will present more discussion of proxy selection. At present, we arbitrarily choose predictors as additional proxies in the monte carlo exercise, use theory-suggested proxies in the asset-pricing exercise, and use only the target as a proxy in the macroeconomic aggregate exercise.

5 Simulation Evidence

Asymptotic arguments are important for understanding the approximate finite sample behavior of estimators. Just as important is the estimator's performance in actual finite samples. This is our mindset as we turn to monte carlo study.

Our simulation study adapts Stock and Watson's (2002a). Our main modification is to additionally consider cases where there exist irrelevant factors that drive variation in the predictors and proxies, but not the target variable.[17] To impose fairness between the 3PRF and PC, we make sure that all predictors as well as proxies are standardized – since we

[17] The current draft is incomplete insofar as we will in the future use at least 2000 draws for all considered specifications. Moreover, we are in process of considering GARCH in the idiosyncratic errors and time variation in the loadings – preliminary results suggest the message does not change under those specifications or for more draws

Table 3: Simulation Set-up

$\boldsymbol{x}_t = \boldsymbol{\phi}_0 + \kappa_{fg}\boldsymbol{\Phi}_1\boldsymbol{f}_t + \boldsymbol{\Phi}_2\boldsymbol{g}_t + \boldsymbol{\varepsilon}_t$	Cross Section is factor-driven: K_f relevant $\boldsymbol{f}$, K_g irrelevant $\boldsymbol{g}$. Loadings possibly time-varying. κ_{fg} controls weakness of relevant factors.
$\boldsymbol{z}_t = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}_1\boldsymbol{f}_t + \boldsymbol{\Lambda}_2\boldsymbol{g}_t + \boldsymbol{\omega}_t$	Proxies are factor-driven, by both relevant and irrelevant factors
$y_{t+1} = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{f}_t + \eta_{t+1} = \tilde{y}_t^{(1)}$	Target driven by relevant factor $\boldsymbol{f}$ plus unforecastable noise, is first proxy
$\boldsymbol{f}_t = A_f\boldsymbol{f}_{t-1} + \boldsymbol{u}_t^f$	Relevant factor evolution
$\boldsymbol{g}_t = A_g\boldsymbol{g}_{t-1} + \boldsymbol{u}_t^g$	Irrelevant factor evolution
$(1 - aL)\varepsilon_{i,t} = (1 + d^2)\nu_{i,t} + d\nu_{i+1,t} + d\nu_{i-1,t}$	Idiosyncracies: a controls serial correlation, d controls cross-correlation
$\boldsymbol{\varepsilon}_t, \boldsymbol{\omega}_t, \boldsymbol{u}_t^f, \boldsymbol{u}_t^g, \nu_{i,t}, \eta_t$ are i.i.d. $N(0,1)\forall i,j,t$.	

default to using the target as a proxy, this means our target variable is also standardized. Table 3 outlines the environment and the requisite notation used in describing the simulation specifications.

The in-sample procedure for PC is to calculate the principal components on the full data sample, then construct forecasts as calculated by a predictive regression of the target on the full-sample principal components. The in-sample 3PRF is the "no-peek" procedure with three observations dropped – see Section 4. The out-of-sample PC procedure reestimates the principal components on all available data as of time t, runs a predictive regression of the target (as known through time t) on these principal components, and then uses this second-stage predictive coefficient along with the estimated loadings (eigenvectors) to forecast the target. The out-of-sample 3PRF procedure is totally recursive – see Section 4.

Our first study imposes that there are only the K_f relevant factors driving the entire system, so that $K_g = 0$: this is shown in Table 4. We make the relevant factors persistent, turn off the idiosyncracies' cross- or serial-correlation, have constant loadings, and no GARCH. We let the number of predictors vary between 5 and 100, and the number of time series observations vary between 50 and 200. When both the 3PRF and PC estimate $k = K_f$ factors, their performance is quite comparable. However, when the 3PRF and PC use $M = 1$ factors when there are actually $K_f = 3$ factors driving the system, we see that the 3PRF quickly identifies the linear combination of these factors that is driving the target, especially when

Table 4: Only Relevant Factors, Varying T, N

% of Infeasible Best Possible R^2												
	In-Sample						Out-of-Sample					
K_f/M	1/1		3/1		3/3		1/1		3/1		3/3	
	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF
N=5,T=50	96.5	95.8	45.6	63.8	94.4	83.0	74.3	77.7	15.3	45.7	66.3	60.2
N=5,T=100	94.5	94.3	37.6	70.4	95.2	88.9	83.3	85.2	23.9	59.7	73.4	69.2
N=10,T=50	98.6	99.1	42.4	78.0	100.0	94.6	84.4	86.6	17.3	56.0	72.8	67.3
N=10,T=100	98.2	98.2	42.0	79.7	100.0	97.5	92.1	93.8	29.0	65.6	86.0	85.6
N=10,T=200	97.9	97.7	26.1	83.5	99.1	97.6	94.5	94.7	23.1	79.8	92.0	93.3
N=50,T=100	100.7	100.7	34.9	88.1	101.8	100.0	94.9	95.6	28.8	80.8	88.7	88.3
N=50,T=200	99.8	99.7	41.8	92.5	100.9	100.3	97.3	98.0	26.3	86.0	95.5	95.6
N=100,T=200	100.4	100.3	33.8	93.5	101.3	100.5	97.7	98.4	18.7	86.1	95.6	97.1

Notes: No irrelevant factors, $K_g = 0$. Factors persistent, $A_f = 0.9$. No idiosyncratic cross- or serial-correlation, $a = 0, d = 0$. Equal strength, $\kappa_{fg} = 1$. Median across 5000 simulations.

Table 5: Equal Number Relevant and Irrelevant Factors, Varying T, N

% of Infeasible Best Possible R^2												
	In-Sample						Out-of-Sample					
K_f/M	1/1		3/1		3/3		1/1		3/1		3/3	
	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF
N=5,T=50	68.4	80.1	–	–	–	–	21.8	55.0	–	–	–	–
N=5,T=100	61.8	84.4	–	–	–	–	44.3	68.2	–	–	–	–
N=10,T=50	64.1	87.4	33.6	63.0	84.1	81.9	47.8	65.4	-6.5	33.8	22.8	32.2
N=10,T=100	71.3	91.4	26.7	62.7	71.6	77.6	54.3	83.4	8.5	41.7	39.9	56.8
N=10,T=200	60.2	90.2	12.3	65.3	65.4	78.8	54.4	84.3	6.3	54.5	53.8	66.7
N=50,T=100	34.9	88.1	22.5	75.3	80.3	88.4	28.8	80.8	4.4	59.7	45.5	67.6
N=50,T=200	62.3	95.5	17.1	82.5	71.9	91.1	51.0	93.9	7.9	70.0	47.3	80.0
N=100,T=200	56.6	96.8	19.3	82.4	72.2	91.6	44.5	93.5	9.2	73.3	54.2	81.7

Notes: Factors persistent, $A_f, A_g = 0.9$. No idiosyncratic cross- or serial-correlation, $a = 0, d = 0$. Equal strength, $\kappa_{fg} = 1$. Median across 5000 simulations.

N is larger. For instance, when $N = 100, T = 200$ the first principal component achieves about 34 percent of the infeasible best possible forecast in-sample, whereas the first 3PRF factor achieves almost 94 percent.

Our second study in Table 5 allows irrelevant factors to exist, and for convenience sets $K_f = K_g$. Other specifications are the same as before, the irrelevant factors also have a persistence of 0.9, and we investigate different values for N and T. There are notable differences between the PC and 3PRF when both estimate the correct number of relevant factors. When $K_f = 1$, the PC achieves 35–71 percent of the infeasible best in-sample, while

Table 6: Varying Factor Persistence

% of Infeasible Best Possible R^2

	In-Sample						Out-of-Sample					
K_f/M	1/1		3/1		3/3		1/1		3/1		3/3	
	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF
					N=10,T=50							
$A_f = 0.9$	98.6	99.1	42.4	78.0	100.0	94.6	84.4	86.6	17.3	56.0	72.8	67.3
$A_f = 0.3$	92.6	92.5	25.4	69.1	94.8	87.3	84.1	81.9	11.8	51.3	71.8	64.2
$A_f = 0.9, A_g = 0.9$	64.1	87.4	33.6	63.0	84.1	81.9	47.8	65.4	-6.5	33.8	22.8	32.2
$A_f = 0.9, A_g = 0.3$	94.7	95.8	30.3	68.9	85.6	86.8	74.3	76.9	-0.4	36.4	45.9	54.8
$A_f = 0.3, A_g = 0.9$	9.0	62.3	6.1	29.1	42.3	48.4	-2.2	52.5	-9.6	14.5	8.4	22.1
$A_f = 0.3, A_g = 0.3$	43.5	82.0	13.1	48.0	52.0	64.7	40.3	73.1	-1.8	34.6	23.3	39.7
					N=100,T=200							
$A_f = 0.9$	100.4	100.3	33.8	93.5	101.3	100.5	97.7	98.4	18.7	86.1	95.6	97.1
$A_f = 0.3$	99.5	99.5	35.6	95.2	100.4	99.5	97.6	98.0	26.2	91.9	95.7	95.4
$A_f = 0.9, A_g = 0.9$	56.6	96.8	19.3	82.4	72.2	91.6	44.5	93.5	9.2	73.3	54.2	81.7
$A_f = 0.9, A_g = 0.3$	99.8	99.8	37.3	92.0	99.8	99.8	96.9	97.9	23.1	85.5	93.0	94.3
$A_f = 0.3, A_g = 0.9$	0.7	88.3	0.9	57.8	5.9	71.3	-3.2	83.6	-2.9	57.2	-3.7	71.1
$A_f = 0.3, A_g = 0.3$	56.6	97.5	11.5	90.4	61.8	95.2	55.8	95.6	6.2	87.8	52.0	92.2

Notes: No idiosyncratic cross- or serial-correlation, $a = 0, d = 0$. Equal strength, $\kappa_{fg} = 1$. Median across 100 simulations.

3PRF achieves 80–97 percent. When $K_f = 3$, the PC achieves 65–84 percent and the 3PRF achieves 78–92 percent, in-sample. More striking is the out-of-sample performance, where the $N = 100, T = 200$ is somewhat representative: PC achieves 9 percent when $k = 1$ and 54 percent when $M = 3$, whereas 3PRF achieves 73 percent and 82 percent, respectively.

The third study in Table 6 allows the relevant and irrelevant factors to show different persistence, either 0.3 or 0.9, and hereafter we focus on the $N = 10, T = 50$ and $N = 100, T = 200$ cases. In general, PC's performance is hurt when the relevant factors are not very persistent while there are irrelevant factors in the data. For instance, if relevant and irrelevant factors have low persistence (0.3), PC achieves 57 to 62 percent of the infeasible best in-sample when estimating the correct number of relevant factors; on the other hand, 3PRF achieves 98 to 95 percent in those same circumstances. If the irrelevant factors are more persistent, PC achieves about 1 percent in-sample and -3 percent out-of-sample, whereas 3PRF achieves 71-88 percent in-sample and 72-84 percent out-of-sample.

In the interest of exposition, we relegate further simulation study results to Appendix D. The main message stays roughly the same across various specifications: the 3PRF effectively extracts information when confounding influences (stronger irrelevant factors, correlated idiosyncratic errors) exist, while PC's performance suffers in those situations. Often, the single factor forecaster of the 3PRF (coming from using only the target variable as a proxy) works surprisingly well, sometimes better than multifactor forecaster coming from PC. As Section 3 has argued, the 3PRF is making better use of the available information than PC when the data fail to adhere to a strong, single factor representation – that is certainly evident in the simulation studies we have presented.

With these lessons in mind, we now turn to two empirical applications.

6 Empirical Results

Our empirical applications have to do with forecasting in either a financial or a macroeconomic context. The first application follows Kelly and Pruitt (2010a) in considering the predictive content contained in the price-dividend ratios of Size and Book-to-Market disaggregated portfolios of CRSP stocks. We quickly sketch out the requisite theory and present some results. The second application follows from Stock and Watson's (2002a,2006,2009) studies involving a well-known set of macroeconomic indicators made publicly available by the authors. To maintain comparability to Stock and Watson's most recent study, the second out-of-sample forecasting exercise follows from their cross-validation-type procedure.

6.1 Asset Pricing

We follow Kelly and Pruitt (2010a) where more details are presented, and so the cross-sectional present value system is constructed as follows: Using the Campbell and Shiller (1988) approximation, an asset's log price-dividend ratio at time t is linearly related to the conditional expectations of future return and dividend growth:

$$pd_{i,t} \quad = \quad \frac{\kappa_i}{1-\rho_i} + \sum_{j=0}^{\infty} \rho_i^j \mathbb{E}_t(-r_{i,t+j+1} + \Delta d_{i,t+j+1}). \tag{16}$$

We assume that expected returns and dividend growth are linear in common factors $\boldsymbol{f}_t$ $(K \times 1)$ across assets and over all horizons

$$\begin{aligned} \mathbb{E}_t[r_{i,t+j}] \quad &= \quad {}_j a_{i,0,r} + {}_j \boldsymbol{a}'_{i,r} \boldsymbol{F}_t \\ \mathbb{E}_t[\Delta d_{i,t+j}] \quad &= \quad {}_j a_{i,0,d} + {}_j \boldsymbol{a}'_{i,d} \boldsymbol{F}_t + {}_j \varepsilon_{i,t}. \end{aligned}$$

Table 7: Stock Market Predictability

	In-Sample			Out-of-Sample		
	LP	PCR	3PRF	LP	PCR	3PRF
			6 PORTFOLIOS			
R^2 (%)	41.03	20.89	31.35	10.57	6.53	29.58
Test	*22.76*	*10.31*	*10.51*	*14.53*	*3.21*	*8.18*
p-val	*< 0.001*	*< 0.001*	*< 0.001*	*< 0.01*	*< 0.05*	−−
			12 PORTFOLIOS			
R^2 (%)	46.11	12.90	26.43	11.32	−5.63	24.35
Test	*19.99*	*5.65*	*9.84*	*12.69*	*1.13*	*5.63*
p-val	*< 0.001*	*0.002*	*< 0.001*	*< 0.01*	−−	−−
			25 PORTFOLIOS			
R^2 (%)	65.34	15.89	29.95	−79.37	−4.60	22.60
Test	*19.55*	*5.73*	*9.50*	−−	*2.74*	*6.27*
p-val	*< 0.001*	*0.002*	*< 0.001*	−−	*< 0.05*	−−

Notes: Annual CRSP data, 1946–2009; Out-of-sample forecasts begin 1981. In-sample *Test* is F-test calculated by circular block bootstrap. Out-of-sample *Test* is Clark and McCracken's (2001) forecast encompassing test against historical mean null forecast. Using 25 Size/Book-to-Market Portfolios. LP stands for naïve linear projection of returns on all portfolios' price-dividend ratios, with coefficients estimated by OLS. For PC and 3PRF, 3 factor forecasters: the first three principal components for PC; the aggregate (future) return, aggregate (future) dividend growth, and aggregate contemporaneous price-dividend ratio as proxies for 3PRF

Substituting the above into (16), we obtain

$$pd_{i,t} = \phi_{i,0} + \boldsymbol{\phi}_i' \boldsymbol{F}_t + \varepsilon_{i,t}. \tag{17}$$

Meanwhile, we find proxies by noting that by definition

$$r_{t+1} = \mathbb{E}_t(r_{t+1}) + \omega_{r,t+1} = \beta_{0,r} + \boldsymbol{\beta}_r' \boldsymbol{F}_t + \omega_{r,t+1} \tag{18}$$

$$\Delta d_{t+1} = \mathbb{E}_t(\Delta d_{t+1}) + \omega_{d,t+1} = \lambda_{0,d} + \boldsymbol{\lambda}_d \boldsymbol{F}_t + \omega_{d,t+1} \tag{19}$$

$$pd_t = \lambda_0 + \boldsymbol{\lambda}' \boldsymbol{F}_t \tag{20}$$

for the aggregate variables $r_{t+1}, \Delta d_{t+1}, pd_t$. We use these as our three proxies for the purposes of forecasting the market return r_{t+1}. Our data include the entire CRSP universe of stocks – see Kelly and Pruitt (2010a) for more details.

Table 7 shows the results of forecasting aggregate market returns using LP, PC, and 3PRF, from a cross section of six, twelve, or twenty-five portfolios sorted by Size and Book-to-Market characteristics. LP displays the obvious signs of overfit – a notable distinction between in-sample and out-of-sample predictability. This is specially evident when we use twenty-five portfolios on our sixty-four annual data points as the in-sample R^2 is 65% but the out-of-sample R^2 is *negative* 79%.[18] Turning to PC, it appears the procedure can miss the stable population relationship. For six portfolios, the in-sample R^2 is 21% and the out-of-sample R^2 is 7%, which is strong: however, as we consider further levels of disaggregation, the in-sample R^2 drops to 16% and the out-of-sample R^2 falls to -5%. Contrast this with the results for the 3PRF: the in-sample R^2 ranges from 26 to 31% while the out-of-sample R^2 ranges between 23 to 30%. The strength and similarity of the in-sample and out-of-sample predictability are suggestive of the 3PRF's precise estimation of the underlying risk factors driving predictable market returns. From a pure forecasting perspective, the 3PRF results in a stable and strong forecaster regardless of the size of the cross section.

6.2 Macroeconomic Forecasting

Following Stock and Watson (2009), we perform a cross-validation-type out-of-sample forecasting procedure, which aims to describe out-of-sample forecasting performance over the whole sample instead of just the most recent portion. We target key macroeconomic aggregates, and use the well-known cross section of $N = 134$ quarterly predictors running 1959:I–2009:I as collected and transformed by Stock and Watson.

The forecasting exercise is as follows: First, we *partial* out all variables (target and predictors) by projecting them on four lags of the target variable. Hence, we aim to forecast the part of the target that is left unexplained by four lags of itself; to do so, we use the parts of the predictors that are *also* left unexplained by four lags of the target. As mentioned in Section 2.2, this is an example of how predictive factor can be combined with the target's own lags to capture dynamic predictive information. Second, we conduct a pseudo out-of-sample forecasting procedure that does not overly reflect the greater- or lesser predictability of macroeconomic variables in the second half of the sample. To do this, we aim to forecast each observation of the sample using all data outside a window surrounding that forecasting time period. That is, for each $t \in \{1, \ldots, T\}$:

1 Let $\mathrm{T}_{-t} = \{1, \ldots, t-4, t+4, \ldots, T\}$ denote the T_{-t}-subsample index set for use with predictors and proxies – the T_{-t}-subsample index set for use with the target is naturally $\{2, \ldots, t-3, t+5, \ldots, T+1\}$. The T_{-t}-subsample does not include t and nearby observations

[18] Recall, the out-of-sample R^2 takes the range $(-\infty, 1)$ – see Harvey (1987).

2 PC: Find PCs for $\boldsymbol{X}$ on the T_{-t}-subsample

2 3PRF: Run first-stage regressions of $\mathbf{x}_i$ on $\boldsymbol{y}$ on the T_{-t}-subsample, for each $i = 1, 2, \ldots, N$; run second-stage regressions using these first-stage coefficients to get second-stage predictive factors

3 PC: Run second-stage predictive regression on the T_{-t}-subsample to get the second-stage coefficients

3 3PRF: Run third-stage predictive regression on second-stage predictive factors on the T_{-t}-subsample

4 PC: Construct the time t pseudo out-of-sample forecast using loadings from **2 PC**, second-stage coefficients from **3 PC**, using $\boldsymbol{x}_t$

4 3PRF: Construct the time t out-of-sample forecast by regressing $\boldsymbol{x}_t$ on the first-stage coefficients from **2 3PRF** to get the second-stage predictive factor, then multiply by the third-stage predictive coefficients from **3 3PRF**

Our modification of the PC procedure from Stock and Watson's (2009) exercise is in step **2 PC**. Whereas they estimate the PCs just once on $\{1, 2, \ldots, T\}$, we re-estimate the PCs on T_{-t} for each t: this makes certain that both PC and 3PRF use the same information set in forecasting the target at time t, and that this information set in no way includes information from time t or nearby periods.[19] For the 3PRF, we use a single proxy – the target variable itself.[20]

The results of the macroeconomic forecasting exercise are reported in Table 8. Following Stock and Watson (2009), we focus on the Root Mean Squared Error (RMSE) relative to the RMSE of the residual coming from just the AR(4): this means that percentages less than 100 are desirable and indicate that the average out-of-sample residual from an AR(4) forecast augmented with a cross-sectional predictive factor or cross-sectional principal component is smaller than the average out-of-sample residual from the AR(4) alone.[21] Without exception, the 3PRF's single predictive factor leads to a larger forecast improvement than does the first principal component. The 3PRF leads noticeable forecast improvements for all variables, notably including 22% for Output (PC: 18%), 15% for Consumption (PC: 5%), 10% for Labor

[19] We abstract from the shrinkage methods explored in Stock and Watson (2009) – these may constitute an interesting and readily accessible extension of the 3PRF, left to future research.

[20] Note that this application of the 3PRF resembles the no-peek version described in Section 4 *except* that the third-stage regression is also run on the sample that leaves out observations near the targeted time period.

[21] Because this is an out-of-sample procedure, percentages less than 100 are not assured.

Table 8: Macro Forecasting

	RMSE Relative to AR(4) Only		
	%		
	Single Factor		
Variable	PC	3PRF	PC – 5 Factors
Output	82.3	78.3	76.9
Consumption	95.2	84.9	87.2
Investment	87.0	80.7	79.3
Imports	88.8	83.6	87.6
Exports	92.0	89.6	91.1
Labor Productivity	95.5	90.0	90.3
Hours	83.7	79.0	79.2
Core PCE	95.4	92.1	95.4
Industrial Production	95.5	93.6	94.4
Capacity Utilization	88.2	83.5	86.0
Employment	94.6	91.9	94.5
Average Hours Worked	96.0	93.4	96.1
Unemployment Rate	95.8	87.5	96.2
Housing Starts	94.1	90.7	92.9

Notes: Data from Stock and Watson (2009), 134 macroeconomic predictors, cross-validation-type out-of-sample forecasting procedure from 1959:I–2009:I.

Productivity (PC: 5%), 8% for Core PCE Inflation (PC: 5%), and 13% for the Unemployment Rate (PC: 4%). Moreover, for most variables the 3PRF's *single* predictive factor outpaces the forecast improvement coming from the first *five* principal components put together:[22] Consumption, Imports, Exports, Labor Productivity, Hours, Core PCE Inflation, Industrial Production, Capacity Utilization, Employment, Average Hours Worked, the Unemployment Rate, and Housing Starts.

These results bear similarity to the simulation study results of Section 5. There monte carlo evidence showed strong performance of the single-proxy 3PRF even when multiple factors existed in the data. One might suppose macroeconomic data is driven by multiple factors. Table 8 shows that the 3PRF effectively extracts information from the factors that are relevant for forecasting each target variable, achieving a blend of parsimony and strong performance.

[22] This is Stock and Watson's (2009) benchmark.

7 Conclusion

This paper has introduced a new statistical technique called the Three Pass Regression Filter (3PRF) for use in applications where a target variable is forecast using a cross section of many predictors. The key to effective forecasting in the many-predictor environment is a dimension-reduction step which proceeds by assuming an approximate factor structure for the data. We show that the 3PRF consistently provides the infeasible best forecast, constructed from knowledge of the true latent factors, for large N and T. This asymptotic efficiency is obtained under a variety of relevant conditions.

We show that the 3PRF reduces the dimension of predictive information via least squares because it is a special restricted case of the Kalman filter. The restrictions imposed on the 3PRF enable it to be estimated in closed form and therefore virtually instantaneously for arbitrarily large N and T, which is a major advantage over the numerical methods that render the Kalman filter practically impossible to estimate in these contexts. We also compare the 3PRF to forecasting via principal components and find that the latter's focus on within-predictor covariation allows the 3PRF to provide superior performance in many cases. Simulation evidence is favorable for the 3PRF, and it seems to quickly and stably identify the relevant predictive content in the cross section. We compare the 3PRF to principal components in forecasting annual aggregate stock market returns and forecasting quarterly macroeconomic variables and find that the 3PRF is a more powerful forecasting method.

References

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.

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

Bernanke, B., J. Boivin, and P. S. Eliasz (2005): "Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) Approach," *The 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.

Clark, T., and M. McCracken (2001): "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105(1), 85–110.

Cochrane, J. (2005): *Asset Pricing*. Princeton University Press, second edn.

de Jong, S., and H. A. L. Kiers (1992): "Principal covariates regression: Part I. Theory," *Chemometrics and Intelligent Laboratory Systems*, 14, 155–164.

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. F. (1976): *Foundations of Finance*. New York: Basic Books.

Fama, E. F., and J. D. MacBeth (1973): "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81(3), 607–36.

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.

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.

Heij, C., P. J. Groenen, and D. van Dijk (2007): "Forecast comparison of principal component regression and principal covariate regression," *Computational Statistics & Data Analysis*, 51(7), 3612–3625.

Horn, R. A., and C. R. Johnson (1985): *Matrix Analysis.* New York: Cambridge University Press.

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

Kelly, B. T., and S. J. Pruitt (2010): "Market Expectations and 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," *The 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.

Onatski, A. (2009): "Testing Hypotheses About the Number of Factors in Large Factor Models," *Econometrica*, 77(5), 1447–1479.

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.

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.

——— (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*vol. 1 of *Handbook of Economic Forecasting*, chap. 10, pp. 515–554. Elsevier.

——— (2009): “Generalized Shrinkage Methods for Forecasting Using Many Predictors,” Working papers, Princeton University.

Thursby, J. G. (1987): “OLS or GLS in the presence of specification error? : An expected loss approach,” *Journal of Econometrics*, 35(2-3), 359–374.

Timmermann, A. (2006): *Forecast Combinations*vol. 1 of *Handbook of Economic Forecasting*, chap. 4, pp. 135–196. Elsevier.

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.

White, H. (2001): *Asymptotic Theory for Econometricians.* Academic Press, second edn.

A Proofs

Our gameplan is as follows: Lemma 1 is available without assumption. Using Assumptions 1, 2, and 3 we are able to prove the remaining lemmas, which ultimately deliver the probability limit of the 3PRF. This means that the 3PRF's plim is proven using only Assumptions 1, 2, and 3. Therefore, the following propositions and corollaries are able to use various combinations of the remaining assumptions to make their statements, taking as given the probability limit shown in Lemma 4.

For convenience, we repeat and introduce the necessary notation here:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_T' \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y_2 \\ y_3 \\ \vdots \\ y_{T+1} \end{bmatrix} \qquad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1' \\ \boldsymbol{z}_2' \\ \vdots \\ \boldsymbol{z}_T' \end{bmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1' \\ \boldsymbol{\varepsilon}_2' \\ \vdots \\ \boldsymbol{\varepsilon}_T' \end{bmatrix} \qquad \boldsymbol{\eta} = \begin{bmatrix} \eta_2 \\ \eta_3 \\ \vdots \\ \eta_{T+1} \end{bmatrix} \qquad \boldsymbol{\omega} = \begin{bmatrix} \boldsymbol{\omega}_1' \\ \boldsymbol{\omega}_2' \\ \vdots \\ \boldsymbol{\omega}_T' \end{bmatrix}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 \end{bmatrix} \qquad \boldsymbol{F}_t = \begin{bmatrix} \boldsymbol{f}_t' & \boldsymbol{g}_t' \end{bmatrix}' \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{\Lambda}_2 \end{bmatrix}$$

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

$$\boldsymbol{X} = \boldsymbol{\iota}\boldsymbol{\phi}_0' + \boldsymbol{F}\boldsymbol{\Phi}' + \boldsymbol{\varepsilon} \qquad \boldsymbol{y} = \boldsymbol{\iota}\beta_0 + \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\eta} \qquad \boldsymbol{Z} = \boldsymbol{\iota}\boldsymbol{\lambda}_0' + \boldsymbol{F}\boldsymbol{\Lambda}' + \boldsymbol{\omega}$$

Let $\boldsymbol{F} \equiv [\ \boldsymbol{f} \quad \boldsymbol{g}\ ]$. Moreover, let $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_T)'$ also be written $\boldsymbol{\varepsilon} = (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_N)$: that is, $\boldsymbol{\varepsilon}_t$ is the time t cross section of idiosyncracies whereas $\boldsymbol{\epsilon}_i$ is the i^{th} predictor's time series of idiosyncratic shocks. Throughout, $\boldsymbol{J}_L \equiv \boldsymbol{I}_L - L^{-1}\boldsymbol{\iota}_L\boldsymbol{\iota}_L'$, $\boldsymbol{I}_L$ is the L-dimensional identity matrix and $\boldsymbol{\iota}_L$ is a L-vector of ones – we omit L when it is clear.

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

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

$$\boldsymbol{X} = \boldsymbol{\iota}\boldsymbol{\phi}_0' + \boldsymbol{F}\boldsymbol{\Phi}' + \boldsymbol{\varepsilon} \qquad \boldsymbol{y} = \boldsymbol{\iota}\beta_0 + \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\eta} \qquad \boldsymbol{Z} = \boldsymbol{\iota}\boldsymbol{\lambda}_0' + \boldsymbol{F}\boldsymbol{\Lambda}' + \boldsymbol{\omega}$$

where $\boldsymbol{F}_t = (\ \boldsymbol{f}_t' \quad \boldsymbol{g}_t'\ )'$, $\boldsymbol{\Phi} = (\ \boldsymbol{\Phi}_1 \quad \boldsymbol{\Phi}_2\ )$, $\boldsymbol{\Lambda} = (\ \boldsymbol{\Lambda}_1 \quad \boldsymbol{\Lambda}_2\ )$, *and* $\boldsymbol{\beta} = (\ \boldsymbol{\beta}_1' \quad \boldsymbol{0}'\ )'$.

$K_f > 0$ *is the dimension of the vector* $\boldsymbol{f}_t$*, and the column dimension of* $\boldsymbol{\Phi}_1, \boldsymbol{\Lambda}_1$ *and the row dimension of* $\boldsymbol{\beta}_1$*.*

$K_g \geq 0$ *is the dimension of the vector* $\boldsymbol{g}_t$*, and the column dimension of* $\boldsymbol{\Phi}_2, \boldsymbol{\Lambda}_2$ *– when* $K_g = 0$*,* $\boldsymbol{g}_t, \boldsymbol{\Phi}_2, \boldsymbol{\Lambda}_2$ *disappear.*

$M > 0$ *is the dimension of the vector* $\boldsymbol{z}_t$*, and the row dimension of* $\boldsymbol{\Lambda}$*.*

Let $K = K_f + K_g$*.*

The unforecastable shock is such that $\mathbb{E}_t(\eta_{t+1}) = 0, \mathbb{E}_t(\eta_{t+1}^2) = \delta_\eta < \infty$ *for all* t*. Hence, the "infeasible best" forecast of* y_{t+1} *conditional on time* t *information is given by* $\beta_0 + \boldsymbol{\beta}'\boldsymbol{F}_t = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{f}_t$*.*

Assumption 2 (Factors and Loadings)**.** *For* $\tilde{\boldsymbol{\phi}}_i \equiv (\phi_{i,0}, \boldsymbol{\phi}_i')'$ *and* $\bar{\phi} < \infty$

1. $T^{-1}\sum_{t=1}^{T}\boldsymbol{F}_t \xrightarrow[T\to\infty]{p} \boldsymbol{\mu}$ *and* $T^{-1}\sum_{t=1}^{T}\boldsymbol{F}_t\boldsymbol{J}_T\boldsymbol{F}_t' \xrightarrow[T\to\infty]{p} \boldsymbol{\Delta}_F$.
2. $|\tilde{\boldsymbol{\phi}}_i| \le \bar{\phi}\ \forall i$.
3. $N^{-1}\sum_{i=1}^{N}\tilde{\boldsymbol{\phi}}_i\boldsymbol{J}_N\tilde{\boldsymbol{\phi}}_i' \xrightarrow[N\to\infty]{p} \begin{bmatrix} B_0 & \boldsymbol{B}_1' \\ \boldsymbol{B}_1 & \boldsymbol{\mathcal{B}} \end{bmatrix}$ *with* $\boldsymbol{\mathcal{B}}$ *nonsingular.*

Assumption 3 (Error Moments)**.** *There exists a constant $A < \infty$ such that*

1. $T^{-1}\sum_{t=1}^{T}\varepsilon_{i,t} \xrightarrow[T\to\infty]{p} \mathbf{0}\ \forall i$ *and* $N^{-1}\sum_{i=1}^{N}\varepsilon_{i,t} \xrightarrow[N\to\infty]{p} \mathbf{0}\ \forall t$
2. $T^{-1}\sum_{t=1}^{T}\boldsymbol{\omega}_t \xrightarrow[T\to\infty]{p} \mathbf{0}$ *and* $T^{-1}\sum_{t=1}^{T}\boldsymbol{\eta}_t \xrightarrow[T\to\infty]{p} \mathbf{0}$.
3. $T^{-1}\sum_{t=1}^{T}\boldsymbol{\varepsilon}_t\boldsymbol{\eta}_t' \xrightarrow[T\to\infty]{p} \mathbf{0}, T^{-1}\sum_{t=1}^{T}\boldsymbol{F}_t\boldsymbol{\eta}_t' \xrightarrow[T\to\infty]{p} \mathbf{0}$ *and* $T^{-1}\sum_{t=1}^{T}\boldsymbol{F}_t\boldsymbol{\omega}_t' \xrightarrow[T\to\infty]{p} \mathbf{0}$.
4. $T^{-1}\sum_{t=1}^{T}\varepsilon_{i,t}\omega_{k,t} \xrightarrow[T\to\infty]{p} \gamma(i,k),$ *and* $\lim_{N\to\infty}\sup_k\sum_{i=1}^{N}|\gamma(i,k)| \le A$.
5. $T^{-1}\sum_{t=1}^{T}\varepsilon_{i,t}\varepsilon_{j,t} \xrightarrow[T\to\infty]{p} \delta(i,j) = \delta(j,i),$ *and* $\lim_{N\to\infty}\sup_j\sum_{i=1}^{N}|\delta(i,j)| \le A$.
6. $N^{-1}\sum_{i=1}^{N}\varepsilon_{i,t}\varepsilon_{i,s} \xrightarrow[N\to\infty]{p} \kappa(t,s) = \kappa(s,t),$ *and* $\lim_{T\to\infty}\sup_s\sum_{t=1}^{T}|\kappa(s,t)| \le A$.

Assumption 4 (Rank Condition)**.** *The matrix $\boldsymbol{\Lambda}$ is nonsingular.*

Assumption 5 (Alternative Rank Condition)**.** $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \mathbf{0})$ *and $\boldsymbol{\Lambda}_1$ is nonsingular.*

Assumption 6 (Normalization)**.** $\boldsymbol{\mathcal{B}} = \boldsymbol{I}$ *and $\boldsymbol{\Delta}_F$ is diagonal.*

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

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

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

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

The second stage regression is

$$\boldsymbol{X} = \boldsymbol{\iota}\ddot{\boldsymbol{\Phi}}_0 + \ddot{\boldsymbol{F}}\hat{\tilde{\boldsymbol{\Phi}}}' + \ddot{\boldsymbol{\epsilon}}$$

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

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

The third stage regression is

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

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

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

and

$$\boldsymbol{\iota}\hat{\breve{\beta}}_0 = T^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\left(\boldsymbol{y}-\hat{\ddot{\boldsymbol{F}}}\hat{\breve{\boldsymbol{\beta}}}\right) = \boldsymbol{\iota}\bar{y} - T^{-1}\boldsymbol{\iota}\boldsymbol{\iota}'\hat{\ddot{\boldsymbol{F}}}\hat{\breve{\boldsymbol{\beta}}}$$

with corresponding Y forecast

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

QED

Lemma 2. *Let Assumptions 1, 2, and 3 hold. Define the matrices $\boldsymbol{\mathcal{K}} = [\kappa(t,s)]$ in position (t,s), $\boldsymbol{\Delta}_\varepsilon = [\delta(i,j)]$ in position (i,j) and $\boldsymbol{\Gamma} = [\gamma(i,k)]$ in position (i,k). Then the following are true:*

1. $N^{-1}\boldsymbol{\varepsilon}\boldsymbol{\phi}_0 \xrightarrow[N\to\infty]{p} \mathbf{0}$
2. $N^{-1}\boldsymbol{\varepsilon}\boldsymbol{\Phi}' \xrightarrow[N\to\infty]{p} \mathbf{0}$
3. $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Delta}_\varepsilon \xrightarrow[N\to\infty]{p} \mathbf{0}$
4. $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Gamma} \xrightarrow[N\to\infty]{p} \mathbf{0}$
5. $N^{-1}\boldsymbol{\Gamma}'\boldsymbol{J}_N\boldsymbol{\Delta}_\varepsilon \xrightarrow[N\to\infty]{p} \mathbf{0}$
6. $T^{-1}\boldsymbol{\iota}'\boldsymbol{\mathcal{K}} \xrightarrow[T\to\infty]{p} \mathbf{0}$
7. $T^{-1}\boldsymbol{\omega}'\boldsymbol{\mathcal{K}} \xrightarrow[T\to\infty]{p} \mathbf{0}$
8. $T^{-1}\boldsymbol{F}'\boldsymbol{\mathcal{K}} \xrightarrow[T\to\infty]{p} \mathbf{0}$

Proof:

Items 1 and 2: By Assumption 2.2, each element of $\boldsymbol{\phi}_0$ is bounded in magnitude by $\bar{\phi}$. The absolute value of the t^{th} element of the $T \times 1$ vector $N^{-1}\boldsymbol{\varepsilon}\boldsymbol{\phi}_0$ is $|N^{-1}\sum_{i=1}^N \boldsymbol{\varepsilon}_t'\boldsymbol{\phi}_0| < |N^{-1}\sum_{i=1}^N \boldsymbol{\varepsilon}_t'\boldsymbol{\iota}|\bar{\phi} \xrightarrow[T\to\infty]{p} \mathbf{0}$ by Assumption 3.1. The same holds for Item 2, replacing $\boldsymbol{\phi}_0$ in the previous argument with any column of $\boldsymbol{\Phi}$.

Items 4 and 3: The matrix $N^{-1}\boldsymbol{\Phi}'\boldsymbol{J}_N\boldsymbol{\Gamma}$ has j,m element equal to

$$N^{-1}\sum_{i=1}^N \phi_{j,i}\delta(i,m) - \left(N^{-1}\sum_{i=1}^N \phi_{j,i}\right)\left(N^{-1}\sum_{i=1}^N \delta(i,m)\right).$$

By Assumptions 2.2 and 3.5, the first term satisfies

$$\left|N^{-1}\sum_{i=1}^N \phi_{j,i}\delta(i,m)\right| \le \left|\bar{\phi}N^{-1}\sum_{i=1}^N \delta(i,m)\right| \le \bar{\phi}N^{-1}\sum_{i=1}^N |\delta(i,m)| \xrightarrow[T\to\infty]{p} \mathbf{0}.$$

Applying similar logic to the second term,

$$\left|\left(N^{-1}\sum_{i=1}^{N}\phi_{j,i}\right)\left(N^{-1}\sum_{i=1}^{N}\delta(i,m)\right)\right| \le \left|\bar{\phi}\left(N^{-1}\sum_{i=1}^{N}\delta(i,m)\right)\right| \le \bar{\phi}N^{-1}\sum_{i=1}^{N}|\delta(i,m)| \xrightarrow[T\to\infty]{p} \mathbf{0}$$

thus the entire expression converges to zero. The same argument applies to Item 3, replacing $\delta(i,m)$ with $\gamma(i,m)$.

Item 5: The matrix $N^{-1}\boldsymbol{\Gamma}'\boldsymbol{J}_N\boldsymbol{\Delta}_\varepsilon$ has j,m element equal to

$$N^{-1}\sum_{i=1}^{N}\gamma(j,i)\delta(i,m) - \left(N^{-1}\sum_{i=1}^{N}\gamma(j,i)\right)\left(N^{-1}\sum_{i=1}^{N}\delta(i,m)\right)$$

Note that both $\gamma(j,i)$ and $\delta(i,m)$ are bounded by $A\ \forall j,i,m$ by Assumptions 2.6 and 3.5, thus both the first and second sums are bounded in magnitude by $N^{-1}A^2$ and therefore converge to zero.

Items 6, 7 and 8: Item 6 follows directly from 2.4. Also by 2.4

$$\left|T^{-1}\sum_{t=1}^{T}\omega_{j,t}\kappa(t,m)\right| \le A\left|T^{-1}\sum_{t=1}^{T}\omega(t,m)\right|$$

which converges to zero in probability by Assumption 3.1. A similar argument applies to Item 8. The j,m element of $T^{-1}\boldsymbol{F}'\boldsymbol{\mathcal{K}}$ is

$$T^{-1}\sum_{t=1}^{t}F_{t,j}\kappa(t,m) = T^{-1}\sum_{t=1}^{t}(F_{t,j}-\mu_j)\kappa(t,m) + \mu_j T^{-1}\sum_{t=1}^{t}\kappa(t,m).$$

The second term on the right-hand side converges to zero by Assumption 2.4. By Assumptions 2.1 and 2.4 the first term satisfies

$$\left|T^{-1}\sum_{t=1}^{t}(F_{t,j}-\mu_j)\kappa(t,m)\right| \quad \le \quad \left|AT^{-1}\sum_{t=1}^{t}(F_{t,j}-\mu_j)\right| \xrightarrow[T\to\infty]{p} 0,$$

thus $T^{-1}\boldsymbol{F}'\boldsymbol{\mathcal{K}} \xrightarrow[T\to\infty]{p} \mathbf{0}$.

QED

Lemma 3. *Let Assumptions 1, 2, and 3 hold. Then the following are true:*

1. $\bar{y} = T^{-1}\boldsymbol{\iota}'\boldsymbol{y} \xrightarrow[T\to\infty]{p} \beta_0 + \boldsymbol{\mu}'\boldsymbol{\beta}$
2. $\boldsymbol{J}_T\boldsymbol{X} \xrightarrow[T\to\infty]{p} (\boldsymbol{F}-\boldsymbol{\mu}')\boldsymbol{\Phi}'$
3. $T^{-1}\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X} \xrightarrow[T\to\infty]{p} \boldsymbol{\Phi}\boldsymbol{\Delta}_F\boldsymbol{\Phi}' + \boldsymbol{\Delta}_\epsilon$
4. $T^{-1}\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z} \xrightarrow[T\to\infty]{p} \boldsymbol{\Phi}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{\Gamma}$
5. $T^{-1}\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y} \xrightarrow[T\to\infty]{p} \boldsymbol{\Phi}\boldsymbol{\Delta}_F\boldsymbol{\beta}$
6. $N^{-1}\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}' \xrightarrow[N\to\infty]{p} B_0\boldsymbol{\iota}\boldsymbol{\iota}' + 2\boldsymbol{\iota}\boldsymbol{B}_1\boldsymbol{F}' + \boldsymbol{F}\boldsymbol{\mathcal{B}}\boldsymbol{F}' + \boldsymbol{\mathcal{K}}$

7. $T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{Z} \xrightarrow[T\to\infty]{p} \boldsymbol{\Delta}_F\boldsymbol{\Lambda}'$

8. $T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{y} \xrightarrow[T\to\infty]{p} \boldsymbol{\Delta}_F\boldsymbol{\beta}'$

9. $T^{-1}\boldsymbol{\mathcal{K}}\boldsymbol{J}_T\boldsymbol{Z} \xrightarrow[T\to\infty]{p} \boldsymbol{0}$

Proof: Immediate from Lemma 2 and Assumptions 1-3.

QED

Lemma 4. *Let Assumptions 1, 2 and 3 hold. Then the three pass regression filter forecast of* $\boldsymbol{y}$ *using cross section* $\boldsymbol{X}$ *and proxies* $\boldsymbol{Z}$ *satisfies*

$$\hat{\boldsymbol{y}} \xrightarrow[T,N\to\infty]{p} \boldsymbol{\iota}\beta_0 + \boldsymbol{\iota}\boldsymbol{\mu}'\boldsymbol{\beta} + (\boldsymbol{F}-\boldsymbol{\iota}\boldsymbol{\mu}')\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right]^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\beta}. \tag{A2}$$

Proof: We consider each case of sequential limits.

Case 1: $T\to\infty$ then $N\to\infty$:

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}\\
&\xrightarrow[T\to\infty]{p} \boldsymbol{\iota}\beta_0 + \boldsymbol{\iota}\boldsymbol{\mu}'\boldsymbol{\beta} + (\boldsymbol{F}-\boldsymbol{\iota}\boldsymbol{\mu}')\boldsymbol{\Phi}'\boldsymbol{J}_N(\boldsymbol{\Phi}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'+\boldsymbol{\Gamma})\\
&\quad\times\left[(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Phi}'+\boldsymbol{\Gamma}')\boldsymbol{J}_N(\boldsymbol{\Phi}\boldsymbol{\Delta}_F\boldsymbol{\Phi}'+\boldsymbol{\Delta}_\epsilon)\boldsymbol{J}_N(\boldsymbol{\Phi}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'+\boldsymbol{\Gamma})\right]^{-1}(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\Phi}'+\boldsymbol{\Gamma}')\boldsymbol{J}_N\boldsymbol{\Phi}\boldsymbol{\Delta}_F\boldsymbol{\beta}\\
&\xrightarrow[N\to\infty]{p} \boldsymbol{\iota}\beta_0 + \boldsymbol{\iota}\boldsymbol{\mu}'\boldsymbol{\beta} + (\boldsymbol{F}-\boldsymbol{\iota}\boldsymbol{\mu}')\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right]^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\beta}
\end{aligned}$$

The first equality is due to Lemma **??**. Convergence in T is due to Lemma 3 Items 2-5. Subsequent convergence in N is due to Lemma 2 Items 1-5.

Case 2: $N\to\infty$ then $T\to\infty$:

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'\boldsymbol{J}_T\boldsymbol{y}\\
&\xrightarrow[N\to\infty]{p} \boldsymbol{\iota}\bar{y} + \boldsymbol{J}_T\left(\boldsymbol{F}\boldsymbol{\mathcal{B}}\boldsymbol{F}'+\boldsymbol{\mathcal{K}}\right)\boldsymbol{J}_T\boldsymbol{Z}\left[\boldsymbol{Z}'\boldsymbol{J}_T\left(\boldsymbol{F}\boldsymbol{\mathcal{B}}\boldsymbol{F}'+\boldsymbol{\mathcal{K}}\right)\boldsymbol{J}_T\left(\boldsymbol{F}\boldsymbol{\mathcal{B}}\boldsymbol{F}'+\boldsymbol{\mathcal{K}}\right)\boldsymbol{J}_T\boldsymbol{Z}\right]^{-1}\\
&\quad\times\boldsymbol{Z}'\boldsymbol{J}_T\left(\boldsymbol{F}\boldsymbol{\mathcal{B}}\boldsymbol{F}'+\boldsymbol{\mathcal{K}}\right)\boldsymbol{J}_T\boldsymbol{y}\\
&\xrightarrow[T\to\infty]{p} \boldsymbol{\iota}\beta_0 + \boldsymbol{\iota}\boldsymbol{\mu}'\boldsymbol{\beta} + (\boldsymbol{F}-\boldsymbol{\iota}\boldsymbol{\mu}')\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right]^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\beta}
\end{aligned}$$

Convergence in N is due to Lemma 3 Item 6. Note that $\boldsymbol{J}_T\boldsymbol{\iota}=0$, thus the first two terms of the probability limit for $N^{-1}\boldsymbol{X}\boldsymbol{J}_N\boldsymbol{X}'$ are annihilated when multiplied by adjacent $\boldsymbol{J}_T$ matrices. Subsequent convergence in T is due to Lemma 3 Items 7-9.

QED

Proposition 1. *Let Assumptions 1, 2, 3 and 4 hold. The three pass regression filter forecaster of* $\boldsymbol{y}$ *using cross section* $\boldsymbol{X}$ *and proxies* $\boldsymbol{Z}$ *is asymptotically efficient.*

Proof: Given Assumptions 1, 2 and 3, Lemma 4 holds and we can therefore manipulate (A2). By Assumption 4, $\boldsymbol{\Lambda}$ is invertible, therefore the probability limit (A2) reduces to $\boldsymbol{\iota}\beta_0+\boldsymbol{F}\boldsymbol{\beta}$.

QED

Proposition 2. *Let Assumptions 1, 2, 3, 5 and 6 hold. The three pass regression filter forecaster of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and proxies $\boldsymbol{Z}$ is asymptotically efficient.*

Proof: Given Assumptions 1, 2 and 3, Lemma 4 holds and we can therefore manipulate (A2). Partition $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\Delta}_F$ as

$$\boldsymbol{\mathcal{B}} = \begin{bmatrix} \boldsymbol{\mathcal{B}}_{11} & \boldsymbol{\mathcal{B}}_{12} \\ \boldsymbol{\mathcal{B}}_{12}' & \boldsymbol{\mathcal{B}}_{22} \end{bmatrix} \quad , \quad \boldsymbol{\Delta}_F = \begin{bmatrix} \boldsymbol{\Delta}_{11} & \boldsymbol{\Delta}_{12} \\ \boldsymbol{\Delta}_{12}' & \boldsymbol{\Delta}_{22} \end{bmatrix}$$

such that the block dimensions of $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\Delta}_F$ coincide. By Assumption 6, the off-diagonal blocks $\boldsymbol{\mathcal{B}}_{12}$ and $\boldsymbol{\Delta}_{12}$ are zero. As a result, the first diagonal block of the term $\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F$ in Equation A2 is $\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{B}}_1\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{B}}_1\boldsymbol{\Delta}_{F,1}$. By Assumption 5, pre- and post-multiplying by $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1, \boldsymbol{0}]$ reduces the term in square brackets to $\boldsymbol{\Lambda}_1\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{B}}_1\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{B}}_1\boldsymbol{\Delta}_{F,1}\boldsymbol{\Lambda}_1$. Similarly, $\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' = [\boldsymbol{\Lambda}_1\boldsymbol{\mathcal{B}}_1\boldsymbol{\Delta}_{F,1}, \boldsymbol{0}]'$ and $\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F = [\boldsymbol{\Lambda}_1\boldsymbol{\Delta}_{F,1}\boldsymbol{\mathcal{B}}_1\boldsymbol{\Delta}_{F,1}, \boldsymbol{0}]$. By Assumption 5, $\boldsymbol{\Lambda}_1$ is invertible and therefore the expression for $\hat{\boldsymbol{y}}$ reduces to $\boldsymbol{\iota}\beta_0 + \boldsymbol{F}\boldsymbol{\beta}$.

QED

Corollary 1. *Let Assumptions 1, 2, 3, 6 hold. Additionally, assume that $K_f = 1$. The target-proxied three pass regression filter forecaster of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and proxy $\boldsymbol{y}$ is asymptotically efficient, for any value of K_g.*

Proof: This follows as a special case of Proposition 2 due to the fact that our only proxy is the target itself (ie. $\boldsymbol{\Lambda} = \boldsymbol{\beta}'$) and the fact that there is only one relevant factor (ie. $\boldsymbol{\beta}' = (\beta_1, \boldsymbol{0}')$, where the relevant factor is the first factor without loss of generality.

QED

Proposition 3. *Let Assumptions 1, 2, 3, 5 and 6 hold. Further assume that the factors $\boldsymbol{f}_t$ have equal variance σ_f^2. Then the target-proxied three pass regression filter forecaster of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and proxy $\boldsymbol{y}$ is asymptotically efficient, for any number of factors K_f and K_g.*

Proof: The equivariance of $\boldsymbol{f}_t$ implies that the expressions in the proof of Proposition 2 are $\boldsymbol{\Lambda} = \boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{0})$, $\boldsymbol{B}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' = \sigma_f^2\boldsymbol{\beta}$, $\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' = \sigma_f^6\boldsymbol{\beta}'\boldsymbol{\beta}$, and $\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}_F\boldsymbol{\beta} = \sigma_f^4\boldsymbol{\beta}'\boldsymbol{\beta}$. The expression for $\hat{\boldsymbol{y}}$ therefore collapses to $\boldsymbol{\iota}\beta_0 + \boldsymbol{F}\boldsymbol{\beta}$.

QED

Proposition 4. *Let Assumptions 1, 2, and 3 hold. Define the matrix*

$$\boldsymbol{\Omega} = \boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F^3\boldsymbol{\Lambda}'\right]^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F - \boldsymbol{S}'\left[\boldsymbol{S}\boldsymbol{\Delta}_F\boldsymbol{S}'\right]^{-1}\boldsymbol{S}$$

where $\boldsymbol{S}$ is a $M \times K$ ($M < K$) matrix with ones on the main diagonal and zeros elsewhere. Then the three pass regression filter forecast of $\boldsymbol{y}$ using cross section $\boldsymbol{X}$ and any set of M proxies $\boldsymbol{Z}$ is (weakly) asymptotically more efficient than the forecast using the first M principal components of $\boldsymbol{X}$ if and only if $\boldsymbol{\Omega}$ is positive (semi-)definite.

Proof: Let $\hat{\boldsymbol{y}}^{PC}$ denote the forecast of $\boldsymbol{y}$ using the first M principal components of $\boldsymbol{X}$. This converges to $\boldsymbol{F}\boldsymbol{S}'\left[\boldsymbol{S}\boldsymbol{\Delta}_F\boldsymbol{S}'\right]^{-1}\boldsymbol{S}\boldsymbol{\Delta}_F\boldsymbol{\beta}$ as $N, T \to \infty$ under Assumptions 1, 2, and 3 (Stock and Watson 2002a). Under the same assumptions, the three pass regression filter forecast $\hat{\boldsymbol{y}}^{3PRF}$ converges to $\boldsymbol{F}\boldsymbol{\Lambda}'\left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F^3\boldsymbol{\Lambda}'\right]^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F^2\boldsymbol{\beta}$. Comparing asymptotic forecast error variance matrices, we see that

$$\lim_{N,T\to\infty} Var(\boldsymbol{y}-\hat{\boldsymbol{y}}^{PC}) - Var(\boldsymbol{y}-\hat{\boldsymbol{y}}^{3PRF}) = \boldsymbol{\Omega}.$$

QED

B The Kalman Filter

This system is defined by the state space

$$\boldsymbol{\Pi}_t = \boldsymbol{M}_0 + \boldsymbol{M}\boldsymbol{\Pi}_{t-1} + \textbf{error}_t^F, \qquad \textbf{error}_t^F \sim N(\boldsymbol{0},\boldsymbol{Q}) \tag{A3}$$

$$\boldsymbol{\Upsilon}_t = \boldsymbol{\Psi}_0 + \boldsymbol{\Psi}\boldsymbol{\Pi}_t + \textbf{error}_t^\Upsilon, \qquad \textbf{error}_t^\Upsilon \sim N(\boldsymbol{0},\boldsymbol{R}) \tag{A4}$$

$$\boldsymbol{\Pi}_t = \begin{pmatrix} \boldsymbol{F}_t \\ \boldsymbol{F}_{t-1} \end{pmatrix} \tag{A5}$$

$$\boldsymbol{\Upsilon}_t = \begin{pmatrix} \tilde{\boldsymbol{z}}_t \\ \boldsymbol{x}_t \end{pmatrix} \tag{A6}$$

Here we assume the k-dimensional "as observed at time t" proxy vector $\tilde{\boldsymbol{z}}_t$ includes the target variable to be forecasted Hence, y_t is an element of $\tilde{\boldsymbol{z}}_t$ (whereas it the time $t+1$ variable y_{t+1} that is an element of $\boldsymbol{z}_t$ as used in Section 2). $\boldsymbol{\Pi}_t$ 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 KF 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 equations (see Hamilton 1994) are:

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{M}\boldsymbol{P}_{t-1|t-1}\boldsymbol{M}' + \boldsymbol{Q} \tag{A7}$$

$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1}\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1} \tag{A8}$$

$$\boldsymbol{K}_t = \boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}'\left(\boldsymbol{\Psi}\boldsymbol{P}_{t|t-1}\boldsymbol{\Psi}' + \boldsymbol{R}\right)^{-1} \tag{A9}$$

$$\boldsymbol{\Pi}_{t|t-1} = \boldsymbol{M}_0 + \boldsymbol{M}\boldsymbol{\Pi}_{t-1|t-1} \tag{A10}$$

$$\boldsymbol{\Upsilon}_{t|t-1} = \boldsymbol{\Psi}_0 + \boldsymbol{\Psi}\boldsymbol{\Pi}_{t|t-1} \tag{A11}$$

$$\boldsymbol{\Pi}_{t|t} = \boldsymbol{\Pi}_{t|t-1} + \boldsymbol{K}_t\left(\boldsymbol{\Upsilon}_t - \boldsymbol{\Upsilon}_{t|t-1}\right) \tag{A12}$$

We first note that the matrix inversion lemma lets us rewrite (A8) 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) (A9) can be rewritten in a form similar to (9) 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 matters, assume $\boldsymbol{\Psi}_0=\boldsymbol{0}$. Hence, the MLE of $\boldsymbol{\Psi}$ satisfies the following

$$vec\hat{(\boldsymbol{\Psi})}=\left(\boldsymbol{\Pi}'\boldsymbol{\Pi}\otimes\hat{\boldsymbol{R}}^{-1}\right)^{-1}\left(\boldsymbol{\Pi}'\boldsymbol{\Upsilon}\otimes\hat{\boldsymbol{R}}^{-1}\right) \tag{A13}$$

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

$$\hat{\boldsymbol{R}}=\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\Upsilon}_t-\boldsymbol{\Psi}\boldsymbol{\Pi}_t. \tag{A14}$$

Equations (A13) and (A14) make it clear that the MLE of $\boldsymbol{\Psi}$ is obtained by a GLS regression of the observable variables $\boldsymbol{\Upsilon}$ on the true factors $\boldsymbol{f}$.

Finally, let us state the optimal linear prediction of $\boldsymbol{\Upsilon}_{t+1}$ on the basis of $\{\boldsymbol{\Upsilon}_t,\boldsymbol{\Upsilon}_{t-1},\ldots\}$ in the case that we ignore the KF's temporal pooling of information. Again for simplicity assume $\boldsymbol{\Psi}_0=\boldsymbol{M}_0=\boldsymbol{0}$. We do this by considering $\boldsymbol{M}$, which we can partition into four square submatrices. Given $\boldsymbol{M}$'s augmented nature, $\boldsymbol{M}_{21}=\boldsymbol{I}$ and $\boldsymbol{M}_{22}=\boldsymbol{0}$ to just shift the location of $\boldsymbol{F}_t$ in $\boldsymbol{\Pi}_t$ and $\boldsymbol{\Pi}_{t+1}$. Then ignoring the temporal pooling is equivalent to restricting $\boldsymbol{M}_{11}=\boldsymbol{M}_{12}=\boldsymbol{0}$. Therefore

$$\begin{aligned}
\boldsymbol{\Upsilon}_{t+1|t} &= \boldsymbol{\Psi}\boldsymbol{\Pi}_{t+1|t}\\
&= \boldsymbol{\Psi}\boldsymbol{M}\boldsymbol{\Pi}_{t|t}\\
&= \boldsymbol{\Psi}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix}\boldsymbol{\Pi}_{t|t}
\end{aligned} \tag{A15}$$

Recall that y is an element of $\boldsymbol{\Upsilon}$. The proxy assumption means that y's row has 0s in the first K_f+K_g columns. Letting the first row of $\boldsymbol{\Psi}$ therefore be $(\boldsymbol{0}',\boldsymbol{\beta}')$, (A15) says

$$y_{t+1|t}=\boldsymbol{\beta}'\boldsymbol{F}_{t|t} \tag{A16}$$

which is the expression used in the text.

C Linear Projection, Principal Components, and the 3PRF

For exposition's sake, assume all variables are time-demeaned.

Linear Projection (LP) OLS consistently estimates the requisite LP coefficients (see Hamilton 1994). Hence we have the LP forecast given by

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{A17}$$

$$\hat{y}_{t+1}^{LP} = \hat{\boldsymbol{\alpha}}'\boldsymbol{x}_t \tag{A18}$$

In particular what does (14) tell us about the optimal weight on the i^{th} predictor? Denote $\boldsymbol{\Sigma}_X = (\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_N)$ with $\boldsymbol{\sigma}_i$ the i^{th} column (or row) of the within-predictor covariance matrix, and $\boldsymbol{\Sigma}_X^{-1} = (\ddot{\boldsymbol{\sigma}}_1, \ldots, \ddot{\boldsymbol{\sigma}}_N)$ so that $\ddot{\boldsymbol{\sigma}}_j'$ is the j^{th} column (or row) of the within-predictor information matrix. Write $\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{s}$, the vector of with-target covariances. Thus the LP weights the i^{th} predictor by

$$\hat{\alpha}_i = \boldsymbol{s}'\ddot{\boldsymbol{\sigma}}_i, \tag{A19}$$

where $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_N)$. This weight is a linear combination of two terms. The first term is the vector of with-target covariances. The second term is the i^{th} column of the within-predictor information matrix. Therefore, LP's weight on the i^{th} predictor is a combination of every predictors' covariance with the target, where the combination is according to information from the within-predictor information matrix.

If the predictors are orthogonal *in-sample*, then $\boldsymbol{\Sigma}_X$ is a diagonal matrix with positive entries on the main diagonal, and therefore $\boldsymbol{\Sigma}_X^{-1}$ is also diagonal with the reciprocal of each nonzero element of $\boldsymbol{\Sigma}_X$ on its main diagonal. Then (A19) reduces to

$$\hat{\alpha}_i = s_i\ddot{\sigma}_{i,i} = s_i\sigma_{i,i}^{-1}.$$

Clearly, this is identically equal to the bivariate regression estimate coming from

$$\left(\sum_{t=1}^{T} x_{i,t}^2\right)^{-1}\left(\sum_{t=1}^{T} x_{i,t}y_{t+1}\right).$$

The standard econometric result (e.g. White 2001) is that this bivariate regression estimate is the population object plus a term that is $o_p(T)$, and therefore the estimate is more precise as T gets large. Therefore, if the predictors are orthogonal to one another, the estimate given by (A19) gets more precise as $T \to \infty$, regardless of how close N is to T as this limit is taken.

Principal Components Forecasts (PC) Following Anderson (2003), the i^{th} principal component of $(T \times N)$ predictor matrix $\boldsymbol{X}$ is given by a linear combination of the N predictors given by the normalized eigenvector corresponding to the j^{th} largest eigenvalue of $\boldsymbol{\Sigma}_X = \boldsymbol{X}'\boldsymbol{X}$. That is, suppose e_j is the j^{th} largest eigenvalue of $\boldsymbol{\Sigma}_X$. Then the j^{th} eigenvector $\tilde{\boldsymbol{l}}_i$ satisfies the equation

$$(\boldsymbol{\Sigma}_X - e_j \boldsymbol{I})\tilde{\boldsymbol{l}}_j = \mathbf{0} \tag{A20}$$

$$\text{s.t. } \tilde{\boldsymbol{l}}_j' \tilde{\boldsymbol{l}}_j = 1. \tag{A21}$$

The j^{th} principal component is thus given by $\boldsymbol{X}\tilde{\boldsymbol{l}}_j = \mathbf{u}_j$. These eigenvectors have the following properties. Letting M denote the number of principal components used in forecasting, denote $\tilde{\boldsymbol{L}} = \begin{pmatrix} \tilde{\boldsymbol{l}}_1 & \cdots & \tilde{\boldsymbol{l}}_M \end{pmatrix}$. Then

$$\tilde{\boldsymbol{L}}'\tilde{\boldsymbol{L}} = \boldsymbol{I} \tag{A22}$$

$$\tilde{\boldsymbol{L}}'\boldsymbol{\Sigma}_X\tilde{\boldsymbol{L}} = \begin{bmatrix} e_1 & 0 & \cdots & 0 \\ 0 & e_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & e_M \end{bmatrix} = \boldsymbol{E}_k. \tag{A23}$$

We first find the second-stage predictive regression coefficient (following Stock and Watson 2002a,b) by regressing $\boldsymbol{y}$ on $\boldsymbol{U} = \begin{bmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_{N_{PC}} \end{bmatrix} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)'$:

$$\hat{\boldsymbol{\delta}}^{PC} = \left(\boldsymbol{U}'\boldsymbol{U}\right)^{-1}\boldsymbol{U}'\boldsymbol{y}.$$

Then the PC forecast is

$$y_{t+1}^{PC} = \hat{\boldsymbol{\delta}}^{PC'}\boldsymbol{u}_t \tag{A24}$$

$$= \boldsymbol{y}'\boldsymbol{X}\tilde{\boldsymbol{L}}\left(\tilde{\boldsymbol{L}}'\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{L}}\right)^{-1}\tilde{\boldsymbol{L}}'\boldsymbol{x}_t \tag{A25}$$

$$= \boldsymbol{s}'\left(\tilde{\boldsymbol{L}}\boldsymbol{E}_{N_{PC}}^{-1}\tilde{\boldsymbol{L}}'\right)\boldsymbol{x}_t \tag{A26}$$

$$= \boldsymbol{s}'\left(e_1^{-1}\tilde{\boldsymbol{L}}_1 + e_2^{-1}\tilde{\boldsymbol{L}}_2 + \cdots + e_M^{-1}\tilde{\boldsymbol{L}}_M\right)\boldsymbol{x}_t \tag{A27}$$

where we have defined $\tilde{\boldsymbol{L}}_j = \tilde{\boldsymbol{l}}_j\tilde{\boldsymbol{l}}_j' = \begin{bmatrix} \boldsymbol{l}_{j,1} & \cdots & \boldsymbol{l}_{j,N} \end{bmatrix}$.

3PRF As noted before, the 3PRF forecast can be rewritten

$$y_{t+1}^{3PRF} = \boldsymbol{s}'\hat{\boldsymbol{\Phi}}'\left[\hat{\boldsymbol{\Phi}}\boldsymbol{\Sigma}_X\hat{\boldsymbol{\Phi}}'\right]^{-1}\hat{\boldsymbol{\Phi}}\boldsymbol{x}_t$$

for any number of proxies.

PC Suppose we use the first PC of $\boldsymbol{X}$ for the purposes of forecasting y_{t+1}. Following Stock and Watson (2002a,b) we find the PC forecast by first regressing y_{t+1} on $u_{1,t}$

$$\hat{\delta}^{PC} = \left(\mathbf{u}_1'\mathbf{u}_1\right)^{-1}\left(\mathbf{u}_1'\boldsymbol{y}\right).$$

Then this OLS estimate multiplies $u_{1,t}$ to deliver the time t PC forecast of y_{t+1}.

$$
\begin{aligned}
\hat{y}^{PC}_{t+1} &= \hat{\delta}^{PC} u_{1,t} \\
&= \left((\mathbf{u}_1'\mathbf{u}_1)^{-1} (\mathbf{u}_1'\boldsymbol{y}) \right) \left(\tilde{\boldsymbol{l}}_1' \boldsymbol{x}_t \right) \\
&= \left(\left(\tilde{\boldsymbol{l}}_1' \boldsymbol{\Sigma}_X \tilde{\boldsymbol{l}}_1 \right)^{-1} \left(\tilde{\boldsymbol{l}}_1' \boldsymbol{X}'\boldsymbol{y} \right) \right) \left(\tilde{\boldsymbol{l}}_1' \boldsymbol{x}_t \right) \\
&= e_1^{-1} \left(\boldsymbol{y}'\boldsymbol{X}\tilde{\boldsymbol{l}}_1 \right) \tilde{\boldsymbol{l}}_1' \boldsymbol{x}_t \\
&= e_1^{-1} \begin{bmatrix} \sum_t x_{1,t} y_{t+1} & \cdots & \sum_t x_{N,t} y_{t+1} \end{bmatrix} \left(\tilde{\boldsymbol{l}}_1 \tilde{\boldsymbol{l}}_1' \right) \boldsymbol{x}_t.
\end{aligned}
$$

Then we can say that the PC forecaster places weight $\hat{\alpha}_i^{PC}$ on predictor i

$$
\hat{\alpha}_i^{PC} = e_1^{-1} \boldsymbol{s}' \boldsymbol{l}_{1,i} \tag{A28}
$$

3PRF When $M = 1$ and $\boldsymbol{Z} = \boldsymbol{y}$, parts of the above become scalars – this allows us to do the following manipulations

$$
\begin{aligned}
y^{3PRF}_{t+1} &= (\boldsymbol{y}'\boldsymbol{X}) \left(\boldsymbol{X}'\boldsymbol{y}(\boldsymbol{y}'\boldsymbol{y})^{-1} \boldsymbol{\Sigma}_X (\boldsymbol{y}'\boldsymbol{y})^{-1} \boldsymbol{y}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{y}(\boldsymbol{y}'\boldsymbol{y})^{-1} \boldsymbol{x}_t & \text{(A29)} \\
&= \boldsymbol{y}'\boldsymbol{X} \left[(\boldsymbol{y}'\boldsymbol{y})^{-1} \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\Sigma}_X \boldsymbol{X}'\boldsymbol{y} (\boldsymbol{y}'\boldsymbol{y})^{-1} \right]^{-1} (\boldsymbol{y}'\boldsymbol{y})^{-1} \boldsymbol{y}'\boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{y} \right) (\boldsymbol{y}'\boldsymbol{y})^{-1} \boldsymbol{x}_t & \text{(A30)} \\
&= \boldsymbol{y}'\boldsymbol{X} \left[\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\Sigma}_X \boldsymbol{X}'\boldsymbol{y} \right]^{-1} \left(\boldsymbol{y}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{y} \right) \boldsymbol{x}_t & \text{(A31)} \\
&= \boldsymbol{s}' \left[\boldsymbol{s}'\boldsymbol{\Sigma}_X \boldsymbol{s} \right]^{-1} \left(\boldsymbol{s}'\boldsymbol{s} \right) \boldsymbol{x}_t. & \text{(A32)}
\end{aligned}
$$

Therefore the 3PRF gives the i^{th} predictor the weight

$$
\hat{\alpha}_i^{3PRF} = s_i \left(\left[\boldsymbol{s}'\boldsymbol{\Sigma}_X \boldsymbol{s} \right]^{-1} \left(\boldsymbol{s}'\boldsymbol{s} \right) \right), \tag{A33}
$$

which we use in the text. Statement (A33) says that the equal weight given to all forecasts $s_i x_{i,t}$ is found by the reciprocal of the Rayleigh-Ritz ratio[23] composed of the with-target covariance vector $\boldsymbol{s}$ and the within-predictor covariance matrix $\boldsymbol{\Sigma}_X$. The maximum Rayleigh-Ritz ratio given $\boldsymbol{\Sigma}_X$ is the eigenvalue e_1, found using the first eigenvector. Therefore $\left(\left[\boldsymbol{s}'\boldsymbol{\Sigma}_X \boldsymbol{s} \right]^{-1} \left(\boldsymbol{s}'\boldsymbol{s} \right) \right)$ is weakly larger than e_1^{-1}.

D Further Simulation Study

[23] See Horn and Johnson (1985).

Table A1: Varying Correlation of Idiosyncratic Errors

	% of Infeasible Best Possible R^2											
	In-Sample						Out-of-Sample					
K_f/M	1/1		3/1		3/3		1/1		3/1		3/3	
	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF
	N=10,T=50											
$a = 0, d = 0$	98.6	99.1	42.4	78.0	100.0	94.6	84.4	86.6	17.3	56.0	72.8	67.3
$a = 0.3, d = 0$	95.9	94.9	47.9	73.8	101.6	96.6	85.9	85.6	16.9	55.4	70.9	66.0
$a = 0.9, d = 0$	93.1	93.2	46.7	73.8	97.5	91.6	77.4	78.8	5.7	38.4	48.9	54.9
$a = 0, d = 1$	94.8	95.2	47.3	77.2	98.6	95.5	76.5	81.8	22.6	68.8	66.4	67.7
$a = 0.3, d = 1$	94.8	96.3	33.9	71.0	95.8	89.3	80.9	77.0	10.8	50.2	64.9	56.9
$a = 0.9, d = 1$	71.5	82.2	40.2	69.1	91.4	88.0	42.0	73.9	16.0	51.5	51.1	60.0
	N=100,T=200											
$a = 0, d = 0$	100.4	100.3	33.8	93.5	101.3	100.5	97.7	98.4	18.7	86.1	95.6	97.1
$a = 0.3, d = 0$	99.9	99.9	34.5	92.5	100.9	99.8	97.0	97.8	21.5	87.7	95.6	95.8
$a = 0, d = 1$	75.0	97.3	13.4	83.1	71.3	88.1	61.2	93.6	8.0	76.8	51.7	83.1
$a = 0.3, d = 1$	51.7	97.4	19.7	84.3	74.1	91.5	42.4	93.9	8.3	70.2	48.3	80.5
$a = 0.9, d = 1$	51.3	95.1	17.7	82.6	67.6	90.3	40.2	89.3	7.9	72.1	48.1	79.0

Notes: No irrelevant factors, $K_g = 0$. Factors persistent, $A_f = 0.9$. Equal strength, $\kappa_{fg} = 1$. Median across 100 simulations

Table A2: Varying Strength of Factors

	% of Infeasible Best Possible R^2											
	In-Sample						Out-of-Sample					
K_f/M	1/1		3/1		3/3		1/1		3/1		3/3	
	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF	PC	3PRF
	N=10,T=50											
$\kappa_{fg} = \frac{1}{10}$	8.9	18.5	8.9	20.5	32.6	34.1	-8.8	-10.3	-6.9	-19.2	-36.4	-55.2
$\kappa_{fg} = \frac{1}{2}$	15.8	56.9	11.5	39.2	57.9	55.8	-6.8	36.2	-11.5	2.1	-8.8	5.2
$\kappa_{fg} = 1$	64.1	87.4	33.6	63.0	84.1	81.9	47.8	65.4	-6.5	33.8	22.8	32.2
$\kappa_{fg} = 2$	97.9	98.1	37.8	75.8	96.7	92.5	80.2	87.0	18.7	49.7	52.3	60.5
$\kappa_{fg} = 10$	100.7	100.5	58.1	80.1	104.2	98.8	92.7	92.1	19.3	56.6	73.6	75.8
	N=100,T=200											
$\kappa_{fg} = \frac{1}{10}$	2.3	6.8	2.4	11.1	13.0	16.0	0.1	4.7	-2.2	-5.1	-7.7	-6.2
$\kappa_{fg} = \frac{1}{2}$	2.4	75.9	3.8	51.7	23.5	68.1	-2.5	69.2	1.1	36.2	0.9	49.1
$\kappa_{fg} = 1$	56.6	96.8	19.3	82.4	72.2	91.6	44.5	93.5	9.2	73.3	54.2	81.7
$\kappa_{fg} = 2$	100.0	99.9	46.5	91.9	99.4	98.7	96.8	97.8	26.0	87.7	92.4	93.7
$\kappa_{fg} = 10$	100.3	100.4	37.8	93.7	101.2	100.3	97.8	98.5	27.4	90.1	95.3	96.0

Notes: Equal number of relevant and irrelevant factors, $K_f = K_g$. Factors persistent, $A_f, A_g = 0.9$. No idiosyncratic cross- or serial-correlation, $a = 0, d = 0$. Median across 100 simulations.