Economic Forecasts Using Many Noises*

Yuan Liao
Department of Economics
Rutgers University

Xinjie Ma
NUS Business School
National University of Singapore

Andreas Neuhierl
Olin School of Business
Washington University in St. Louis

Zhentao Shi
Department of Economics
The Chinese University of Hong Kong

December 9, 2023

Abstract

This paper addresses a key question in economic forecasting: does pure noise truly lack predictive power? Economists typically conduct variable selection to eliminate noises from predictors. Yet, we prove a compelling result that in most economic forecasts, the inclusion of noises in predictions yields greater benefits than its exclusion. Furthermore, if the total number of predictors is not sufficiently large, intentionally adding more noises yields superior forecast performance, outperforming benchmark predictors relying on dimension reduction. The intuition lies in economic predictive signals being densely distributed among regression coefficients, maintaining modest forecast bias while diversifying away overall variance, even when a significant proportion of predictors constitute pure noises. One of our empirical demonstrations shows that intentionally adding $300 \sim 6,000$ pure noises to the Welch and Goyal (2008) dataset achieves a noteworthy 10% out-of-sample R^2 accuracy in forecasting the annual U.S. equity premium. The performance surpasses the majority of sophisticated machine learning models.

Key words: machine learning, factor model, double descent, dense signals

*Acknowledgement.

1 Introduction

In the realm of economic forecasting, the forecast outcome is typically shaped by a concise set of low-dimensional economic factors summarizing the state of the macroeconomy, financial markets, and policy-related indices. These factors are however latent economic variables so cannot be directly used for economic forecasts. Instead, economists often rely on a set of high-dimensional predictors that are considered to hold predictive information regarding the outcome variable, given that they are largely influenced by the latent factors. The informativeness of these predictors varies, with some providing robust predictive signals and others serving as mere noise, exhibiting minimal conditional predictive power up to zero. Consequently, it is customary for economists to employ dimension reduction techniques, including Lasso, Ridge, and principal components analysis (PCA), among others, to enhance the precision of economic forecasts. See Ng (2013) for an excellent review for variable selections in economic forecasts.

The objective of this paper is to address a key question in economic forecasts: does pure noise truly lack predictive power? Consider a hypothetical scenario, where an economist has collected a large number of predictors. Some of them carry informative potential for forecasting the desired outcome variable. Assuming the economist has a priori knowledge distinguishing informative predictors from mere noise, it is then widely accepted that the economist should initiate variable selection by excluding all noises from the predictor set. In this hypothetical scenario, variable selection is a straightforward task because she knows the identities of informative predictors and noises.

Strikingly, we argue that in most economic forecasting scenarios, the economist should retain the noises in her set of predictors. Formally, we shall prove a compelling result that the inclusion of noises in predictions yields greater benefits than its exclusion. Furthermore, if the total number of predictors is not sufficiently large, she should intentionally add more noises. Then the overall forecast performance will surpass that of many benchmark predictors reliant on dimension reduction techniques such as Lasso, Ridge, and PCA, even if these methods use truly informative predictors without any noises.

The prediction method we recommend is the pseudoinverse ordinary least squares (pseudo-OLS), which simply replaces the inverse matrix in the usual definition of OLS by the Moore–Penrose generalized inverse, and is always well defined regardless of the dimensionality. This estimator is often known as "ridgeless regression" in the machine learning literature (e.g., Hastie et al. (2022)), as it equals the limit of the ridge estimator with the penalty goes to zero. But unlike the regular ridge regression that intentially

introduces a nontrivial penalty to prevent the model from overfitting, we emphasize the *anti-regularizion* perspective of the pseudo-OLS, in the sense that the in-sample data are perfectly interpolated: the fitted in-sample outcomes exactly equal the true outcomes.

Contrary to the conventional statistical wisdom, which asserts that overfitting significantly undermines forecast performance by inflating out-of-sample variance, we show that this is no longer the case when a substantial number of additional predictors are included, even if these predictors constitute pure noises. The crucial insight lies in the observation that, with a sufficiently large total number of predictors, denoted as p, the overall variance is *diversified away*. Consequently, out-of-sample variance decays as p approaches to infinity. This phenomenon is well connected to the double descent phenomenon in the machine learning literature, e.g., (Belkin et al., 2019; Hastie et al., 2022; Arora et al., 2019). That is, as the model complexity exceeds the sample size and continues increasing, a second descent of the prediction occurs in the extremely overparametrized regime.

While our model shares similarities with recent theoretical developments in linear forecast models, we adopt a fundamentally distinct perspective by concentrating on the prevalent context of economic forecasts. In this context, predictive information is embedded within a few latent factors and is conveyed through a collection of informative predictors, along with a substantial number of spurious predictors—essentially, pure idiosyncratic noises. Our analysis reveals that in most economic forecasting scenarios, the model's predictive signal is densely distributed among the regression coefficients of a large number of predictors. This provides a theoretical underpinning for recent empirical insights highlighted in Giannone et al. (2021), stating that "*the empirical support for low-dimensional models is generally weak... economic data are not informative enough to uniquely identify the relevant predictors when a large pool of variables is available to the researcher.*" Expanding on this, we explore the intuition that the dense feature of economic forecasting models diversifies away overall variance, even when a significant proportion of predictors constitute pure noises.

However, a potential limitation arises when an excessive number of noises are intentionally included, as this may inflate the bias. Therefore, the total number of predictors (the sum of informative and uninformative predictors) p can be regarded as a tuning parameter. We address this concern by asserting that the benefits derived from reducing variance outweigh the costs associated with bias inflation. As such, the procedure remains relatively robust to the choice of p, and this parameter is straightforward to tune. We also demonstrate choosing p using standard cross-validations for both cross-sectional predictions and time-series predictions.

We concentrate on the forecast perspective, and acknowledge that adding noises may reduce the interpretability of the model, compared to classic linear methods such as Lasso and PCA. Analyzing the feature importance of informative predictors in this context would be an interesting topic which we leave for future research.

In one of the empirical illustrations, we apply the pseudo-OLS with intentionally added noises to forecast the annual U.S. equity premium, using the extensively utilized dataset presented by Welch and Goyal (2008). We find that the addition of $300 \sim 6,000$ noises into the original set of sixteen predictors yields a noteworthy 10% out-of-sample R^2 accuracy. Remarkably, this finding remains highly robust to the number of included noises. The performance surpasses all linear models and the majority of sophisticated nonlinear machine learning models on forecasting the equity premium.

Theoretical analysis on extremely overparametrized regime has received extensive attentions in the recent statistical and econometric literature. Hastie et al. (2022); Lee and Lee (2023); Chinot et al. (2022) studied the pseudo-OLS regressions in the overparametrized regime for linear forecasting models. Besides, Mei and Montanari (2019) studied the bias-variance tradeoff in random feature regressions. Other related papers in the economic literature include Spiess et al. (2023) for treatment effects studies, and Kelly et al. (2022); Fan et al. (2022) for asset pricing. Didisheim et al. (2023) showed that the Sharpe ratio monotonically increases as the model's complexity grows in the overparametrized regime. Most of the existing works, however, are concerned with weakly correlated designs, whereas informative predictors in many economic forecasts are strongly correlated due to the common economic factors.

We adopt the following notations. Let $\|A\|$ denote the ℓ_2 norm if A is a vector, and the operator norm if A is a matrix. Let $\sigma_j(\cdot)$ be the jth largest singular value of a matrix; and let $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ respectively denote the minimum and maximum nonzero singular values of the matrix. For two sequences $a_{p,n}$ and $b_{p,n}$, we denote $a_{p,n} \ll b_{p,n}$ (or $b_{p,n} \gg a_{p,n}$) if $a_{p,n} = o(b_{p,n})$. Also, denote by $a_{p,n} \asymp b_{p,n}$ if $a_{p,n} \ll b_{p,n}$ and $a_{p,n} \gg b_{p,n}$.

2 The Model

2.1 The oracle model

The objective is to forecast an outcome variable y_t. We assume that the true data generating process (DGP) for y_t is as follows:

$$y_t = \rho' f_t + \epsilon_{y,t}, \quad t = 1, ..., n \quad \text{(true DGP)} \tag{2.1}$$

where f_t is a vector of low-dimensional ($\dim(f_t) = K$) latent factors. The model admits an intercept term by setting the first component of f_t as one. Note that while observations are indexed by subscript t, we allow both cross-sectional forecasts, in which y_t denotes the outcome of the t th subject; and the time series forecast in which the outcome depends on the lagged factors:

$$Y_{t+1} = \rho' f_t + \eta_{t+1},$$

by setting $y_t := Y_{t+1}$ and $\epsilon_{y,t} := \eta_{t+1}$ as in Stock and Watson (2002).

In addition, economists observe a set of high-dimensional regressors

$$X_t = (x_{1,t}, ..., x_{p,t})', \quad \dim(X_t) = p$$

which potentially carries the predictive information regarding y_t. We assume that X_t depends on the common factors through the following factor model:

$$x_{i,t} = \lambda_i' f_t + u_{i,t}, \quad \mathbb{E}(u_{i,t}|f_t, \epsilon_{y,t}) = 0, \quad i = 1, ..., p \tag{2.2}$$

where λ_i is a vector of loadings for the i th variable. Also, $\mathbb{E}(u_{i,t}|f_t, \epsilon_{y,t}) = 0$ so the predictive power of $x_{i,t}$ on y_t is only through the latent factors. We emphasize that f_t may be weak in the sense that some component of X_t may not depend on f_t. Therefore, we partition X_t as:

$$X_t = \begin{pmatrix} X_{I,t} \\ X_{N,t} \end{pmatrix} = \begin{pmatrix} \Lambda_I \\ 0 \end{pmatrix} f_t + u_t, \tag{2.3}$$

where Λ_I is a $p_0 \times K$ matrix of λ_i that load nontrivially on the factors. Hence

$$X_t : \begin{cases} \text{informative predictors:} & X_{I,t} = \Lambda_I f_t + u_{I,t}, \quad \dim(X_{I,t}) = p_0 \\ \text{noises:} & X_{N,t} = u_{N,t}, \quad \dim(X_{N,t}) = p - p_0. \end{cases}$$

Here for simplicity of presentation, we assume that $X_{I,t}$ are the first p_0 elements of X_t, but we do not explicitly require the identities of $X_{I,t}$ be known in practice.

Therefore, there are two types of predictors: the informative predictors are the collection of $x_{i,t}$ which meaningfully load on the latent factors with nontrivial loading matrix Λ_I, and the pure noises whose loadings are zero, and do not carry predictive information regarding y_t. For ease of presentation, let's assume for now that the factors are strong among the informative predictors, in the sense that

$$c < \sigma_K\left(\frac{1}{p_0}\Lambda_I'\Lambda_I\right) < ... < \sigma_1\left(\frac{1}{p_0}\Lambda_I'\Lambda_I\right) < C,$$

for some constants $C, c > 0$, where $\sigma_i(\cdot)$ denotes the i th largest eigenvalue. (Our general result will weaken this condition.)

We allow the partition (identities of informative predictors and noises) to be either known or unknown, and we do not distinguish the two cases. As will become clear later, our methodology treats them equally, and we refrain from employing any variable selection procedures to screen off the noises. Importantly, the inclusion of pure noises $X_{N,t}$ in the collection of predictors is motivated by the following two scenarios:

Scenario I: weak factor models. In economic forecasts, researchers often collect a large number of predictors which are *potentially* correlated with the outcome variable to forecast. However, the factor-model may be *weak* in the sense that many of these predictors may not depend on the latent factors, but are merely idiosyncratic noises. Their spurious predictive power arises from correlations with informative predictors through the idiosyncratic noises. Consequently, conditioning on the informative predictors, $X_{N,t}$ no longer have predictive power. Meanwhile, economists are typically unaware of which predictors among the collected set meaningfully depend on the factors and which do not. As a result, even though the factors are strong within the informative predictors, their overall influence is weak among all the predictors, because

$$\sigma_1\left(\frac{1}{p}\sum_{i=1}^{p}\lambda_i\lambda_i'\right) = \sigma_1\left(\frac{1}{p}\Lambda_I'\Lambda_I\right) \asymp \frac{p_0}{p} \to 0$$

which may decay fast if p is much larger than p_0. This is the case where the partition $(X_{I,t}', X_{N,t}')$ is unknown.

Scenario II: intentionally included noises. Suppose economists have a priori knowledge that the first p_0 collected predictors are all informative, while the remaining

$p - p_0$ are noises. Traditional statistical wisdom would suggest using only the informative predictors and excluding all the noises. However, as a novel finding in this paper, we shall argue that both the informative and noise predictors should be retained in the model for predictions. Particularly, if $X_{I,t}$ is not sufficiently numerous (in the sense that p_0 is not much larger than the sample size), we should *intentionally* include many pure noises $X_{N,t}$, so that the total number of predictors $X_t = (X_{I,t}, X_{N,t})$ is much larger than the sample size. A key contribution of this paper is to establish the wisdom of artificially including many noises. This is the case where the partition $(X'_{I,t}, X'_{N,t})$ is known.

The asymptotic regime. We require $p_0 \to \infty$, but $p - p_0 = \dim(X_{N,t})$ can be either a bounded constant (or zero) corresponding to the case that most (or all) predictors are informative, or $\dim(X_{N,t}) \to \infty$ much faster than p_0 and n, corresponding to the case of many pure noises. In addition, we explicitly require the total number of predictors p be much larger than the sample size: $p/n \to \infty$. In the case that the number of collected predictors are not that many, this means one can intentionally add pure noises so that $p/n \to \infty$.

2.2 The working model

While (2.1) underlines the true DGP, it is not feasible as the factors are latent. Therefore, a standard approach would first estimate the latent factors from model (2.2) using principal components analysis (PCA) and proceed to forecast using the estimated factors (e.g., Connor and Korajczyk (1988); Stock and Watson (2002); Bai and Ng (2006)).

We shall proceed differently. We directly use the collected predictors X_t to forecast, potentially with the inclusion of many noises $X_{N,t}$, by working with the following working model:

$$y_t = X_t'\beta + e_t = \sum_{i=1}^{p} x_{i,t}\beta_i + e_t. \tag{2.4}$$

The model (2.4) is estimated using pseudoinverse ordinary least squares (pseudo-OLS):

$$\widehat{\beta} = \left(\sum_{t=1}^{n} X_t X_t'\right)^{+} \sum_{t=1}^{n} X_t y_t, \tag{2.5}$$

Let $\mathcal{OOS}$ denote a set of out-of-sample predictors, where we observe $X_{\text{new}} \in \mathcal{OOS}$. We forecast its outcome variable using

$$\widehat{y}_{\text{new}} = X'_{\text{new}}\widehat{\beta}. \tag{2.6}$$

In the one-step-ahead forecast exercise, this corresponds to forecasting Y_{T+1} using $X_T'\widehat{\beta}$.

Recall that $\dim(X_t) = p \gg n$. So in the definition of the pseudo-OLS, $(\sum_{t=1}^n X_t X_t')^+$ denotes the pseudo-inverse of the design matrix[1], which is well defined regardless of the p/n ratio, and becomes the ordinary least squares if $p < n$. This estimator is also known as "ridge-less regression", as is shown by Hastie et al. (2022):

$$\widehat{\beta} = \lim_{\lambda \to 0^+} \left(\sum_{t=1}^n X_t X_t' + \lambda I \right)^{-1} \sum_{t=1}^n X_t y_t.$$

Due to its connection with ridge regression, some scholars also consider $\widehat{\beta}$ as an "implicitly regularized estimator," as the ridge penalty decays to zero. However, we *do not* take a stand on this viewpoint; instead, we regard $\widehat{\beta}$ as an *anti-regularized* estimator. This perspective arises from a crucial fact of complete in-sample interpolation: if $p \geq n$,

$$y_t = X_t'\widehat{\beta}, \quad t = 1, ..., n \quad \text{(all in-sample data)}.$$

This is because $X(X'X)^+X' = I$ if $p \geq n$. Hence in the ultrahigh-dimensional regime, the in-sample data are perfectly interpolated. In contrast, we consider the term "regularization" as a methodology intentionally employed to prevent complete in-sample interpolation.

Computations. To efficiently compute $\widehat{\beta}$, respectively write X and Y as the $n \times p$ and $n \times 1$ matrix and vector of X_t and y_t. Then $\widehat{\beta} = (X'X)^+X'Y$, where $(X'X)^+$ is the pseudo-inverse of a $p \times p$ dimensional matrix. In many numerical studies, we would expect p to be of several thousands or even larger, so it is recommended to use the "reduced singular value decomposition" (reduced-SVD) to efficiently compute the Moore-Penrose pseudoinverse: Let S_n denote the $n \times n$ diagonal matrix of nonzero singular values of X; let U_n denote the $n \times n$ matrix of the left singular vectors, and V_n denote the $p \times n$ matrix of right singular vectors corresponding to the top n singular values. Then

$$\widehat{\beta} = V_n S_n^{-1} U_n Y.$$

This only requires computing the reduced SVD instead of the full-sized SVD, and is much faster than the usual pseudoinverse functions in leading computing software, such

[1]The pseudoinverse (or Moore–Penrose inverse) of a symmetric matrix A is defined as $A^+ = U_1 D_1^{-1} U_1'$, where D_1 is a diagonal matrix consisting of non-zero eigenvalues of A, and U_1 is the matrix whose columns are the eigenvectors corresponding to the nonzero eigenvalues.

as '`pinv`' in Matlab. [2]

2.3 The fundamental question and main results

The objective of this paper is to answer the following fundamental question:

Does pure noise truly lack predictive power?

In other words, even if the identities of $X_{N,t}$ are known, should we really exclude them from the set of predictors? Consider a hypothetical scenario, where an economist has a priori knowledge that the first p_0 of the collected predictors are all informative, and the rest $p - p_0$ are pure noises, and that $p_0 < n$. Then the well-accepted statistical wisdom would naturally guide us to exclude the noises, and only use the informative predictors. Namely to predict y_t using the following model:

$$y_t = X'_{I,t}\beta_I + e_t = \sum_{i=1}^{p_0} x_{i,t}\beta_i + e_t.$$

The coefficients can be estimated using either OLS or the ridge regression. Indeed, we shall show that if the idiosyncratic components in u_t in (2.2) are mutually uncorrelated, then the true latent-factor based DGP induces the following linear regression model:

$$y_t = X'_{I,t}\beta_I + X'_{N,t}\beta_N + e_t, \quad \text{with } \beta_N = 0 \text{ if } \mathrm{Cov}(u_t) \text{ is diagonal.} \tag{2.7}$$

Therefore, since the identities of $X_{I,t}$ and $X_{N,t}$ are known in this hypothetical scenario, it should be an unanimous agreement that one should exclude $X_{N,t}$ from the forecast.

Perhaps surprisingly, we shall argue that for most economic forecasting problems, if $p_0 < n$, noises should be always retained in the forecast model. In fact, the objective of this paper is to argue for the following **practical recommendations**:

I. *If it is believed that X_t is a collection of both informative predictors and pure noises and that the informative predictors have predictive power through the latent factors, then economists should retain all the predictors and use pseudo-OLS (2.5)-(2.6), instead of attempting variable selections.*

II. *If the total number of predictors in X_t is not sufficiently large, then economists should intentionally add more pure noises until p is sufficiently large.*

[2] Computing the reduced SVD can be very fast as long as n is not large. The function is `U, S, V = np.linalg.svd(X, full_matrices=False)` in Python, and is `[U, S, V] = svd(X,'econ')` in Matlab. Alternatively, one can use `beta = np.linalg.svd(X)@Y` in Python, because $(X'X)^+X' = X^+$.

The crucial insight lies in the observation that, supported by a few latent factors, the predictive signals are densely distributed among high-dimensional coefficients. Consequently, the overall variance of the forecast is *diversified away*, even when a significant proportion of predictors constitute pure noises. Meanwhile, the dense predictive signals also maintain the forecast bias at a modest level."

Let p_0 and p respectively denote the number of informative predictors and all predictors (informative and noises). We shall prove the following **main results**:

1. if (p_0, p, n) satisfy (where $p - p_0$ is the number of noises in the predictor set):

$$\frac{n}{p} \to 0, \quad \frac{p}{p_0 n} \to 0,$$

   then the pseudo-OLS can achieve the *oracle predictive risk*, that is, its predictive mean squared error (MSE) asymptotically converges to that of the latent factors f_t, as if the factors were directly used for forecast.

2. In the case where $p_0/n \to c \in (0, 1)$, that is, there are "many but not sufficiently many" informative predictors, the predictive MSE of both OLS and the ridge regression are strictly larger than the oracle predictive risk, therefore are sub-optimal.

These results provide a direct answer to the fundamental question we raised in the beginning of this paper: while the idiosyncratic noises do not have direct predictive power, they actually have *indirect* predictive power, by diversifying away the out-of-sample variances, and hence reducing the predictive MSE. The compelling implication of this result is that the inclusion of noises in predictions yields greater benefits than its exclusion.

Contrary to conventional statistical wisdom, which asserts that overfitting significantly undermines forecast performance by inflating out-of-sample variance, we show that this is no longer the case when a substantial number of additional predictors are included. Instead, the analysis moves into the new regime of the recent *double descent* phenomenon on the prediction risk, which has gained increasing attention in the recent machine learning literature. That is, as the model complexity exceeds the sample size and continues increasing, a second descent of the prediction occurs in the extremely over-parametrized regime. It was first illustrated in the empirical work by (Belkin et al., 2019; Hastie et al., 2022; Arora et al., 2019), and was theoretically studied in linear models with ridge regressions, e.g., (Mei and Montanari, 2019; Belkin et al., 2020).

While our model shares similarities with recent theoretical developments in linear forecast models, we adopt a novel viewpoint by focusing on the effect of noises in double descent. Our approach is well motivated by the context of economic forecasts. In this context, predictive information is embedded within a few latent factors and is conveyed through a collection of informative predictors, along with a substantial number of spurious predictors—essentially, pure idiosyncratic noises.

2.4 Why not Lasso or PCA

When the collection of predictors contains many noises, the Lasso is one of the most popular forecasting methods, as the use of ℓ_1-penalty can often effectively remove the noises thus achieve dimension reductions. But this is no longer the case in many economic forecasting exercises where the informative predictors carry predictive information through latent economic factors, often as macroeconomic variables and state variables. In this case, even though our working model has a seemingly sparse representation as in (2.7), there are two key features that differentiate it from the usual setting of the Lasso forecasts: first, the latent factors introduce strong cross-sectional dependences among the informative predictors, causing strong colinearity and substantially slowing down the statistical rate of convergence for Lasso (e.g., Fan et al. (2020); Hansen and Liao (2018)). Secondly, the latent factors make the predictive signals *densely* distributed among many informative predictors, under which Lasso cannot effectively select variables to well represent the others Chernozhukov et al. (2017); Giannone et al. (2021). In other words, the model is not sparse enough.

Meanwhile, the PCA is another popular forecasting method that is particularly attractive for economic data because it effectively extracts the latent "indices" (or factors) from the big economic data. But its key limitation is that the quality of the estimated factors critically depends on the strength of the factors, in our notation, p_0. When $p_0/p \to 0$ fast, however, it is practically difficult to correctly estimate K, the number of factors. Even if K is correctly specified, factors are still estimated poorly because signals in the informative predictors are contaminated by too many noises.

In contrast, the pseudo-OLS forecast, which we recommend in this paper, does not require variable selections or determining the number of factors, also robust to weak factors. It works relatively well so long as $p_0 \to \infty$ (sufficient informative predictors) and p is large. When p is not large enough, just add noises!

3 Asymptotic Results

This section formally presents the main results, under more general conditions regarding (p_0, p, n) and the factor strength. Specifically, we assume the following DGP on X_t:

Assumption 1 (DGP on X_t)**.** *The p-dimensional X_t can be decomposed as:*

$$X_t = \begin{pmatrix} X_{I,t} \\ X_{N,t} \end{pmatrix} = \begin{pmatrix} \Lambda_I \\ 0 \end{pmatrix} f_t + u_t, \tag{3.1}$$

where $\dim(X_{I,t}) = p_0$ *and* $\dim(f_t) = K$.

(i) The factor-strength among $X_{I,t}$ is denoted by $\psi_{p,n}$. That is, there is a sequence $\psi_{p,n} \to \infty$, $\psi_{p,n} = O(p_0)$, such that

$$\sigma_K\left(\Lambda_I'\Lambda_I\right) \asymp \sigma_1\left(\Lambda_I'\Lambda_I\right) \asymp \psi_{p,n}.$$

(ii) $\mathbb{E}(\rho' f_t | X_t) = \beta' X_t$ for some $\beta \in \mathbb{R}^p$.

(iii) The top K eigenvalues of $\Lambda' \mathbb{E} f_t f_t' \Lambda / \psi_{n,p}$ are distinct.

This assumption allows the more general case under which $\psi_{n,p}/p_0$ may decay to zero, so that the informative predictors may be "semi-strong". In addition, we require $\mathbb{E}(\rho' f_t | X_t)$ be a linear function of X_t. The sufficient condition is that (f_t, u_t) follow a jointly normal distribution.

3.1 The signal distribution of economic forecasting models

Recall that we have defined two models:

$$y_t = \rho' f_t + \epsilon_{y,t}, \quad \text{(oracle model, the true DGP)} \tag{3.2}$$

$$y_t = X_t'\beta + e_t, \quad \text{(working model, the model for practical forecast).} \tag{3.3}$$

While we use the working model to replace the oracle model in practical forecast (because f_t is unknown), the first question to address is the gap between the two models. That is, if all of parameters, (β, ρ, f_t), could have been perfectly learned from the data, can the two models produce the same predictive MSE? Note that the predictive MSE would respectively converge to the marginal variances $\text{Var}(\epsilon_{y,t})$ and $\text{Var}(e_t)$ for the two models, hence the question is essentially asking whether the two residual variances are asymptotically the same.

The answer is affirmative as $p_0 \to \infty$, and the analysis relies on characterizing the coefficient β and the noise e_t using the latent factor model. In fact, we shall show that (3.3) is implied by (3.2) as follows. Let

$$\mathbb{E}X_t X_t' = \Lambda \mathbb{E} f_t f_t' \Lambda' + \mathrm{Cov}(u_t)$$

denote the $p \times p$ covariance matrix of X_t, where $\Lambda = (\Lambda_I', 0')'$.

Theorem 1. *Suppose (3.2) and Assumption 1 hold, with* $\mathbb{E}(\epsilon_{y,t}|f_t, u_t) = 0$, *and* $\mathbb{E}(u_t|f_t, \epsilon_{y,t}) = 0$. *Also suppose the eigenvalues of* $\mathrm{Cov}(u_t)$ *are bounded away from both zero and infinity. In addition,* $\|\rho\| \le C$. *Then (3.3) holds:*

$$y_t = X_t'\beta + e_t, \quad \mathbb{E}(e_t|X_t) = 0,$$

where:

(i) $\beta = \mathrm{Cov}(u_t)^{-1}\Lambda(\Sigma_f^{-1} + \Lambda' \mathrm{Cov}(u_t)^{-1}\Lambda)^{-1}\rho$.

(ii) As $\psi_{n,p} \to \infty$,

$$\mathrm{Var}(e_t) = \mathrm{Var}(\epsilon_{y,t}) + O(\psi_{p,n}^{-1}).$$

Result (ii) of Theorem 1 shows that if there are growing number of informative predictors that nontrivially load on the latent factors, then predicting using the working model (X_t-based) would asymptotically do as well as predicting using the oracle model (f_t-based).

In addition, this theorem also characterizes the high dimensional coefficient β on the collection of predictors. The expression in result (i) yields two implications: First, if $\mathrm{Cov}(u_t)$ is block-diagonal, in the sense that the idiosyncratic $u_{I,t}$ and $u_{N,t}$ are uncorrelated, then the coefficient $\beta_N = 0$, corresponding to the pure idiosyncratic noise. From this perspective, $X_{N,t}$ does not have direct predictive power regarding y_t. Secondly, β is a "dense" signal with a decaying ℓ_2-norm:

$$\|\beta\| \le \|\mathrm{Cov}(u_t)^{-1}\Lambda\| \|\rho\| \sigma_{\min}(\Lambda' \mathrm{Cov}(u_t)^{-1}\Lambda)^{-1} = O(\psi_{p,n}^{-1/2}). \tag{3.4}$$

This theoretically supports the empirical insights in Giannone et al. (2021), highlighting that in many economic forecast problems, signals do not concentrate on a single sparse model, but rather "on a wide set of models that often include many predictors." Therefore, forecasting based on sparse variable selections (e.g., Lasso) that concentrate on a few predictors could not fully capture the predictive power in the forecast coefficients.

3.2 Bias-variance of the pseudo-OLS with many intentional noises

3.2.1 The main results

Let $X_{\text{new}} \in \mathcal{OOS}$ denote a particular out-of-sample predictor, by which we forecast the outcome using the pseudo-OLS as $\widehat{y}_{\text{new}} := X'_{\text{new}}\widehat{\beta}$. Let the true out-of-sample outcome be generated as:

$$y_{\text{new}} = X'_{\text{new}}\beta + e_{\text{new}} = \rho' f_{\text{new}} + \epsilon_{y,\text{new}}.$$

The predictive MSE, conditioning on the in-sample data $X := (X_1, ..., X_n)'$, is given as

$$\mathbb{E}[(y_{new} - \widehat{y}_{\text{new}})^2|X] = \mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})^2|X] + \text{Var}(e_{\text{new}})$$

where we imposed the assumption that e_{new} is independent of the in-sample X.

Our previous theorem shows that working with the X_t-model would achieve the oracle forecast risk, i.e.,

$$\text{Var}(e_{\text{new}}) \to \text{Var}(\epsilon_{y,\text{new}}).$$

Therefore, it suffices to focus on the first component of the MSE, decomposed as:

$$\mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})^2|X] = \text{bias}(\widehat{y}_{\text{new}})^2 + \text{Var}(\widehat{y}_{\text{new}} - X'_{\text{new}}\beta|X).$$

where

$$\begin{aligned} \text{bias}(\widehat{y}_{\text{new}})^2 &= \beta' A_X \mathbb{E}X_t X_t' A_X \beta, \quad \text{where } A_X := (X'X)^+ X'X - I \\ \text{Var}(\widehat{y}_{\text{new}} - X'_{\text{new}}\beta|X) &= \sigma_e^2 \,\text{tr}(X'X)^+ \mathbb{E}X_t X_t' \times O_P(1). \end{aligned} \tag{3.5}$$

Assumption 2. *Let e denote the $n \times 1$ vector of in-sample e_t. Suppose:*
(i) $\mathbb{E}(\epsilon_{new} X_{new}|X, e) = 0$, $\mathbb{E}(e|X_{\text{new}}, X) = 0$.
(ii) $\mathbb{E}(X_{\text{new}}X'_{\text{new}}|X) = \mathbb{E}X_t X_t'$, *and* $\text{Var}(e|X_{\text{new}}, X) = \sigma_e^2 I$ *for some* $\sigma_e^2 > 0$.
(iii) $\text{Var}(e_{\text{new}}) = \text{Var}(e_t)$ *and* $\text{Var}(\epsilon_{y,\text{new}}) = \text{Var}(\epsilon_{y,t})$.
(iv) $\|\mathbb{E}(ee'|X, X_{\text{new}})\| = O_P(\sigma_e^2)$ *for some* $\sigma_e^2 > 0$.

In the appendix we prove that Assumption 2 (iii) is satisfied if $(f_t, \epsilon_{y,t}, u_t)$ are i.i.d. jointly normal.

Assumption 3. *Recall that $u_{i,t}$ is the idiosyncratic noise in $x_{i,t} = \lambda_i' f_t + u_{i,t}$, and U is the $n \times p$ matrix of $u_{i,t}$.*

(i) $u_{i,t}$ is independent and identically distributed (i.i.d.) across both (i, t)

(ii) $\mathbb{E}u_{i,t}^4 < C$, *and* $c < \min_{j \le p_0} \text{Var}(u_{j,t}) \le \max_{j \le p_0} \text{Var}(u_{j,t}) < C$ *for some* $C, c > 0$.

Assumption 3 simplifies the technical arguments by assuming the idiosyncratic components are i.i.d. over both cross-sectional and times. This yields a fast rate of convergence for the prediction MSE. In the appendix, we present a more general results by weakening the i.i.d. assumption.

We have the following theorem.

Theorem 2. *Suppose the assumptions of Theorem 1 and Assumption 2 hold. In addition, suppose* $p > n$. *Suppose* $p = o(\psi_{p,n} n)$ *and* $n = o(p)$.

(i) The forecast bias and variance:

$$\begin{aligned} \text{bias}(\widehat{y}_{\text{new}})^2 &= O_P\left(\frac{p}{\psi_{p,n} n}\right) \\ \text{Var}(\widehat{y}_{\text{new}} - X'_{\text{new}}\beta | X) &= O_P\left(\frac{1}{n} + \frac{n}{p}\right). \end{aligned}$$

(ii) Also,

$$\mathbb{E}[(y_{new} - \widehat{y}_{\text{new}})^2 | X] \to^P \text{Var}(\epsilon_{y,t}).$$

Our main results imply the following deep insights:

1. As the predictive MSE converges to $\text{Var}(\epsilon_{y,t})$, this theorem shows that we can achieve the oracle prediction MSE, as if the latent factors were used directly to forecast.

2. The bias goes to zero as long as $p = o(\psi_{n,p} n)$, which requires sufficient predictive power carried by the informative predictors. In particular, if all the informative predictors are strong, i.e., $\psi_{p,n} \asymp p_0$. Then the pseudo-OLS prediction is asymptotically unbiased as long as the number of informative predictors satisfies:

$$p_0 \gg \frac{p}{n}.$$

   Meanwhile, the variance decays as $n/p \to 0$. This means, we can reduce the variance by intentionally adding $\kappa_p := p - p_0$ noises, which is still consistent as long as

$$\kappa_p \ll p_0 n.$$

3. Unlike the traditional wisdom that the variance is amplified as p diverges, here the variance goes to zero as long as $p \gg n$, that is, the total number of predictors

grows faster than the sample size. Importantly, the decay of variance itself does not require conditions on the predictive power, i.e., it does not matter whether the predictors are mostly noises or informative. As we keep adding predictors so that p increases, the variance would continue decreasing until the first term $1/n$ becomes dominating, even if most of the added predictors are pure noises.

4. Therefore, we have reached a reasonable explanation of our striking result that including high-dimensional noises is more beneficial than removing the noises. Once the in-sample data are perfectly interpolated, the role of adding pure noises is to diversify away the variance, while the role of keeping informative predictors (e.g., informative macroeconomic variables) is to reduce the bias.

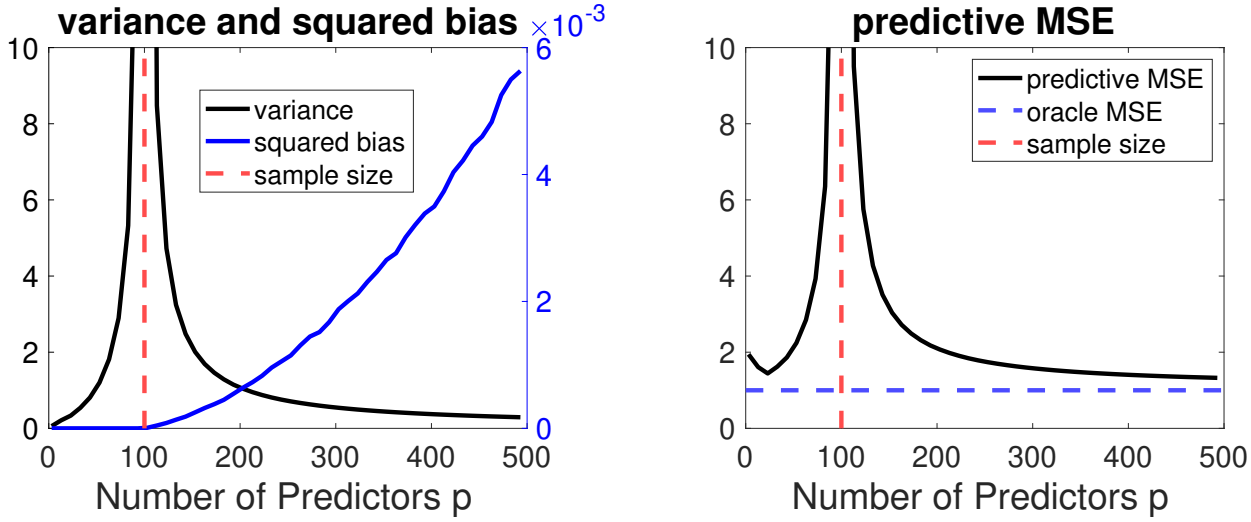

Figure 1: Theoretical predictive variance and squared bias (left panel) and MSE (right panel), averaged over 500 replications. The horizontal axis is the number of predictors increasing from 3 to 500, and we fix $n = 100$. The first $p_0 = \min\{p, 0.9n\}$ are informative predictors, generated using a 3-factor model of strong factors. The remaining $p - p_0$ are i.i.d. Gaussian noises. The vertical dashed line is where p equals n, and the horizontal dashed line on the right panel refers to $\text{Var}(\epsilon_{y,t})$, the oracle predictive MSE.

Figure 1 plots the theoretical curves of bias-variance (left panel) and predictive MSE $\mathbb{E}[(y_{new} - \widehat{y}_{\text{new}})^2|X]$ (right panel) based on (3.5), as the total number of predictors increases, in a 3-factor model. Here the first $p_0 = \min\{p, 0.9n\}$ predictors are informative, while the remaining $p - p_0$ predictors are i.i.d. white noises generated from $N(0, 1)$. As is clearly illustrated, the variance monotonically increases as p increases even though the first p_0 added predictors are all informative, and peaks at $p = n$ where the in-sample data are perfectly interpolated. Meanwhile, after $p > 0.9n$, the added predictors are pure

noises, and the variance starts to decay after $p > n$. This is consistent with our theory: as $p \to \infty$, the variance would continue decreasing until $1/n$ becomes the dominating term.

In addition, the squared bias is constantly zero until $p = n$ and starts to increase, but in a much smaller magnitude. This is also consistent with what the theory predicts. In this case

$$\text{bias}(\widehat{y}_{\text{new}})^2 = O_P\left(\frac{p}{0.9n^2}\right).$$

The bias depicted on the left panel of Figure 1 does not seem diminishing because in this design $\psi_{p,n} \sim 0.9n$ is fixed ($n = 100$) and we only vary p in the plot.

Overall, the predictive MSE (right panel) illustrates a double-descent phenomenon, where the first descent occurs before $p < 20$, due to the decay of the gap $\text{Var}(e_t) - \text{Var}(\epsilon_{y,t})$. The second descent occurs after $p > n$, due to the decay of variance, and eventually the MSE decays to the oracle MSE as if the latent factors were used for prediction.

3.2.2 The intuition

We now provide the intuitions of the surprising result that both bias and variance diminish when artificial noises are added as predictors.

We start with discussing the bias. Recall that

$$\text{bias}(\widehat{y}_{\text{new}})^2 = \|(\mathbb{E}X_t X_t')^{1/2} V_{-n} V_{-n}' \beta\|^2$$

where V_{-n} is $p \times (p-n)$ with columns being the eigenvectors corresponding to the $p-n$ zero-eigenvalues of $X'X$. Let

$$(\mathbb{E}X_t X_t')^{1/2} = \bar{V}_K D_K \bar{V}_K' + \bar{V}_{-K} D_{-K} \bar{V}_{-K}'$$

be the SVD of $(\mathbb{E}X_t X_t')^{1/2}$, decomposed into two parts corresponding to the first K eigenvectors/values $(\bar{V}_K, D_K)$ and last $p-K$ eigenvectors/values $(\bar{V}_{-K}, D_{-K})$. We then decompose: $\text{bias}(\widehat{y}_{\text{new}})^2 \leq 2A_1 + 2A_2$, where

$$A_1 := \underbrace{\|\bar{V}_K D_K \bar{V}_K' V_{-n} V_{-n}' \beta\|^2}_{\text{factor-model design}}, \quad A_2 := \underbrace{\|\bar{V}_{-K} D_{-K} \bar{V}_{-K}' V_{-n} V_{-n}' \beta\|^2}_{\text{i.i.d. design}}.$$

The intuitions for both terms decreasing to zero correspond to two scenarios of the design matrix, respectively being the factor-model design and the classic i.i.d. design.

As for term A_1, by the Sin-theta theorem, the strong eigenvalues ensure that $\bar{V}_K$ can be well estimated by the top K eigenvectors of the sample covariance $X'X$, and the latter are orthogonal to V_{-n}. Therefore, the Sine-Theta theorem implies

$$\|\bar{V}_K' V_{-n}\|^2 = O_P(\frac{p}{\psi_{p,n} n}).$$

This implies

$$A_1 = O_P(\frac{p}{\psi_{p,n} n}).$$

Meanwhile, A_2 involves the remaining bounded eigenvalues. Even though $(\bar{V}_{-K}, D_{-K})$ cannot be consistently estimated when p is large, the intuition is aligned with the classic case of i.i.d. design: the fact that $\|D_{-K}\|$ is bounded and the dense β implies

$$\|A_2\|^2 \leq O(1)\|\beta\|^2 = O(\psi_{p,n}^{-1}).$$

This ensures that the bias decreases as long as $p = o(\psi_{p,n} n)$. In the case $p_0 \sim \psi_{p,n}$, i.e., factors are strong among the informative predictors, this means the bias diminishes as long as the number of added noises $\kappa = p - p_0$ should not exceed $p_0 n$.

As for the variance, consider a special case that *all predictors are noises*, in which case $\Lambda = 0$, so the variance becomes:

$$\mathrm{Var}(\widehat{y}_{\mathrm{new}} - X_{\mathrm{new}}'\beta | X) = \sigma_e^2 \sigma_u^2 \,\mathrm{tr}[(U'U)^+],$$

where U is the $n \times p$ matrix of idiosyncratic and artificial noises. Essentially, this is a problem in the classic asymptotic regime if we switch the role between p, n by noting that if $n/p \to 0$, then

$$\mathrm{tr}[(U'U)^+] = \mathrm{tr}[(UU')^{-1}].$$

The high-dimensional $p \times p$ matrix $(U'U)^+$ is replaced by the "relatively high-dimensional" $n \times n$ matrix $(UU')^{-1}$. Hence by switching the role of n and p, we can analyze this term using the usual asymptotic theory and reach $\mathrm{tr}[(U'U)^+] = O_P(n/p)$. [3] Consequently, the high-dimensionality p helps to diversify away the variance even if all predictors are noises.

[3] Applying the Marchenko-Pastur law Marchenko and Pastur (1967) or Theorem 2 of Bai and Yin (1993), when $u_{i,t}$ are i.i.d in both dimensions the nonzero eigenvalues of $\left(\frac{U'U}{p}\right)$ is bounded away from zero if $p, n \to \infty$. The order of the variance is hence determined by $\frac{n}{p}$.

3.3 Discussions

3.3.1 The bias from many intentional noises

Importantly, both Figure 1 and our theory raise a warning flag about the rising bias when p becomes too large. That is, the bias increases with p as:

$$\text{bias}(\widehat{y}_{\text{new}})^2 = O_P\left(\frac{p}{\psi_{p,n}n}\right),$$

where $\psi_{p,n}$ denotes the factor strength within those p_0 informative predictors. We therefore give the following practical recommendations to address this issue.

1. First, in some applications most of the collected predictors are actually informative, and the number of informative predictors are much larger than the sample size. For instance, when forecasting assets' returns using individual level stock returns, almost all predictors depend on the common factors (e.g., the market index and other nontraded factors). In this case, $p \asymp p_0 \asymp \psi_{p,n} \gg n$, and hence $\text{bias}(\widehat{y}_{\text{new}})^2 = O_P\left(\frac{1}{n}\right)$ which does not increase with p. Hence the bias is not a concern. In addition, as $p \gg n$, there is no need to intentionally add pure noises, as there are sufficiently many informative predictors to diversify away the variance.

2. Secondly, when p_0 is only mildly large (i.e., $p_0/n \to c \in (0,1)$), but it is believed that most of the informative predictors load strongly on the common factors, then $p_0 \asymp \psi_{p,n}$, we then recommend intentionally add $p - p_0$ pure noises, and merge them fore prediction. Then

$$\mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})^2|X] = O_P\left(\frac{p}{p_0 n} + \frac{n}{p}\right) + O_P(n^{-1}).$$

   We keep adding noises until the order of p minimizes the sum of the first two terms:

$$p = C \times n\sqrt{p_0} = \arg\min_p \left(\frac{p}{p_0 n} + \frac{n}{p}\right),$$

   where C is a constant to be chosen. This would also yield the optimal rate of convergence $\mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})^2|X] = O_P(p_0^{-1/2} + n^{-1})$. In finite sample, we recommend choosing $C > 2p_0^{-1/2}$ so that $p > 2n$ and by cross-validation. In practice, tuning the constant C is straightforward due to its insensitivity. In the numerical studies section, we demonstrate how to tune it through cross-validations.

3. Instead of adding pure noises, alternative approaches to intentionally increasing the number of predictors involve adding transformations to the informative predictors $X_{I,t}$, such as polynomials, lags, and interactions. These transformations aim to include nonlinear predictive information, potentially mitigating bias inflation while still diversifying the variance. While this strategy may improve the performance in practice, we recommend this procedure with reservations. From an economic interpretation standpoint, it is hard to explain whether the forecast improvement is due to the nonlinear predictability or simply because of the variance diversification. Consequently, we perceive the addition of nonlinear transformations as being *more artificial* than including pure noises.

3.3.2 Comparison with linear double descent literature

Our study distinguishes itself from related linear models in the statistical literature in three crucial aspects.

Firstly, many economic objectives for forecasts are inherently linked to a few latent factors determining the economic status of the forecasting environment. Inspired by this, we assume that model (3.2) is the true DGP, while treating model (3.3) as a working model that is induced by the true DGP due to a linear representation assumption. Therefore, the observed high-dimensional predictors carry the predictive power from their explainability by these economic factors, as opposed to causing the outcome variable. A direct consequence is the prediction coefficients being densely distributed across predictors, invalidating classical forecasting methods based on dimension reductions. This observation is also closely aligned with empirical findings in Giannone et al. (2021).

Secondly, the informative predictors that are well explainable by the latent factors are highly mutually correlated, causing a distinctive approximate low-rank representation in the predictor covariance matrix, (e.g., Bai (2003); Fan et al. (2013)). There are a few eigenvalues growing fast to infinity, and the remaining eigenvalues are bounded. This structure is typically excluded by the random matrix theory in the recent statistical literature addressing the extreme overparametrization (e.g., Hastie et al. (2022)). [4]

Lastly, we specifically focus on the impact of including a substantial number of noises on economic forecasts and arrive at seemingly surprising results—adding noises proves to

[4] Hastie et al. (2022) derived the formula of predictive mean squared errors, under a case which they called "latent space model". Their model appears similar to the latent-factor model being considered in this paper, but with a critical difference that eigenvalues of the covariance matrix X_t are assumes to be bounded. Such an assumption is hardly satisfied in economic forecasting models with latent factors, where the top K eigenvalues of $\text{Cov}(X_t)$ should increase at order $\sqrt{n\psi_{p,n}}$.

be advantageous rather than detrimental.

3.4 Sub-optimality using only informative predictors

To further shed light on the benefits of including uninformative/pure noises as predictors for variance reductions, we now analyze the behavior of benchmark forecast methods when the informative predictors are perfectly known but 'not sufficiently many'. The analysis is guided by the traditional statistical wisdom of the bias-variance tradeoff.

Suppose it is perfectly known to the economists that only the first p_0 of the predictors, $X_{I,t}$, load nontrivially on the latent factors, and that the remaining $p - p_0$ predictors are pure noises, that is, the following working model is practically feasible:

$$y_t = X'_{I,t}\beta_I + e_t, \quad \dim(X_{I,t}) = p_0 < n. \tag{3.6}$$

Then excluding the noises, but just using $X_{I,t}$ to forecast, is seemingly the "right way" to go. Consider the setting where $p_0/n \to c \in (0,1)$. Because $p_0 < n$, both the OLS and ridge regression are well defined:

$$\begin{aligned} \text{OLS}: &\quad \widehat{y}_{\text{new,ols}} = X'_{I,\text{new}}(X'_I X'_I)^{-1} X'_I Y \\ \text{Ridge}: &\quad \widehat{y}_{\text{new,ridge}} = X'_{I,\text{new}}(X'_I X'_I + \lambda I)^{-1} X'_I Y \end{aligned}$$

where $X_{I,new}$ is a p_0-dimensional out-of-sample observation of only the informative predictors.

We now show that in this setting, neither OLS nor the ridge regression could achieve the oracle forecast. In contrast, using the pseudo-OLS with intentionally added noises would achieve the optimal forecast.

Theorem 3 (Only Informative predictors)**.** *Suppose economists directly forecast y_t using model (3.6) by either OLS or ridge regression with ℓ_2- penalty (λ). Suppose assumptions of Theorem 2 hold, and $p_0/n \to c \in (0,1)$.*

(i) The forecast MSE of OLS and ridge:

$$\liminf_{n,p_0} \mathbb{E}[(y_{\text{new}} - \widehat{y}_{\text{new,ols}})^2 | X] \quad > \quad \text{Var}(\epsilon_{y,t})$$

$$\liminf_{n,p_0} \inf_{\lambda \in \mathbb{R}} \mathbb{E}[(y_{\text{new}} - \widehat{y}_{\text{new,ridge}})^2 | X] \quad > \quad \text{Var}(\epsilon_{y,t}).$$

(ii) In contrast, the pseudo-OLS $\widehat{y}_{\text{new}}$ satisfies:

$$\mathbb{E}[(y_{\text{new}} - \widehat{y}_{\text{new}})^2|X] \to^P \text{Var}(\epsilon_{y,t}).$$

Result (ii) is simply a restatement of Theorem 2, presented here to contrast with the suboptimality of the OLS and ridge regression. The fundamental issue with the latter methods is that when p_0 is not sufficiently large, even if all the predictors are informative, the predictor behaves as in the traditional asymptotic regime, which would suffer from the classic overfitting issue. Specifically, the bias and variance of the ridge regression have the following lower bounds:

$$\begin{aligned}
\text{bias}(\widehat{y}_{\text{new,ridge}})^2 &\geq \left(\frac{\lambda}{\sigma_{K+1}(X_I'X_I)+\lambda}\right)^2 \frac{\|\rho\|_2^2}{\max_{j\leq p_0}\text{Var}(u_{j,t})} \\
\text{Var}(\widehat{y}_{\text{new,ridge}} - X_{\text{new}}'\beta|X) &\geq \text{Var}(e_t)\min_{j\leq p_0}\text{Var}(u_{j,t})\,(p_0-K)\min_{K<j\leq p_0}\frac{\sigma_j(X_I'X_I)}{(\sigma_j(X_I'X_I)+\lambda)^2}.
\end{aligned}$$

While the ridge regression attempts to properly choosing the penalty λ to balance the bias and variance, we show that there is no λ that make both bias and variance simultaneously decay to zero. In addition, the OLS is a special case by setting $\lambda = 0$, whose variance is inflated by p_0:

$$\text{Var}(\widehat{y}_{\text{new,ols}} - X_{\text{new}}'\beta|X) \geq \text{Var}(e_t)\min_{j\leq p_0}\text{Var}(u_{j,t})\frac{(p_0-K)}{\sigma_{K+1}(X_I'X_I)} \gtrsim \frac{p_0}{n} > c.$$

He (2023) considered a dense factor augmented model in the asymptotic regime $p_0/n \to c \in (0,1)$, and showed that the ridge regression is optimal among a set of regularized estimators. Theorem 3 complements his results by showing that despite being optimal, in this regime Ridge is out-of-sample inconsistent, and cannot reduce the prediction risk to the oracle level. In contrast, if we jump out of the regime by intentionally adding many noises so that $p/n \to \infty$, the simple pseudo-OLS achieves the oracle risk.

4 Simulations

We demonstrate the performance of intentional inclusion of white noises for forecasting using Monte Carlo experiments. The outcome variable is generated using a 3-factor model: $y_t = \rho' f_t + \epsilon_{y,t}$. In addition, we generate p_0 informative predictors and $p - p_0$

noises:

$$x_{i,t} = \begin{cases} \lambda_i' f_t + u_{i,t}, & i = 1, ..., p_0. \\ u_{i,t}, & i = p_0 + 1, ..., p \end{cases}, \quad \lambda_i = \lambda_{i,0} \times p_0^{-\tau},$$

where $(f_t, \epsilon_{y,t}, \lambda_{i,0}, u_{i,t})$ are all standard normal. Here $\tau \in [0, 1/2]$ determines the strength of the factors within the informative predictors, so that $\Lambda_I'\Lambda_I \asymp \psi_{p,n} \asymp p_0^{1-2\tau}$, the larger τ, the weaker the factors. We set p to take values on a grid that are evenly spaced from 1 to $p_{\max} = 1000$. These generated $x_{i,t}$ are to be used to fit forecasting models, and evaluated at additional 50 testing predictors X_{new} to predictor their out-of-sample outcomes.

Let us consider two scenarios in the simulation study, where the identities of informative predictors are known in one scenario and unknown in the other.

4.1 Unknown identities of informative predictors

Suppose the economist does not know which predictors are informative and which are not, so she decides to use all the collected predictors (including both informative ones and the $p - p_0$ noises). We set two values for $\tau \in \{1/2, 1/4\}$, where $\tau = 1/2$ corresponds to very weak factors (i.e., $\Lambda_I'\Lambda_I \asymp 1$), and $\tau = 1/4$ corresponds to relatively strong factors (i.e., $\Lambda_I'\Lambda_I \asymp p_0^{1/2}$).

Three methods are compared in this study: (i) The pseudo-OLS; (ii) The PCA, where the number of factors and the factors are estimated using all the p predictors. Using the in-sample estimated $\widehat{\lambda}_i$, we estimate the out-of-sample factors $\widehat{f}_{\text{new}}$ by regressing X_{new} on $\widehat{\lambda}_i$, and forecast the outcome variables using $\widehat{f}_{\text{new}}$; (iii) The Lasso, whose penalty is chosen by 10-fold cross-validation.

Figure 8 plots the predictive MSE, averaged over 50 replications, as the number of predictors p increases. This means if $p \leq p_0$, all predictors being used are informative, whereas $p - p_0$ noises are included when $p > p_0$. The red dashed vertical line in each panel indicates the number of informative predictors. We present more cases in the appendix. Summarizing results in Figure 8 we can conclude the following numerical findings:

1. We observe a typical double descent phenomenon on the predictive MSE curve for pseudo-OLS when the factors are relatively strong ($\tau = 1/4$). The first descent occurs when $p < n$, which lies in the traditional wisdom of bias-variance tradeoff. As p keeps increasing, the in-sample data become complete interpolating (perfectly over-fitting). Then the predictive MSE for the pseudo-OLS starts decaying. In most of the presented cases, the pseudo-OLS is the best or one of the best among

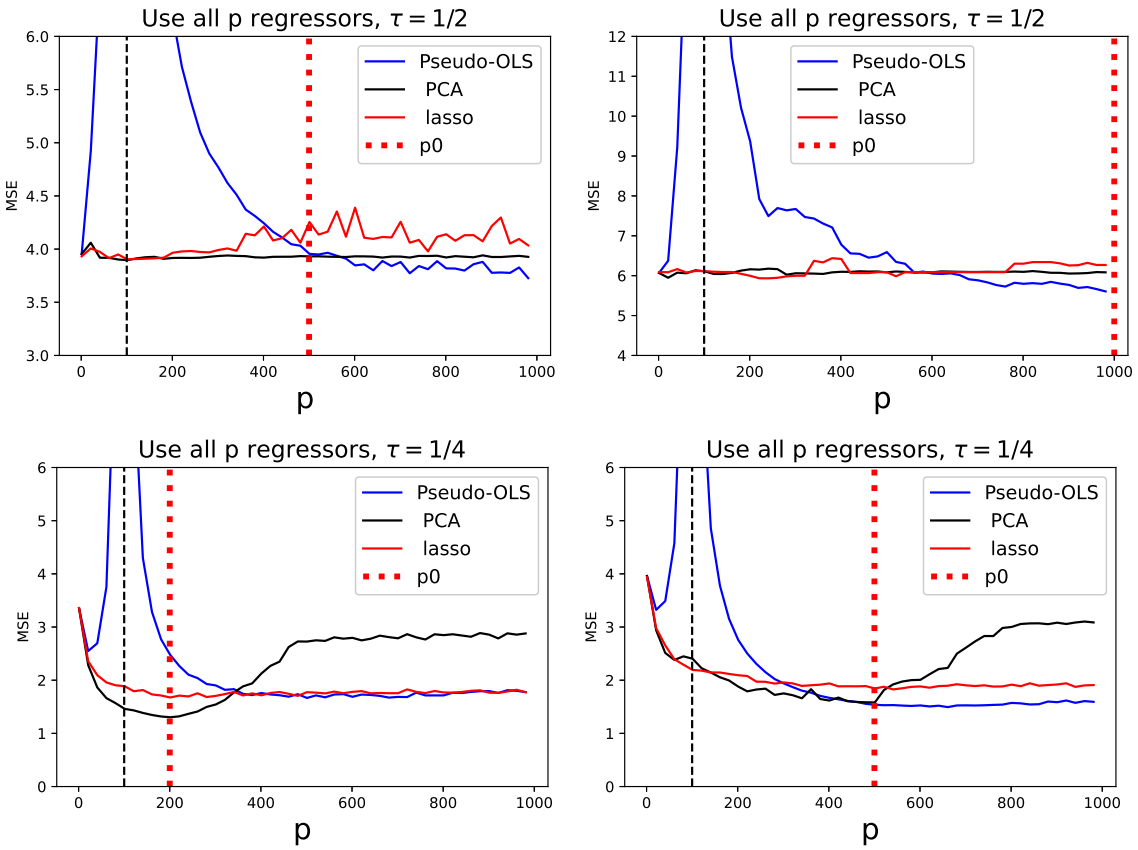

Figure 2: Predictive MSE $\sum_{j=1}^{50}(y_j - \widehat{y}_j)^2$ averaged from 50 replications as the number of predictors p increases. The vertical red dashed line indicates the number of informative predictors p_0; the black dashed line indicates the sample size $n = 100$.

the three methods when p is very large.

2. In the second panel where all predictors are informative ($p_0 = 1000$) but also very weak ($\tau = 1/2$), the pseudo-OLS continuously benefits from the inclusion of informative predictors even though the predictors are very weakly informative. The trend of decay in its MSE does not vanish even when $p = 1000$, and is the best among the three. In contrast, the factors are so weak that the PCA does not benefit from the large p, whose predictive MSE curve is basically flat. In all other three panels, where p_0 stops increasing at some points but p continues increasing, the MSE of pseudo-OLS flats out.

3. The PCA works reasonably well when $p \leq p_0$, but it starts to become worse as noises are included in the predictors. This is noticeably pronounced when factors

are relatively strong: the MSE for PCA starts to increase after $p > p_0$, and becomes the worse among the three methods.

4. Lasso does well when the number of informative predictors are relatively small. As shown in the third panel of Figure 8 ($p_0 = 200, \tau = 1/4$), Lasso is slightly worse than PCA when $p < p_0$, but as more noises are added to the prediction, the working model becomes more sparse, from which Lasso benefits and eventually performs almost indistinguishable from the pseudo-OLS. But when $p_0 = 500, 1000$, the common dependences within predictors are stronger, leading to denser regression coefficients in the working model. Lasso then becomes worse. This is particularly evident in the first two panels, where Lasso performs the worst among the three methods as p reaches to 1000.

4.2 Known identities of informative predictors

We now consider the "striking case" where the economist knows which predictors are informative and which are not, and nevertheless she decides to keep all the predictors and intentionally add many white noises to implement the pseudo-OLS.

We compare four methods in this case: (i) The pseudo-OLS which uses all p predictors (when $p > p_0$, the additional predictors are noises); (ii) The PCA; (iii) The Lasso, (iv) The ridge. Except for the pseudo-OLS, all the other three methods are "oracle" in the sense that they use only (and all of the) informative predictors, without any noises.

Figure 3 plots the predictive MSE as the number of predictors increases in this case. Each of the horizontal dashed lines represents the MSE of one of the oracle forecasting methods, and the blue solid line is the MSE of the pseudo-OLS. The MSE of the oracle factor prediction is basically one in all cases, which is the residual variance of the true DGP. We plot for $p_0 \in \{200, 500\}$ and for selected cases for τ as these cases are representative. We can also conclude the following numerical findings:

1. The first two panels respectively fix $p_0 = 500$ and compare the cases of weak factors with relatively strong factors. Starting from $p = p_0$, the pseudo-OLS performs the best when $\tau = 1/2$, and is on par with ridge when $\tau = 1/4$. As in the previous study, when factors are very weak the pseudo-OLS continuously benefits from the reduced variance as noises are added into predictions, even though new direct predictive information is no longer available after $p > p_0$.

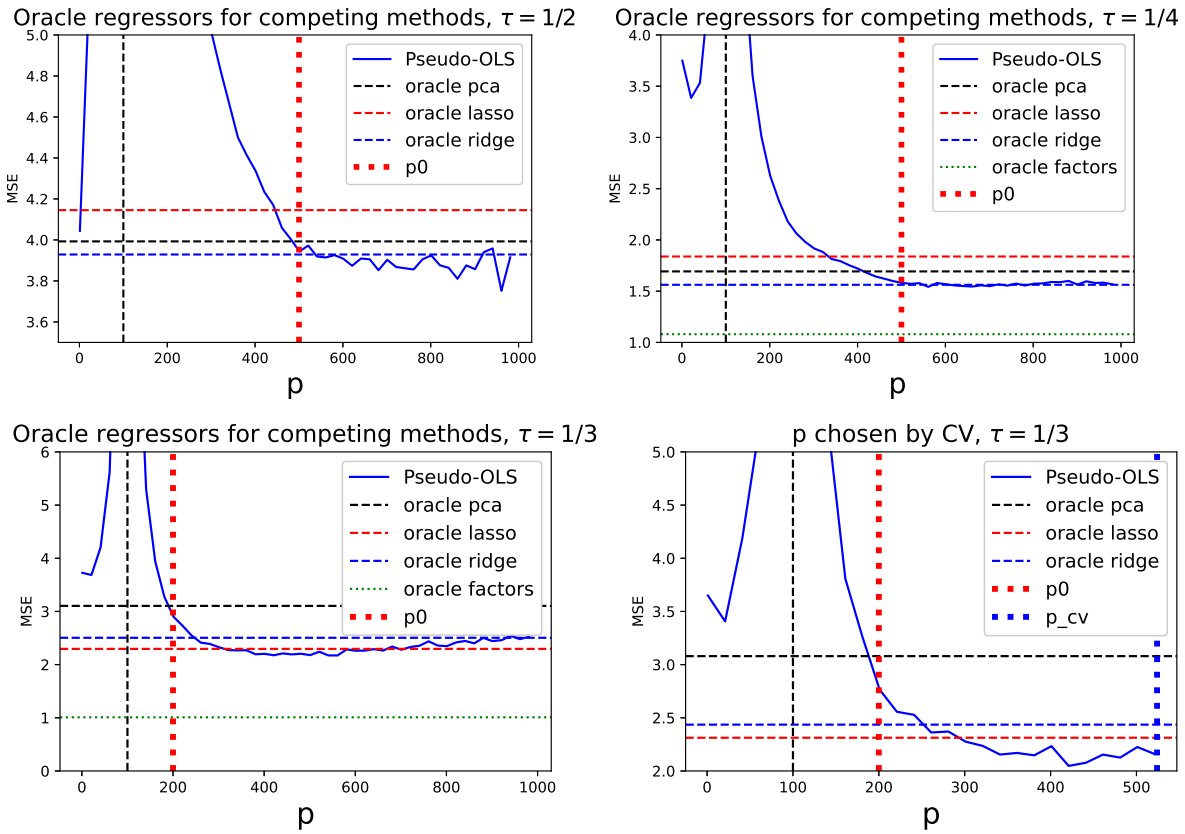

Figure 3: Predictive MSE $\sum_{j=1}^{50}(y_j - \widehat{y}_j)^2$ averaged from 50 replications as the number of predictors p increases. The vertical red dashed line indicates the number of informative predictors p_0; the vertical black dashed line indicates the sample size $n = 100$. The vertical blue dashed line in the last panel indicates the averaged p chosen by the cross validation.

2. When p_0 is moderate as in the third panel, the pseudo-OLS has a U-shaped predictive MSE after $p > n$, which reaches the minimum MSE around $p \in (400, 600)$ and starts to increase again. The is also well connected with our theory that the moderate informative signal causes the raising bias to dominate the decaying variance. Hence it is desirable to properly choose p in this case.

3. Panel 4 in Figure 3 plots the predictive MSE, under the same setting as in panel 3 for (τ, p_0), when the maximum amount of added noises p_{cv} is chosen by the cross-validation (CV). As guided by the theory, we set

$$p_{\text{cv}} = C \times n\sqrt{p_0}$$

and apply 10-fold CV to choose C on a grid evenly spanning on the range $\mathcal{C} =$

[0.2, 1]. The range is chosen so that the smallest possible value for p_{cv} is not much larger than $p_0 = 200$, whereas the largest possible value is not much larger than the original limit 1000. We then divide the training data into 10 folds using the 80-20 splitting rule, and obtain the optimal $p_{\text{cv}} \approx 520$, averaged over 50 replications. This is plotted as the vertical blue dashed line in panel 4. We observe that the pseudo-OLS yields the best out-of-sample predictions when noises are added up to p_{cv}, and CV suggests stop adding new noises at this point.

4. In contrast, the oracle forecasting methods, PCA, Lasso, and Ridge, even though only using the informative predictors, do not predict as well as the pseudo-OLS with many artificial noises in many cases. The PCA mainly suffers from weak factor issues, whereas the ridge does not have sufficiently diversified variance if p_0 is not large enough. In addition, in the first two panels, Lasso exhibits the poorest performance due to the model's high density, with half of the predictors being informative. Conversely, when p_0 is moderate as in the last two panels, Lasso outperforms PCA and ridge because the model is significantly denser.

5 Forecast Applications

Our empirical illustrations include four economic forecast exercises, respectively being forecasting S&P firms earnings, U.S. equity premium, U.S. unemployment rate, and countries' GDP growth rate.

5.1 Earnings forecasts

In capital market research, predicting corporate accounting earnings holds considerable relevance for fundamental analysis and equity valuation. Essentially, accounting earnings are a fundamental economic variable and precise predictions of earnings are crucial in evaluating the intrinsic value of a company's stock. This stance is underpinned by both analytical and empirical evidence. Analytically, the accounting-based valuation framework proposed by Ohlson (1995) and Feltham and Ohlson (1995) states the forecasted earnings as direct inputs into the valuation formula.[5] Empirically, extensive research indi-

[5]Our emphasis on the residual income valuation model doesn't imply it's the sole or superior method for equity valuation. Penman (1998) demonstrated that both dividend and cash-flow methods yield valuations akin to those of the residual income approach under specific conditions. The residual income model, rooted in accrual accounting, is especially useful for analyzing financial statements based on accrual accounting. However, since cash flows and dividends are directly linked to accrual figures through

cates that the accounting earnings are the payoff that investors forecast when estimating equity value (Penman and Sougiannis, 1998; Ball and Nikolaev, 2022).

Early research in time-series forecasting indicates the superiority of the random walk model. However, obtaining a stationary series long enough for accurate ARIMA-model parameter estimation is often impractical due to the infrequent nature of annual earnings reports. Researchers thus shift towards panel-data approaches, employing a broad set of predictors such as financial statement information that are potentially informative (Fairfield et al., 1996; Nissim and Penman, 2001; So, 2013). Recent work utilizes machine learning techniques to forecast accounting earnings, acknowledging the nonlinear relationships between predictors and future earnings, acknowledging the nonlinear relationships between predictors and future earnings (Chen et al., 2022).

Following Chen et al. (2022), we use high-dimensional detailed financial data as predictors. Since 2012, U.S. public companies have been obligated to utilize eXtensible Business Reporting Language (XBRL) tags for the presentation of quantitative data in their 10-K filings submitted SEC. Our analysis incorporates both current and preceding year data, normalized by total assets, and computes annual percentage changes. The focus is on financial data consistently reported by a minimum of 10 percent of the firms over our sample period, yielding a total of 1,316 predictors. Furthermore, we use pro forma Earnings Per Share (EPS) data sourced from I/B/E/S as the target variable. We merge data from SEC XBRL documents and I/B/E/S, emphasizing on companies possessing available share price information from the Center for Research in Security Prices (CRSP), nonzero total assets, and XBRL document filing promptly after the fiscal year-end. Consequently, our dataset encompasses 1,237 firm-year observations (829 for training and 408 for testing) for companies listed in the S&P 500 index, spanning the years 2013 to 2015.

Figure 4 plots the predictive MSE of the pseudo-OLS, CV-Ridge and CV-Lasso. For the pseudo-OLS, we set the maximum value for p as $p_{\max} = C \times n\sqrt{p_0}$, and choose C using cross-validation in a range so that $p_{\max}$ varies from 2,000 to 15,037 (equals $0.5 \times n\sqrt{p_0}$). The result shows that the MSE of pseudo-OLS starts to decrease when the total number of predictors are over 1250, and surpasses that of Ridge and Lasso when $p = 2,000$. The MSE for PCA is not plotted because it is twice as worse as the CV-Lasso. In addition, it is noteworthy that the prediction MSE exhibits a considerable magnitude. The unreported MSE for the naive sample mean is notably high, amounting to $279.36 \times 1e-3$. This observation aligns qualitatively with the recognition that predicting earnings poses a

basic accounting principles, forecasting accrual accounting figures also facilitates the projection of free cash flows and dividends (Nissim and Penman, 2001).

formidable challenge, primarily attributable to the limited information in the predictors. Despite the pre-screening conducted for feature selection, a substantial portion of the predictors still comprises only 10 percent reported firm-year data. Nevertheless, the pseudo-OLS produces one of the best predictions among the compared methods.

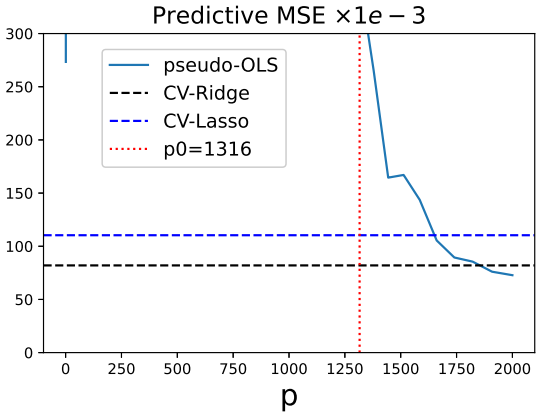

Figure 4: Predictive MSE using earnings data. Data comprises $n = 829$ training sample of firm-year observations for SP500 companies from 2013 to 2015, including $p_0 = 1,316$ predictors. The vertical axis is $\sum_n (y_{n+1} - \widehat{y}_{n+1})^2$, the horizontal axis is $\log(p)$, and the horizontal tick is p. Regardless of p, CV-Lasso and CV-Ridge use the p_0 macrovariables, whereas the psedo-OLS uses additional $p - p_0$ intentionally generated $N(0, 1)$ noises if $p > p_0$. PCA is not plotted as it is twice as worse as CV-Lasso.

5.2 U.S. equity premium prediction

Predicting the U.S. equity premium stands as a pivotal task in asset pricing, marked by the formidable challenges posed by the substantial volatility and instabilities inherent in the market index. Welch and Goyal (2008) conducted a comprehensive examination of prevailing working models at that time, ultimately asserting that "The evidence suggests that most models are unstable or even spurious." Since the publication of this seminal work, academic research in forecasting time-varying future equity premia has significantly advanced (e.g., Hirshleifer et al. (2009); Atanasov et al. (2020); Bekaert and Hoerova (2014); Chava et al. (2015); Chen et al. (2018); Colacito et al. (2016); Huang et al. (2015); Jondeau et al. (2019); Jones and Tuzel (2013); Kelly and Pruitt (2013); Martin (2017); Møller and Rangvid (2015); Rapach et al. (2016)). Many of them introduced new informative predictors, alongside innovative methodologies such as Lasso, PCA, and nonlinear machine learning. In light of these advancements, Goyal et al. (2023) conducted a new comprehensive review of recently proposed prominent predictive mod-

els, yet arriving at conclusions qualitatively consistent with their 2008 study. Notably, in the context of an annual forecast horizon, the majority of models exhibit bad predicting performance, with R^2 either negative or only marginally positive.

As an empirical illustration, we employ the 16 macroeconomic variables described by Welch and Goyal (2008) to forecast the equity premium. Following the same exercise of Giannone et al. (2021), we use data spanning from 1948 to 2015 with $p_0 = 16$ original predictors, and use annual moving windows with sample size $n = 17$. The first prediction occurs for the 1965 observation, and is repeated 51 times, each time expanding the training sample by one year and shifting the evaluation sample accordingly.

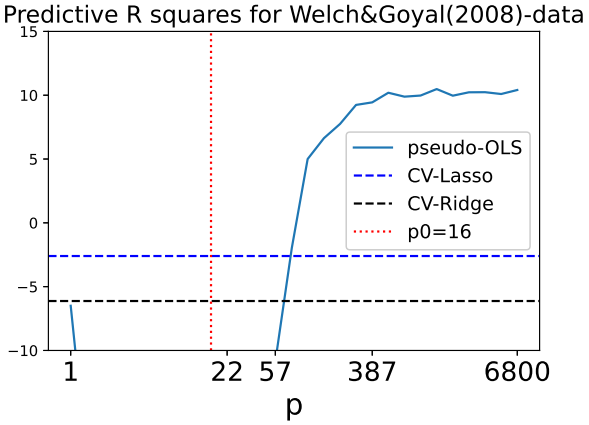

Figure 5: Out-of-sample R^2 for predicting the U.S. equity premium, using the dataset described by Welch and Goyal (2008), and updated on the webpage by Amit Goyal. The yearly data spans from 1948 to 2015, with $p_0 = 16$ original predictors. We use rolling windows of $n = 17$ year for one-year horizon forecast. The vertical axis is OOS R^2. The horizontal axis is plotted as $\log(p)$, and ticked using p. Regardless of p, both CV-Lasso and CV-Ridge use the p_0 macrovariables, whereas the psedo-OLS uses additional $p - p_0$ intentionally generated $N(0, 1)$ noises if $p > p_0$.

Methodologically, we intentionally add additional noises and merge with the original 16 macroeconomic variables to implement the pseudo-OLS. The number of added noises is chosen following the guidance of our theory, by setting $p = C \times n\sqrt{p_0}$, where C is chosen such that p takes values on a grid that are evenly spaced on a logarithmic scale. This leads to in total $300 \sim 6,000$ included noises as predictors. We also compare with Lasso and Ridge, which use only the 16 variables, and tuned by the cross-validation.

Figure 5 graphs the out-of-sample R^2, defined as

$$R^2 = 1 - \frac{\sum_{t+1}(y_{t+1} - \widehat{y}_{t+1})^2}{\sum_{t+1}(y_{t+1} - \bar{y}_t)^2}$$

where $\bar{y}_t$ denotes the in-sample mean of the t th rolling window. As both CV-Lasso and CV-Ridge fix regressors being 16 macroeconomic variables, their R^2 are depicted as dashed horizontal lines, and evaluated as -2.60 (CV-Lasso) and -6.12 (CV-Ridge). We also implemented PCA on the 16 variables whose R^2 is too bad to be depicted on the plot. In sharp contrasts, the pseudo-OLS with intentionally added noises performs strikingly well: it deteriorates when $p < 16$, but quickly comes back and outperforms the benchmark (sample mean prediction) and the other methods when 200 noises are added. After 400 noises are added its out-of-sample R^2 reaches to 10%, and becomes very stable even after 6800+ noises are added to prediction. [6]

5.3 Macroeconomic forecast

The second empirical illustration is in macroeconomic forecast where we use the FRED-MD dataset of McCracken and Ng (2016) to forecast the U.S. unemployment rate. The data contains $p_0 = 123$ macroeconomic predictors, ranging from 1960-June to 2019-December. We use 120-month moving windows to estimate the forecast model and conduct one-month-ahead forecast of the unemployment rate.

This macroeconomic dataset is widely recognized for its inherent challenge of relatively weak factors. Data-driven techniques for determining the number of factors, e.g., Bai and Ng (2002), typically suggest 8-10 factors, explaining 50-62% of the total variations. As such, adopting cross-validation becomes desirable to determine the optimal number of introduced noises, guarding against biases due to insufficient predictive information.

Therefore, we implement the pseudo-OLS as follows: if $p \leq p_0$, we use the first p components of the macroeconomic variables. [7] When $p > p_0$, we generate $p - p_0$ white noises from the standard normal distribution as artificial predictors, and merge with the original p_0 macroeconomic variables. We let p take values on a grid that are evenly spaced on a logarithmic scale from p_0 to $p_{\max}$ where $p_{\max} = C \times n\sqrt{p_0}$. To determine the optimal C, the full data of 715 months is split into training and validation samples. The

[6] We only use the original 16 macroeconomic variables from the Welch and Goyal (2008) dataset plus noises. So the pseudo-OLS is essentially still a linear predictor, and the gained predictability is mainly from the diversification of the out-of-sample variance. In comparison, Gu et al. (2020) exhaustively examined many nonlinear machine learning methods. Using up to 94 firm level characteristics, they found most of the predictors they examined, e.g., random forest and gradient boosting, reach less than ten percent annual R^2. Their most prominent machine learning predictor is neural networks, whose R^2 ranges from 10 to 15 percent.

[7] While variables in McCracken and Ng (2016) do not have a particular order, we simply take the first p columns according to the natural order in the original `FRED-MD` dataset, downloaded from `https://research.stlouisfed.org/wp/more/2015-012`.

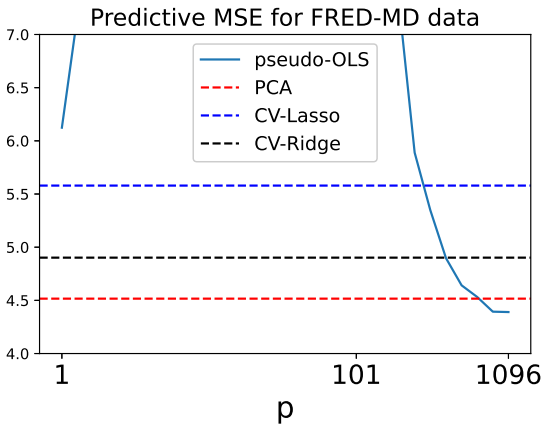

Figure 6: Predictive MSE using 123 Macroeconomic data from McCracken and Ng (2016). Data spans from 1960-May to 2019-December with $p_0 = 123$ predictors. We use rolling windows of $n = 120$ months for one-month horizon forecast. The vertical axis is $\sum_n (y_{n+1} - \widehat{y}_{n+1})^2$, the horizontal axis is $\log(p)$, and the horizontal tick is p. Regardless of p, the PCA, CV-Lasso and CV-Ridge use the p_0 macrovariables, whereas the psedo-OLS uses additional $p - p_0$ intentionally generated $N(0, 1)$ noises if $p > p_0$.

first 500 months are used as the training sample on which the constant C is determined via cross-validation, and the remaining 215 months are used for forecasts. Specifically, we conduct moving window forecasts on the training sample, regenerating white noises $R = 20$ times. Let $\widehat{y}_{t+1,r}(C)$ denote the predicted y_{t+1} using $C \times n\sqrt{p_0}$ predictors (including both macrovariables and noises) from the r th repetition. The optimal C is then determined by

$$C^* = \arg\min_{\mathcal{C}} \sum_{r=1}^{R} \sum_{t+1=121}^{500} (y_{t+1} - \widehat{y}_{t+1,r}(C))^2$$

where $\mathcal{C}$ is a predetermined grid. The optimal number of predictors determined in this way is approximately $p_{\max} \approx 1096$. With the determined $p_{\max}$, we conduct moving window forecasts of the unemployment rate on the validation sample, and compute the predictive MSE. In addition, we implemented PCA, CV-Lasso, and CV-Ridge on the validation sample, each uses all the p_0 macroeconomic variables regardless of p. The number of "factors" for PCA is determined using the PC_{p1} criterion from Bai and Ng (2002), and the tuning parameters for CV-Lasso and CV-Ridge are determined using 5-fold cross-validation.

Figure 6 plots the predictive MSE as p increases from 1 to $p_{\max}$. The MSE for the other three methods hold constant at approximately 5.5, 5.0 and 4.5. Meanwhile, the pseudo-OLS stops when the total number of predictors reaches 1096, corresponding to 123 macrovariables plus 973 added noises, and reaches MSE approximately 4.3, being the

lowest among the comparing methods.

5.4 Growth forecasts

We use the data of Barro and Lee (1994) to predict the GDP growth rates. This well known dataset consists of $p_0 = 60$ socio-economic and geographical characteristics from 90 countries spanning from 1960 to 1985. We estimate the model on a randomly selected sample of $n = 45$ countries, evaluating its predictions for the remaining 45 countries. We repeat this exercise 100 times and compute the predictive MSE averaged over the 100 random repetitions.

To implement pseudo-OLS, we generate $p-p_0$ white noises and merge to the original 60 predictors. As in the previous exercises, p takes values on a grid spaced on a logarithmic scale from p_0 to $p_{\max} = C \times n\sqrt{p_0}$. The predictors are known to have strong predictability of the GDP growth rate, so we set a large C, which makes $p_{\max} \approx 34,857$. As usual, the competing methods, PCA, CV-Lasso, CV-Ridge only use the original 60 predictors.

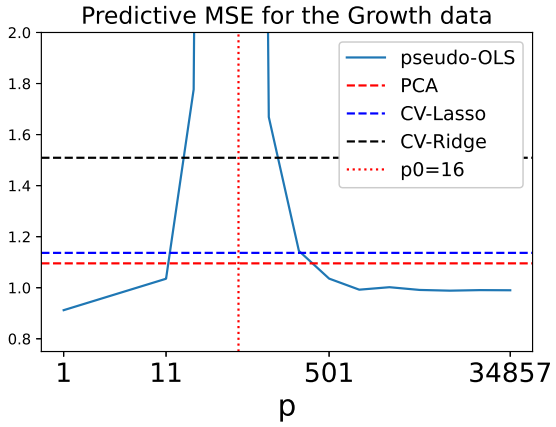

Figure 7: Predictive MSE using 60 socio-economic and geographical characteristics from Barro and Lee (1994). Data for the growth rate of GDP from 90 countries. We estimate the model on a randomly selected sample of $n = 45$ countries, evaluating its predictions for the remaining 45 countries. We repeat this exercise 100 times. The vertical axis is $\sum_n (y_{n+1} - \widehat{y}_{n+1})^2$, the horizontal axis is $\log(p)$, and the horizontal tick is p. Regardless of p, the PCA, CV-Lasso and CV-Ridge use the p_0 macrovariables, whereas the psedo-OLS uses additional $p - p_0$ intentionally generated $N(0, 1)$ noises if $p > p_0$.

Figure 7 plots the predictive MSE of different methods. The pseudo-OLS outperforms other methods before and after the complete interpolation. Notably, its predictive MSE is smaller than minimum value of the second descent when only the first 11 predictors are used, showcasing the informativeness of the original predictors. Furthermore, it exhibits

strong robustness to the number of added noises, stabilizing after 500 additions and maintaining an MSE around 1.0 across p varying from 501 to 34,857.

6 Conclusion

In the realm of economic forecasts, economists frequently collect a large number of predictors, and undertake variable selection processes to eliminate noises from the predictor set before proceeding forecasts. In contrast, we prove a compelling result that in most economic forecasts, the inclusion of noises in predictions yields greater benefits than its exclusion. Furthermore, if the total number of predictors is not sufficiently large, economists should intentionally add more noises, resulting in superior forecast performance that outperforms many benchmark predictors relying on dimension reduction techniques. The intuition is that economic predictive signals often densely distributed among the regression coefficients, maintaining a modest level of forecast bias while diversifying away the overall variance, even when a significant proportion of these predictors constitute pure noises.

A Additional Simulations

The previous simulation studies have set fixed number of informative predictors. We now examine the dynamic case by setting $p_0 = 0.5p$, and still let p vary from 1 to 1000. As before, we compare the pseudo-OLS with three other methods: PCA, Lasso and Ridge. Except for pseudo-OLS, all the other three use only the p_0 informative predictors, excluding noises.

From Figure 8 we see that first when factors are strong, all models are close to the oracle risk being one. Secondly, in most cases oracle Ridge is comparable with pseudo-OLS. This is not surprising because as p_0 increases Ridge is also superior with smaller variance. Finally, with growing p_0 and τ increases, the model becomes denser and factors become weaker. Then the pseudo-OLS dominates either Lasso or PCA or both.

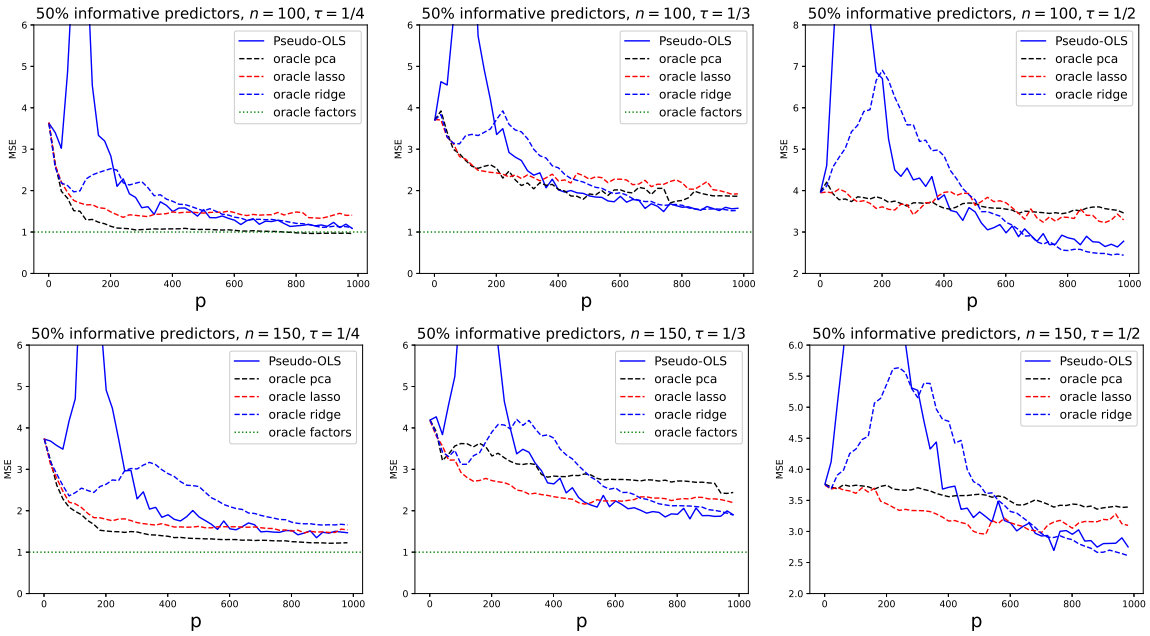

Figure 8: Predictive MSE $\sum_{j=1}^{50}(y_j - \widehat{y}_j)^2$ averaged from 10 replications as the number of predictors p increases. The number of informative predictors $p_0 = 0.5p$.

B Proofs

B.1 Proof of Bias-variance expression

The objective is to prove

$$\mathbb{E}[(y_{new} - \widehat{y}_{\text{new}})^2|X] = \text{bias}(\widehat{y}_{\text{new}})^2 + \text{Var}(\widehat{y}_{\text{new}} - X'_{\text{new}}\beta|X) + \text{Var}(e_{\text{new}})$$

where

$$\begin{aligned} \text{bias}(\widehat{y}_{\text{new}})^2 &= \beta' A_X \mathbb{E} X_t X_t' A_X \beta, \quad \text{where } A_X := (X'X)^+ X'X - I \\ \text{Var}(\widehat{y}_{\text{new}} - X'_{\text{new}}\beta|X) &= \sigma_e^2 \,\text{tr}(X'X)^+ \mathbb{E} X_t X_t' O_P(1), \end{aligned}$$

for some $\sigma_e^2 > 0$.

Proof. Let $M := \mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})^2|X]$. Then

$$\mathbb{E}[(y_{\text{new}} - \widehat{y}_{\text{new}})^2|X] = M + \mathbb{E}[e_{\text{new}}^2|X] + 2\mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})e_{\text{new}}|X].$$

The second term is $\mathbb{E}[e_{\text{new}}^2|X] = \text{Var}(e_t)$ by assumption. The third term is

$$\begin{aligned} & \mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})e_{\text{new}}|X] = \mathbb{E}[X'_{\text{new}}(\beta - \widehat{\beta})e_{\text{new}}|X] \\ = \;& \mathbb{E}[\beta' X_{\text{new}} e_{new}|X] - \mathbb{E}[e_{\text{new}} X'_{\text{new}}\widehat{\beta}|X] = -\mathbb{E}[e_{\text{new}} X'_{\text{new}}(X'X)^+ X'e|X] \\ = \;& -\mathbb{E}\{\mathbb{E}[e_{\text{new}} X'_{\text{new}}|X, e](X'X)^+ X'e|X\} = 0 \end{aligned}$$

by the assumption $\mathbb{E}(e_{\text{new}} X_{\text{new}}|X, e) = 0$. We now decompose M. Let

$$R = \mathbb{E}(\widehat{\beta}|X) = (X'X)^+ X'X\beta + \mathbb{E}[(X'X)^+ X'e|X] = (X'X)^+ X'X\beta,$$

then $R - \beta = A_X\beta$ and $\widehat{\beta} - R = (X'X)^+ X'e$. So

$$\begin{aligned} M &= \mathbb{E}[(X'_{\text{new}}(\widehat{\beta} - \beta))^2|X] \\ &= \mathbb{E}[(X'_{\text{new}}(\widehat{\beta} - R))^2|X] + \mathbb{E}[(X'_{\text{new}}(R - \beta))^2|X] + 2\mathbb{E}[(R - \beta)' X_{\text{new}} X'_{\text{new}}(\widehat{\beta} - R)|X]. \end{aligned}$$

By the assumption $\mathbb{E}(e|X, X_{\text{new}}) = 0$, the third term is

$$\mathbb{E}[(R - \beta)' X_{\text{new}} X'_{\text{new}}(\widehat{\beta} - R)|X] = \beta' A_X \mathbb{E}[X_{\text{new}} X'_{\text{new}}(X'X)^+ X'e|X] = 0.$$

The second term is "squared bias": by the assumption $\mathbb{E}(X_{\text{new}}X_{\text{new}}'|X) = \mathbb{E}(X_tX_t')$,

$$\text{bias}(\widehat{y}_{\text{new}})^2 \;:=\; \mathbb{E}[(X_{\text{new}}'(R-\beta))^2|X] = \beta' A_X \mathbb{E}(X_{\text{new}}X_{\text{new}}'|X)A_X\beta = \beta' A_X \mathbb{E}(X_tX_t')A_X\beta.$$

The first term is variance. The assumption that $\mathbb{E}[ee'|X, X_{\text{new}}]\| = O_P(\sigma_e^2)$ for some $\sigma_e^2 > 0$ implies (verified by Lemma 4):

$$\begin{aligned}
& \text{Var}(\widehat{y}_{\text{new}} - X_{\text{new}}'\beta|X) = \mathbb{E}[(X_{\text{new}}'(\widehat{\beta}-R))^2|X] = \mathbb{E}[(X_{\text{new}}'(X'X)^+X'e)^2|X] \\
= \;& \mathbb{E}\{X_{\text{new}}'(X'X)^+X'\mathbb{E}[ee'|X, X_{\text{new}}]X(X'X)^+X_{\text{new}}|X\} \\
\leq \;& \mathbb{E}\{\|X_{\text{new}}'(X'X)^+X'\|^2|X\}\|\mathbb{E}[ee'|X, X_{\text{new}}]\| \\
= \;& \sigma_e^2\mathbb{E}\{X_{\text{new}}'(X'X)^+X_{\text{new}}|X\}O_P(1) = \sigma_e^2 \,\text{tr}(X'X)^+\mathbb{E}X_tX_t'O_P(1).
\end{aligned}$$

□

Lemma 4. *Suppose at least one of the two cases holds:*

(i) $(f_t, \epsilon_{y,t}, u_t, X_{\text{new}}, t = 1...n)$ *are i.i.d., and jointly normal,* $\mathbb{E}(Y|X, X_{\text{new}}) = \mathbb{E}(Y|X)$,

(ii) $n = O(\psi_{p,n})$, $\|\,\text{Cov}(u_t)\| = O(1)$, $\mathbb{E}(\epsilon_y|U, F, X_{\text{new}}) = 0$ *and* $\|\mathbb{E}(\epsilon_y\epsilon_y'|X, X_{\text{new}})\| = O_P(1)$.

Then $\|\mathbb{E}[ee'|X, X_{\text{new}}]\| = O_P(1)$.

Proof. (i) We have, (e, X, X_{new}) are jointly normal, as they can be written as linear combinations of X_{new} and (F, ϵ_y, U), the matrices of $(f_t, \epsilon_{y,t}, u_t)$. Also, by the assumption $\mathbb{E}(Y|X, X_{\text{new}}) = \mathbb{E}(Y|X)$,

$$\mathbb{E}(e|X, X_{\text{new}}) = \mathbb{E}(Y|X, X_{\text{new}}) - X\beta = \mathbb{E}(F\rho|X) - X\beta = 0$$

This implies e is independent of (X, X_{new}). Hence $\mathbb{E}[ee'|X, X_{\text{new}}] = \mathbb{E}ee' = \text{Cov}(e)$. Also, $\text{Cov}(e_t, e_s) = 0$ for any $t \neq s$ because of the i.i.d. assumption (i.e., (y_t, X_t) and (y_s, X_s) are independent). Thus $\mathbb{E}[ee'|X, X_{\text{new}}] = \sigma_e^2 I_n$, where $\sigma_e^2 = \text{Var}(e_t)$.

(ii) Alternatively, let $A := \epsilon_y((\rho - \Lambda'\beta)'F' + \beta'U')$. Let $C := U\beta$, and $D := F(\rho - \Lambda'\beta)$. Then

$$ee' = \epsilon_y\epsilon_y' + CC' + DD' + CD' + DC' + A + A'.$$

We have $\|\mathbb{E}(ee'|X, X_{\text{new}})\| \leq \|\mathbb{E}(\epsilon_y\epsilon_y'|X, X_{\text{new}})\| + 2\|\mathbb{E}(A|X, X_{\text{new}})\| + 2\mathbb{E}(\|C\|^2 + \|D\|^2|X, X_{\text{new}})$. Our assumption ensures $\mathbb{E}(\epsilon_y|U, F, X_{\text{new}}) = 0$ hence $\mathbb{E}(A|X, X_{\text{new}}) = 0$ and $\|\mathbb{E}(\epsilon_y\epsilon_y'|X, X_{\text{new}})\| = O_P(1)$. In addition,

$$\mathbb{E}\|C\|^2 = O(n)\|\beta\|^2 = O(n/\psi_{p,n}), \quad \mathbb{E}\|D\|^2 = O(n)\|\rho - \Lambda'\beta\|^2 = O(n/\psi_{p,n}^2).$$

Hence $\|\mathbb{E}(ee'|X, X_{\text{new}})\| \leq O_P(1 + n/\psi_{p,n}) = O_P(1)$.

□

B.2 Proof of Theorem 1

Proof. The condition $\mathbb{E}(\epsilon_{y,t}|f_t, u_t) = 0$ implies that in the oracle model $\mathbb{E}(\epsilon_{y,t}|X_t) = 0$ as X_t follows the factor model (1). As a results, in the working model

$$\mathbb{E}(y_t|X_t) = \mathbb{E}(\rho' f_t|X_t) + \mathbb{E}(\epsilon_{y,t}|X_t) = \beta' X_t$$

by the condition $\mathbb{E}(\rho' f_t|X_t) = \beta' X_t$. The error term $\mathbb{E}(e_t|X_t) = \mathbb{E}(y_t - X_t'\beta|X_t) = 0$.

(i) Pre-multiplying both side of the condition $\mathbb{E}(\rho' f_t|X_t) = \beta' X_t$ by X_t and take unconditional expectation, we have $\mathbb{E}\left(X_t \mathbb{E}(f_t'\rho|X_t)\right) = \mathbb{E}\left(X_t X_t'\right)\beta$. The factor model (1) implies that $\mathbb{E}X_t X_t' = \Lambda\Sigma_f\Lambda' + \text{Cov}(u_t)$, where $\Sigma_f := \mathbb{E}f_t f_t'$, is invertible. Thus we solve

$$\beta = \mathbb{E}(X_t X_t')^{-1}\mathbb{E}X_t f_t'\rho = \mathbb{E}(X_t X_t')^{-1}\Lambda\Sigma_f\rho.$$

Substituting the Woodbury matrix identity

$$\mathbb{E}(X_t X_t')^{-1} = \text{Cov}(u_t)^{-1} - \text{Cov}(u_t)^{-1}\Lambda(\Sigma_f^{-1} + \Lambda'\,\text{Cov}(u_t)^{-1}\Lambda)^{-1}\Lambda'\,\text{Cov}(u_t)^{-1}$$

into the above expression yields

$$\begin{aligned}
\beta &= \text{Cov}(u_t)^{-1}\Lambda\left[I - (\Sigma_f^{-1} + \Lambda'\,\text{Cov}(u_t)^{-1}\Lambda)^{-1}\Lambda'\,\text{Cov}(u_t)^{-1}\Lambda\right]\Sigma_f\rho \\
&= \text{Cov}(u_t)^{-1}\Lambda\left[(\Sigma_f^{-1} + \Lambda'\,\text{Cov}(u_t)^{-1}\Lambda)^{-1}\Sigma_f^{-1}\right]\Sigma_f\rho \\
&= \text{Cov}(u_t)^{-1}\Lambda(\Sigma_f^{-1} + \Lambda'\,\text{Cov}(u_t)^{-1}\Lambda)^{-1}\rho.
\end{aligned}$$

(ii) By definition,

$$e_t = y_t - X_t'\beta = (f_t'\rho + \epsilon_{y,t}) - (f_t'\Lambda + u_t')\,\beta = \epsilon_{y,t} + f_t'(\rho - \Lambda'\beta) - u_t'\beta,$$

and thus

$$\text{Var}(e_t) = \text{Var}(\epsilon_{y,t}) + \text{Var}(f_t'(\rho - \Lambda'\beta)) + \text{Var}(u_t'\beta)$$

by the condition $\mathbb{E}(\epsilon_{y,t}|f_t, u_t) = 0$ and the orthogonality between f_t and u_t. It remains to bound $\text{Var}(f_t'(\rho - \Lambda'\beta)) + \text{Var}(u_t'\beta)$.

We define $\Phi := \Lambda'\,\text{Cov}(u_t)^{-1}\Lambda/\psi_{n,p}$. Since the eigenvalues of $\text{Cov}(u_t)$ is bounded away

from 0 and ∞, under Assumption 1 we have $\sigma_{\max}(\Phi) \asymp \sigma_{\max}(\Lambda_I'\Lambda_I)/\psi_{p,n} \asymp 1$ and similarly $\sigma_{\min}(\Phi) \asymp 1$.

The explicit expression of β in (i) gives $\Lambda'\beta = \rho - \Sigma_f^{-1}(\Sigma_f^{-1} + \Lambda' \operatorname{Cov}(u_t)^{-1}\Lambda)^{-1}\rho$. We have

$$\begin{aligned}
\operatorname{Var}(f_t'(\rho - \Lambda'\beta)) &= \rho(\Sigma_f^{-1} + \Lambda' \operatorname{Cov}(u_t)^{-1}\Lambda)^{-1}\Sigma_f(\Sigma_f^{-1} + \Lambda' \operatorname{Cov}(u_t)^{-1}\Lambda)^{-1}\rho \\
&= \rho(\Sigma_f^{-1} + \Phi\psi_{p,n})^{-1}\Sigma_f(\Sigma_f^{-1} + \Phi\psi_{p,n})^{-1}\rho \\
&\leq \left\|(\Sigma_f^{-1/2} + \Sigma_f^{1/2}\Phi\psi_{p,n})^{-1}\right\|^2 \|\rho\|^2 \\
&\leq \|\Sigma_f\|\|\rho\|^2\|\Sigma_f^{-1}\|^2\sigma_{\min}(\Phi\psi_{p,n})^{-2} = O(\psi_{p,n}^{-2}).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\operatorname{Var}(u_t'\beta) &= \beta' \operatorname{Cov}(u_t)\beta = \left\|\operatorname{Cov}(u_t)^{-1/2}\Lambda(\Sigma_f^{-1} + \Lambda' \operatorname{Cov}(u_t)^{-1}\Lambda)^{-1}\rho\right\|^2 \\
&\leq \left\|(\Sigma_f^{-1} + \Phi\psi_{p,n})^{-1}\rho\right\|^2 \|\psi_{p,n}\Phi\| \\
&\leq \psi_{p,n}\sigma_{\min}(\Sigma_f^{-1} + \Phi\psi_{p,n})^{-2}\|\Phi\| \|\rho\|^2 \\
&\leq \psi_{p,n}^{-1}\sigma_{\min}(\Phi)^{-2}\|\Phi\| \|\rho\|^2 = O(\psi_{p,n}^{-1}).
\end{aligned}$$

We thus conclude $\operatorname{Var}(e_t) - \operatorname{Var}(\epsilon_{y,t}) = \operatorname{Var}(f_t'(\rho - \Lambda'\beta)) + \operatorname{Var}(u_t'\beta) = O(\psi_{p,n}^{-1})$.

□

B.3 Proof of Theorem 2

By the i.i.d. assumption for $u_{i,t}$, we can write $\operatorname{Cov}(u_t) = \sigma_u^2 I_p$ where $\sigma_u^2 = \operatorname{Var}(u_{i,t})$.

B.3.1 Bias

Recall $\operatorname{bias}(\widehat{y}_{\text{new}})^2 = \beta' A_X \Sigma_X A_X \beta$, where $\Sigma_X := \mathbb{E}X_t X_t'$ and $A_X := (X'X)^+X'X - I$. Using the notations laid out in the main text, we apply the singular decomposition $X = U_n S_n V_n'$ (notice V_n is a $p \times n$ matrix), and thus

$$X'X = V_n S_n^2 V_n', \quad (X'X)^+ = V_n S_n^{-2} V_n'$$

and $A_X = V_n V_n' - I_p = -V_A V_A'$ where V_A is a $p \times (p-n)$ matrix, columns being eigenvectors of the $p \times p$ matrix $X'X$ corresponding to the $p-n$ eigenvalues. Therefore, $V_A' V_n = 0$.

We rewrite

$$\text{bias}(\widehat{y}_{\text{new}})^2 = \beta' V_A V_A' \Sigma_X V_A V_A' \beta \leq \|V_A' \Sigma_X V_A\| \, \|V_A' \beta\|^2 \tag{B.1}$$

bounding $\|V_A' \Sigma_X V_A\|$

We focus on the first factor of (B.1). Using matrix notation, the factor model of the $n \times p$ matrix X is

$$X = F\Lambda' + U, \tag{B.2}$$

where F is $n \times K$ and Λ is $p \times K$. Then

$$X'X = n\Lambda\Sigma_f\Lambda' + E + U'U$$

where $E := -\Lambda \left(F'F - n\Sigma_f\right)\Lambda' + \Lambda F'U + U'F\Lambda'$.

We first check the orders of $\|E\|$. Under Assumption 3, $\|U'U\| = O_P(p)$. Since $\|F'F/n - \Sigma_f\| = O_p(n^{-1/2})$, for all $j \leq K$ we have

$$\|\Lambda \left(F'F - n\Sigma_f\right)\Lambda'\| \leq \sigma_{\max}\left(\Lambda\Lambda'\right) \|F'F - n\Sigma_f\| = O_p\left(\sqrt{n}\psi_{p,n}\right).$$

and the cross term

$$\begin{aligned} \|U'F\Lambda'\| &\leq \sqrt{\sigma_{\max}\left(\Lambda F'F\Lambda'\right)\sigma_{\max}\left(U'U\right)} \\ &= \sqrt{\sigma_{\max}\left(\Lambda\Lambda'\right) O_P\left(n\right)\sigma_{\max}\left(U'U\right)} = O\left(\sqrt{\psi_{p,n} np}\right). \end{aligned}$$

We obtain

$$\|E\| = O_P(\sqrt{n}\psi_{p,n} + \sqrt{\psi_{p,n}pn}) = O_P(\sqrt{\psi_{p,n}pn})$$

where the order follows as $\frac{\sqrt{n}\psi_{p,n}}{\sqrt{\psi_{p,n}pn}} = \frac{\sqrt{\psi_{p,n}}}{\sqrt{p}}$ is bounded away from ∞ given $\psi_{p,n} = O\left(p_0\right)$ and $p_0 \leq p$. We thus have

$$\begin{aligned} V_A'\Sigma_X V_A &= V_A'\left(\frac{X'X}{n} - \left(\frac{X'X}{n} - \Sigma_X\right)\right)V_A \\ &= V_A'V_n S_n^2 V_n' V_A - V_A'\left(\frac{X'X}{n} - \Sigma_X\right)V_A \\ &= -V_A'\left(\frac{X'X}{n} - \Sigma_X\right)V_A = V_A'\left(\text{Cov}\left(u_t\right) - \frac{E}{n} - \frac{U'U}{n}\right)V_A \end{aligned}$$

where the third line follows by the fact that the columns of V_A is orthogonal to the

columns of V_n, and thus we find the order of the first factor in (B.1)

$$\begin{aligned}\|V_A'\Sigma_X V_A\| &\leq \left\|V_A'\left(\mathrm{Cov}\,(u_t) - \frac{E}{n} - \frac{U'U}{n}\right)V_A\right\| \\ &\leq \|\mathrm{Cov}\,(u_t)\| + \left\|\frac{U'U}{n}\right\| + \left\|\frac{E}{n}\right\| \leq O_P\left(\sqrt{\frac{\psi_{p,n}p}{n}} + \frac{p}{n}\right).\end{aligned}$$

bounding $\|V_A'\beta\|^2$. Recall $\beta = \sigma_u^{-2}\Lambda(\Sigma_f^{-1} + \Lambda'\,\mathrm{Cov}(u_t)^{-1}\Lambda)^{-1}\rho$. First,

$$X'X = n\Lambda\Sigma_f\Lambda' + E + U'U$$

By Davis-Khan theorem, the top K eigenvalues of $X'X/n$ and $\Lambda\Sigma_f\Lambda'$, respectively denoted by $\sigma_{j,X}$ and σ_j for $j \leq K$, are bounded by

$$\max_{j\leq K}|\sigma_{j,X} - \sigma_j| \leq \left\|\frac{U'U}{n}\right\| + \left\|\frac{E}{n}\right\| \leq O_P\left(\sqrt{\frac{\psi_{p,n}p}{n}} + \frac{p}{n}\right).$$

Hence $\psi_{p,n} \gg p/n$ implies $\sigma_{j,X} \asymp \psi_{p,n}$ for $j \leq K$. There is $K \times K$ matrix H so that columns of ΛH are eigenvectors of $\Lambda\Sigma_f\Lambda'$, and $\|H^{-1}\| = O_P(\psi_{p,n}^{1/2})$. Let V_K denotes the first K eigenvectors of $X'X/n$. By Sin-theta inequality,

$$\|V_K - \Lambda H\| \leq O_P(\psi_{p,n}^{-1})\left[\left\|\frac{U'U}{n}\right\| + \left\|\frac{E}{n}\right\|\right] \leq O_P(\frac{p}{\psi_{p,n}n} + \sqrt{\frac{p}{\psi_{p,n}n}}) = O_P(\sqrt{\frac{p}{\psi_{p,n}n}}).$$

Also note that $V_A'V_K = 0$ because of the orthogonality. Hence

$$\|V_A'\Lambda\|^2 \leq \|V_A'\Lambda HH'\Lambda V_A\|\|H^{-1}\|^2 \leq O_P(\psi_{p,n})\|\Lambda H - V_K\| = O_P(\sqrt{\frac{\psi_{p,n}p}{n}}).$$

Hence

$$\|V_A'\beta\|^2 \leq O(1)\|V_A'\Lambda\|^2\|\Sigma_f^{-1} + \Lambda'\,\mathrm{Cov}(u_t)^{-1}\Lambda)^{-1}\|^2 = O_P(\sqrt{\frac{\psi_{p,n}p}{n}})\psi_{p,n}^{-2}.$$

We thus conclude

$$\mathrm{bias}(\widehat{y}_{\mathrm{new}})^2 \leq O_P\left(\sqrt{\frac{\psi_{p,n}p}{n}} + \frac{p}{n}\right)O_P(\sqrt{\frac{\psi_{p,n}p}{n}})\psi_{p,n}^{-2} = O_P\left(\frac{p}{\psi_{p,n}n}\right).$$

B.3.2 Variance

The variance is

$$\mathrm{Var}(\widehat{y}_{\mathrm{new}} - X'_{\mathrm{new}}\beta|X) = \sigma_e^2 \,\mathrm{tr}\left((X'X)^+\Sigma_X\right) O_P(1), \quad \text{where } \sigma_e^2 := \mathrm{Var}(e_t|X_n, X).$$

Since $\Sigma_X = \Lambda\Sigma_f\Lambda' + \sigma_u^2 I_p$, we have

$$\mathrm{tr}\left((X'X)^+\Sigma_X\right) = \mathrm{tr}(\Lambda'(X'X)^+\Lambda\Sigma_f) + \sigma_u^2 \,\mathrm{tr}\left((X'X)^+\right) \tag{B.3}$$

We focus on $\mathrm{tr}(\Lambda'(X'X)^+\Lambda\Sigma_f)$ first. Then

$$\begin{aligned}
\|\Lambda'\left(X'X\right)^+\Lambda\| &= \|\Lambda'\left(X'X\right)^+\left(X-U\right)'F\left(F'F\right)^{-1}\| \\
&= \|\Lambda'\left(X'X\right)^+\left(X-U\right)'\frac{F}{n}\|O_P(1) \\
&\leq \left\|\left(\left(X'X\right)^+\right)^{1/2}\Lambda\right\| \left\|\left(\left(X'X\right)^+\right)^{1/2}\left(X-U\right)'\frac{F}{n}\right\| O_P(1)
\end{aligned}$$

where the first line by the expression $\Lambda = \left(X-U\right)'F\left(F'F\right)^{-1}$ from the factor model; the second line follows by $(F'F/n)^{-1} = O_P(1)$, and the last inequality by the Cauchy-Schwarz inequality. Rearrange the above inequality,

$$\begin{aligned}
&\|\Lambda'\left(X'X\right)^+\Lambda\| \\
\leq\ & \left\|\left(\left(X'X\right)^+\right)^{1/2}\left(X-U\right)'\frac{F}{n}\right\|^2 O_P(1) \\
\leq\ & 2\left\{\left\|\left(\left(X'X\right)^+\right)^{1/2}X'\frac{F}{n}\right\|^2 + \left\|\left(\left(X'X\right)^+\right)^{1/2}U'\frac{F}{n}\right\|^2\right\} O_P(1) \\
\leq\ & 2\left\{\sigma_{\max}\left(X\left(X'X\right)^+X'\right) + \sigma_{\max}\left(U\left(X'X\right)^+U'\right)\right\}\frac{F'F}{n^2}O_P(1) \\
=\ & \frac{2}{n}\left\{1 + \sigma_{\max}\left(U\left(X'X\right)^+U'\right)\right\} O_P(1)
\end{aligned} \tag{B.4}$$

where the last line follows by the fact that $X\left(X'X\right)^+X'$ is idempotent. In the curly bracket,

$$\sigma_{\max}\left(U\left(X'X\right)^+U'\right) = \sigma_{\max}\left(\left(\frac{X'X}{p}\right)^+\right)\sigma_{\max}\left(\frac{UU'}{p}\right)$$

$$\leq C_u \sigma_{\max}\left(\left(\frac{X'X}{p}\right)^{+}\right) = \frac{C_u}{\sigma_{\min}\left(XX'/p\right)}$$

as $\|U\|^2 = O_P(p)$ when $p > n$. Let $v \in \mathbb{R}^n$ with $\|v\| = 1$ be the eigenvector of the $n \times n$ matrix XX' corresponding to its n th eigenvalue. Then

$$\begin{aligned}
\sigma_{\min}\left(XX'/p\right) &= v'\frac{XX'}{p}v = v'F\frac{\Lambda'\Lambda}{p}F'v + 2v'F\frac{\Lambda'U'}{p}v + v'\frac{UU'}{p}v \\
&\geq v'F\frac{\Lambda'\Lambda}{p}F'v - \left|2v'F\frac{\Lambda'U'}{p}v\right| + c_u \\
&\geq v'F\frac{\Lambda'\Lambda}{p}F'v - 2\left\|\left(\frac{\Lambda'\Lambda}{p}\right)^{1/2}F'v\right\|\left\|\left(\frac{\Lambda'\Lambda}{p}\right)^{-1/2}\frac{\Lambda'}{p}U'v\right\| + c_u \\
&\geq \alpha'\alpha - 2\|\alpha\|\|M\| + c_u \\
\alpha &:= \left(\frac{\Lambda'\Lambda}{p}\right)^{1/2}F'v, \quad M := \left(\frac{\Lambda'\Lambda}{p}\right)^{-1/2}\frac{\Lambda'}{p}U'v
\end{aligned} \tag{B.5}$$

where the first inequality follows under Assumption 3. Also, $\sigma_{\min}(U'U/p) > c_u$.

In addition,

$$\|\operatorname{Var}(M)\| = \left\|\left(\frac{\Lambda'\Lambda}{p}\right)^{-1/2}\frac{\Lambda'\operatorname{Cov}(u_t)\Lambda}{p^2}\left(\frac{\Lambda'\Lambda}{p}\right)^{-1/2}\right\| \leq \frac{C}{p},$$

and therefore the event $\left\{\|M\| \leq \frac{\sqrt{c_u}}{2}\right\}$ holds w.p.a.1. as $p, n \to \infty$. Under this event, we continue (B.5):

$$\begin{aligned}
\sigma_{\min}\left(XX'/p\right) &\geq \|\alpha\|^2 - \sqrt{c_u}\,\|\alpha\| + c_u \\
&= \left(\|\alpha\| - \frac{\sqrt{c_u}}{2}\right)^2 + \frac{3}{4}c_u \geq \frac{3}{4}c_u.
\end{aligned} \tag{B.6}$$

Hence $\sigma_{\max}\left(U\left(X'X\right)^{+}U'\right) \leq C_u/(0.75c_u)$. Substituting it into (B.4) we have

$$\|\Lambda'\left(X'X\right)^{+}\Lambda\| \leq \frac{2}{n}\left(1 + \frac{C_u}{0.75c_u}\right)O_P(1) = O_P\left(1/n\right).$$

Hence the first term in (B.3) is bounded by $\operatorname{tr}\left(\Lambda'\left(X'X\right)^{+}\Lambda\right) = O_P\left(1/n\right)$.

For the second term in (B.3), we have

$$\begin{aligned}\operatorname{tr}\left((X'X)^{+}\right) &= \frac{1}{p}\operatorname{tr}\left(\left(\frac{X'X}{p}\right)^{+}\right) = \frac{1}{p}\sum_{j=1}^{n}\frac{1}{\sigma_j\left(X'X/p\right)} \\ &\leq \frac{1}{p}\cdot\frac{n}{\sigma_n\left(XX'/p\right)} = O_P\left(\frac{n}{p}\right)\end{aligned}$$

by (B.6). We conclude that the variance is

$$\operatorname{Var}(\widehat{y}_{\text{new}} - X'_{\text{new}}\beta|X) = O_P\left(\frac{1}{n}+\frac{n}{p}\right).$$

B.4 Proof of Theorem 3

The closed-form solution of the ridge estimator with the regressors in the set I is

$$\widehat{\beta}_I(\lambda) = (X_I'X_I + \lambda I)^{-1}X_I'Y,$$

where $\lambda \geq 0$ is the tuning parameter. The population counterpart of the coefficient is

$$\beta_I = \Sigma_{X,I}^{-1}\Lambda_I'\rho \tag{B.7}$$

where $\Sigma_{X,I} = \Lambda_I\Lambda_I' + \sigma_u^2 I_{p_0}$ is a $p_0 \times p_0$ matrix, Λ_I is a $K \times p_0$ matrix, and the K-dimensional vector ρ is the same as the full model. In the restricted model on the set I we have

$$\operatorname{bias}_I(\lambda)^2 = \beta_I'A_I\Sigma_{X,I}A_I\beta_I \tag{B.8}$$

where $A_I = (X_I'X_I + \lambda I)^{-1}X_I'X_I - I_{p_0}$ and

$$\operatorname{Var}_I(\lambda) = \sigma_e^2\operatorname{tr}\left[\Sigma_{X,I}(X_I'X_I+\lambda I)^{-1}X_I'X_I(X_I'X_I+\lambda I)^{-1}\right].$$

To simplify notation, we denote by $\sigma_j = \sigma_j(X'X)$ as the jth eigenvalue of $X'X$.

Bias. Let S_I be the diagonal matrix of the singular values of X_I, and thus $S_I^2 = \operatorname{diag}(\sigma_1,\ldots,\sigma_{p_0})$. We diagonalize $X_I'X_I = V_IS_I^2V_I'$ and thus

$$(X_I'X_I+\lambda I)^{-1} = V_I\left(S_{I,p_0}^2+\lambda I_{p_0}\right)^{-1}V_I' = V_I\operatorname{diag}\left(\left\{(\sigma_j+\lambda)^{-1}\right\}_{j\leq p_0}\right)V_I'$$

and then

$$A_I = V_I \left(S_{I,p_0}^2 + \lambda I_{p_0}\right)^{-1} S_{I,p_0}^2 V_I' - I_{p_0} = -V_I \text{diag}\left(\left\{\lambda/(\sigma_j+\lambda)\right\}_{j\le p_0}\right) V_I'. \tag{B.9}$$

Substitute (B.7) and (B.9) into (B.8):

$$\text{bias}_I\left(\lambda\right)^2 = \rho'\Lambda_I'\Sigma_{X,I}^{-1} V_I \text{diag}\left(\left\{\frac{\lambda}{\sigma_j+\lambda}\right\}_{j\le p_0}\right) V_I'\Sigma_{X,I} V_I \text{diag}\left(\left\{\frac{\lambda}{\sigma_j+\lambda}\right\}_{j\le p_0}\right) V_I'\Sigma_{X,I}^{-1}\Lambda_I\rho$$

If we drop the first K elements $\left\{\lambda/(\sigma_j+\lambda)\right\}_{j\le K}$ from $\text{diag}\left(\left\{\lambda/(\sigma_j+\lambda)\right\}_{j\le p_0}\right)$ to produce $\text{diag}\left(\left\{\frac{\lambda}{\sigma_j+\lambda}\right\}_{-K}\right) := \text{diag}\left(\left\{\frac{\lambda}{\sigma_j+\lambda}\right\}_{K<j\le p_0}\right)$, in the above expression the associated eigenvectors are also eliminated, which reduces V to V_I. The bias becomes

$$\begin{aligned}
&\text{bias}_I\left(\lambda\right)^2 \\
\ge\ & \rho'\Lambda_I'\Sigma_{X,I}^{-1} V_{I,-K} \text{diag}\left(\left\{\frac{\lambda}{\sigma_j+\lambda}\right\}_{-K}\right) V_{I,-K}'\Sigma_{X,I} V_{I,-K} \text{diag}\left(\left\{\frac{\lambda}{\sigma_j+\lambda}\right\}_{-K}\right) V_{I-K}'\Sigma_{X,I}^{-1}\Lambda_I\rho \\
\ge\ & \left(\frac{\lambda}{\sigma_{K+1}+\lambda}\right)^2 \rho'\Lambda_I'\Sigma_{X,I}^{-1}\Lambda_I\rho = \left(\frac{\lambda}{\sigma_{K+1}+\lambda}\right)^2 \rho'\Lambda_I'\left(\Lambda_I\Lambda_I' + \sigma_u^2 I_{p_0}\right)^{-1}\Lambda_I\rho,
\end{aligned}$$

where the second inequality holds as $\frac{\lambda}{\sigma_j+\lambda} \ge \frac{\lambda}{\sigma_{K+1}+\lambda}$ for all $K < j \le p_0$. Then we have

$$\rho'\Lambda_I'\left(\Lambda_I\Lambda_I' + \Sigma_{I,u}\right)^{-1}\Lambda_I'\rho \ge \rho'\rho \frac{\sigma_K(\Lambda_I'\Lambda_I)}{\sigma_K(\Lambda_I'\Lambda_I) + \sigma_u^2} \ge \frac{\rho'\rho}{2\sigma_u^2}$$

where the last inequality holds for sufficiently large sample size as $\sigma_u^2/\sigma_K(\Lambda_I'\Lambda_I) \to 0$. We conclude that bias

$$\text{bias}_I\left(\lambda\right)^2 \ge \left(\frac{\lambda}{\sigma_{K+1}+\lambda}\right)^2 \frac{\|\rho\|^2}{2\sigma_u^2}.$$

The computation of the **variance** is straightforward.

$$\begin{aligned}
\text{Var}_I(\lambda) &= \sigma_e^2\ \text{tr}\left[\Sigma_{X,I} V_I \left(S_I^2 + \lambda I_{p_0}\right)^{-1} S_I^2 \left(S_I^2 + \lambda I_{p_0}\right)^{-1} V_I'\right] \\
&\ge \sigma_e^2\ \text{tr}\left[\text{Cov}\left(u_{I,t}\right) V_I \left(S_I^2 + \lambda I_{p_0}\right)^{-1} S_I^2 \left(S_I^2 + \lambda I_{p_0}\right)^{-1} V_I'\right] \\
&= \sigma_e^2\sigma_u^2\ \text{tr}\left[\left(S_I^2 + \lambda I_{p_0}\right)^{-1} S_I^2 \left(S_I^2 + \lambda I_{p_0}\right)^{-1}\right] \\
&= \sigma_e^2\sigma_u^2 \sum_{j=1}^{p_0} \frac{\sigma_j}{(\sigma_j+\lambda)^2} \ge \sigma_e^2\sigma_u^2 \sum_{j=K+1}^{p_0} \frac{\sigma_j}{(\sigma_j+\lambda)^2}
\end{aligned}$$

$$
\begin{aligned}
&\geq \sigma_e^2\sigma_u^2\,(p_0-K)\min_{K<j\leq p_0}\frac{\sigma_j}{(\sigma_j+\lambda)^2}\\
&\geq \sigma_e^2\sigma_u^2\,(p_0-K)\frac{\sigma_{p_0}}{(\sigma_{K+1}+\lambda)^2}
\end{aligned}
$$

where in the first inequality we used: if $A-B$ is semipositive definite, then for any matrix V, $\mathrm{tr}(AVV')-\mathrm{tr}(BVV')=\mathrm{tr}(V'(A-B)V)\geq 0$ beause $V'(A-B)V$ is semipositive definite.

Summary. Given the lower bounds of the bias and variance, we have

$$
\mathrm{bias}_I(\lambda)^2+\mathrm{Var}_I(\lambda)\geq \underbrace{\left(\frac{\lambda}{\sigma_{K+1}+\lambda}\right)^2\frac{\|\rho\|^2}{2\sigma_u^2}}_{\mathrm{LB}}+\underbrace{\sigma_e^2\sigma_u^2\,(p_0-K)\frac{\sigma_{p_0}}{(\sigma_{K+1}+\lambda)^2}}_{\mathrm{LV}}.
$$

It is governed by σ_{K+1} and σ_{p_0}. We now show the order of these two eigenvalues.

For σ_{K+1}, let v_{K+1} be the p_0-dim sample eigenvector associated with it . The quadratic form

$$
\begin{aligned}
v_{K+1}'\frac{X_I'X_I}{n}v_{K+1} &= v_{K+1}'\Lambda_I\frac{F'F}{n}\Lambda_I'v_{K+1}+2v_{K+1}'\Lambda_I\frac{F'U'}{n}v_{K+1}+v_{K+1}'\frac{U_I'U_I}{n}v_{K+1}\\
&\leq 2\left(v_{K+1}'\Lambda_I\frac{F'F}{n}\Lambda_I'v_{K+1}+v_{K+1}'\frac{U_I'U_I}{n}v_{K+1}\right)\\
&\leq 2v_{K+1}'\Lambda_I'\frac{F'F}{n}\Lambda_Iv_{K+1}+2C_{u,p_0},
\end{aligned}
$$

where $\sigma_{\max}(\frac{U_I'U_I}{n})\leq C_{u,p_0}$ because $n\asymp p_0$.

Also, there is $K\times K$ matrix H_I so that columns of Λ_IH_I are eigenvectors of $\Lambda_I\frac{F'F}{n}\Lambda_I'$, and $\|H_I^{-1}\|=O_P(\psi_{p,n}^{1/2})$. Let V_K denote the $p_0\times K$ matrix whose columns are the top K eigenvectors of $X_I'X_I$. We have $v_{K+1}'V_K=0$. The Sin-theta inequality guarantees that

$$
\begin{aligned}
\left\|v_{K+1}'\Lambda_I\right\| &\leq \left\|v_{K+1}'\Lambda_IH_I\right\|\|H_I^{-1}\|\leq\left\|v_{K+1}'(\Lambda_IH_I-V_K)\right\|O_P(\psi_{p,n}^{1/2})\\
&\leq \sqrt{\psi_{p_0,n}}O_p(\sqrt{p_0/(\psi_{p_0,n}n)})=O_P(1).
\end{aligned}
\tag{B.10}
$$

A lower bound of the smallest eigenvalue σ_{p_0} can be derived in a similar fashion as that in the proof of Theorem 2. Let v_{p_0} be the eigenvector associated with σ_{p_0}, We have

$$
\begin{aligned}
\frac{\sigma_{p_0}}{n} &= v_{p_0}'\frac{X_I'X_I}{n}v_{p_0}\\
&= v_{p_0}'\Lambda_I'\frac{FF'}{n}\Lambda_Iv_{p_0}+2v_{p_0}'\Lambda_I'\frac{FU}{n}v_{p_0}+v_{p_0}'\frac{U'U}{n}v_{p_0}
\end{aligned}
$$

$$\geq \|\Lambda_I v_{p_0}\|^2 (1 + o_p(1)) - 2 \|\Lambda_I v_{p_0}\| \left\| \frac{FU}{n} v_{p_0} \right\| + c_{u,p_0}$$
$$\geq \|\Lambda_I v_{p_0}\|^2 - 2 \|\Lambda_I v_{p_0}\| \left\| \frac{FU}{n} v_{p_0} \right\| + c_{u,p_0} + o_p(1)$$

where the last line holds given $\|\Lambda_I v_{p_0}\| = O_p(1)$ as well. Conditional on this event $\left\{ \left\| \frac{FU}{n} v_{p_0} \right\| \leq \frac{\sqrt{c_{u,p_0}}}{2} \right\}$, which occurs with w.p.a.1. asymptotically, we continue the above display expression

$$\frac{\sigma_{p_0}}{n} \geq \|\Lambda_I v_{p_0}\|^2 - \sqrt{c_{u,p_0}} \|\Lambda_I v_{p_0}\| + c_{u,p_0} + o_p(1)$$
$$\geq \frac{3}{4} c_{u,p_0} + o_p(1) \geq \frac{1}{2} c_{u,p_0}$$

for sufficiently large sample size.

The above computation shows that there are two absolute constant $c_X, C_X \in (0, \infty)$ such that the event

$$c_X n \leq \sigma_{p_0} \leq \sigma_{K+1} \leq C_X n$$

holds w.p.a.1. In other words, all eigenvalues $\{\sigma_j\}_{K<j\leq p_0}$ are of order n. Hence

$$\text{bias}_I(\lambda)^2 + \text{Var}_I(\lambda) \geq \text{LB} + \text{LV}$$

where

$$\text{LB} \geq \left(\frac{\lambda}{\sigma_{K+1} + \lambda} \right)^2 \frac{\|\rho\|^2}{2\sigma_u^2} \geq \left(\frac{\lambda}{C_X n + \lambda} \right)^2 \frac{\|\rho\|^2}{2\sigma_u^2}$$

and

$$\text{LV} \geq \sigma_e^2 \sigma_u^2 (p_0 - K) \frac{c_X n}{(C_X n + \lambda)^2}.$$

Fix any constant $\bar{C} > 0$.

- If $\lambda \in [0, n\bar{C}]$, then $\text{LV} \geq \sigma_e^2 \sigma_u^2 \frac{c_X}{(C_X + \bar{C})^2} \frac{p_0 - K}{n} \geq c_0$ for some $c_0 > 0$.
- If $\lambda > n\bar{C}$, then $\text{LB} \geq \left(\frac{\bar{C}}{C_X + \bar{C}} \right)^2 \frac{\|\rho\|^2}{2\sigma_u^2} > c_0$ for some $c_0 > 0$.

This implies $\inf_{\lambda \geq 0} [\text{bias}_I(\lambda)^2 + \text{Var}_I(\lambda)] \geq c_0 > 0$.

The OLS estimator is a special case of ridge regression with $\lambda = 0$, under which $A_I = (X_I' X_I)^{-1} X_I' X_I - I_{p_0} = 0$, bias is zero. But the variance does not vanish.

References

Arora, S., N. Cohen, W. Hu, and Y. Luo (2019). Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems 32*, 7413–7424.

Atanasov, V., S. V. Møller, and R. Priestley (2020). Consumption fluctuations and expected returns. *The Journal of Finance 75*(3), 1677–1713.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*, 135–171.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*, 191–221.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*(4), 1133–1150.

Bai, Z. and Y. Yin (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability 21*(3), 1275–1294.

Ball, R. and V. V. Nikolaev (2022). On earnings and cash flows as predictors of future cash flows. *Journal of Accounting and Economics 73*(1), 101430.

Barro, R. J. and J.-W. Lee (1994). Sources of economic growth. In *Carnegie-Rochester conference series on public policy*, Volume 40, pp. 1–46. Elsevier.

Bekaert, G. and M. Hoerova (2014). The vix, the variance premium and stock market volatility. *Journal of econometrics 183*(2), 181–192.

Belkin, M., D. Hsu, S. Ma, and S. Mandal (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences 116*(32), 15849–15854.

Belkin, M., D. Hsu, and J. Xu (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science 2*(4), 1167–1180.

Chava, S., M. Gallmeyer, and H. Park (2015). Credit conditions and stock return predictability. *Journal of Monetary Economics 74*, 117–132.

Chen, X., Y. H. Cho, Y. Dou, and B. Lev (2022). Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research 60*(2), 467–515.

Chen, Y., G. W. Eaton, and B. S. Paye (2018). Micro (structure) before macro? the predictive power of aggregate illiquidity for stock returns and economic activity. *Journal of Financial Economics 130*(1), 48–73.

Chernozhukov, V., C. Hansen, and Y. Liao (2017). A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics 45*(1), 39–76.

Chinot, G., M. Löffler, and S. van de Geer (2022). On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *The Annals of Statistics 50*(4), 2306–2333.

Colacito, R., E. Ghysels, J. Meng, and W. Siwasarit (2016). Skewness in expected macro fundamentals and the predictability of equity returns: Evidence and theory. *The Review of Financial Studies 29*(8), 2069–2109.

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

Didisheim, A., S. B. Ke, B. T. Kelly, and S. Malamud (2023). Complexity in factor pricing models. Technical report, National Bureau of Economic Research.

Fairfield, P. M., R. J. Sweeney, and T. L. Yohn (1996). Accounting classification and the predictive content of earnings. *Accounting Review*, 337–355.

Fan, J., Y. Ke, and K. Wang (2020). Factor-adjusted regularized model selection. *Journal of Econometrics 216*(1), 71–85.

Fan, J., Z. T. Ke, Y. Liao, and A. Neuhierl (2022). Structural deep learning in conditional asset pricing. *Available at SSRN 4117882*.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B 75*, 603–680.

Feltham, G. A. and J. A. Ohlson (1995). Valuation and clean surplus accounting for operating and financial activities. *Contemporary accounting research 11*(2), 689–731.

Giannone, D., M. Lenza, and G. E. Primiceri (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica 89*(5), 2409–2437.

Goyal, A., I. Welch, and A. Zafirov (2023). A comprehensive 2021 look at the empirical performance of equity premium prediction ii. *Swiss Finance Institute Research Paper* (21-85).

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Hansen, C. and Y. Liao (2018). The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications. *Econometric Theory*, 1–45.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics 50*(2), 949.

He, Y. (2023). Ridge regression under dense factor augmented models. *Journal of the American Statistical Association*, 1–13.

Hirshleifer, D., K. Hou, and S. H. Teoh (2009). Accruals, cash flows, and aggregate stock returns. *Journal of Financial Economics 91*(3), 389–406.

Huang, D., F. Jiang, J. Tu, and G. Zhou (2015). Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies 28*(3), 791–837.

Jondeau, E., Q. Zhang, and X. Zhu (2019). Average skewness matters. *Journal of Financial Economics 134*(1), 29–47.

Jones, C. S. and S. Tuzel (2013). New orders and asset prices. *The Review of Financial Studies 26*(1), 115–157.

Kelly, B. and S. Pruitt (2013). Market expectations in the cross-section of present values. *The Journal of Finance 68*(5), 1721–1756.

Kelly, B. T., S. Malamud, and K. Zhou (2022). The virtue of complexity in return prediction. Technical report, National Bureau of Economic Research.

Lee, S. and S. Lee (2023). The mean squared error of the ridgeless least squares estimator under general assumptions on regression errors. *arXiv preprint arXiv:2305.12883*.

Marchenko, V. A. and L. A. Pastur (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik 114*(4), 507–536.

Martin, I. (2017). What is the expected return on the market? *The Quarterly Journal of Economics 132*(1), 367–433.

McCracken, M. W. and S. Ng (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Mei, S. and A. Montanari (2019). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*.

Møller, S. V. and J. Rangvid (2015). End-of-the-year economic growth and time-varying expected returns. *Journal of Financial Economics 115*(1), 136–154.

Ng, S. (2013). Variable selection in predictive regressions. *Handbook of economic forecasting 2*, 752–789.

Nissim, D. and S. H. Penman (2001). Ratio analysis and equity valuation: From research to practice. *Review of accounting studies 6*, 109–154.

Ohlson, J. A. (1995). Earnings, book values, and dividends in equity valuation. *Contemporary accounting research 11*(2), 661–687.

Penman, S. H. (1998). A synthesis of equity valuation techniques and the terminal value calculation for the dividend discount model. *Review of accounting studies 2*, 303–323.

Penman, S. H. and T. Sougiannis (1998). A comparison of dividend, cash flow, and earnings approaches to equity valuation. *Contemporary accounting research 15*(3), 343–383.

Rapach, D. E., M. C. Ringgenberg, and G. Zhou (2016). Short interest and aggregate stock returns. *Journal of Financial Economics 121*(1), 46–65.

So, E. C. (2013). A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics 108*(3), 615–640.

Spiess, J., G. Imbens, and A. Venugopal (2023). Double and single descent in causal inference with an application to high-dimensional synthetic control. *arXiv preprint arXiv:2305.00700*.

Stock, J. and M. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies 21*(4), 1455–1508.