# Does Noise Hurt Economic Forecasts?*

Yuan Liao† Xinjie Ma ‡ Andreas Neuhierl§ Zhentao Shi ¶

**Abstract**

No, it can help. To support our argument, we show that economic forecast models driven by latent factors are not sparse. This fact allows us to establish a compelling result that including noise in predictions yields greater benefits than excluding it, which contradicts the common practice of removing noise from predictors via variable selection techniques. Empirically, we apply a pseudo-OLS approach to four real-data applications including forecasting the U.S. inflation rate. The performance of our method that interpolates the in-sample data surpasses many commonly used models that rely on dimension reduction.

*The authors are grateful to Joachim Freyberger, Christian Hansen, Benjamin Holcblat, Wenxin Jiang, Oliver Linton, Alberto Martin-Utrera, Semyon Malamud, Ulrich Müller, Andrew Patton, Peter C.B. Phillips, Michael Pollmann, Seth Pruitt, Tom Severini, Youngki Shin, Yao Zheng, seminar participants at Aalto, Duke, Tsinghua, Peking, McMaster and conference participants at the SFS Cavalcade 2024 for valuable comments. Shi acknowledges the partial financial support from the Research Grants Council of Hong Kong No. 14617423.

†Department of Economics, Rutgers University, yuan.liao@rutgers.edu

‡NUS Business School, National University of Singapore, xinjiema@nus.edu.sg

§Olin Business School, Washington University in St. Louis, andreas.neuhierl@wustl.edu

¶Department of Economics, The Chinese University of Hong Kong, zhentao.shi@cuhk.edu.hk

# 1 Introduction

In the realm of economic forecasting, the forecast outcome is typically shaped by a parsimonious set of low-dimensional economic factors summarizing the state of the macroeconomy, financial markets and policy-related indices. These factors are, however, unobserved (latent) economic variables which cannot be directly used for economic forecasts. Instead, economists often rely on high-dimensional predictors that are considered to contain predictive information regarding the outcome variable, implicitly assuming that they are driven by the latent factors. The informativeness of these predictors often varies, with some providing robust predictive signals and others serving as mere noise, exhibiting minimal conditional predictive power. Consequently, it is customary for economists to attempt variable selection and employ dimension reduction techniques to enhance the precision of economic forecasts, including Lasso, Ridge, principal components analysis (PCA), and partial least squares, among others (Sala-I-Martin, 1997; Connor and Korajczyk, 1988; Stock and Watson, 2002; Bai and Ng, 2006; Belloni et al., 2014; Kelly and Pruitt, 2013; Jurado et al., 2015). See Ng (2013) for an excellent review for variable selections in economic forecasts and the references therein.

However, variable selection is often not innocuous and comes with at least three challenges to empirical studies. First, as shown by many researchers, consistent variable selection is difficult to achieve (Leeb and Pötscher, 2008; Belloni et al., 2012, 2014). Secondly, recent empirical findings provide evidence that the sparsity assumption is fragile in many economic settings, e.g., Giannone et al. (2021); Kozak et al. (2020); Kolesár et al. (2023). Several recent empirical studies in economic forecasting have illuminated the fact that optimal forecasting behavior is typically not achieved by sparse models. For instance, using four empirical forecast exercises, Giannone et al. (2021) concluded 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.*" Kolesár et al. (2023) documented that the sparsity assumption is fragile in empirical studies and rejected it in three empirical applications when comparing with the ordinary least squares estimator. A parallel observation is also supported by Kozak et al. (2020) in the context of asset pricing.

The third and perhaps most surprising challenge is that, variable selection, even if perfectly conducted, may not yield optimal performance for predictions. Consider a hypothetical scenario in which an economist has collected a large number of predictors — some of them are informative for forecasting the outcome variable of interest, whereas

others are pure noise. Assuming the economist has *a priori* knowledge distinguishing informative predictors from noise, it is then widely accepted that the economist should initiate variable selection by excluding all noise from the predictor set. In this hypothetical scenario, variable selection is a straightforward task because she knows the identities of the informative predictors and noise.

Strikingly, we argue that in most economic forecasting scenarios, the economist should retain noise in her set of predictors. Formally, we prove a surprising result — under some conditions including noise in predictions yields greater benefits than its exclusion. Furthermore, if the total number of predictors is not sufficiently large, she should intentionally add more noise. In doing so, 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 the informative predictors only and discard the uninformative features.

This paper makes three contributions. First, we formally show that a class of important economic forecast models is better approximated by dense models rather than sparse models (Theorem 1). We consider a large number of economic variables driven by a low-dimensional factor structure, and are interested in forecasting one of the variables by the others. As the low-dimensional latent variables (factors) are unobservable, in empirical econometric applications the linear regression is the "go-to model" for such predictions; it serves as a reduced-form "working model" for approximation. This is the scenario of economic forecasting we consider throughout the paper. In response to the empirical findings elucidated in previous studies, we provide a theoretical underpinning by showing that the predictive signal is densely distributed among the high-dimensional regression coefficients of the working model, instead of sparsely concentrating on few predictors. This theoretical explanation sheds light on the non-sparse nature of economic forecasting models as observed in the recent empirical literature.

As the second contribution, we show how economic forecasts can benefit from the blessing of overfitting in the presence of many potentially uninformative predictors and noise, given the non-sparse nature of economic forecasting models (Theorem 2). Rather than attempting variable selection, we recommend intentionally increasing the dimension by adding noise and using 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. This estimator is always well defined regardless of the dimensionality. Pseudo-OLS is often referred to 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

approaching zero.

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 noise. The insight lies in the crucial observation that, with a sufficiently large total number of predictors, denoted by $p$, the overall variance is *diversified away*. Consequently, out-of-sample variance decays as $p$ approaches infinity. This phenomenon is the embodiment of the "double descent/benign overfitting" phenomenon in the machine learning literature, e.g., e.g., Belkin et al. (2019) and Arora et al. (2019). That is, as the model complexity exceeds the sample size and continues increasing, a second descent of the prediction risk occurs in the extremely over-parametrized regime. Contrary to the aforementioned statistical theories, in which all features are important and intuitively each additional feature adds information, we work in setting where the true model is low dimensional and the additional variables are pure noise. This require the development of new techniques to establish a benign overfitting results.[1] We explore the intuition that the dense features of economic forecasting models diversify away overall variance, even when a significant proportion of predictors are completely uninformative about the outcome, i.e. pure noise.

In practice, we must determine the total number of predictors (the sum of informative and uninformative predictors) $p$ as a tuning parameter. We propose a way to determine the total number of predictors in a way such that the benefits derived from reducing variance outweigh the costs associated with bias inflation. At the same time, it turns out that the procedure remains quite robust with respect to the choice of $p$. We suggest choosing $p$ using standard cross-validation for both cross-sectional predictions and time-series predictions and illustrate that it works remarkably well in many applications.

In one of the empirical illustrations, we apply pseudo-OLS with intentionally added noise 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 $300 \sim 10,000$ noise 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 noise. The performance surpasses many sophisticated machine learning models for forecasting the U.S. equity premium.

The last contribution presents a compelling finding that forecasting relying on "perfect variable selections" does not lead to optimal forecasting performance (Theorem 3).

[1] We present a more elaborate discussion of the formal differences in Section 2.4.

Specifically, we explore an ideal scenario when noise has been eliminated from the predictor set and all informative predictors have been retained. In this case, Ridge regression fails to achieve the optimal forecast mean squared error when predictive signals are densely distributed among the informative predictors.[2]

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 Fan et al. (2022) for asset pricing. Kelly et al. (2024) and Didisheim et al. (2023) observe an interesting phenomenon, which they dub "the virtue of complexity". They show that the Sharpe ratio monotonically increases as model complexity grows in the overparametrized regime. Most of the existing work, however, are concerned with weakly correlated designs, whereas informative predictors in many economic forecasts are strongly correlated due to the common economic factors. Finally, our result is related to the "training with noise" literature (Sietsma and Dow, 1991; Bishop, 1995), who showed that addition of random noise to the input data during training can lead to improvements in neural network generalization. The difference is that in "training with noise", random vectors are being added to each input predictor, who "smear out" the importance of each single data point, making the network just capture the overall pattern but be less affected by individual data points. In contrast, we add noise as additional predictors and merge with existing features thus increasing the dimension. In our setting, the increased forecasting performance is mainly due to the variance diversification.

We adopt the following notation. Let $\|A\|$ denote the $\ell_2$ norm if $A$ is a vector, or the operator norm if $A$ is a matrix. Let $\sigma_j(\cdot)$ be the $j$th 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]We focus on Ridge regression as the benchmark for our theoretical investigations, complemented by comparisons with Lasso and PCA in our numerical analyses.

# 2 The Economic Forecast 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:

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

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 $f_t$ to one. Note that while observations are indexed by subscript $t$, we allow cross-sectional forecasts, in which $y_t$ denotes the outcome of the $t$-th subject, or the time series forecast where observations follow a temporal order.

In addition, the economist observes 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 about $y_t$. 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 $\lambda_i$ is a vector of loadings for the $i$-th variable. The mean independence condition $\mathbb{E}(u_{i,t}|f_t, \epsilon_{y,t}) = 0$ entails that the predictive power of $x_{i,t}$ stems from the latent factors only. Linking $y_t$ and $X_t$ via a common factor structure is common in economic forecasting, e.g. Forni and Reichlin (1998) and De Mol et al. (2008).

We emphasize that $f_t$ may be weak in the sense that some components of $X_t$ may not depend on $f_t$. In this case $X_t$ can be partitioned 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 $\Lambda_I$ is a $p_0 \times K$ matrix of nonzero $\lambda_i$ that loads on the factors. Hence

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

Here for ease of exposition we place $X_{I,t}$ into the first $p_0$ elements of $X_t$; in practice the forecast outcomes are invariant to re-ordering of the regressors. We allow the identities of informative predictors and noise to be either known or unknown. Our methodology treats the two cases equally, and we refrain from employing variable selection procedures to screen off the noise.

Our paper considers the following scenarios that are routinely encountered in empirical studies:

**Scenario I: Many Predictors With Noise.** In economic forecasting, researchers often collect a large number of predictors which are *potentially* correlated with the outcome variable of interest. But many of the collected predictors may contain little predictability. Meanwhile, economic theory typically provides no clear guidance about which predictors signifies the factors and which do not. As a result, even though the factors are strong within the informative predictors, the strength can be diluted over all predictors due to the presence of many noise variables. This is the case where the partition $(X'_{I,t}, X'_{N,t})$ is unknown.

**Scenario II: Intentionally Included Noise.** Suppose researchers have *a priori* knowledge that the collected $p_0$ predictors are all informative, but $p_0$ is not much larger than the sample size. Traditional statistical wisdom suggests using only the informative predictors. However, as a novel finding in this paper, we shall argue that many pure noise variables should be *intentionally* included to make the total number of predictors $X_t = (X_{I,t}, X_{N,t})$ much larger than the sample size. A key contribution of this paper is revealing the benefit of including many noise variables in the economic forecast context. 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 noise variables. 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 noise so that $p/n \to \infty$.

## 2.2 The Working Model

While (2.1) underlines the true DGP, it is infeasible in applications as the factors are latent. One standard approach is first estimating the latent factors from model (2.2)

using principal components analysis (PCA), dynamic factors, or partial least squares, and then conducting forecast based on the estimated factors.

We proceed differently. We use the collected predictors $X_t$ to forecast, potentially with the inclusion of many noise variables, $X_{N,t}$, in 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):

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

where $(\sum_{t=1}^{n} X_t X_t')^{+}$ denotes the pseudoinverse of the design matrix.[3] This estimator $\widehat{\beta}$ is exactly the ordinary least squares if $p < n$, whereas the pseudo-inverse makes sure that it is well defined regardless of the $p/n$ ratio, including $\dim(X_t) = p > n$. This estimator is also known as "ridgeless regression", as is shown by Hastie et al. (2022):

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

We would like to emphasize that $\widehat{\beta}$ perfectly interpolates the in-sample data in the sense that if $p \geq n$,

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

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

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

Before we conclude this subsection, we add a remark about computation. To efficiently compute $\widehat{\beta}$ in high-dimensions, 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 expect $p$ to be of

[3] 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.

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. When the sample size is moderate, it is much faster than the usual pseudoinverse functions in leading computing software, such as '`pinv`' in Matlab.[4]

## 2.3 The Fundamental Questions and Main Results

The objective of this paper is to answer the following two fundamental questions: *Is the economic forecast model, as defined in Section 2.1, better approximated by a sparse or a dense model? In addition, should economists attempt variable selection before conducting forecasts?*

### 2.3.1 Practical recommendations

Consider a hypothetical scenario where it is known that the first $p_0$ (with $p_0 < n$) predictors are all informative and the other $p - p_0$ are pure noise. Then the well-accepted statistical wisdom will naturally guide us to exclude the noise and use the informative predictors only, 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 Ridge regression. Indeed, we will show that if the idiosyncratic components in $u_t$ in (2.2) are mutually uncorrelated, then

[4]The reduced SVD computes fast when $n$ is not very 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 $(X'X)^+X' = X^+$. When the sample size is also large, however, it is faster than reduced-SVD by directly solving the system of equations $A\beta = B$, where $A = X'X + \epsilon I$ for a very small $\epsilon > 0$ and $B = X'Y$. In Python, this can be done via `beta= np.linalg.solve(A, B)`.

the true latent-factor-based DGP induces the following linear regression model:

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

Since the identities of $X_{I,t}$ and $X_{N,t}$ are known, the researcher may want to exclude $X_{N,t}$ from the forecast.

Surprisingly, we shall argue that for the economic forecast problems under consideration, if $p_0 < n$ it is better to retain the noise 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 the informative predictors are driven by a set of latent economic state variables (factors), then economists should retain all the predictors and use pseudo-OLS (2.5)-(2.6), instead of attempting variable selections.*

II. *If the number of predictors is not sufficiently large, then the economist should intentionally add more features until p is sufficiently large. This can be beneficial even if the added features have little or no predictability, e.g., pure noise.*

The insight lies in the observation that, driven 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 consist of pure noise. Meanwhile, the dense predictive signals maintain the forecast bias at a modest level.

### 2.3.2 An improvement: Noise-denoise procedure

As one of the practical recommendations, if the number of informative predictors in the dataset is not sufficiently large relative to the sample size, we recommend adding noise to intentionally increase the overall dimensionality. This rationalizes the idea that a substantially overfit model reduces the out-of-sample variance. In the meantime, inspired by the classical Rao-Blackwell argument in the statistical literature, this predictor can be further improved.

Recall that $(X_I, Y)$ denotes the in-sample data, where $X_I$ is the data matrix of the original predictors. Write the out-of-sample features as

$$X_{\text{new}} = (X'_{\text{new},I}, X'_{\text{new},N})'$$

where $X_{\text{new},I}$ denotes the set of pre-determined informative predictors that come with the data, and $X_{\text{new},N}$ is the intentionally added noise. Accordingly, we partition $\widehat{\beta} = (\widehat{\beta}_I', \widehat{\beta}_N')'$ as the coefficients corresponding to $(X_{\text{new},I}', X_{\text{new},N}')'$.

**Noise-denoise forecast**: After intentionally adding noise and obtaining $\widehat{\beta}$, we simply forecast using

$$\widehat{y}_{\text{new},I} := (X_{\text{new},I}', 0')' \widehat{\beta} = X_{\text{new},I}' \widehat{\beta}_I.$$

That is, we only pair the out-of-sample informative predictor $X_{\text{new},I}$ with the trained coefficient $\widehat{\beta}_I$. Alternatively, we can also compute this $\widehat{y}_{\text{new},I}$ multiple times with different sets of generated noise $X_N$ and then average the forecasts, as described in the following algorithm:

**Algorithm 1.** *Denoised estimator*

**Step 1 (adding noise):** *Generate $n \times (p - p_0)$ i.i.d. $N(0,1)$ matrix as the noise $X_N$. Merge with $X_I$ so that $X = (X_I, X_N)$, and $\widehat{\beta} = (X'X)^+ X'Y$. Let $\widehat{\beta}_I$ be the subvector of $\widehat{\beta}$ corresponding to $X_I$ and $\widehat{y}_{\text{new},I} := X_{\text{new},I}' \widehat{\beta}_I$.*

**Step 2 (denoising):** *Repeat Step 1 for $B$ times to obtain $\widehat{y}^1_{\text{new},I}, ..., \widehat{y}^B_{\text{new},I}$. Forecast using $\widehat{y}^*_{\text{new}} := \frac{1}{B} \sum_{b=1}^B \widehat{y}^b_{\text{new},I}$.*

The above algorithm is designed to approximate

$$\mathbb{E}(\widehat{y}_{\text{new},I} | X_I, Y, X_{\text{new},I}),$$

which is a forecast insensitive to the realization of $X_N$ as its randomness is averaged out via the $B$ repetitions.

### 2.3.3 Overview of main results

We will establish the following **three main results**:

1. (Theorem 1) In the predictive model

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

the coefficient $\beta$ is *dense* in the sense that $\|\beta\| \to 0$.

2. (Theorem 2) If $(p_0, p, n)$ satisfies (recall $p - p_0$ is the number of noise variables in the predictor set):

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

then 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.

3. (Theorem 3) In the case where $p_0/n \to c \in (0, 1)$, that is, there are moderately many informative predictors and only informative predictors are used without adding noise, the predictive MSE of both OLS and Ridge are strictly larger than the oracle predictive risk and thus are sub-optimal.

These results provide a clear answer to the fundamental questions we raised in the beginning of this paper: Economic forecasts can be better approximated by non-sparse models, and moreover the inclusion of noise in predictions yields greater benefits than its exclusion.

## 2.4 Relation with the linear double descent literature

Contrary to conventional statistical wisdom, which asserts that overfitting 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 *double descent* phenomenon on the prediction risk, which has gained increasing attention in the machine learning community. That is, as the model complexity exceeds the sample size and continues to grow, a second descent of the prediction occurs in the extremely overparametrized regime. It was first illustrated in the empirical work by Belkin et al. (2019), Hastie et al. (2022), and Arora et al. (2019), and its theory has been explored in linear models with Ridge regressions, e.g., Mei and Montanari (2019); Belkin et al. (2020); Dobriban and Wager (2018); Lee and Lee (2023).

Our work differs from the latest statistical literature on overparametrized models. The latter treats the high-dimensional regression model with random coefficient $\beta$ drawn from a *prior distribution* as the true underlying DGP of the outcome data. In addition, the predictor design matrix is assumed to have bounded eigenvalues, implying that predictors are nearly mutually independent. Our study distinguishes itself in three crucial aspects.

Firstly, many economic objectives for forecasts are inherently linked to only a few latent factors determining the economic status of the forecasting environment. Inspired

by this, we assume that model (2.1) is the true DGP, while treating model (2.4) as a working model. Given this specification, the observed high-dimensional predictors $X_t$ do not "cause" the outcome variable $y_t$ but instead their predictive power is inherited from the latent economic factors. As a direct consequence, we prove that the prediction coefficients are densely distributed across predictors ($\|\beta\| \to 0$), which invalidates classical forecasting methods based on dimension reductions. This observation is closely aligned with empirical findings in Giannone et al. (2021).

Secondly, the latent factors deliver high mutual correlations among the informative predictors, causing a distinctive approximate low-rank representation in the predictor covariance matrix: there are a few eigenvalues growing fast to infinity. This structure is typically excluded by the random matrix theory in the recent statistical literature of the extreme overparametrization. For example, Hastie et al. (2022) derived the formula of predictive mean squared errors under what they called "latent space model", where the eigenvalues of the covariance matrix $X_t$ are assumes to be bounded. In sharp contrast, this paper considers economic forecasting models with latent factors, where the top $K$ eigenvalues of $\text{Cov}(X_t)$ diverge to infinity at an order $\sqrt{n\psi_{p,n}}$, where $\psi_{p,n}$ measures the strength of factors.

Lastly, we specifically focus on the impact of including a substantial amount of noise on economic forecasts and arrive at surprising results — adding noise transpires to be advantageous rather than detrimental. In the regime of moderately many informative predictors when $p_0$ is proportional to $n$, in particular, using only the informative predictors without intentionally added noise does not achieve the optimal forecast mean squared error.

## 2.5 Why not Lasso or PCA?

When the collection of predictors contains many genuinely uninformative variables (noise), the Lasso is one of the most popular forecasting methods, as the use of $\ell_1$-penalty can often remove the noise and thereby achieving dimension reduction. This is no longer the case, however, in economic forecasting exercises where the informative predictors carry predictive information through latent economic factors, for instance as macroeconomic variables and state variables. In addition, there are two key features that differentiate (2.7) from the usual setting of the Lasso forecasts: first, the latent factors introduce strong collinearity and substantially slowing down the statistical rate of convergence for the Lasso (e.g., Hansen and Liao (2018)). Secondly, the latent factors make the predictive

signals *densely* distributed among many informative predictors, under which Lasso cannot select variables with satisfactory representation (Chernozhukov et al., 2017; Giannone et al., 2021). In other words, the model is not sparse enough.

Meanwhile, PCA is another popular method to effectively extract the latent "indices" (or factors) from the large set of predictors. 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 very difficult to correctly estimate $K$, the number of factors. Even if $K$ is correctly specified, the factors are still estimated poorly. Above all, the presence of weak factors poses fundamental challenges, and recent work in the literature aims at feature selections to remove noise and therefore enhance the factor strength, e.g., Giglio et al. (2023) and Chao and Swanson (2022).

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

# 3 Economic Forecasts are Non-Sparse: A Theoretical Perspective

Both the target of prediction $y_t$ and the predictors $X_t$ in this paper's model are driven by low-dimensional latent variables:

$$y_t = \rho' f_t + \epsilon_{y,t} \tag{3.1}$$

$$X_t = \begin{pmatrix} X_{I,t} \\ X_{N,t} \end{pmatrix} = \begin{pmatrix} \Lambda_I \\ 0 \end{pmatrix} f_t + u_t, \tag{3.2}$$

where $\dim(X_t) = p$, $\dim(X_{I,t}) = p_0$ and $\dim(f_t) = K$. As the true factors are unobservable, the economist carries out forecasts using

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

Recently, a multitude of influential empirical studies have documented that competitive economic forecast results are often aligned with non-sparse models. This section aims to provide a theoretical understanding of this finding, by answering two questions:

*Question 1: Is $\beta$ in (3.3) a dense or sparse coefficient vector?*

*Question 2: Can the working model generate forecasts with the same accuracy as if the true factors were known?*

In particular, Question 2 addresses the gap between the two models: if $(\beta, \rho, f_t)$, could have been perfectly learned from the data, can the two models produce the same predictive MSE? Note that MSE would respectively converge to the marginal variances $\text{Var}(\epsilon_{y,t})$ and $\text{Var}(e_t)$ for the two models, hence this question is essentially asking whether the two residual variances are asymptotically the same.

The answers to both questions are affirmative as $p_0 \to \infty$. There are economic theories that imply models depending on low dimensional state variables (factors), e.g. Merton (1973) or Lucas (1978). Empirically, these unobserved state variables can be approximated by high dimensional observable variables $(X_t)$ collected by the economist. While none of the observable variables perfectly substitute for the latent state variables, their combination provides a strong approximation. This, in turn, implies that the empirical economic model is better approximated by dense models. This approximation is analogous to using many weak instruments to achieve strong identification power, where each individual instrument alone offers weak identification power.

Let us assume the following DGP for $X_t$:

**Assumption 1** (DGP of $X_t$)**.** *(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 (Recall that $\sigma_i(A)$ denotes the $i$-th largest singular value of matrix $A$)*

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

*(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, where $\Lambda = (\Lambda_I', 0')'$.*

*(iv)* $\mathbb{E}(\epsilon_{y,t} | f_t, u_t) = 0$, *and* $\mathbb{E}(u_t | f_t, \epsilon_{y,t}) = 0$.

This assumption allows $\psi_{n,p}/p_0$ to decay to zero, so that the informative predictors may be "semi-strong". In addition, to make the theoretical derivation transparent in closed-forms $\mathbb{E}(\rho' f_t | X_t)$ is assumed to be a linear function of $X_t$, which is commonly imposed in classical regression theory.

The following theorem provides answers to both Questions 1 and 2 raised in this section.

**Theorem 1.** *Suppose (3.1) and Assumption 1 holds. Also suppose the eigenvalues of* $\mathrm{Cov}(u_t)$ *are bounded away from both zero and infinity. In addition,* $\|\rho\| \leq C$ *for some absolute constant* $C$*. Then (3.3) holds:*

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

*where:*

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

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

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

This theorem characterizes the high dimensional coefficient $\beta$ on the collection of predictors.[5] The expression in result (i) yields an important insight: the working model for economic forecasts is non-sparse. This theoretically supports the empirical observation 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." As a result, forecasting based on sparse variable selection (e.g., Lasso) will fail to fully capture the predictive power in the forecast coefficients.

In addition, Result (ii) of Theorem 1 shows that if the strength of the factors, indexed by $\psi_{n,p}$, diverges as $n, p \to \infty$, then predicting using the working model ($X_t$-based) will be as good as using the oracle model with the unknown factors $f_t$.

Below we provide an example, where Theorem 1 directly answers a critical question in the asset pricing literature.

**Example 1** (Stochastic Discount Factors)**.** *The stochastic discount factor (SDF) plays a pivotal role in asset pricing theory (e.g., Cochrane (2009)). In an unconditional asset pricing model, the SDF,* $M_t$*, satisfies*

$$\mathbb{E}M_tR_t = 0$$

*for any asset's excess return* $R_t$*. If asset returns are explained by a set of risk factors,*

[5] Models (3.1)-(3.3) are also widely used in program evaluations using panel data, e.g., Hsiao et al. (2012). Also see Shi and Huang (2023) for a variable selection approach for post-selection inference when $\beta$ is dense.

*$f_t$, with factor risk premium $\lambda$, then it can be shown that*

$$M_t = \lambda' \operatorname{Cov}(f_t)^{-1} f_t + \epsilon_t + o_P(1), \quad \dim(R_t) \to \infty.$$

*where $\epsilon_t$ depends on the idiosyncratic errors in $R_t$. The above expression decomposes the SDF into a systematic component that is explained by the risk factors, and an unsystematic component $\epsilon_t$. One fundamental question in asset pricing is whether $M_t$ is explainable by assets' characteristics. Let $X_t$ be the high-dimensional returns of sorted portfolios, which are asset returns constructed using firm characteristics. They depend on the risk factors through*

$$X_t = \Lambda\lambda + \Lambda f_t + u_t$$

*where $\Lambda$ is the matrix of exposures to systematic risk ("betas" of sorted portfolios). Explaining $M_t$ using $X_t$ through a linear model leads to the working model*

$$M_t = X_t'\beta + e_t.$$

*By applying Theorem 1, $\beta$ is not a sparse vector. This answers the empirical question raised by Kozak et al. (2020), who observed that SDF sparsely formed by characteristics "cannot adequately summarize the cross-section of expected stock returns."*

# 4 The Blessing of intentional overfitting

## 4.1 Main results

The economist collects the in-sample data $(X, Y)$ where the $p$ columns of $X$ are partitioned as:

$$X = (\underbrace{X_I}_{p_0}, \underbrace{X_N}_{p-p_0})$$

corresponding to informative features and noise, determined by whether its $\lambda_i$ is zero or not. She aims to forecast an out-of-sample outcome $y_{\text{new}}$ using its feature $X_{\text{new}}$, which also includes both informative features and noise.

In a unified framework we allow Scenarios I and II in Section 2.1. In either scenario, the economist makes forecasts using pseudo-OLS:

$$\widehat{y}_{\text{new}} := X_{\text{new}}'\widehat{\beta}, \quad \text{where } \widehat{\beta} = (X'X)^+ X'Y.$$

Let the true out-of-sample outcome be generated by

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

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 $e_{\text{new}} := y_{\text{new}} - X'_{\text{new}}\beta$. Theorem 1 shows that $\text{Var}(e_{\text{new}}) \to \text{Var}(\epsilon_{y,\text{new}})$. Therefore, it suffices to focus on the first component of the MSE, which can be decomposed as:

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

where for any generic forecast, $\widehat{y}$, the bias is defined as

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

We impose several technical assumptions below to establish the benefit of adding noise for forecasts. While these assumptions facilitate proofs, they may be stronger than necessary. As this paper aims to introduce new research on benign overfitting in economic forecasting, we anticipate that these conditions can be relaxed in future research.

**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)* $\mathbb{E}(X_{\text{new}}X'_{\text{new}}|X) = \mathbb{E}X_tX'_t$, *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* $\sigma_e^2 > 0$.

Assumption 2 (i)–(iii) assume invariance of the training data and the new out-of-sample data. (iv) is a simplifying condition to regularize the conditional covariance matrix of the error term in the working model.[6]

**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 absolute constants* $C, c > 0$.

[6] As a demonstrative case, it is provable that Assumption 2 (iv) holds if $(f_t, \epsilon_{y,t}, u_t)$ are i.i.d. jointly normal (Lemma 4 (i) in the Appendix.)

Assumption 3 simplifies the technical arguments by assuming the idiosyncratic components are i.i.d. over both the cross sectional and time dimensions, which yields a fast rate of convergence for the prediction MSE.[7] We have the following theorem.

**Theorem 2.** *Suppose the assumptions of Theorem 1 and Assumptions 2 and 3 hold. In addition, suppose* $p > n$, $p = o(\psi_{p,n} n)$ *and* $n = o(p)$. *Then:*

*(i) The forecast bias and variance:*

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

*(ii) As* $n, p \to \infty$,

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

Theorem 2 implies that we can achieve the oracle predictive MSE, as if the latent factors were revealed by an oracle and used directly in forecasting. In addition, contrary to conventional wisdom suggesting that the variance is amplified as $p$ diverges, here the variance diminishes. The reduction of variance requires no condition on the predictive power, i.e., it does not matter whether the predictors are mostly noise or informative.

In the case that many components in $X$ are added noise, the prediction can be further improved by using either $\widehat{y}_{\text{new},I} = X'_{\text{new},I}\widehat{\beta}_I$ or its average conditioning on the data:

$$\widehat{y}^*_{\text{new}} = \mathbb{E}(\widehat{y}_{\text{new},I}|X_I, Y, X_{\text{new},I}).$$

This estimator, in terms of predictive MSE, is inspired by the classical Rao-Blackwell argument in the statistical literature. To see this, note that

$$X'_{\text{new}}\beta - \widehat{y}^*_{\text{new}} = \mathbb{E}(X'_{\text{new}}\beta - \widehat{y}_{\text{new}}|X_I, Y, X_{\text{new},I}).$$

Using the law of iterated expectations we have

$$\begin{aligned} \mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}^*_{\text{new}})^2|X_I] &= \mathbb{E}\left\{\left[\mathbb{E}(X'_{\text{new}}\beta - \widehat{y}_{\text{new}}|X_I, Y, X_{\text{new},I})\right]^2 \middle| X_I\right\} \\ &\leq \mathbb{E}\left\{\mathbb{E}\left[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})^2|X_I, Y, X_{\text{new},I}\right] \middle| X_I\right\} \end{aligned}$$

[7] In the appendix we consider the more general case to allow cross-sectional heteroskedasticity and dependencies among $u_{i,t}$ (Lemma 5 in the appendix).

$$= \mathbb{E}\left[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})^2 | X_I\right], \tag{4.1}$$

which shows that $\widehat{y}^*_{\text{new}}$ enjoys a smaller predictive MSE than that of $\widehat{y}_{\text{new}}$.

In practice if $\widehat{y}^*_{\text{new}}$ is computationally intensive as it involves repeated pseudo-OLS fitting given the realized noise in each round, we recommend the simple denoise predictor $\widehat{y}_{\text{new},I}$. The following theorem shows the improvement of $\widehat{y}^*_{\text{new}}$ and $\widehat{y}_{\text{new},I}$ upon $\widehat{y}_{\text{new}}$.

**Proposition 1** (Noise-denoise forecast)**.** *Consider Scenario II where the economist knows the identity of informative predictors and intentionally adds noise. Suppose the assumptions of Theorem 2 hold.*

*(i) For the denoise predictor* $\widehat{y}_{\text{new},I}$*,*

$$\text{bias}(\widehat{y}_{\text{new},I})^2 \leq \text{bias}(\widehat{y}_{\text{new}})^2,$$

$$\text{Var}[\widehat{y}_{\text{new},I} - \mathbb{E}(\widehat{y}_{\text{new},I}|X, X_{\text{new}})|X] \leq \text{Var}[\widehat{y}_{\text{new}} - \mathbb{E}(\widehat{y}_{\text{new}}|X, X_{\text{new}})|X].$$

*(ii) For the averaged denoise predictor* $\widehat{y}^*_{\text{new}}$*,*

$$\mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}^*_{\text{new}})^2|X_I] \leq \mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})^2|X_I].$$

## 4.2 Discussion

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) in a 3-factor model. Here the first $p_0 = \min\{p, n\}$ predictors are informative, while the remaining $p - p_0$ predictors are i.i.d. white noise variables 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 $p > n$, the added predictors are pure noise, and the variance starts to decay. This is consistent with our theory: as $p \to \infty$, the variance continues to decrease until the $1/n$ term becomes dominant.

In addition, the squared bias remains zero until $p = n$. Though after passing this threshold it starts to increase, it is in a much smaller magnitude than that of the variance. This is also consistent with what the theory predicts. The bias depicted on the left panel of Figure 1 does not diminish because here we fix $\psi_{p,n} \asymp n$ at $n = 100$ while we vary $p$ (up to 1000).

Overall, the predictive MSE (right panel) illustrates a double-descent phenomenon,

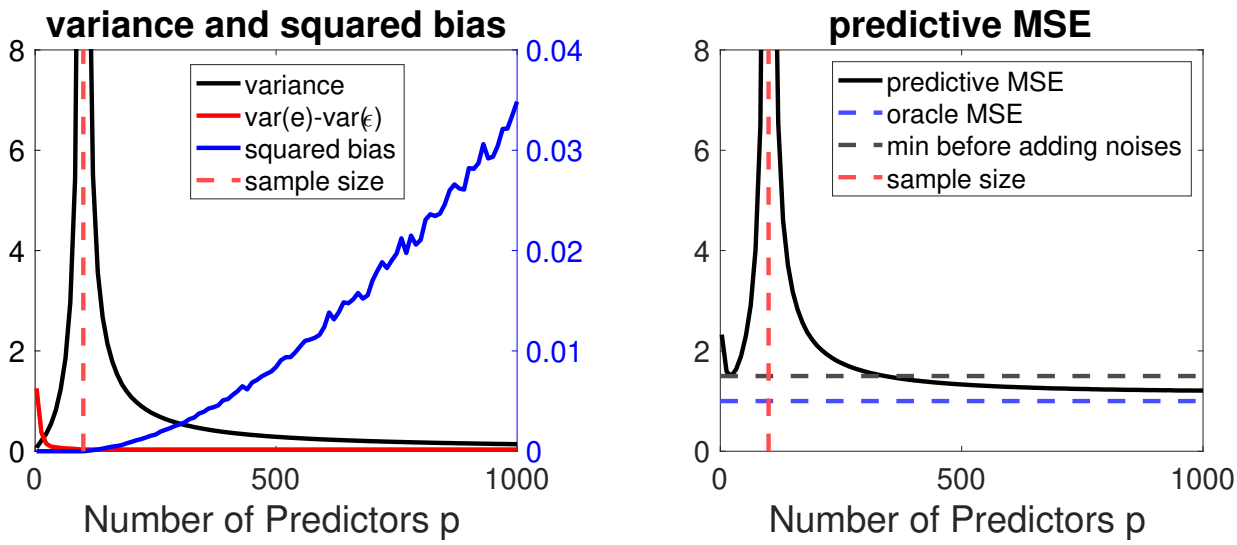


Figure 1: Bias-Variance Tradeoff

Notes: 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 $p_0 = \min\{p, n\}$ are informative predictors, generated using a 3-factor model of strong factors. The remaining $p - p_0$ are i.i.d. Gaussian noise. 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. The red curve on the left panel plots $\text{Var}(e_t) - \text{Var}(\epsilon_{y,t})$.

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 approaches the oracle MSE as if the latent factors were used for prediction.

## 4.3 How much noise to add?

Importantly, both Figure 1 and our theory indicate bias inflation when $p$ becomes excessively large. We therefore propose a data-driven way to choose the number of added noise. If it is believed that most of the informative predictors load strongly on the common factors, then $p_0 \asymp \psi_{p,n}$ and our theory shows:

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

The first term $\frac{p}{p_0 n}$ arises from the order of the bias, which inflates when too many noise variables are added. We can choose the optimal number of predictors $p$ by minimizing

the rate of convergence:

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

where $C$ is a constant to be chosen. This would also yield the optimal rate of convergence $\mathbb{E}[(X'_{\text{new}}\beta - \widehat{y}_{\text{new}})^2|X] = O_P(p_0^{-1/2} + n^{-1})$. A data-driven choice of $C$ is available via commonly used tuning strategies, which we now discuss.

For cross-sectional predictions where the ordering of data is unimportant, the "leave-one-out" is a computationally attractive method. Suppose $C$ is a candidate choice, then denote the total number of predictors, $p(C) := C \times n\sqrt{p_0}$ (after adding noise). For each $t \leq n$, let $\widehat{\beta}_{-t}(C)$ denote the pseudo-OLS estimator using $p(C)$ predictors and all data except the $t$-th observation. Then the optimal $C$ can be chosen by minimizing:

$$\text{leave-one-out:} \quad C^* := \arg\min_C \sum_{t=1}^{n} \left( y_t - X_t' \widehat{\beta}_{-t}(C) \right)^2.$$

Appealingly, we do not have to compute $n$ leave-one-out $\widehat{\beta}_{-t}(C)$ for this procedure, thanks to an elegant analytic formula given by Shen et al. (2023). They showed that the leave-one-out procedure is equivalent to minimizing:

$$C^* = \min_C \| \operatorname{Diag}(G(C))^{-1} G(C) Y \|^2, \quad G(C) = (X(C)X(C)')^{-1} \tag{4.2}$$

where $X(C)$ denotes the $n \times p(C)$ matrix of $p(C)$ predictors, and $\operatorname{Diag}(G)$ takes the diagonal elements of $G$. Here $G(C)$ is invertible because $\operatorname{rank}(X(C)) = n < p(C)$, whose computation is manageable so long as $n$ is not very large.

For time-series predictions where there is a natural ordering of the data, we recommend the usual training-testing tuning. Reserve a portion of observations at the start of the time dimension, based on which we conduct one-period ahead forecast, and the optimal $C^*$ is chosen to minimize the aggregated forecast error. After that, we carry $C^*$ over to the testing data for forecasting.

To illustrate the effectiveness of data-driven choices of $p_{cv} = p(C^*)$, Figure 2 plots the predictive MSE of two simulated designs, where the DGP is generated as in one of the simulation designs in Section 6.1 with weak factors. The first $p_0 = 200$ predictors are informative predictors (i.e. they load nontrivially on factors), whereas the rest $p - p_0$ predictors are intentionally added noise variables. In both cases the data-driven choice $p_{cv}$ suitably identifies a proper number of added noise.

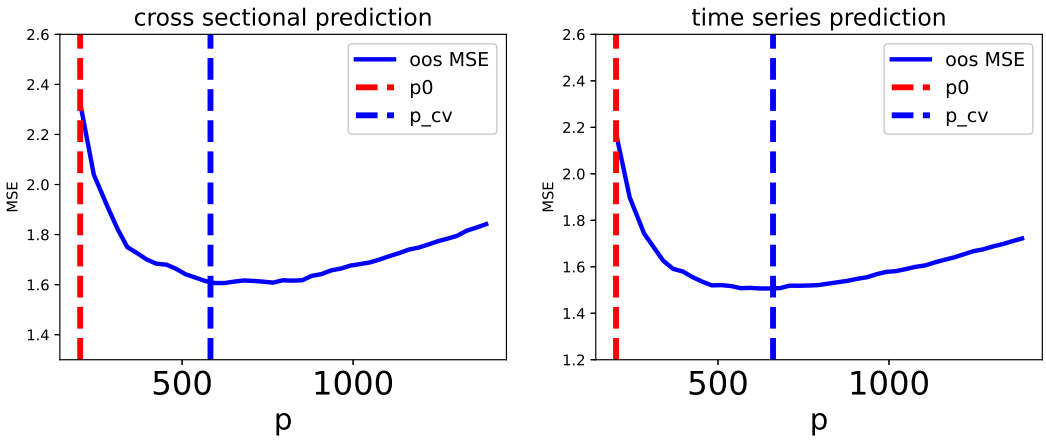


Figure 2: Predictive MSE and Data-driven Tuning

Notes: $p_{cv} = p(C^*)$ is averaged from 50 replications as the number of predictors increases. The first $p_0 = 200$ are informative predictors that generated from a factor model, whereas the rest $p - p_0$ predictors are i.i.d. $N(0, 1)$ random variables (noise). The sample size is fixed to 100. We use $\widehat{y}_{\text{new}} = X'_{\text{new}}\widehat{\beta}$ where $X_{\text{new}}$ include both informative predictors and added noise. The left panel is a cross-sectional prediction where factors are i.i.d. sequences. The right panel is a time series prediction where factors satisfy a stationary autoregressive AR(1) model; the first fifty percent of the data are used as tuning period. The dashed vertical line "$p_{cv}$" is the average over 50 replications.

## 4.4 When $n$ is very large: mini-batch

In applications of panel data models, economists often pool the data of many individuals over several time periods, making the overall sample size $n$ very large. Meanwhile, the Pseudo-OLS requires $n \ll p$ to embrace the *benign overfitting*. When $n$ and $p$ are so large that numerical calculation of the generalized inverse goes beyond our computing capacity, one convenient solution is to *randomly drop* a fraction of the data to artificially reduce the sample size.

The efficacy of this solution is indicated by the asymptotic result. Recall that the proved rate of convergence of the out-of-sample forecast is

$$R(n) := \frac{p}{p_0 n} + \frac{n}{p}.$$

For instance, if $n/p$ converges to a positive constant asymptotically, then $R(n)$ does not diminish to zero. In contrast, if we reduce the sample size to $\widetilde{n} = p/\sqrt{p_0}$ by randomly

dropping many samples, it will yield a more desirable rate (by replacing $n$ by $\widetilde{n}$)

$$R(\widetilde{n}) = \frac{1}{\sqrt{p_0}} + \frac{\sqrt{p_0}}{p} \to 0$$

which implies consistency.

Note that *randomly dropping* part of the data does not mean to waste data. Instead, this procedure can be repeated several times and aggregated in a distributed manner. The algorithm is summarized as follows.

**Algorithm 2.** *Determine a reduced sample size $\widetilde{n}$ that is only moderately large. Then*

**Step 1** *Randomly sample $\widetilde{n}$ rows (samples) from $(X, Y)$ without replacement, obtain a reduced data set $(\widetilde{X}, \widetilde{Y})$.*

**Step 2** *Apply pseudo-OLS on $(\widetilde{X}, \widetilde{Y})$ to obtain a forecast $\widehat{y}_{\text{new}}$.*

**Step 3** *Repeat Step 2 for $B$ times to obtain $\widehat{y}^1_{\text{new}}, ..., \widehat{y}^B_{\text{new}}$, and then forecast using $\frac{1}{B}\sum_{b=1}^{B} \widehat{y}^b_{\text{new}}$.*[8]

In machine learning this method is referred to as *mini-batching*, where the full dataset is divided into several smaller batches. The model is then fitted separately on each batch and the results are aggregated. This boosts computational efficiency. Our theory also covers this method by replacing the sample size with $\widetilde{n}$. Therefore, our results explicitly interpret the success of the mini-batching procedure as a means of regularization in the linear model: by dropping portions of the data, a substantially overfitting model embraces benign overfitting in each batch.

# 5 Using only informative predictors is not optimal

To further shed light on the benefits of including idiosyncratic and artificial noise as predictors for variance reductions, we now contrast the result to 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 trade-off.

Suppose it is revealed to the economist that only the first $p_0$ of the predictors, $X_{I,t}$, load on the latent factors, and that the remaining $p - p_0$ predictors are pure noise, that

[8] In time series forecasts where the natural ordering is important, we can sequentially divide the time series into several non-overlapping batches (blocks). One can separately apply pseuod-OLS on each batch and the results are aggregated (by taking the average of pseudo-OLS coefficients then multiplied with the out-of-sample predictor).

is, the following model is practically feasible:

$$y_t = X'_{I,t}\beta_I + e_t, \quad \dim(X_{I,t}) = p_0 < n. \tag{5.1}$$

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

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

where $X_{\text{new},I}$ 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 achieves the oracle forecast. In contrast, Theorem 2 has established the optimality of using the pseudo-OLS with intentionally added noise.

**Theorem 3** (Only Informative predictors)**.** *Suppose the economist forecasts using model (5.1) by either OLS or Ridge with $\ell_2$-penalty ($\lambda$). Suppose the assumptions in Theorem 2 hold, and $p_0/n \to c \in (0,1)$.*

*(i) The predictive MSE of OLS and Ridge:*

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

$$\liminf_{n,p_0} \inf_{\lambda \geq 0} \mathbb{E}[(y_{\text{new}} - \widehat{y}^I_{\text{new,ridge}}(\lambda))^2 | X] \;>\; \text{Var}(\epsilon_{y,t}).$$

*(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 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. While Ridge regression attempts to properly choose the penalty to balance the bias and variance, Result (i) shows that there is no $\lambda \geq 0$ that makes both bias and variance simultaneously decay to zero.

In a very recent contribution to the properties of Ridge regression, He (2023) considered a dense factor augmented model in the asymptotic regime $p_0/n \to c \in (0,1)$, and showed that Ridge regression is optimal among a set of regularized estimators. Theorem 3 complements his results by showing in this regime Ridge cannot reduce the prediction risk all the way to the oracle level. If we jump out of the regime, nevertheless, by intentionally adding enough noise so that $p/n \to \infty$, the simple pseudo-OLS achieves the oracle risk.

# 6 Simulation

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

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

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 is $\tau$, the weaker are 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 $X_{\text{new}}$ to predictor their out-of-sample outcomes.

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

## 6.1 Unknown identities of informative predictors

Suppose the economist does not know which predictors are informative, so she decides to use all the collected predictors (including both informative ones and the $p - p_0$ noise). 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) Pseudo-OLS; (ii) Principal component analysis (PCA), where the number of factors and the factors are estimated using all the $p$ predictors. Based on the in-sample estimated $\widehat{\lambda}_i$, we estimate the out-of-sample factors

$\widehat{f}_{\text{new}}$ by regressing $X_{\text{new}}$ 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.

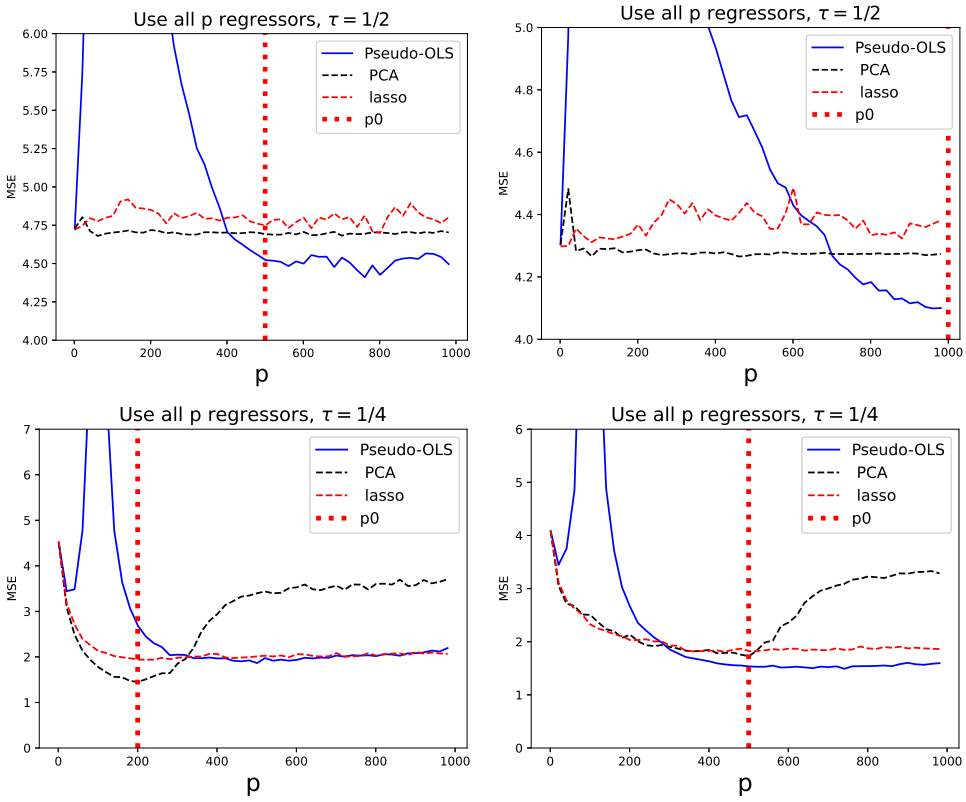


Figure 3: Simulation in Scenario I

Predictive MSE $\frac{1}{50}\sum_{j=1}^{50}(y_j - \widehat{y}_j)^2$ 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 $n = 100$.

Figure 3 plots the predictive MSE, averaged over 50 replications, as the number of predictors $p$ increases. This means all predictors are informative when $p \leq p_0$, whereas $p - p_0$ noise variables are included when $p > p_0$. The red dashed vertical line in each panel indicates the number of informative predictors. Figure 3 conveys the following numerical findings:

1. In the second panel, where all predictors are informative ($p_0 = 1000$) but very weak ($\tau = 1/2$), the pseudo-OLS continuously benefits from the inclusion of these predictors, even though they are weakly informative. The trend of decay in its MSE persists even when $p = 1000$, outperforming the other methods. In contrast,

the factors are so weak that PCA does not gain from the large $p$, resulting in a predictive MSE curve that remains essentially flat. In the remaining three panels, where $p_0$ stops increasing at some point while $p$ continues to grow, the MSE of pseudo-OLS flattens out.

2. PCA works reasonably well when $p \leq p_0$, but its performance deteriorates as more noise is introduced into the predictors. The MSE for PCA begins to increase after $p > p_0$, making it the worst among the three methods. Lasso performs well when the number of informative predictors is relatively small. But when $p_0 = 500$ or 1000, the common dependencies among predictors are stronger, leading to denser regression coefficients in the working model. Consequently, Lasso's performance deteriorates.

## 6.2 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 noise variables to implement pseudo-OLS.

We compare four methods in this case: (i) The proposed pseudo-OLS which uses all $p$ predictors (when $p > p_0$, the additional predictors are noise), and the denoised version as described in Section 2.3 (labeled as "pseudo-OLS-denoise"); (ii) PCA; (iii) the Lasso, (iv) Ridge regression. 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 noise.

Figure 4 plots the predictive MSE as the number of predictors increases. 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. We plot for $p_0 \in \{200, 500\}$ and for selected $\tau$ as these cases are representative. We observe 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 noise are added into predictions, even though new direct predictive information is no longer available after $p > p_0$. In addition, the pseudo-OLS-denoise improves the pseudo-OLS.

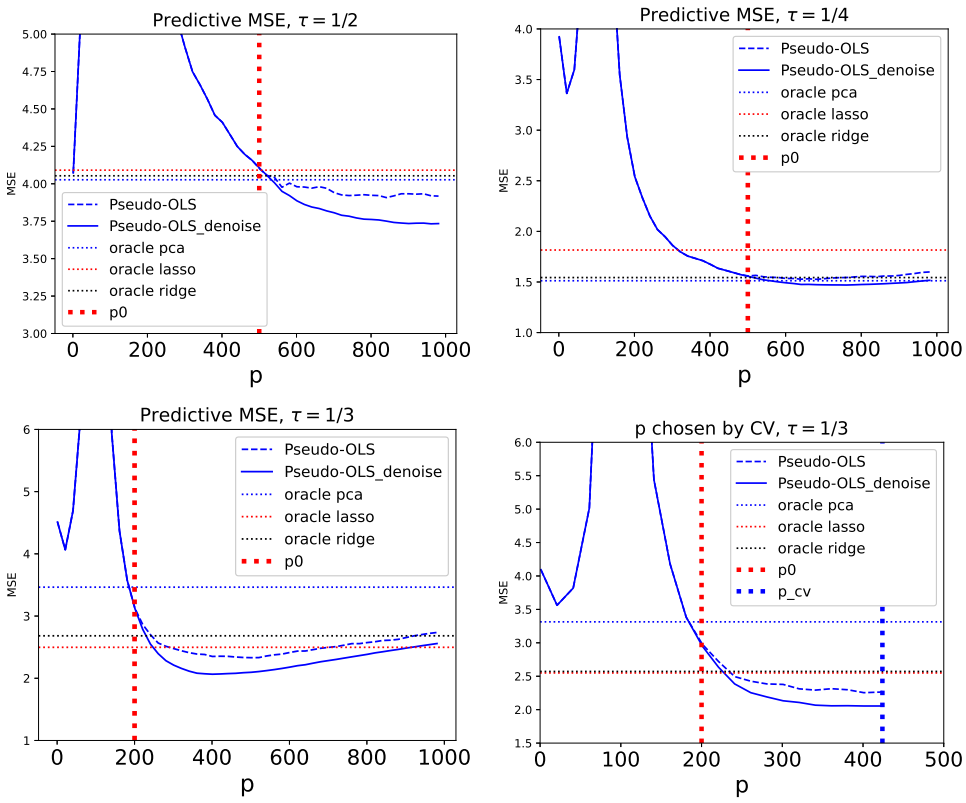


Figure 4: Simulation in Scenario II

Predictive MSE $\frac{1}{50}\sum_{j=1}^{50}(y_j - \widehat{y}_j)^2$ 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 blue dashed line $p_{cv}$ in the last panel indicates the averaged $p$ chosen by the leave-one-out cross validation.

2. When $p_0$ is moderate as in the third panel, pseudo-OLS and pseudo-OLS-denoise exhibit U-shaped predictive MSE after $p > n$, which reaches the minimum MSE around $p \in (400, 600)$ and starts to increase again. This is implied by our theory that the moderate informative signal causes the rising bias to dominate the decaying variance. Hence it is necessary to properly choose $p$ in this case. Panel 4 in Figure 4 plots the predictive MSE, under the same setting as in Panel 3 for $(\tau, p_0)$, but the maximum amount of added noise $p_{\text{cv}}$ is chosen by the leave-one-out CV.

3. In contrast, even though the oracle forecasting methods — PCA, Lasso, and Ridge — only use the informative predictors, they do not predict as well as the pseudo-OLS with many artificial noise in many cases. PCA mainly suffers from weak factor issues, whereas 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.

# 7 Empirical Applications

Our empirical illustrations include four economic forecasting applications: forecasting the U.S. inflation rate, a group of countries' GDP growth rates, the U.S. equity premium, and the S&P firms earnings.

Throughout this section, we employ the noise-denoise forecast procedure as described in Section 2.3, labeled as "pseudo-OLS-denoise" in all empirical applications.

## 7.1 U.S. inflation forecast

The Federal Reserve relies on inflation forecasts to guide monetary policy decisions, while businesses, investors, and governments depend on these predictions for planning, investment, and budgeting. After recovering from the 2020 downturn, the U.S. economy experienced its highest rate of inflation since the 1980s. As inflation has begun to slow down more recently, the Fed is considering lowering interest rates again. Accurate inflation forecasts are essential for the effective functioning of the economy.

We use the FRED-MD dataset of McCracken and Ng (2016) to forecast the change of the U.S. inflation rate, constructed based on the Consumer Price Index (CPI):

$$\Delta\text{inflation}_{t+1} = \log(\text{CPI}_{t+1}/\text{CPI}_t) - \log(\text{CPI}_t/\text{CPI}_{t-1}).$$

The data contains $p_0 = 103$ macroeconomic predictors, ranging from 1959-June to 2024-January. We use 120-month moving windows to estimate the forecast model and conduct one-month-ahead forecast.

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

We implement the pseudo-OLS as follows: generate $p - p_0$ white noise variables from the standard normal distribution as artificial predictors, and merge them with the original

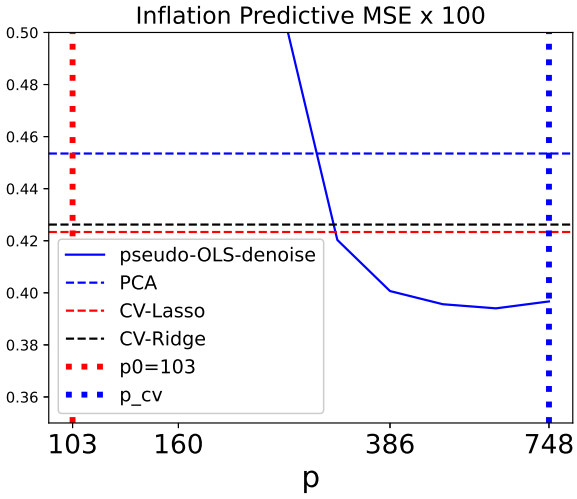


Figure 5: Inflation Predictive MSE

Inflation Predictive MSE using FRED-MD McCracken and Ng (2016). Data spans from 1959-June to 2024-January with $p_0 = 103$ predictors and $p - p_0$ added noise. We use rolling windows of $n = 120$ months for one-month horizon forecast. The vertical axis is $100 \times \frac{1}{n}\sum_n (y_{n+1} - \widehat{y}_{n+1})^2$, 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 pseudo-OLS uses additional $p - p_0$ intentionally generated $N(0, 1)$ noise, and averaged over 50 times.

$p_0$ 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 776 months is split into training and validation samples. The first 200 months are used as the training sample on which the constant $C$ is determined via cross-validation, and the remaining 576 months are used for forecasts fixing the chosen $C$.[9]

The optimal number of predictors determined in this way is approximately $p_{\max} \approx 748$. With the determined $p_{\max}$, we conduct moving window forecasts of inflation on the validation sample, and compute the predictive MSE. In addition, we implement PCA, CV-Lasso, and CV-Ridge on the validation sample, where each uses all the $p_0$ macroeconomic variables regardless of $p$. The number of "factors" for PCA is determined using the $PC_1$ criterion from Bai and Ng (2002), and the tuning parameters for CV-Lasso and CV-Ridge are determined using 5-fold cross-validation.

Figure 5 plots the predictive MSE as $p$ increases from 1 to $p_{\max}$. Meanwhile, the

[9]On the training sample, for each candidate $C$ we conduct rolling-window based one-month-ahead forecasts and compute the overall out-of-sample MSE. Then choose the optimal $C$ yielding the smallest MSE on the training sample.

pseudo-OLS stops when the total number of predictors reaches $p_{\max}$, corresponding to 103 macro variables plus 645 added noise variables, and reaches lowest predictive MSE among the comparing methods.

## 7.2 Growth forecasts

Economic growth is a fundamental issue that directly impacts human welfare and freedom. Nobel laureate Robert Lucas famously stated, "Once one starts to think about [economic growth], it is hard to think about anything else" (Lucas, 1988, p.5). While Lucas provided a theoretical framework through parsimonious neoclassical models, real-world economic growth involves a multitude of factors and was one of the earliest areas in economics that benefited from big data analysis, e.g., Sala-I-Martin (1997).

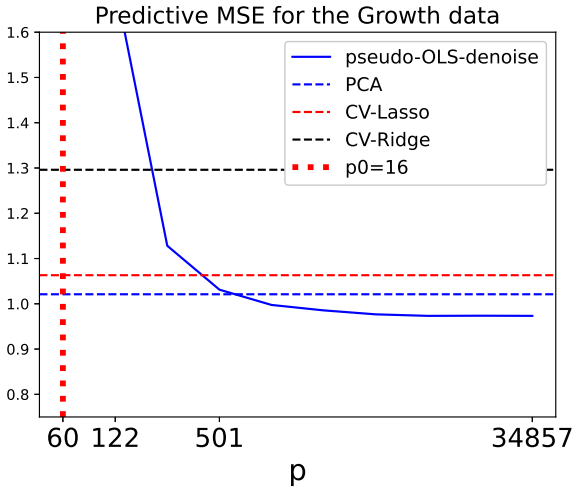


Figure 6: Growth Predictive MSE

Predictive MSE using $p_0 = 60$ socio-economic and geographical characteristics from Barro and Lee (1994), and $p - p_0$ added noise. 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 $\frac{1}{n}\sum_n (y_{n+1} - \widehat{y}_{n+1})^2$, the horizontal axis is $\log(p)$, and the horizontal tick is $p$. Regardless of $p$, the PCA, CV-Lasso and CV-Ridge use the $p_0$ socio-economic variables, whereas the pseudo-OLS uses additional $p - p_0$ intentionally generated $N(0, 1)$ noise variables. The MSE equals 20.24 with the original 60 predictors.

In this application we use the data of Barro and Lee (1994) to predict the GDP growth rates across countries. 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 noise variables and merge them with 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 6 plots the predictive MSE of different methods. The MSE equals 20.24 when we use only the original 60 predictors, which is not depicted on the plot. Pseudo-OLS outperforms other methods after complete interpolation. Furthermore, it exhibits a high degree of robustness to the number of added noise variables, stabilizing after 500 additions and maintaining an MSE around 1.0 across $p$ varying from 501 to 34,857.

### 7.3 U.S. equity premium prediction

Predictability of the U.S. equity premium is a central question in asset pricing research. Many macro finance models imply time-varying discount rates as reviewed in Cochrane (2011) and by now there is ample evidence that aggregate discount rates (expected returns) are indeed time-varying. It is however of considerable dispute to what extent this variation can be predicted. 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 its publication, academic research in forecasting time-varying future equity premia has significantly advanced (e.g., Hirshleifer et al. (2009); Atanasov et al. (2020); Chava et al. (2015); Jondeau et al. (2019); Jones and Tuzel (2013); Kelly and Pruitt (2013)). Many of them introduced new informative predictors, alongside innovative methodology 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 models, yet arriving at conclusions qualitatively consistent with their 2008 study. Notably, in the context of an annual forecast horizon, the majority of models exhibit discouraging predicting performance, with $R^2$ either negative or only marginally positive.

As an empirical illustration, we employ the 16 main variables described by Welch and Goyal (2008) to forecast the equity premium. Following the same exercise as Giannone et al. (2021), we use annual data spanning from 1948 to 2015 with $p_0 = 16$ original

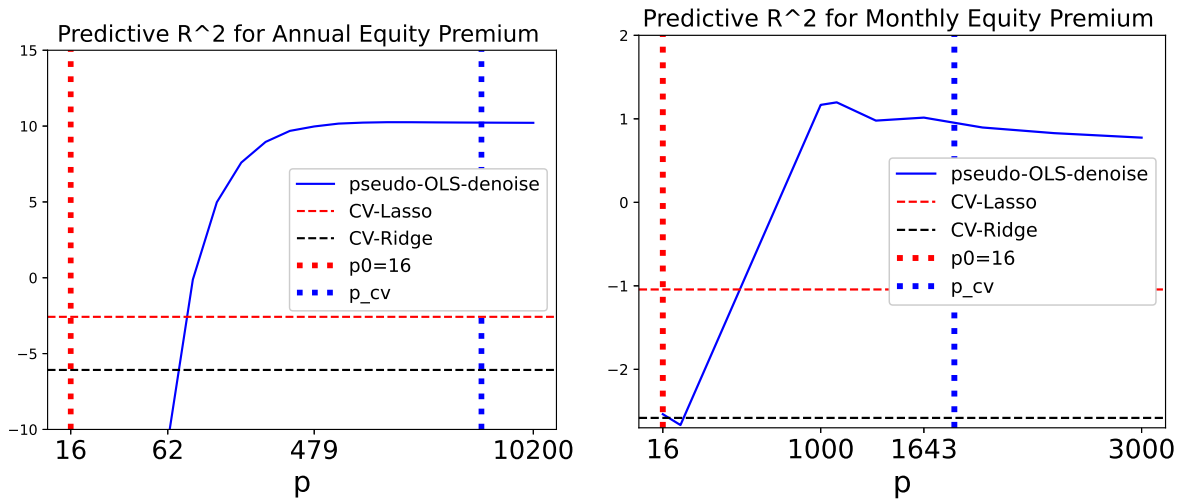


Figure 7: Out-of-sample $R^2$ for predicting the U.S. equity premium

We use the $p_0 = 16$ predictors described by Welch and Goyal (2008) and $p - p_0$ added noise. Left panel: annual prediction; Right panel: monthly prediction. $p_{cv}$ denotes the number of added noise selected by time series cross-validation. The vertical axis is out-of-sample $R^2$. For the annual data we use rolling windows of $n = 17$ year for one-year horizon forecast and the horizontal axis is plotted as $\log(p)$, and ticked using $p$; for the monthly data we use expanding windows for one-month horizon forecast, and the horizontal axis is plotted as $p$. Regardless of $p$, both CV-Lasso and CV-Ridge use the $p_0$ macro variables, whereas the pseudo-OLS uses additional $p - p_0$ intentionally generated $N(0, 1)$ noise, followed by the noise-denoise procedure as described in Section 2.3. The PCA performs too poorly to be depicted along the y-axis.

predictors, and use moving windows with sample size $n = 17$. The first prediction occurs for the 1965 observation, and is rolled over 51 times, each time expanding the training sample by one year and shifting the evaluation sample accordingly.[10]

We intentionally add additional noise and merge it with the original 16 variables to implement the denoise pseudo-OLS. We compare it with Lasso and Ridge, which use the 16 variables only and are tuned by the cross-validation. Figure 7 (left panel) 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. The number of added noise variables $p$ is chosen following the guidance of our theory by setting $p = C \times$

[10]In order to facilitate an easy comparison, we use the same sample as in the original studies, i.e. Welch and Goyal (2008). For the annual data, we use the data from the *Econometrica* website to maintain consistency with Giannone et al. (2021).

$n\sqrt{p_0}$, 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 $\sim$ 10,000 included noise for annual prediction and 1,000 $\sim$ 3,000 included noise for monthly prediction as predictors. Then we determine the data-driven $p_{cv}$ as the optimal choice of $p$ on the range, which is also plotted as the dotted blue vertical lines. Moreover, to study the robustness to the number of added noise variables, we nevertheless calculate and plot the $R^2$ up to the maximum candidate $p$ in the range (which is 10,200 for annual prediction, and 3,000 for monthly prediction), even though $p_{cv}$ suggested early stopping.

As CV-Lasso and CV-Ridge use all 16 financial and macroeconomic variables, their $R^2$'s are depicted as dashed horizontal lines, and evaluated as $-2.60$ (CV-Lasso) and $-6.12$ (CV-Ridge). We also implement PCA on the 16 variables, whose $R^2$ is too low to be depicted on the plot. In sharp contrast, the pseudo-OLS-denoise with intentionally added noise performs strikingly well: when $p$ goes beyond 16, it quickly comes back and outperforms the benchmark (sample mean prediction) and the other methods when $p$ is around 200. After 400 noise are added its out-of-sample $R^2$ reaches to 10%, and becomes very stable even after more than 10,000 noise are accumulated to annual prediction.[11]

The survey by Welch and Goyal (2008) shows that forecasting the U.S. equity premium at the monthly frequency is even harder than at the annual frequency and hardly any model in their study achieves a positive out-of-sample $R^2$. We again use the original predictors in Welch and Goyal (2008) variables, ranging from 1959-January to 2002-December (downloaded from Ivo Welch's data website), to conduct monthly forecast of the equity premium using expanding windows. Data of the first 200 months are used for tuning the number of intentionally added noise variables. We then implement the noise-denoise procedure for the remaining data to conduct one-month ahead forecast. The right panel of Figure 7 plots the predictive $R^2$ (in percentage), which peaks at $R^2 = 1.22\%$ when about 1,000 noise variables are added to the model. Overall, the predictive $R^2$ is steady and robust to the number of added noise variables, despite a slight dip at the end. When up to 3,000 noise variables are added, the out-of-sample $R^2$ is approximately 0.75%. The cross-validation method suggests stops at $p_{cv} = 2,006$ (depicted as the vertical blue line on the Figure).

[11] We use the original 16 macroeconomic and financial variables only from the Welch and Goyal (2008) dataset plus noise, maintaining a linear predictive model; the improved predictability from pseudo-OLS is mainly due to the diversification of the out-of-sample variance. In comparison, Gu et al. (2020) examined a variety of nonlinear machine learning methods with additional features. Using up to 94 firm level characteristics, they found a majority of the methods they examined, including 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.

## 7.4 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) employs the expected (predicted) earnings as direct inputs into the valuation formula.[12] Empirically, extensive research indicates 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, as discussed by Monahan (2018), the random walk model's prediction is misleading as it is inconsistent with standard assumptions about economics and accounting.[13] 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 (Chen et al., 2022).

Following Chen et al. (2022), we use high-dimensional detailed financial data as predictors. Since 2012, the U.S. public companies have been obligated to utilize a new reporting format, the so-called eXtensible Business Reporting Language (XBRL) tags, for the presentation of quantitative data in their 10-K filings submitted to the 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

[12] Our emphasis on the residual income valuation model does not imply it is 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 basic accounting principles, forecasting accrual accounting figures also facilitates the projection of free cash flows and dividends (Nissim and Penman, 2001).

[13] As argued by Monahan (2018), "changes in a firm's expected future earnings that accompany successful innovations by its managers (competitors) are not permanent. This, in turn, implies that earnings will not follow a random walk." In addition, due to delayed recognition of economic news and the nature of conservativeness of accounting, the random-walk model is not a proper economic model for earnings prediction.

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.

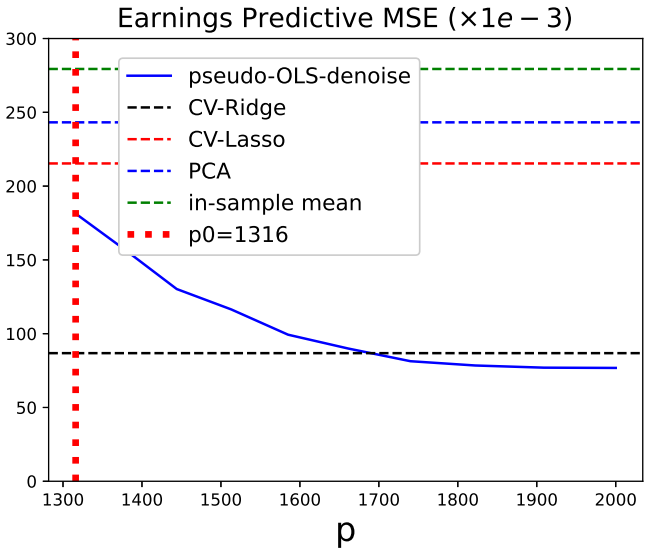


Figure 8: Predictive MSE for predicting changes in annual earnings

We use $n = 829$ training samples of firm-year observations for S&P 500 companies from 2013 to 2015, including $p_0 = 1,316$ predictors and $p - p_0$ added noise variables. The vertical axis is $\frac{1}{N}\sum_n (y_{n+1} - \widehat{y}_{n+1})^2$, the horizontal axis is showing the dimension of the features, $p$. PCA, CV-Lasso and CV-Ridge use always $p_0$ features, whereas pseudo-OLS-denoise uses additional $p - p_0$ intentionally generated $N(0, 1)$ noise variables if $p > p_0$.

Figure 8 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 leave-one-out cross-validation. The result shows that the MSE of pseudo-OLS starts to decrease as we add noise to the original $p_0$ predictors, and surpasses that of Ridge and Lasso when $p = 2,000$. In addition, it is noteworthy that the predictive MSE is large in magnitude. This observation aligns qualitatively with the recognition that predicting earnings poses a formidable challenge, primarily attributable to the limited information in the predictors and to the high persistence of earnings. Nevertheless, the pseudo-OLS produces one of the best predictions among the competing methods.

# 8 Conclusion

We provide a theoretical justification that economic forecast models are not sparse. In addition, we prove a compelling result that including noise in predictions yields greater benefits than its exclusion. Furthermore, if the total number of predictors is not sufficiently large relative to the sample size, intentionally adding noise yields superior forecast performance, outperforming benchmark predictors relying on dimension reduction. Therefore, economic forecasts can significantly benefit from benign overfitting even if a significant proportion of predictors are pure noise.

Our empirical applications illustrate this principle in four areas of economic forecasting. For inflation forecasts, adding generated noise variables can improve over the universe of variables in the FRED-MD database. In forecasting economic growth across countries, our procedure helps surpasses the original socio-economic variables in the well-studied Barro and Lee (1994) dataset. In financial economics, adding noise again improves over the financial and macroeconomic variables and in the case of monthly data helps to achieve a positive out-of-sample $R^2$, which is not available directly from the original predictors. We finally revisit the classic accounting problem of earnings prediction. Even though many features are available in this domain, moving the problem to "even higher dimensions" by including generated noise further boosts the empirical performance.

# A Proofs

## A.1 Proof of Bias-variance expressions

The objective is to prove

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

with

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

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 focus on $M$. 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 $\widehat{\beta} - R = (X'X)^+ X'e$. We decompose $M$ into

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

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", as

$$X'_{\text{new}}R = X'_{\text{new}}\mathbb{E}(\widehat{\beta}|X) = X'_{\text{new}}\mathbb{E}(\widehat{\beta}|X, X_{\text{new}}) = \mathbb{E}(\widehat{y}_{\text{new}}|X, X_{\text{new}}).$$

By the assumption $\mathbb{E}(X_{\text{new}}X'_{\text{new}}|X) = \mathbb{E}(X_tX'_t)$,

$$\begin{aligned}\text{bias}(\widehat{y}_{\text{new}})^2 &= \mathbb{E}[(X'_{\text{new}}\beta - \mathbb{E}(\widehat{y}_{\text{new}}|X, X_{\text{new}}))^2|X] \\ &= \mathbb{E}[(X'_{\text{new}}(R-\beta))^2|X] = \beta' A_X \mathbb{E}(X_{\text{new}}X'_{\text{new}}|X)A_X\beta = \beta' A_X \mathbb{E}(X_tX'_t)A_X\beta.\end{aligned}$$

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 below):

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

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

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

*(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) Under the condition $(e, X, X_{\text{new}})$ are jointly normal, as they can be written as linear combinations of $X_{\text{new}}$ and $(F, \epsilon_y, U)$, the matrices of $(f_t, \epsilon_{y,t}, u_t)$. Moreover, the assumption $\mathbb{E}(Y|X, X_{\text{new}}) = \mathbb{E}(Y|X)$ yields

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

which implies, under joint normality, that $e$ is independent of $(X, X_{\text{new}})$; Hence $\mathbb{E}[ee'|X, X_{\text{new}}] = \mathbb{E}ee' = \text{Cov}(e)$. Furthermore, because of the i.i.d. assumption (i.e., $(y_t, X_t)$ and $(y_s, X_s)$ are independent) we have $\text{Cov}(e_t, e_s) = 0$ for any $t \neq s$. In this case we have an exact expression $\mathbb{E}[ee'|X, X_{\text{new}}] = \sigma_e^2 I_n$, where $\sigma_e^2 = \text{Var}(e_t)$.

(ii) Under the second set of conditions, let $B := \epsilon_y((\rho - \Lambda'\beta)'F' + \beta'U')$, $C := U\beta$,

and $D := F(\rho - \Lambda'\beta)$. Then the decomposition

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

yields

$$\|\mathbb{E}(ee'|X, X_{\text{new}})\| \leq \|\mathbb{E}(\epsilon_y\epsilon_y'|X, X_{\text{new}})\| + 2\|\mathbb{E}(B|X, X_{\text{new}})\| + 2\mathbb{E}(\|C\|^2 + \|D\|^2|X, X_{\text{new}}).$$

Our assumption ensures $\mathbb{E}(\epsilon_y|U, F, X_{\text{new}}) = 0$ hence $\mathbb{E}(B|X, X_{\text{new}}) = 0$ and $\|\mathbb{E}(\epsilon_y\epsilon_y'|X, X_{\text{new}})\| = O_P(1)$. In addition,

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

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

## A.2 Proof of Theorem 1

*Proof.* The condition $\mathbb{E}(\epsilon_{y,t}|f_t, u_t) = 0$ implies that in the oracle model (2.1) $\mathbb{E}(\epsilon_{y,t}|X_t) = 0$ as $X_t$ follows the factor model (2.2). As a result, in the working model (3.3):

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

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

Part (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 (2.2) 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

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

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

Given the explicit expression of $\beta$, its $l_2$ norm is bounded by

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

under Assumption 1.

Part (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

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

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 $\operatorname{Var}(f_t'(\rho - \Lambda'\beta)) + \operatorname{Var}(u_t'\beta)$.

We define $\Phi := \Lambda' \operatorname{Cov}(u_t)^{-1}\Lambda/\psi_{n,p}$. Since the eigenvalues of $\operatorname{Cov}(u_t)$ is bounded away from 0 and $\infty$, 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 $\beta$ in Part (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,

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

$$\leq \quad \psi_{p,n}^{-1}\sigma_{\min}(\Phi)^{-2}\|\Phi\|\,\|\rho\|^2 = O(\psi_{p,n}^{-1}).$$

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

## A.3 Proof of Theorem 2

### A.3.1 Bias

The bias rate of convergence then follows from the following lemma.

**Lemma 5.** *Suppose* $\|\mathrm{Cov}(u_t)\| < C$. *Then*

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

*If in addition,* $\mathrm{Cov}(u_t) = \sigma_u^2 I$ *for some* $\sigma_u^2 > 0$, *then*

$$\mathrm{bias}(\widehat{y}_{\mathrm{new}})^2 \leq O_P\left(\frac{p}{n\psi_{p,n}} + (\frac{p}{n\psi_{p,n}})^{3/2}\right).$$

*Proof.* Recall $\mathrm{bias}(\widehat{y}_{\mathrm{new}})^2 = \beta' A_X \Sigma_X A_X \beta$, 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

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

**Bounding** $\|V_A'\Sigma_X V_A\|$. We focus on the first factor of (A.1). Using matrix notation, the factor model of the $n \times p$ matrix $X$ is

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

where $F$ is $n \times K$ and $\Lambda$ 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 $\infty$ given $\psi_{p,n} = O\left(p_0\right)$ and $p_0 \leq p$. We thus have

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

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 (A.1):

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

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

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

This implies

$$\|V_A'\beta\|^2 \leq O(1)\|V_A' \operatorname{Cov}(u_t)^{-1}\Lambda\|^2\|\Sigma_f^{-1}+\Lambda' \operatorname{Cov}(u_t)^{-1}\Lambda)^{-1}\|^2 = O\left(\psi_{p,n}^{-2}\|V_A' \operatorname{Cov}(u_t)^{-1}\Lambda\|^2\right).$$

If $\operatorname{Cov}(u_t) \neq \sigma_u^2 I$ for some $\sigma_u^2 > 0$, then $\|V_A' \operatorname{Cov}(u_t)^{-1}\Lambda\|^2 = O(\psi_{p,n})$. This implies

$$\operatorname{bias}(\widehat{y}_{\text{new}})^2 \leq O_P\left(\frac{1}{\psi_{p,n}}\right) O_P\left(\sqrt{\frac{\psi_{p,n}p}{n}} + \frac{p}{n}\right) = O_P\left(\sqrt{\frac{p}{n\psi_{p,n}}} + \frac{p}{n\psi_{p,n}}\right).$$

On the other hand, if $\operatorname{Cov}(u_t) = \sigma_u^2 I$ for some $\sigma_u^2 > 0$, then the bias rate can be improved. Note

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

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

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

Hence $\psi_{p,n} \gg p/n$ implies $\sigma_{j,X} \asymp \psi_{p,n}$ for all $j \leq K$. There exists an $K \times K$ matrix $H$ such that columns of $\Lambda H$ 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 the Sin-theta inequality,

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

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

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

Hence $\|V_A'\beta\|^2 \leq O(1)\|V_A'\Lambda\|^2\|\Sigma_f^{-1} + \Lambda' \operatorname{Cov}(u_t)^{-1}\Lambda)^{-1}\|^2 = O_P\left(\sqrt{\frac{\psi_{p,n}p}{n}}\right)\psi_{p,n}^{-2}$. We thus conclude,

$$\begin{aligned}\operatorname{bias}(\widehat{y}_{\text{new}})^2 &\leq O_P\left(\sqrt{\frac{\psi_{p,n}p}{n}} + \frac{p}{n}\right) O_P\left(\sqrt{\frac{\psi_{p,n}p}{n}}\right)\psi_{p,n}^{-2} \\ &= O_P\left(\left(\frac{p}{\psi_{p,n}n}\right) + \left(\frac{p}{\psi_{p,n}n}\right)^{3/2}\right) = O_P\left(\frac{p}{\psi_{p,n}n}\right). \qquad \square\end{aligned}$$

### A.3.2 Variance

The variance is

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

Since $\Sigma_X = \Lambda\Sigma_f\Lambda' + \sigma_u^2 I_p$ under Assumption 3, we have

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

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

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

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

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

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

$$\begin{aligned}
\sigma_{\max}\left(U\left(X'X\right)^+U'\right) &= \sigma_{\max}\left(\left(\frac{X'X}{p}\right)^+\right)\sigma_{\max}\left(\frac{UU'}{p}\right) \\
&\leq C_u\sigma_{\max}\left(\left(\frac{X'X}{p}\right)^+\right) = \frac{C_u}{\sigma_{\min}\left(XX'/p\right)}
\end{aligned}$$

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

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

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 with probability approaching one (w.p.a.1.) as $p, n \to \infty$. Under this event, we continue (A.5):

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

Hence $\sigma_{\max}\left(U\left(X'X\right)^{+}U'\right) \leq C_u/(0.75c_u)$. Substituting it into (A.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 (A.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 (A.3), we have

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

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

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

## A.4 Proof of Proposition 1

*Proof.* Part (i). We write the denoise estimator as $\widehat{y}_{\text{new},I} = X'_{\text{new},I}\widehat{\beta}_I = X'_{\text{new}}Q_I\widehat{\beta}$, where $Q_I := \text{diag}\left((\mathbf{1}'_{p_0}, \mathbf{0}'_{p-p_0})'\right)$. The expression of the bias becomes

$$\text{bias}\left(\widehat{y}_{\text{new},I}\right)^2 = \beta' A_X Q_l \Sigma_X Q_I A_X \beta$$

which implies the difference between $\text{bias}\left(\widehat{y}_{\text{new}}\right)^2$ and $\text{bias}\left(\widehat{y}_{\text{new},I}\right)^2$ is

$$\text{bias}\left(\widehat{y}_{\text{new}}\right)^2 - \text{bias}\left(\widehat{y}_{\text{new},I}\right)^2 = \beta' A_X \left(\Sigma_X - Q_l \Sigma_X Q_I\right) A_X \beta. \tag{A.7}$$

Under the assumptions of Theorem 2, we have

$$\Sigma_X = \begin{pmatrix} \Lambda_I \Sigma_F \Lambda'_I + \sigma_u^2 I_{p_0} & 0 \\ 0 & \sigma_u^2 I_{p-p_0} \end{pmatrix}, \quad Q_I \Sigma_X Q_I = \begin{pmatrix} \Lambda_I \Sigma_F \Lambda'_I + \sigma_u^2 I_{p_0} & 0 \\ 0 & 0 \end{pmatrix}$$

and therefore the term in the parenthesis of (A.7) is

$$\Sigma_X - Q_l \Sigma_X Q_I = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_u^2 I_{p-p_0} \end{pmatrix} = \sigma_u^2 \left(I_p - Q_I\right), \tag{A.8}$$

which is semi-positive definite. Substitute it into (A.7) and we obtain

$$\text{bias}\left(\widehat{y}_{\text{new}}\right)^2 - \text{bias}\left(\widehat{y}_{\text{new},I}\right)^2 \geq 0.$$

In terms of the variance, we denote

$$\Sigma_{Xe} := X'\mathbb{E}\left[ee'|X\right]X = \begin{pmatrix} \underset{(p_0 \times p_0)}{\Sigma_{Xe,II}} & \underset{(p_0 \times (p-p_0))}{0} \\ \underset{((p-p_0)\times p_0)}{0} & \underset{((p-p_0)\times(p-p_0))}{\Sigma_{Xe,NN}} \end{pmatrix}$$

where the off-diagonal blocks are 0 because under Scenario II $X_{N,t}$ is completely independent of $X_{I,t}$ and $\epsilon_{y,t}$. As a result, $Q_I \Sigma_{Xe} Q_I = \begin{pmatrix} \Sigma_{Xe,II} & 0 \\ 0 & 0 \end{pmatrix}$ keep the upper-left block

only. The variance

$$\begin{aligned}
\mathrm{Var}\left[\widehat{y}_{\mathrm{new},I}|X\right] &= \mathrm{tr}\left(\left(X'X\right)^{+} Q_I \Sigma_{Xe} Q_I \left(X'X\right)^{+}\right) \\
&= \mathrm{tr}\left(\left(X'X\right)^{+}\left(\Sigma_{Xe} - \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{Xe,NN} \end{pmatrix}\right)\left(X'X\right)^{+}\right) \\
&\leq \mathrm{tr}\left(\left(X'X\right)^{+} \Sigma_{Xe} \left(X'X\right)^{+}\right) \\
&= \mathrm{Var}\left[\widehat{y}_{\mathrm{new}}|X\right],
\end{aligned}$$

where the inequality follows as $\begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{Xe,NN} \end{pmatrix}$ is semi-positive definite.

Part (ii). Conditional on $X$, we have proved in Part (i) that the squared bias and variance of $\widehat{y}_{\mathrm{new},I}$ are both no larger than that counterparts of $\widehat{y}_{\mathrm{new}}$. Therefore we obtain the second inequality in the statement

$$\mathbb{E}[(X'_{\mathrm{new}}\beta - \widehat{y}_{\mathrm{new},I})^2|X_I] \leq \mathbb{E}[(X'_{\mathrm{new}}\beta - \widehat{y}_{\mathrm{new}})^2|X_I]$$

by the fact $X_I$ is a subset of $X$.

Next we show the first inequality. Let $G := (X_I, Y, X_{\mathrm{new},I})$. It is easy to see $X'_{\mathrm{new}}\beta = X'_{\mathrm{new},I}\beta_I = \mathbb{E}(X'_{\mathrm{new},I}\beta_I|G)$, and thus

$$X'_{\mathrm{new}}\beta - \widehat{y}^{*}_{\mathrm{new},I} = \mathbb{E}(X'_{\mathrm{new},I}\beta_I - y_{\mathrm{new},I}|G).$$

It follows that

$$\begin{aligned}
\mathbb{E}[(X'_{\mathrm{new}}\beta - \widehat{y}^{*}_{\mathrm{new}})^2|X_I] &= \mathbb{E}\left\{\left[\mathbb{E}(X'_{\mathrm{new},I}\beta_I - \widehat{y}_{\mathrm{new},I}|G)\right]^2 \middle| X_I\right\} \\
&\leq \mathbb{E}\left\{\mathbb{E}\left[(X'_{\mathrm{new},I}\beta_I - \widehat{y}_{\mathrm{new},I})^2|G\right] \middle| X_I\right\} \\
&= \mathbb{E}\left[(X'_{\mathrm{new},I}\beta_I - \widehat{y}_{\mathrm{new},I})^2|X_I\right]. \qquad \square
\end{aligned}$$

## A.5 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

$$\beta_I = \Sigma_{X,I}^{-1} \Lambda_I' \rho \tag{A.9}$$

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

$$\text{bias}_I(\lambda)^2 = \beta_I' A_I \Sigma_{X,I} A_I \beta_I \tag{A.10}$$

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

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

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

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

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

and then

$$A_I = V_I \left(S_{I,p_0}^2 + \lambda I_{p_0}\right)^{-1} S_{I,p_0}^2 V_I' - I_{p_0} = -V_I \text{diag}\left(\left\{\lambda/(\sigma_j + \lambda)\right\}_{j \leq p_0}\right) V_I'. \tag{A.11}$$

Substitute (A.9) and (A.11) into (A.10):

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\leq 2v'_{K+1}\Lambda'_I\frac{F'F}{n}\Lambda_I v_{K+1} + 2C_{u,p_0},$$

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

Also, there is $K \times K$ matrix $H_I$ so that columns of $\Lambda_I H_I$ 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 $X'_IX_I$. We have $v'_{K+1}V_K = 0$. The Sin-theta inequality guarantees that

$$\begin{aligned}\left\|v'_{K+1}\Lambda_I\right\| &\leq \left\|v'_{K+1}\Lambda_I H_I\right\|\|H_I^{-1}\| \leq \left\|v'_{K+1}(\Lambda_I H_I - V_K)\right\| O_P(\psi_{p,n}^{1/2}) \\ &\leq \sqrt{\psi_{p_0,n}}O_p(\sqrt{p_0/(\psi_{p_0,n}n)}) = O_P(1). \end{aligned} \tag{A.12}$$

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

$$\begin{aligned}\frac{\sigma_{p_0}}{n} &= v'_{p_0}\frac{X'_IX_I}{n}v_{p_0} \\ &= v'_{p_0}\Lambda'_I\frac{FF'}{n}\Lambda_I v_{p_0} + 2v'_{p_0}\Lambda'_I\frac{FU}{n}v_{p_0} + v'_{p_0}\frac{U'U}{n}v_{p_0} \\ &\geq \|\Lambda_I v_{p_0}\|^2(1+o_p(1)) - 2\|\Lambda_I v_{p_0}\|\left\|\frac{FU}{n}v_{p_0}\right\| + c_{u,p_0} \\ &\geq \|\Lambda_I v_{p_0}\|^2 - 2\|\Lambda_I v_{p_0}\|\left\|\frac{FU}{n}v_{p_0}\right\| + c_{u,p_0} + o_p(1)\end{aligned}$$

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

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

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 $A_I = (X_I'X_I)^{-1}X_I'X_I - I_{p_0} = 0$ leads to zero bias. 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. 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.

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.

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.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Bishop, C. M. (1995). Training with noise is equivalent to tikhonov regularization. *Neural Computation 7*(1), 108–116.

Chao, J. C. and N. R. Swanson (2022). Selecting the relevant variables for factor estimation in favar models. *Available at SSRN 4308280*.

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.

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.

Cochrane, J. (2009). *Asset pricing: Revised edition*. Princeton university press.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance 66*(4), 1047–1108.

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

De Mol, C., D. Giannone, and L. Reichlin (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics 146*(2), 318–328.

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

Dobriban, E. and S. Wager (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics 46*(1), 247–279.

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

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.

Forni, M. and L. Reichlin (1998). Let's get real: a factor analytical approach to disaggregated business cycle dynamics. *The Review of Economic Studies 65*(3), 453–473.

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

Giglio, S., D. Xiu, and D. Zhang (2023). Prediction when factors are weak. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2023-47).

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.

Hsiao, C., S. H. Ching, and S. K. Wan (2012). A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics 27*(5), 705–740.

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.

Jurado, K., S. C. Ludvigson, and S. Ng (2015). Measuring uncertainty. *American Economic Review 105*(3), 1177–1216.

Kelly, B., S. Malamud, and K. Zhou (2024). The virtue of complexity in return prediction. *The Journal of Finance 79*(1), 459–503.

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

Kolesár, M., U. K. Müller, and S. T. Roelsgaard (2023). The fragility of sparsity. *arXiv preprint arXiv:2311.02299*.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

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*.

Leeb, H. and B. M. Pötscher (2008). Sparse estimators and the oracle property, or the return of hodges' estimator. *Journal of Econometrics 142*(1), 201–211.

Lucas, R. E. (1978). Asset prices in an exchange economy. *Econometrica*, 1429–1445.

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics 22*(1), 3–42.

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*.

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, 867–887.

Monahan, S. J. (2018). Financial statement analysis and earnings forecasting. *Foundations and Trends in Accounting 12*(2), 105–215.

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.

Sala-I-Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, 178–183.

Shen, D., D. Song, P. Ding, and J. S. Sekhon (2023). Algebraic and statistical properties of the ordinary least squares interpolator. *arXiv preprint arXiv:2309.15769*.

Shi, Z. and J. Huang (2023). Forward-selected panel data approach for program evaluation. *Journal of Econometrics 234*(2), 512–535.

Sietsma, J. and R. J. Dow (1991). Creating artificial neural networks that generalize. *Neural Networks 4*(1), 67–79.

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.