Benign Overfitting in Economic Forecasting via Noise Regularization*

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

March 21, 2026

Abstract

This paper studies linear overparameterized models in economic forecasting and highlights the benefit of including predictors with no predictive power (noise variables) as a means of regularization. We consider a setting where both the outcome variable and the high-dimensional predictors are driven by a small number of latent factors, and provide a theoretical justification that the linear forecast model is dense rather than sparse. We forecast the economic outcome without estimating the factors and reach the same asymptotic accuracy as if the true latent factors were known and directly used, avoiding concerns about estimating weak factors. This is achieved by the inclusion of many noise predictors, which serves as a regularization that shrinks the eigenvalues of the design matrix and reduces the out-of-sample variance. In contrast, perfect variable selection that removes noise variables can worsen forecasts when the number of retained predictors is comparable to the sample size. Empirically, we apply this approach to forecasting U.S. inflation, international GDP growth, and the U.S. equity risk premium, finding that noise regularization improves and stabilizes predictive performance.

*The authors are grateful to Tim Christensen, Joachim Freyberger, Christian Hansen, Benjamin Holcblat, Wenxin Jiang, Stefano Giglio, Oliver Linton, Alberto Martin-Utrera, Semyon Malamud, Michael McCracken, Ulrich Müller, Andrew Patton, Peter C.B. Phillips, Michael Pollmann, Seth Pruitt, Tom Severini, Youngki Shin, Yao Zheng, and seminar participants Aalto, Duke, Tsinghua, Peking, Purdue, McMaster, Montreal, Wisconsin and conference participants at the SFS Cavalcade 2024 for valuable comments.

†Department of Economics, University of Iowa, yuan-liao@uiowa.edu

‡Tippie College of Business, University of Iowa, xinjie-ma@uiowa.edu

§Mitch Daniels School of Business, Purdue University, aneuhier@purdue.edu

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

1 Introduction

Economic outcomes arise from complex, high-dimensional processes. Output growth, inflation, and asset returns are shaped by many interacting forces that are not directly observed. Economists have nonetheless sought parsimonious representations through variable selection or dimensionality reduction. But when predictors share latent factors, predictive information is diffuse: each variable carries only partial information, and no small subset captures what the full panel provides collectively. This appears to be the empirical reality. Giannone et al. (2021) show that economic data rarely favor sparse predictive representations, and Kolesár et al. (2023) document the fragility of sparsity assumptions. This paper takes the denseness of economic predictive models seriously and examines its implications for the practice of forecasting, by addressing two questions: First, under what economic models can the denseness be justified? Second, what methods are appropriate, what conventional approaches fail, and why?

In economic forecasting, observable variables are driven by a small number of latent factors representing macroeconomic, financial, or policy conditions. Because factors are unobserved, empirical practice relies on large panels of observable predictors as proxies, and estimates linear regressions as reduced-form working models. This factor structure has a fundamental implication for the reduced-form working model: the population regression coefficients are *dense*. Predictive information is spread across many predictors, and no small subset captures the full signal. This contrasts with the sparsity assumption underlying Lasso and related methods. Yet when the number of predictors p is comparable to the sample size n, the high dimensionality imposes challenges to estimating the working model for forecasts. The standard response, namely variable selection by Lasso or dimension reduction by PCA (Stock and Watson, 2002b; Bai and Ng, 2002), rests on the belief that eliminating noise improves out-of-sample accuracy. This belief can be misleading. Reliable factor estimation requires sufficiently strong eigenvalue separation. When factors are weak, or when many predictors dilute factor strength across the full panel, determining the number of factors and recovering their span poses significant challenges (Onatski, 2012; Carrasco and Rossi, 2016).

Our paper develops a new approach that avoids estimating any factors, whether strong or weak, while achieving the same first order optimality as if the true factors were used. The central insight is that forecast accuracy depends not on which predictors are retained, but on how many. When predictors substantially outnumber observations, the design matrix self regularizes. In this regime, the Ridgeless estimator, that is, the minimum

norm least squares solution computed via the Moore–Penrose pseudoinverse, benefits from implicit regularization. Many predictors collectively dilute each variable's influence on the fitted values, driving out-of-sample variance toward zero, a manifestation known as *benign overfitting* (Belkin et al., 2019; Mei and Montanari, 2019; Hastie et al., 2022). Variable selection however, disrupts this mechanism. Even perfect selection keeps the predictor-to-observation ratio near one, where self-regularization fails. The failure is not about omitted signal, but a consequence of insufficient predictor dimension. The empirical relevance is immediate: the FRED-MD dataset offers roughly 130 macroeconomic predictors for rolling estimation windows of approximately 120 monthly observations (McCracken and Ng, 2016), so even aggressive screening leaves the ratio near one, precisely the regime where variance fails to decay.

The appropriate response is to expand the feature space rather than contract it. We propose a simple forecasting approach, which we call "noise-regularization": applying Ridgeless regression while retaining uninformative predictors; when economic theory does not provide a sufficiently rich set of informative predictors, we artificially augment the predictor set with randomly generated noise variables independent of the data. The key insight is that benign overfitting operates through regularization of the eigenvalue spectrum rather than predictor informativeness. Even though noise variables contain no signal, they inflate the eigenvalues of the design matrix and reduce out-of-sample variance. Unlike Ridge, which imposes uniform deterministic shrinkage, noise-regularization induces random shrinkage primarily in directions without signal. Because the shrinkage function is concave, Jensen's inequality implies that this random shrinkage yields a smaller shrinkage bias in finite samples.

We formalize these insights in four main results. Let p_0 denote the number of informative predictors that load on the latent factors. We consider a setting in which a large number of economic variables, including the forecast target, are driven by a few latent factors. Our first result establishes a *dense representation* of the forecasting working model (Theorem 1): the population regression coefficients are diffuse, with their squared norm decaying as p_0 and the factor strength increase. As dimension grows, the working model forecast converges to the oracle prediction: the same asymptotic mean squared error (MSE) obtained using the true factors. The second result quantifies the *cost of perfect variable selection* (Proposition 1). Even if an oracle correctly identifies the p_0 informative predictors, applying Ridgeless regression to them alone fails to achieve first-order optimality when p_0/n converges to a positive constant. Post-Lasso, OLS on the correctly specified model, or interpolation using only informative predictors all remain

suboptimal. Meanwhile, we acknowledge that our focus is forecast accuracy; for structural interpretation or causal inference, variable selection may still be appropriate. The third result establishes *benign overfitting* (Theorems 2 and 3). The noise-regularization achieves the oracle MSE asymptotically, without estimating the factors or their number, accommodating weak factors and temporal dependence. When noise predictors are added deliberately and forecasts are averaged over repeated noise draws via Rao–Blackwell, the variance rate improves strictly. Because factor estimation is bypassed, the required factor strength conditions are mild, and the results apply to both cross-sectional and time-series settings.

Overparameterization does not universally produce benign overfitting. In particular, for the autoregressive model AR(p), adding more lagged predictors does not deliver similar gains. In the fourth result, we establish an AR(p) *symmetry property* (Proposition 2): for a time series of length T, the Ridgeless forecast from AR(p) numerically coincides with that from AR($T-p$) for every lag order p. Increasing the lag order to the overfitting regime is therefore equivalent to fitting a shorter autoregression. The intuition is that each additional lag reduces the effective sample by one observation. Lag expansion thus fails to induce eigenvalue inflation. This result clarifies that the gains from noise regularization require expanding the predictor set with genuinely exogenous variables.

Together, these four results identify a unified mechanism: benign overfitting in dense, factor-driven forecasting environments is achieved by expanding the feature space, not by selecting within it.

Our analysis provides a structural economic foundation for benign overfitting and extends the double descent literature (Belkin et al., 2019; Bartlett et al., 2020; Hastie et al., 2022; Lee and Lee, 2023) to factor-driven environments. We show that the dense coefficient vector in the reduced-form working model arises naturally from the oracle factor structure, rather than being imposed as a technical condition. We further demonstrate that the second descent in forecast MSE reflects eigenvalue shrinkage, which can be induced even by adding noise variables, a distinction particularly relevant in economic studies when p_0/n is not large. Finally, whereas existing analyses typically assume bounded eigenvalues, factor models feature a few rapidly diverging spikes, requiring new technical arguments.[1]

Empirically, we apply our approach to three settings: monthly U.S. inflation forecast-

[1]In a contemporaneous study, Meng et al. (2025) analyze double descent in portfolio allocation with spiked eigenvalues. Among economic applications of overparameterization, Kelly et al. (2024) and Didisheim et al. (2023) document the benefits of rich nonlinear models.

ing using FRED-MD (McCracken and Ng, 2016), cross-country GDP growth forecasting using Barro and Lee (1994), and the U.S. equity risk premium forecasting (Welch and Goyal, 2008). In each application, we apply the proposed noise-regularization forecast by augmenting the predictor set with artificially generated noise variables. Across all settings and forecast horizons, predictive accuracy is stable or improves as noise predictors are added, consistent with eigenvalue inflation reducing variance without meaningful additional bias. For U.S. inflation, noise regularization achieves the lowest predictive MSE among all methods considered. For cross-country GDP growth, the method stabilizes at an MSE near 1.0, compared with 20.24 using the original 60 predictors alone, and remains robust over a wide range of noise counts. For the equity premium, where competing methods yield negative out-of-sample R^2, noise regularization achieves a positive R^2 of approximately 1.2%. Noise regularization outperforms Ridge, Lasso, and PCA across all three applications, and exhibits substantially lower sensitivity to the tuning parameter.

We adopt the following notation. Let $\|A\|$ denote the ℓ_2 norm if A is a vector, or the operator norm if A is a matrix. For two positive 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 $a_{p,n} = O(b_{p,n})$ and $b_{p,n} = O(a_{p,n})$.

2 Model and Dense Representation

2.1 Factor-driven forecasting

The objective is to forecast a scalar outcome y_t, driven by K latent factors f_t in the true data generating process

$$y_t = \rho' f_t + \epsilon_{y,t}, \quad \mathbb{E}(\epsilon_{y,t}|f_t) = 0, \quad \text{(true DGP)} \tag{2.1}$$

where ρ is a $K \times 1$ vector of factor loadings on the outcome. The model admits an intercept by setting the first component of f_t to unity. In addition to the target, the economist observes a p-dimensional vector of predictors $x_t = (x_{1,t}, \ldots, x_{p,t})'$ that share the latent factors through

$$x_t = \Lambda f_t + u_t, \quad \mathbb{E}(u_t|f_t, \epsilon_{y,t}) = 0, \tag{2.2}$$

where $\Lambda = (\lambda_1, \ldots, \lambda_p)'$ is a $p \times K$ matrix of factor loadings and $u_t = (u_{1t}, \ldots, u_{pt})'$ is an idiosyncratic error. The mean independence condition $\mathbb{E}(u_t|f_t, \epsilon_{y,t}) = 0$ ensures that all

predictive power of x_t for y_t operates through the common factors; Remark 1 discusses why this condition is central to the dense forecasting setting.

We allow both time series and cross-sectional forecasts. In the time series context, the primitive processes $(f_t, \epsilon_{y,t}, u_t)$ are stationary over $t = 1, \ldots, n$, and f_t may be dynamic, containing finitely many lags of a baseline factor vector as in Stock and Watson (2002a); for cross-sectional applications, t indexes individuals rather than time periods.

Not all predictors need be informative. We partition x_t into components that load on the factors and components that do not:

$$x_t = \begin{pmatrix} x_{I,t} \\ x_{N,t} \end{pmatrix} = \begin{pmatrix} \Lambda_I \\ 0 \end{pmatrix} f_t + u_t, \quad \begin{cases} \text{informative predictors:} & \dim(x_{I,t}) = p_0 \\ \text{noise:} & \dim(x_{N,t}) = p - p_0, \end{cases} \tag{2.3}$$

where Λ_I is a $p_0 \times K$ matrix with all rows nonzero. The ordering is without loss of generality. When $p - p_0$ is large relative to p_0, the factors are diluted across the full predictor set and become *weak* in the sense of the factor-strength literature, even if they are strong within the informative block alone. Our methodology treats the identities of $x_{I,t}$ and $x_{N,t}$ as potentially unknown and does not require variable selection to screen off the noise.

Two empirically relevant scenarios motivate the analysis in the presence of many noise predictors in (2.3).

Scenario I: Unknown noise identities. Economists routinely assemble large predictor sets such as macroeconomic indicators, financial variables, survey data that are *potentially* correlated with the target, but economic theory may not exactly identify which variables carry the factor signal. Many collected predictors may be pure noise, diluting factor strength across the full set. The partition in (2.3) is unknown to the forecaster.

Scenario II: Intentionally added noise. Suppose instead that the economist has *a priori* knowledge that all p_0 collected predictors are informative, but p_0 is not much larger than n (formally, $p_0/n \to c \in (0, \infty)$). Proposition 1 below shows that this regime is insufficient for first-order forecast optimality, regardless of how accurately the informative variables are identified. Traditional wisdom would retain only these predictors. A central finding of this paper is that deliberately augmenting the predictor set with pure noise variables, i.e. making p much larger than n improves forecast accuracy through eigenvalue regularization.

Remark 1. The mean independence condition in (2.2) rules out direct predictability

through the idiosyncratic errors, so that all forecasting power operates through the common factors as in Forni and Reichlin (1998) and De Mol et al. (2008). If instead $\epsilon_{y,t}$ is correlated with u_t, predictability then arises from both factors and idiosyncratic components, and the regression coefficient may be sparse: in the extreme case $y_t = x_{1,t}$, the coefficient vector is $\beta = (1, 0, \ldots, 0)'$. We exclude this channel so that predictive information is spread across many variables.

2.2 Dense reduced-form representation

Given the data $(y_t, x_t)_{t=1}^n$, the linear working model

$$y_t = x_t'\beta + e_t, \quad \mathbb{E}(x_t e_t) = 0, \quad \text{(working model)} \tag{2.4}$$

with $\beta = \mathbb{E}(x_t x_t')^{-1}\mathbb{E}(x_t y_t)$, is the standard reduced-form approach to forecasting with many predictors. Economic theory implies that many outcomes depend on low-dimensional state variables (consumption growth, productivity, or risk factors) that are not directly observed (Merton, 1973; Lucas, 1978). While no single predictor perfectly proxies the latent factors, their linear combination provides a close approximation. This suggests that the coefficient vector β should not be sparse but dense: many small, nonzero entries rather than few large ones.

Theorem 1 formalizes this intuition. To state it, we introduce notation and the asymptotic regime. Let $\sigma_j(\cdot)$ denote the jth largest singular value of a matrix, and let $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the smallest and largest nonzero singular values. The factor strength is indexed by a sequence $\psi_{p,n} \to \infty$. Our asymptotic framework requires $p, p_0, \psi_{p,n} \to \infty$ as $n \to \infty$, with $p_0 \asymp n$ and $n = O(p)$.

Assumption 1. *(i) There is a sequence $\psi_{p,n} \to \infty$ such that $\sigma_1(\Lambda_I'\Lambda_I) \asymp \sigma_K(\Lambda_I'\Lambda_I) \asymp \psi_{p,n} = O(p_0)$.*

(ii) There exist finite constants $c_u, C_u \in (0, \infty)$ such that $c_u \le \sigma_{\min}(\Sigma_u) \le \sigma_{\max}(\Sigma_u) \le C_u$, where $\Sigma_u = \mathbb{E}(u_t u_t')$.

(iii) $\|\rho\|$ is bounded.

Condition (i) requires that all K eigenvalues of $\Lambda_I'\Lambda_I$ grow at a common rate $\psi_{p,n}$, which is imposed for notational convenience. Crucially, we allow $\psi_{p,n}/p_0 \to 0$: the factors may be weak even within the informative block. The divergence of $\psi_{p,n}$ is necessary to

separate the factor signal from the idiosyncratic component. Conditions (ii) and (iii) are standard.

We now characterize the coefficient vector β and the residual variance of the working model. Let $\sigma_\epsilon^2 = \mathbb{E}(\epsilon_{y,t}^2)$ denote the oracle forecast variance under the true factors, and $\sigma_e^2 = \mathbb{E}(e_t^2)$ the residual variance of (2.4).

Theorem 1. *Suppose (x_t, y_t) are generated by models (2.1) and (2.2), and let Assumption 1 hold. Then the coefficient of the working model (2.4) satisfies:*

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

(ii) $\sigma_e^2 - \sigma_\epsilon^2 = O(\psi_{p,n}^{-1}) = o(1)$.

Result (i) establishes that the working model is dense: as factor strength grows, the ℓ_2-norm of β decays to zero. This contrasts with sparse models, in which at least one coefficient remains bounded away from zero regardless of dimension. The conventional sparse view assumes that only a small subset of predictors carries predictive power. However, this perspective has been increasingly questioned by empirical evidence, including Giannone et al. (2021) and many related studies. Perhaps a more convincing interpretation is that predictors are not collected arbitrarily, but are selected under the guidance of economic theory and empirical experience. As a result, many predictors are informative because they share exposure to latent economic factors. When such predictors are used directly in a reduced-form regression, each variable carries a small amount of signal through these common factors. As the cross-sectional dimension grows, this signal is distributed across many regressors, leading to a dense representation in which $\|\beta\| \to 0$. Result (i) formalizes this intuition and is closely aligned with the empirical findings in Giannone et al. (2021).

Result (ii) quantifies the gap between the working model (2.4) and the oracle one (2.1): the residual variances σ_e^2 and σ_ϵ^2 coincide as $\psi_{p,n} \to \infty$. In principle, the economist can achieve the same forecast accuracy using the observed predictors x_t directly, without estimating the latent factors or specifying their number.

2.3 Forecast under variable selection

Given the dense nature of the working model established in Theorem 1, a natural question is whether dimension reduction or noise removal remains desirable. Consider a hypothetical oracle who knows that only the first p_0 predictors load on the latent factors, while

the remaining $p - p_0$ are pure noise. This oracle retains only the informative predictors and estimates the following "ideal" model:

$$y_t = x'_{I,t}\beta_I + e_t, \qquad \dim(x_{I,t}) = p_0. \tag{2.5}$$

With the known identity of informative predictors, retaining only $x_{I,t}$ and discarding the noise appears to be the natural course of action. The objective is to forecast out-of-sample data y_{new} satisfying:

$$y_{\text{new}} = \rho' f_{\text{new}} + \epsilon_{y,\text{new}}, \qquad x_{I,\text{new}} = \Lambda_I f_{\text{new}} + u_{\text{new}}.$$

We focus on a simple tuning-free dense forecasting method, the *Ridgeless regression*

$$\widetilde{\beta}_I = \left(\sum_{t=1}^{n} x_{I,t}x'_{I,t}\right)^{+} \sum_{t=1}^{n} x_{I,t}y_t, \tag{2.6}$$

to estimate the dense coefficient in (2.5), where $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse. This estimator coincides with OLS when $p_0 < n$ and interpolates the data when $p_0 \geq n$ (meaning $y_t = x'_{I,t}\widetilde{\beta}_I$ for all in-sample data). It forecasts using:

$$\widetilde{y}^I_{\text{new}} = x'_{I,\text{new}}\widetilde{\beta}_I.$$

Throughout we assume $\sigma^2_\epsilon = \mathbb{E}(y_t - \rho' f_t)^2 = \mathbb{E}(y_{\text{new}} - \rho' f_{\text{new}})^2$, so that the in-sample and out-of-sample oracle MSE coincide.

Proposition 1 (Informative predictors only)**.** *If Assumption 1 holds and* $u_{\text{new}}, \epsilon_{y,\text{new}}, X_I = (x'_{I,1}, \ldots, x'_{I,n})'$*, and* $\epsilon_y = (\epsilon_{y,t})_{t=1}^n$ *are mutually independent, then*

$$\liminf_{n,p_0\to\infty} \mathbb{E}[(y_{\text{new}} - \widetilde{y}^I_{\text{new}})^2 | X_I] \quad > \quad \sigma^2_\epsilon$$

as $p_0/n \to \gamma \in (0, \infty)$*.*

The imposed assumptions in Proposition 1 requires regression error terms $(\epsilon_{y,\text{new}}, u_{\text{new}})$ be independent of in-sample data, but we allow correlations between f_{new} and in-sample f_t, so that dynamic factors in time series settings are allowed. The proposition establishes that even perfect variable selection fails to achieve first-order forecast optimality. The failure is not due to omitted signal: all retained predictors are informative. Rather, when p_0 is proportional to n, the dimension is not sufficiently large to regularize the

eigenvalues of the design matrix $X_I'X_I$, which is crutial for reducing the out-of-sample variance. Hence Ridgeless cannot benefit from benign overfitting.

This result should be interpreted broadly as an extension of the classic post–variable selection phenomenon to the overparameterized regime. Suppose a dimension reduction method, such as Post-Lasso (Belloni and Chernozhukov, 2013), identifies the informative predictors without error, and OLS is then applied to the selected set. When $p_0 < n$, the oracle Ridgeless estimator $\widetilde{\beta}_I$ coincides with OLS, making it equivalent to Post-Lasso under perfect selection. Proposition 1 implies that even in this idealized setting, the resulting forecast fails to achieve first-order optimality when $p_0/n \to \gamma \in (0, \infty)$.

Symmetry of AR(p). If eigenvalue inflation cannot be achieved by retaining only informative predictors, one might instead seek overparameterization by expanding model complexity in a different direction. In time series forecasting, a natural candidate is to increase the lag order in an autoregressive model. Consider a time series of length T, where the objective is to forecast $y_{\text{new}} = y_{T+1}$. Let $y_{a:b} = (y_t)_{t=a}^{b}$ denote the column vector of y_t for periods a to b. Consider the AR(p) model with no intercept, estimated by Ridgeless regression, which is equivalent to running AR(p) with the demeaned data. The Ridgeless AR(p) coefficient and one-step-ahead forecast are

$$\widehat{\beta}_p^{AR} = \left(\sum_{t=p+1}^{T} y_{t-p:t-1} y_{t-p:t-1}' \right)^{+} \sum_{t=p+1}^{T} y_{t-p:t-1} y_t, \tag{2.7}$$

$$\widehat{y}_{T+1}^{AR}(p) = y_{T-p+1:T}' \widehat{\beta}_p^{AR}. \tag{2.8}$$

For the AR(p), the effective sample size is $n = T - p$. So (2.7) coincides with OLS when $p < T/2$ and interpolates when $p \geq T/2$.

If benign overfitting were driven simply by increasing the number of regressors, one would expect gains from choosing $p > T/2$. The following proposition, however, shows that this intuition is incorrect by establishing what we call the "symmetry" property of AR(p).

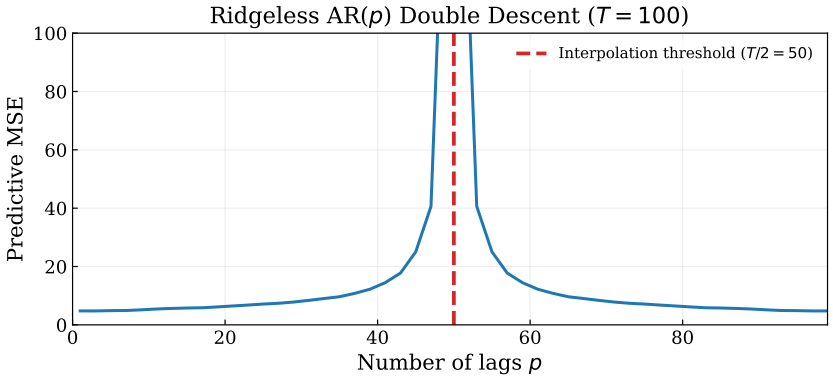
Proposition 2 (Symmetry of AR(p))**.** *Let $\widehat{y}_{T+1}^{AR}(p)$ be the Ridgeless AR(p) forecast defined in (2.7)–(2.8). Then for any $p = 1, \ldots, T-1$,*

$$\widehat{y}_{T+1}^{AR}(p) = \widehat{y}_{T+1}^{AR}(T-p).$$

Proposition 2 reveals a novel symmetry property: an overparameterized AR(p) model

with $p > T/2$ is numerically identical to an underparameterized AR$(T-p)$ model. As p increases, the effective sample size $n = T - p$ decreases proportionally, so the eigenvalues of the design matrix do not grow in a manner that induces spectral regularization. Lag expansion therefore does not generate the eigenvalue inflation necessary for benign overfitting. Figure 3 illustrates this symmetry.

Figure 1: The symmetry property of AR(p) forecast using Ridgeless.



Notes: The blue curve plots the predictive MSE. The time series length is $T = 100$. The interpolation threshold is $T/2 = 50$. When $p < T/2$, the predictor is equivalent to the conventional OLS.

Taken together, Propositions 1 and 2 convey a unified message: benign overfitting in dense forecasting environments requires eigenvalue growth in the design matrix, and neither oracle variable selection nor lag expansion delivers it. Under oracle selection, p_0 proportional to n leaves eigenvalues insufficient; under AR(p), the effective sample size shrinks as the lag order grows. The central lesson is that the spectrum, not informativeness alone, governs forecast accuracy in the overparameterized regime. This motivates the approach developed in Section 3: rather than estimating factors or selecting variables, we deliberately expand the predictor dimension to induce the spectral regularization that neither oracle selection nor lag expansion can provide.

Why not Lasso or PCA? Theorem 1 establishes that β is dense, which immediately limits the scope of the Lasso (Chernozhukov et al., 2017; Giannone et al., 2021). In addition, Proposition 1 covers the Post-Lasso under perfect variable selection as a special case, which also yields suboptimal forecast. PCA offers an alternative by estimating the factors directly, but the quality of the estimated factors depends critically on factor strength. When many noise variables are present, the factors appear weak across the

full predictor set, and correctly estimating the number of factors becomes difficult in practice; recent work by Giglio et al. (2023) and Chao and Swanson (2022) addresses this via feature selection to enhance factor strength. Furthermore, weak factors need not arise solely from noise contamination; factors can also be weak within informative predictors. If the eigenvalues of $\Lambda_I'\Lambda_I/p_0$ decay to zero as $p_0 \to \infty$ (so that $\psi_{p,n}$ grows slowly relative to p_0), the factor signal is intrinsically weak and PCA continues to suffer from poor factor recovery regardless of how successfully noise variables are screened.

3 Noise Regularization

Under the conditions of Theorem 1, the working model is dense, meaning that a non-negligible fraction of predictors contributes to predictive performance. In practice, the predictor set comprises informative variables X_I $(n \times p_0)$, those loading on the latent factors, and noise variables X_N $(n \times (p - p_0))$, independent of the outcome and the factors. Regardless of whether the economist knows the identities of informative predictors, Proposition 1 shows that working with only informative predictors is insufficient for first order forecast optimality when the number of informative predictors p_0 satisfies $p_0/n \to c < \infty$. This is empirically relevant because economic theory may provide only a limited number of informative predictors.

This motivates our proposed forecast approach: rather than eliminating noise variables, one should retain all available predictors and, when necessary, even intentionally add pure noise variables to expand the predictor set. Importantly, our approach does not distinguish between the two scenarios described in Section 2. Formally, let $X = [X_I, X_N]$ be the $n \times p$ matrix pooling all predictors. In the first scenario, X_N consists of the noise variables already present in the dataset and in the second, X_N consists of artificially added noise columns. In both cases we apply Ridgeless regression on predictors with noise:

$$\widehat{\beta} = (X'X)^{+}X'y, \tag{3.1}$$

where $y = (y_1, \ldots, y_n)'$. We call (3.1) as "noise-regularized estimator". A central contribution of this paper is to show that high-dimensional noise variables serve as an implicit regularization device: they inflate the non-factor eigenvalues of $X'X$ (those not corresponding to the factors), thereby reducing out-of-sample variance. The next subsection explains this regularization mechanism in detail.

3.1 Noise-Regularization as a Shrinkage Mechanism

Given that the added noise contains no economic information, how does it facilitate prediction? It transpires that adding noise can be viewed as a regularization method, which inflates the eigenvalues of the design matrix. To see this, define the following $p_0 \times 1$ vector-valued shrinkage function:

$$\mathcal{A}(M) = X_I'(X_I X_I' + M)^{-1} Y,$$

where M is an $n \times n$ symmetric semi-positive definite matrix, provided that the inverse exists. When $p > n$, the definition of pseudo-inverse gives $(X'X)^+ X = X'(XX')^{-1}$. Therefore the subvector of the noise-regularized estimator (3.1) corresponding to the informative predictors can be written as

$$\widehat{\beta}_I = \iota_I (X'X)^+ X'Y = X_I'(XX')^{-1} Y = \mathcal{A}(X_N X_N')$$

where the first p_0 predictors are assumed informative without loss of generality; $\iota_I = (I_{p_0}, 0)$ is a $p_0 \times p$ selection matrix, and the last equality decomposes the $n \times n$ matrix $XX' = X_I X_I' + X_N X_N'$, which is invertible when $p > n$. As a comparison, perfect variable selection sets $M = 0$, and Ridge regression with tuning parameter λ sets $M = \lambda I_n$. The choice of M governs how aggressively the eigenvalues of $X_I X_I'$ are inflated; a larger M makes $(X_I X_I' + M)^{-1}$ smaller in the semidefinite order, directly controlling out-of-sample variance.

To see the effect of shrinkage using a noise matrix, take the singular value decomposition (SVD) of $X_I = U_I D_I V_I'$, and define

$$\mathcal{S}(M, \Sigma) = \operatorname{tr}\left[V_I D_I' (D_I D_I' + U_I' M U_I)^{-2} D_I V_I' \Sigma\right].$$

Consider the case when $x_{\text{new},I}$ is independent of the training data, and it follows that the out-of-sample variance is

$$\operatorname{Var}(x_{\text{new},I}'(\widehat{\beta}_I - \beta_I)|X) = \mathcal{S}(X_N X_N', \Sigma_{X_I}).$$

When $M = X_N X_N'$, it inflates the eigenvalues from $D_I D_I'$ along a non-factor direction to $D_I D_I' + U_I' M U_I$. Meanwhile, it is well-known from the random matrix theory that the nonzero eigenvalues of the noise matrix $X_N X_N'$ grow proportionally to p if $p/n \to \infty$. As a consequence, the order of the non-spiked (the largest $(K+1)$-th to the n-th) eigenvalues

change from p_0 to the order of p. Also for factor models, we have $\Sigma_{X_I} = \Lambda_I \Sigma_f \Lambda_I' + \Sigma_{u_I}$, which yields a further decomposition of the variance into the factor-related component and the idiosyncratic component, and for now let us focus on the latter. This yields

$$\mathcal{S}(X_N X_N', \Sigma_{u_I}) \leq C_u \operatorname{tr}\left[(D_I D_I' + U_I' W U_I)^{-2} D_I D_I'\right] = O_P\left(\frac{n + \psi_{p,n}}{p^2}\right),$$

which has a fast converging out-of-sample forecast variance.

In comparison, if we do not shrink the eigenvalues but only use X_I after perfect variable selection, the Ridgeless estimator (2.6) would take the form:

$$\widetilde{\beta}_I = (X_I' X_I)^+ X_I' Y = \mathcal{A}(0 \cdot I_n),$$

whose forecast variance is

$$\operatorname{Var}(x_{\text{new},I}'(\widetilde{\beta}_I - \beta_I)|X) = \mathcal{S}(0 \cdot I_n, \Sigma_X).$$

Here $M = 0 \cdot I_n$ does not shrink the eigenvalues of $D_I D_I'$ and thus is of order p_0. Then there exists a $0 < c_u \leq \sigma_{\min}(\Sigma_{u_I})$ such that

$$\mathcal{S}(0 \cdot I_n, \Sigma_{u_I}) \geq c_u \operatorname{tr}((D_I D_I')^{-1}) \asymp \frac{n}{p_0}.$$

The variance does not converge when $n/p_0 \nrightarrow 0$.

Therefore, the inclusion of many random noise in essence inflates the eigenvalues to a much larger magnitude to reduce the out-of-sample variance. In this sense, it is similar to the Ridge regression with a large choice of tuning parameter, which also shrinks the eigenvalues. It is therefore important to see how the noise-regularized estimator compares with the Ridge estimator.

The Ridge estimator, when applied to X_I with tuning parameter λ, takes the form

$$\widetilde{\beta}_{\text{ridge}}(\lambda) = (X_I' X_I + \lambda \cdot I_n)^{-1} X_I' Y = \mathcal{A}(\lambda \cdot I_n).$$

In addition, since $\mathbb{E} X_N X_N' = (p - p_0) I_n$, a natural benchmark of the Ridge tuning parameter is $\lambda = (p - p_0)$. Hence the noise-regularized estimator uses random shrinkage $X_N X_N'$ whereas Ridge uses the deterministic shrinkage $\lambda \cdot I_n = \mathbb{E} X_N X_N'$. Both estimators introduce shrinkage bias. Here we heuristically compare this bias. Conditioning on the informative predictors, the estimation bias of any estimator taking the form $\widehat{\beta} = \mathcal{A}(M)$

is given by

$$\mathbb{E}(\widehat{\beta}|X_I) - \beta = \mathbb{E}[\mathcal{A}(M)|X_I] - \beta = \mathbb{E}[f(M)|X_I]\beta,$$

where $f(M) = X_I'(X_I X_I' + M)^{-1} X_I - I_{p_0}$. Note that $f(M)$ is a convex matrix-valued function, and is a negative semidefinite matrix whenever M is positive semidefinite. Applying the matrix-valued Jensen's inequality (e.g., Tropp (2015)),

$$\mathbb{E}[f(X_N X_N')|X_I] \succcurlyeq f(\mathbb{E}(X_N X_N'|X_I)) = f(\lambda \cdot I_n),$$

where we write $B \succcurlyeq A$ (equivalently $A \preccurlyeq B$) if $A - B$ is negative semidefinite. Hence, the shrinkage effects to β have the following relationship:

$$\underbrace{\mathbb{E}[f(\lambda \cdot I)|X_I]}_{\text{Ridge bias}} \preccurlyeq \underbrace{\mathbb{E}[f(X_N X_N')|X_I]}_{\text{Noise-regularized bias}} \preccurlyeq 0 \cdot I_n.$$

Intuitively speaking, this means the shrinkage bias introduced by the random matrix $X_N X_N'$ is less than that of the deterministic matrix $\mathbb{E}X_N X_N' = \lambda \cdot I_n$ with $\lambda = p - p_0$, yielding a smaller finite-sample shrinkage bias than that of the corresponding Ridge.

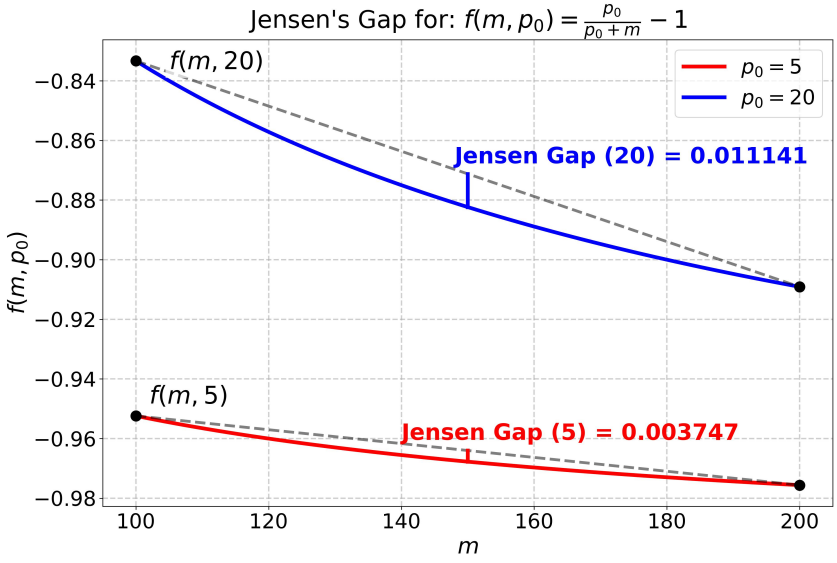


Figure 2: Jensen's Gap

Notes: This plots the function $f(w, p_0)$ for two values of $p_0 \in \{5, 10\}$. It illustrates that the Jensen gap is larger as p_0 increases.

Interestingly, the bias improvement over Ridge becomes more substantial as p_0 in-

creases, this can be seen by examining the Jensen's gap, i.e., the difference between the upper and lower side of the Jensen's inequality. Consider the scalar version of $f(M)$, say $f(m, p_0) = \frac{p_0}{p_0+m} - 1$. The Jensen gap is defined as

$$\mathbb{E}f(m, p_0) - f(\mathbb{E}m, p_0).$$

As shown in Figure 2, when m is much larger than p_0, the Jensen gap is larger for larger value of p_0. In the matrix-valued case, this corresponds to the case that when $\dim(X_N) > \dim(X_I)$, larger dimension of X_I yields a larger reduction of the shrinkage bias compared to the Ridge.

To summarize, all three estimators admit shrinkage representations through $\mathcal{A}(M)$, but differ in the structure of regularization. Perfect variable selection sets $M = 0$ and provides no eigenvalue inflation, leaving the variance bounded away from zero. Ridge regression sets $M = \lambda I_n$ and imposes uniform, deterministic shrinkage. Noise regularization sets $M = X_N X_N'$ and induces random shrinkage that, by the convexity of f, produces less bias than its Ridge counterpart while achieving the same variance reduction.

Below, we formalize this regularization intuition, by establishing the first-order forecast optimality: The Ridgeless regularized by sufficiently many noise achieves the same asymptotic MSE as if the true factors were directly used for prediction. We shall present the results separately for the two scenarios.

3.2 Scenario I: identities of noise are unknown

In the first scenario, the economist does not know the identities of informative and noise variables. She uses all predictors in $X = [X_I, X_N]$, and predicts y_{new} by

$$\widehat{y}_{\text{new}} = x_{\text{new}}' \widehat{\beta}.$$

We impose technical assumptions below to establish the benefit of forecast with noises. Recall the data generating process in (2.1) and (2.2) shared by the training data and the test data indexed by for $t \in \{1, \ldots, n\} \cup \{\text{new}\}$.

Assumption 2. *(i) The innovation satisfies*

$$(U', u_{\text{new}}) = \Sigma_u^{1/2}(W', w_{\text{new}}),$$

where $(W', w_{\text{new}})'$ is an $(n+1) \times p$ matrix of i.i.d. entries with zero mean, unit vari-

ance, and finite 4th moment. In addition, $(W', w_{\text{new}})'$ *and* $\{(f_t, \epsilon_{y,t})_{t\in\{1,\dots,n\}\cup\{\text{new}\}}\}$ *are independent.*

(ii) Conditioning on (F, f_{new})*, the error terms* $(\epsilon_{y,t})_{t\in\{1,\dots,n\}\cup\{\text{new}\}}$ *are homoskedastic, zero mean, and independent over time.*

(iii) $\|\mathbb{E}\left[f_{\text{new}} f_{\text{new}}' | F\right]\| \leq C_f$ *for a positive constant* C_f*.*

(iv) The top K *eigenvalues of* $\Lambda_I \Sigma_f \Lambda_I'/\psi_{p,n}$ *are distinct.*

Assumption 2 (i) imposes that (U', u_{new}) are linear transformations of the two-way i.i.d. matrix with a finite 4th moment, which allows to invoke results from the random matrix theory to reveal the behavior of empirical eigenvalues of the idiosyncratic sample covariance matrix $U'U/n$ of large dimensions. The independence between $(W', w_{\text{new}})'$ and $\{(f_t, \epsilon_{y,t})_{t\in\{1,\dots,n\}\cup\{\text{new}\}}\}$ strengthens the mean independence in (2.2) to simplify the computation of the variance of the estimator. Condition (ii) regulates the unpredictable error in the target variable. Though it is possible to extend to martingale difference sequences, we maintain the independence to keep it in line with the assumption on u_t. Condition (iii) is a technical assumption that bounds the conditional variance of the new factor realizations. Finally, Condition (iv) ensures the uniqueness of the eigenvectors associated with the distinct top K eigenvectors.

Remark 2. Our method allows for both cross-sectional and time series forecasts. In the context of time series forecasting, we emphasize that both y_t and x_t may exhibit serial dependence, through the dynamics in factors. For instance, in dynamic factor models with a finite order lag polynomials, f_t is a finite-dimensional vector of lagged factors, and we follow Stock and Watson (2002a) to treat these lagged factors as if they are static. We also allow the test data $(y_{\text{new}}, x_{\text{new}})$ to be dependent of the training data if the "new" data is "h-period-ahead" of the training data (y, X). What is ruled out by Assumption 2 (i) and (ii) is the serial dependence of the idiosyncratic errors $(\epsilon_{y,t}, u_t)$ over time, to simplify the arguments. This condition resembles what Hirano and Wright (2017, p.621) impose on the regression error (e_t in our notation) when they calculate the prediction risk, and Doz et al. (2012, pp.1016–1017) in quasi-maximum likelihood estimation.

Conditioning on X, the predictive MSE is

$$\mathbb{E}[(y_{new} - \widehat{y}_{\text{new}})^2 | X] = \sigma_\epsilon^2 + \mathbb{E}\left[(\rho' f_{\text{new}} - \widehat{y}_{\text{new}})^2 | X\right]$$

The first term σ_ϵ^2 is the minimum forecast MSE under the true latent factors. The second term, depending on the estimator, can be decomposed as

$$\rho' f_{\text{new}} - \widehat{y}_{\text{new}} = (\rho' f_{\text{new}} - x'_{\text{new}}\beta) - x'_{\text{new}}(\widehat{\beta} - \beta) \tag{3.2}$$

where $\rho' f_{\text{new}} - x'_{\text{new}}\beta$ is the approximation error of the working linear model (2.4). The estimation error $x'_{\text{new}}(\widehat{\beta} - \beta)$ can be further decomposed into

$$\begin{aligned} x'_{\text{new}}(\widehat{\beta} - \beta) &= x'_{\text{new}}\delta_\beta + x'_{\text{new}}B_X\epsilon_y, & (3.3)\\ \text{where } \delta_\beta &= \underbrace{A_X\beta}_{\text{projection}} + \underbrace{B_X\left(F(\rho - \Lambda'\beta) - U\beta\right)}_{\text{working model approx. error}} & (3.4) \end{aligned}$$

with $A_X = (X'X)^+X'X - I_p$ and $B_X = (X'X)^+X'$. The term $x'_{\text{new}}\delta_\beta$ can be viewed as the bias (conditioning on (U, F)). The other term $x'_{\text{new}}B_X\epsilon_y$ constitutes the variance due to the innovation ϵ_y, and it is independent of β. Since

$$\mathbb{E}[[x'_{\text{new}}(\widehat{\beta} - \beta)]^2|X] \leq 2\mathbb{E}[(x'_{\text{new}}\delta_\beta)^2|X] + 2\mathbb{E}[(x'_{\text{new}}B_X\epsilon_y)^2|X],$$

the following theorem presents the convergence of each component on the right-hand side.

Theorem 2. *Suppose Assumptions 1 and 2 hold. If $p = o(\psi_{p,n}n)$ and $n = o(p)$, then:*

(i) The components of the forecast error is:

$$\begin{aligned} \mathbb{E}[(x'_{\text{new}}\delta_\beta)^2|X] &= O_P\left(\frac{p}{\psi_{p,n}n} + \frac{n}{p}\right), \quad \text{(bias)}\\ \mathbb{E}[(x'_{\text{new}}B_X\epsilon_y)^2|X] &= O_P\left(\frac{1}{n} + \frac{n}{p}\right), \quad \text{(variance)}. \end{aligned}$$

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

$$\mathbb{E}[(y_{\text{new}} - \widehat{y}_{\text{new}})^2|X] \xrightarrow{P} \sigma_\epsilon^2.$$

Theorem 2 implies that keeping the noise variables and applying Ridgeless forecast, we can achieve the oracle predictive MSE, as if the latent factors were used directly in forecasting. This result allows factors to be relatively weak among the informative predictors, i.e., $\psi_{p,n} = o(p_0)$ in Assumption 1 (i).

Contrary to conventional wisdom suggesting that the variance is amplified under overfitting, here the variance diminishes as $p \gg n$. The reduction of variance requires no

condition on the predictability, i.e., it does not matter whether the predictors are mostly noise or informative. This is the phenomenon as documented as "benign-overfitting" in the literature. But different from the literature, the overfitting is mainly caused by high-dimensional uninformative predictors X_N. In the presence of many noise variables the out-of-sample variance still decays as $p \to \infty$.

3.3 Scenario II: intentionally added noise

When economic theory and empirical evidence guide data collection, regressors are likely to be informative, but the number of available predictors p_0 may not be very large, say, just proportional to n. This occurs, for instance, in diffusion index forecasts using FRED-MD as mentioned in Introduction.

We recommend adding noise to intentionally increase the overall dimensionality. The inclusion of noise predictors inflates eigenvalues of the design matrix and thus reduces the out-of-sample variance. In the meantime, inspired by the Rao–Blackwellization for risk reduction (Hirano and Wright, 2017), the impact of randomness can be smoothed out by taking average of repeated Ridgeless predictions.

Recall that (X_I, y) denotes the in-sample data in this scenario, and the corresponding out-of-sample predictor is $x_{\text{new},I}$, where both X_I and $x_{\text{new},I}$ are informative predictors.

Algorithm 1. *Noise-regularized forecast.*

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 to get $X = (X_I, X_N)$, and obtain the Ridgeless estimator*

$$\widehat{\beta} = (X'X)^+ X'y.$$

Let $\widehat{\beta}_I$ be the subvector of $\widehat{\beta}$ corresponding to X_I and predict $\widehat{y}_{\text{new},I} := x'_{\text{new},I}\widehat{\beta}_I$.

Step 2 (denoising): *Repeat Step 1 to obtain $\widehat{\beta}_I^b$ and then $\widehat{y}_{\text{new}}^b = x'_{\text{new},I}\widehat{\beta}_I^b$ for an index $b = 1, \ldots, B$. Averaging over the B replications to get a forecast*

$$\widehat{y}_{\text{new}}^* = B^{-1} \sum_{b=1}^{B} \widehat{y}_{\text{new},I}^b = x'_{\text{new},I}\widehat{\beta}_I^*$$

where $\widehat{\beta}_I^* = B^{-1} \sum_{b=1}^{B} \widehat{\beta}_{\text{new},I}^b$.

Step 1 creates artificial pure noise X_N to be merged with X_I, and then uses only the subvector $\widehat{\beta}_I$ for prediction; the forecast depends on the realization of X_N. Step 2

averages out the randomness caused by X_N through the B repetitions. This estimator approximates

$$\mathbb{E}(\widehat{y}_{\text{new},I}|X_I, y, x_{\text{new},I}) \tag{3.5}$$

with a sufficiently large B. This estimator can be connected to Bayesian model average: it arises as an equally weighted average forecasts (conditioning on the data) from B submodels $\widehat{y}^b_{\text{new},I}$, where each submodel is implied by the drawn random noise, whose posterior probability is $1/B$.

As a model averaging estimator, it reduces the mean squared error of submodels. With a slight abuse of notation, we shall use $\widehat{y}^*_{\text{new}}$ to denote (3.5); it is a function of the observed data only, independent of the noise. It follows that

$$x'_{\text{new}}\beta - \widehat{y}^*_{\text{new}} = \mathbb{E}(x'_{\text{new}}\beta - \widehat{y}_{\text{new},I}|X_I, y, x_{\text{new},I}),$$

and the law of iterated expectations gives

$$\begin{aligned}
\mathbb{E}[(x'_{\text{new}}\beta - \widehat{y}^*_{\text{new}})^2|X_I] &= \mathbb{E}\left\{ [\mathbb{E}(x'_{\text{new}}\beta - \widehat{y}_{\text{new},I}|X_I, y, x_{\text{new},I})]^2 \,\middle|\, X_I \right\} \\
&\leq \mathbb{E}\left\{ \mathbb{E}\left[(x'_{\text{new}}\beta - \widehat{y}_{\text{new},I})^2|X_I, y, x_{\text{new},I} \right] \,\middle|\, X_I \right\} \\
&= \mathbb{E}\left[(x'_{\text{new}}\beta - \widehat{y}_{\text{new},I})^2|X_I \right], \qquad (3.6)
\end{aligned}$$

where the inequality follows by the celebrated Rao-Blackwell theorem. It shows that $\widehat{y}^*_{\text{new}}$ enjoys a smaller predictive MSE than that of a single realization $\widehat{y}_{\text{new},I}$.

In this scenario we define $x_{\text{new}} = (x'_{\text{new},I}, 0')'$ and $\beta = (\beta'_I, 0')'$. Then

$$\mathbb{E}[(\widehat{y}^*_{\text{new}} - x'_{\text{new}}\beta)^2|X_I] \leq 2\mathbb{E}[(x'_{\text{new}}\delta^*_\beta)^2|X_I] + 2\mathbb{E}[(x'_{\text{new}}B^*_X\epsilon_y)^2|X_I]$$

where

$$\delta^*_\beta = A^*_X\beta + B^*_X\left(F(\rho - \Lambda'\beta) - U\beta\right). \tag{3.7}$$

with $A^*_X = \mathbb{E}[(X'X)^+X'X|X_I] - I_p$ and $B^* = \mathbb{E}[(X'X)^+X|X_I]$ averaged out the randomness of X_N. We have the following result for the noise-regularized forecast.

Theorem 3. *Under the assumptions in Theorem 2, we have*

(i) The forecast squared bias and variance:

$$\begin{aligned}\mathbb{E}[(x'_{\text{new}}\delta^*_\beta)^2|X_I] &= O_P\left(\frac{p}{\psi_{p,n}n}\right), \quad \text{(bias)}\\ \mathbb{E}[(x'_{\text{new}}B^*_X\epsilon_y)^2|X_I] &= O_P\left(\frac{1}{n}+\frac{np_0}{p^2}\right). \quad \text{(variance).}\end{aligned}$$

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

$$\mathbb{E}[(y_{\text{new}} - \widehat{y}^*_{\text{new}})^2|X_I] \xrightarrow{P} \sigma^2_\epsilon.$$

This theorem shows that adding the noise enables achieving the oracle forecast MSE as if the true factors were used. In addition, the rate of convergence of the forecast variance is faster than that of Scenario I when $p = o(n^2)$, as a result of the Rao–Blackwellization step.

Figure 3 plots the theoretical curves of bias-variance (left panel) and predictive MSE $\mathbb{E}[(y_{new} - \widehat{y}^*_{\text{new}})^2|X_I]$ (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. $N(0, 1)$ white noise variables. 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 the theory. The depicted bias does not diminish because here we fix $\psi_{p,n} \asymp n$ at $n = 100$ while we vary p.

Overall, the predictive MSE (right panel) illustrates the double descent phenomenon, where the first descent occurs before $p < 20$, due to the decay of the gap $\sigma^2_e - \sigma^2_\epsilon$. 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. While double descent has been documented in the recent statistical literature, our result is novel. Section 3.4 below elaborates on the novelty of our contribution relative to that literature.

Both Figure 3 and our theory indicate an inflation of bias 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,

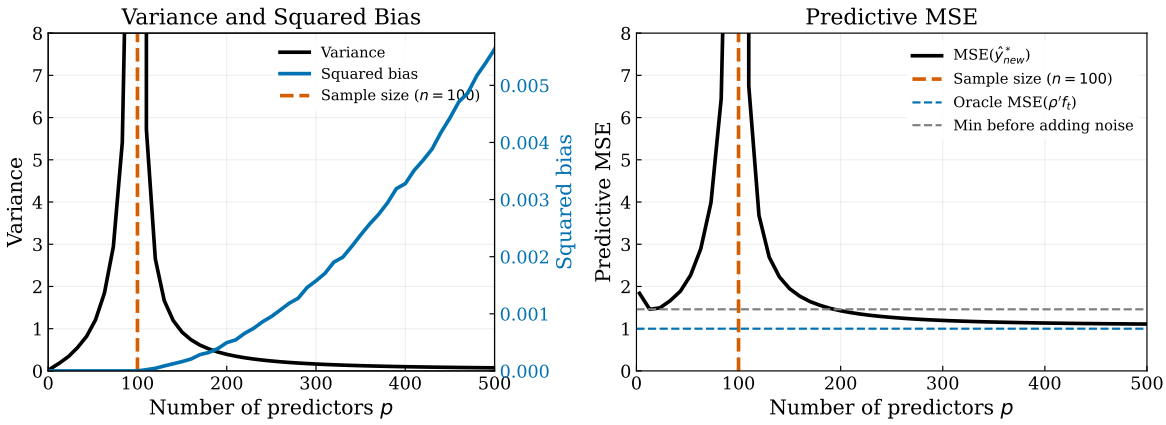


Figure 3: Bias-Variance Tradeoff

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 $\text{MSE}(\rho' f_t) = \text{Var}(\epsilon_{y,t})$, the oracle predictive MSE.

then $p_0 \asymp \psi_{p,n}$ and our theory shows:

$$\mathbb{E}[(x'_{\text{new}}\beta - \widehat{y}^*_{\text{new}})^2 | X] = O_P\left(\frac{p}{p_0 n} + \frac{np_0}{p^2}\right).$$

We can choose the optimal number of predictors p by minimizing the rate of convergence:

$$p = C \times (np_0)^{2/3} \asymp \arg\min_p \left(\frac{p}{p_0 n} + \frac{np_0}{p^2}\right),$$

where C is a constant to be determined. 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, $p(C) = C \times (np_0)^{2/3}$ (after adding noise). For each $t \leq n$, let $\widehat{\beta}_{-t}(C)$ denote the Ridgeless estimator using $p(C)$ predictors and all data except the t-th observation. Then the optimal C can be chosen by minimizing:

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

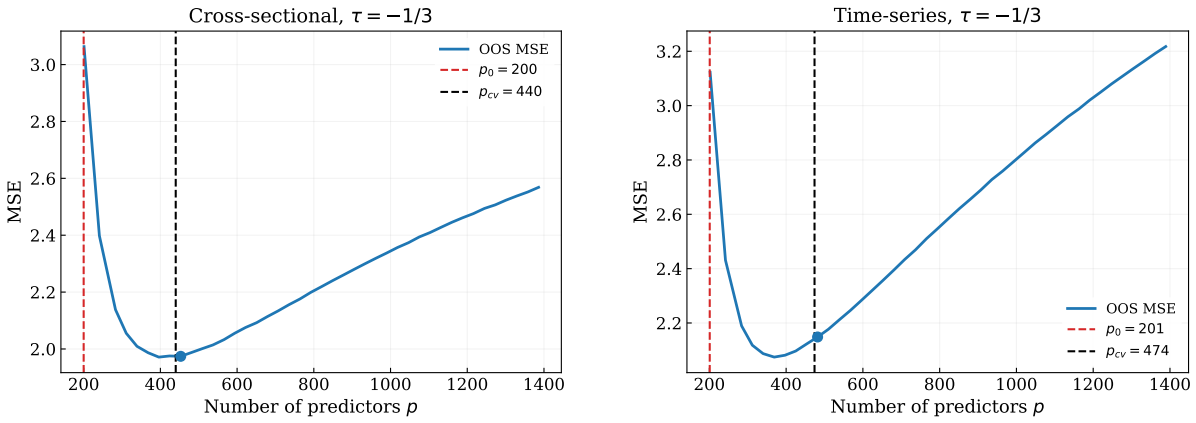
Appealingly, we do not have to compute n leave-one-out $\widehat{\beta}_{-t}(C)$ for this procedure, thanks to an elegant analytic formula given by Shen et al. (2023). They showed that the leave-one-out procedure is equivalent to minimizing:

$$C^* = \arg\min_C \| \operatorname{Diag}(G(C))^{-1} G(C) Y \|^2, \quad G(C) = (X(C)X(C)')^{-1} \tag{3.8}$$

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 rolling window tuning for time series cross-validation: predetermine T_0 and V_0 as the sizes of training and validation sets. Let $\widehat{\beta}_t(C)$ denote the estimated coefficient using T_0 periods of data: $t - T_0 + 1, ..., t$, with tuning parameter C. We then choose C^* by:

$$C^* = \arg\min_C \sum_{l=0}^{L-1} (y_{T_0+l} - x'_{T_0+l} \widehat{\beta}_{T_0+l}(C))^2.$$



(a) Cross-sectional prediction

(b) Time-series prediction

Figure 4: Predictive MSE and Data-driven Tuning

Notes: Out-of-sample MSE of the noise-regularized Rao-Blackwell estimator $\widehat{y}^*_{\text{new}} = x'_{\text{new},I} \widehat{\beta}^*_I$ as the total number of predictors p increases. The first p_0 predictors are informative and generated from a factor model with strength $\tau = -1/3$; the remaining $p - p_0$ are i.i.d. $N(0, 1)$ noise. The sample size is $n = 100$. Panel (a) is a cross-sectional setting where factors are i.i.d. sequences, and p_{cv} is selected by Rao-Blackwell leave-one-out cross-validation (3.8). Panel (b) is a time-series setting where factors follow a stationary AR(1) process, and p_{cv} is selected by rolling-window cross-validation with the first half of the sample reserved for tuning. The dashed vertical line marks p_{cv} averaged over 50 replications.

To illustrate the effectiveness of data-driven choices of $p_{cv} = p(C^*)$, Figure 4 plots the predictive MSE of two simulated designs, where the DGP is generated as in one of the simulation designs in Section 4.1 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.

3.4 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, the double descent phenomenon shows that it is possible to retain predictability by increasing the number of regressors beyond the threshold of interpolation. It has gained increasing attention in the machine learning community. As the model complexity exceeds the sample size and continues to grow, a second descent of the prediction occurs in the excessively 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 Ridgeless regressions, e.g., Mei and Montanari (2019); Belkin et al. (2020); Dobriban and Wager (2018); Lee and Lee (2023).

Our work differs from the recent statistical literature on overparameterized models in three important respects. First, while the literature does not explicitly assume $\|\beta\| \to 0$, this condition is in fact necessary for the forecast consistency. In the i.i.d. setting without presence of strong factors, Hastie et al. (2022) characterize the limiting predictive MSE, whose convergence requires this condition. It turns out that $\|\beta\| \to 0$ remains necessary in the presence of factors, otherwise, the overparametrized bias would not vanish asymptotically. However, the literature has provided little justification for why such a condition should hold in any economically meaningful model.

Our paper provides a formal and economically motivated justification for this condition. We consider a setting in which the true DGP is driven by a low-dimensional factor structure, $y_t = \rho' f_t + \epsilon_{y,t}$, while the econometrician estimates an overparameterized linear regression $y_t = x_t'\beta + u_t$ as a working model. In this environment, the population regression coefficient β is a reduced-form object implied by the underlying factor structure. We show that, as the cross-sectional dimension grows, the factor structure implies $\|\beta\| \to 0$. Hence, the shrinkage condition required for consistency is not imposed, but arises naturally from a standard economic DGP.

Secondly, as explained in the previous subsection, we show that the second descent arises from the shrinkage of the eigenvalue spectrum induced by the presence of many noise variables. Although these noise variables contain no predictive information about the target variable, they cause the high-dimensional sample covariance matrix to have a highly dispersed eigenvalue spectrum. This lowers the out-of-sample variance.

Lastly, from an important technical perspective, our framework explicitly analyzes the impact of a few very large eigenvalues on overparametrized forecasts. In much of the recent work, the design matrix is assumed to have uniformly bounded eigenvalues, which effectively imposes a weak dependence structure across predictors. In contrast, many economic forecasting environments are inherently driven by a small number of latent common factors. These factors generate strong cross-sectional dependence among informative predictors, inducing an approximate low-rank structure in the population covariance matrix. While some recent work on overparameterized models allows for diverging eigenvalues induced by factor structures (e.g., Meng et al. (2025)), the implications of such growing eigenvalues in forecasting models remain unclear.

We provide a careful analysis of this case and show how the interaction between factor-driven spectral spikes and overparameterization shapes the predictive risk. To understand the intuition, start with the matrix form of the factor model

$$\underset{(n\times p)}{X} = \underset{(n\times K)}{F}\underset{(K\times p)}{\Lambda'} + U. \tag{3.9}$$

By applying SVD on both $X'X$ and $\mathbb{E}x_t x_t'$, we can separate the effect of the factor model from its remaining counterparts. First, consider the SVD of $\mathbb{E}x_t x_t'$:

$$\mathbb{E}x_t x_t' = \bar{V}_K \bar{D}_K^2 \bar{V}_K' + \bar{V}_{-K} \bar{D}_{-K}^2 \bar{V}_{-K}'$$

corresponding to the first K eigenvectors/values $(\bar{V}_K, \bar{D}_K)$ and last $p-K$ eigenvectors/values $(\bar{V}_{-K}, \bar{D}_{-K})$. In addition, consider the SVD on $X'X$:

$$X'X = V_K D_K^2 V_K' + V_{-K} D_{-K}^2 V_{-K}' + V_{-n} 0 V_{-n}',$$

where D_K^2 and D_{-K}^2 respectively denote the diagonal matrices of the largest K and the remaining $n-K$ nonzero eigenvalues of $X'X$; columns of V_K and V_{-K} are the corresponding eigenvectors; columns of V_{-n} are the eigenvectors corresponding to the $p-n$ zero-eigenvalues.

Then we decompose the forecast bias and variance of $\widehat{y}_{\text{new}}$ into:

$$\mathbb{E}[(x'_{\text{new}}\delta_\beta)^2|X] \;=\; A_1 + A_2, \quad \mathbb{E}[(x'_{\text{new}}B_X\epsilon_y)^2|X_I] = B_1 + B_2$$

where for $\Sigma_X = \mathbb{E}(x_t x_t')$,

$$A_1 = \underbrace{\|\bar{V}_K \bar{D}_K \bar{V}_K' V_{-n} V_{-n}'\beta\|^2}_{\text{factor-model design}}, \qquad B_1 = \underbrace{\sigma_e^2 \operatorname{tr}\left(V_K D_K^{-2} V_K' \Sigma_X\right)}_{\text{factor-model design}},$$
$$A_2 = \underbrace{\|\bar{V}_{-K} \bar{D}_{-K} \bar{V}_{-K}' V_{-n} V_{-n}'\beta\|^2}_{\text{i.i.d. design}}, \qquad B_2 = \underbrace{\sigma_e^2 \operatorname{tr}\left(V_{-K} D_{-K}^{-2} V_{-K}' \Sigma_X\right)}_{\text{i.i.d. design}}$$

Both bias and variance depend on components that arise from the factor model A_1, B_1, and from the idiosyncratic noise A_2, B_2 that behave similarly to the classic i.i.d. design. In particular, the convergence of A_1 and B_1 rely on three properties of the factor model: (i) the population eigenvectors $\bar{V}_K$ can be well approximated by the sample eigenvectors V_K; (ii) the implied reduced form parameter is dense in that $\|\beta\| \to 0$; and (iii) the top K eigenvalues grow very fast. As for A_1, the sample eigenvectors corresponding to the first K eigenvalues are orthogonal to V_{-n}. Hence (i) implies $\|\bar{V}_K' V_{-n}\| \approx \|V_K' V_{-n}\| \xrightarrow{P} 0$, and the convergence of A_1 is ensured together with (ii). In addition, the behavior of B_1 is driven by the top K sample eigenvalues, and thus B_1 converges as $\|D_K^{-2}\| = O_P((\psi_{p,n} n)^{-1})$.

In summary, when most of the high dimensional predictors are generated by a few common factors, the factors lead to a few spiked eigenvalues facilitating the convergence of variance; they also naturally yield a dense regression coefficient to shrink the bias.

4 Simulation

We run Monte Carlo experiments to check the performance of noise-regularization under the factor model specification in (2.1) and (2.2). The outcome variable is generated from a $K = 4$ factor model $y_t = \rho' f_t + \epsilon_{y,t}$, with p_0 informative predictors and $p - p_0$ noise:

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

and $(f_t, \epsilon_{y,t}, \lambda_{i,0}, u_{i,t})$ are all standard normal, and $\rho \sim N(0, I_K)$. Here $\tau \in [-1/2, 0]$ determines the factor strength so that $\Lambda_I' \Lambda_I \asymp p_0^{1+2\tau}$; the closer τ to zero, the stronger the factors. To capture heterogeneous factor strengths, we allow τ to differ across the

$K = 4$ factors and consider two configurations:

- *Homogeneous:* $\tau = -1/3$ for all 4 factors (moderately weak).
- *Mixed:* $\tau = (-1/4, -1/3, -1/2, -1/2)$, combining one strong, one moderate, and two weak factors.

The mixed case is of particular interest because the two factors with $\tau = -1/2$ are very weak and is also a plausible empirical scenario. The sample size is $n = 100$ throughout, and test MSE is evaluated on $n_{\text{new}} = 500$ new observations. All results are averaged over 100 Monte Carlo replications. We set p to take values on a grid evenly spaced from 1 to $p_{\max} = 1000$.

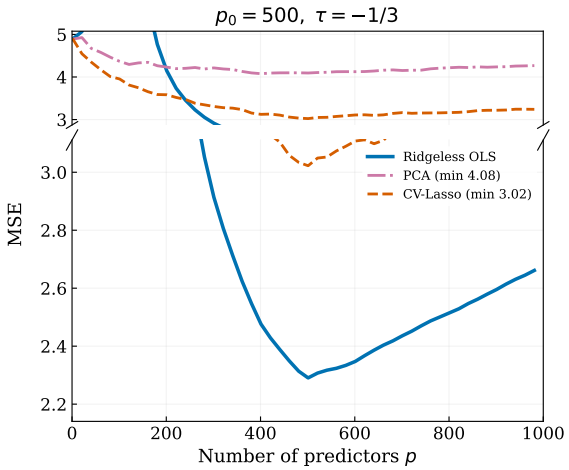
We consider two scenarios as in Sections 3.2 and 3.3, respectively.

4.1 Scenario I: unknown noise identities

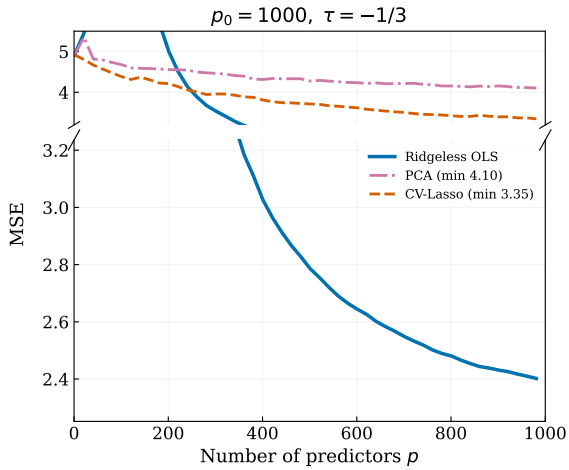
Suppose the economist does not know which predictors are informative, so she uses all p collected predictors. We set $p_0 \in \{500, 1000\}$; all predictors are informative when $p \leq p_0$, and the additional $p - p_0$ are noise when $p > p_0$. Three methods are compared, all using the full set of p predictors: the proposed Noise-regularized estimator, PCA with the number of factors chosen by Bai and Ng (2002), and CV-Lasso where the tuning parameter is chosen by 10-fold cross-validation.

Figure 5 plots the predictive MSE as p increases. All predictors are informative when $p \leq p_0$, whereas $p - p_0$ noise variables enter when $p > p_0$. In all four panels, the noise-regularization dominates both PCA and CV-Lasso, particularly in the overparameterized regime ($p > n$). In panel (b), where all predictors are informative ($p_0 = 1000$) with homogeneous $\tau = -1/3$, including more weak predictors continuously benefits noise-regularization, driving its MSE down to 2.40 at the right edge. In panels (a) and (c), where $p_0 = 500$ and the number of informative predictors stops increasing, the MSE curve flattens after p_0 but remains well below the competitors.

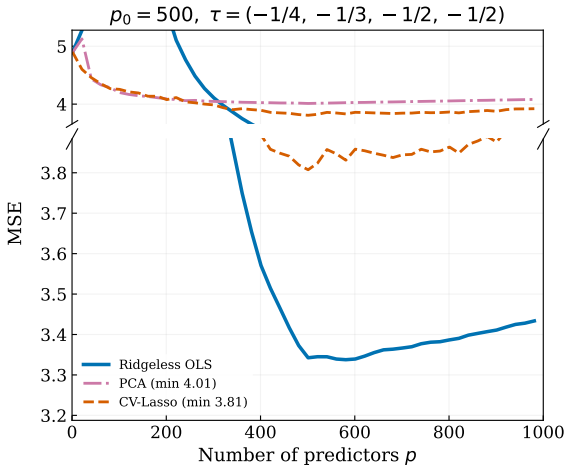
PCA gains little from large p because the factors are weak, which misses a substantial portion of the predictive signal. CV-Lasso struggles for a different reason: the dense factor structure spreads signal across all informative predictors, each carrying a small but nonzero coefficient. Lasso's variable selection, designed for sparse models, selects too few predictors to recover this diffuse signal.



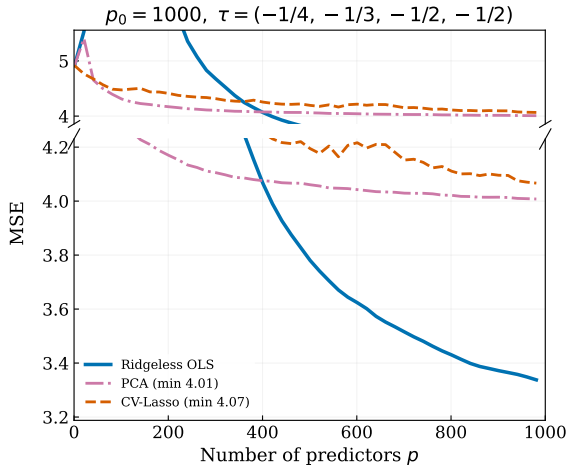
(a) $p_0 = 500$, $\tau = -1/3$



(b) $p_0 = 1000$, $\tau = -1/3$



(c) $p_0 = 500$, mixed τ



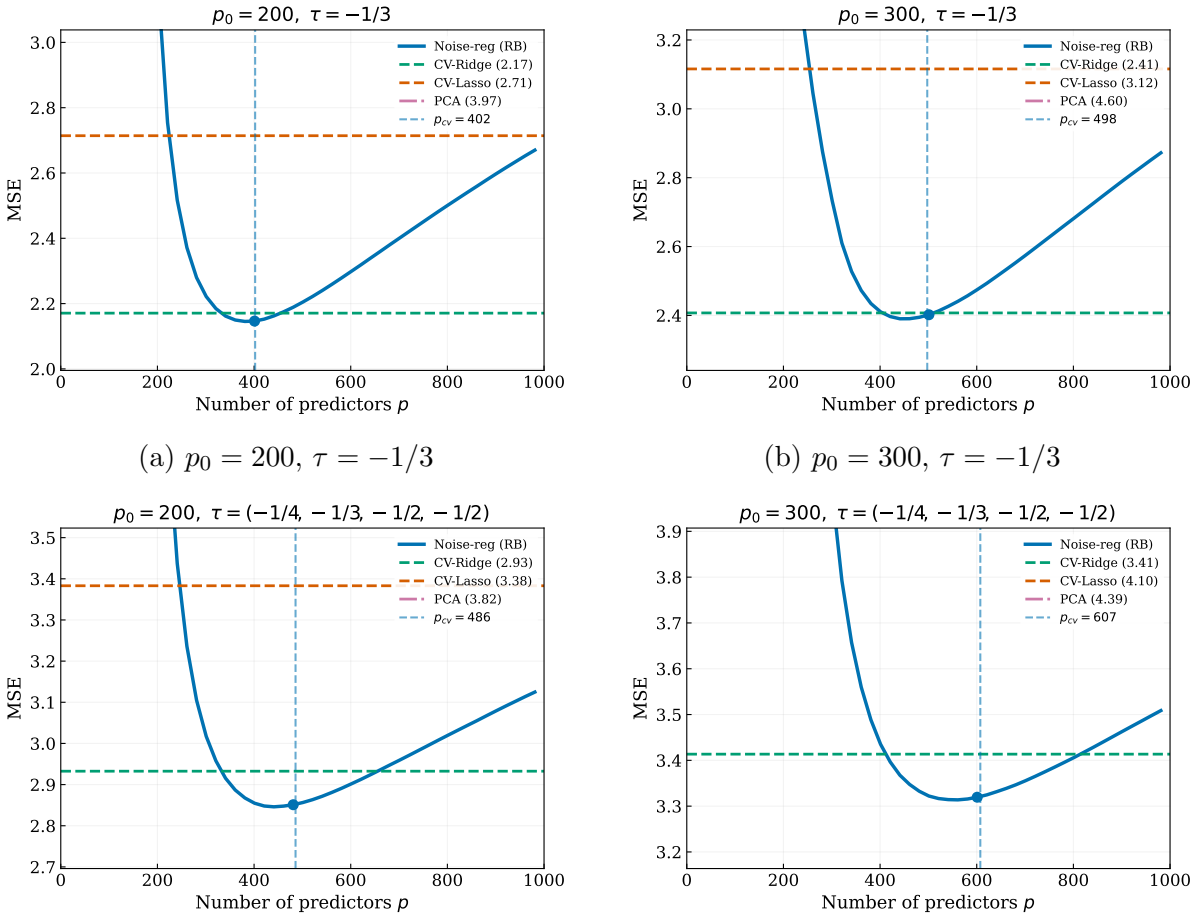
(d) $p_0 = 1000$, mixed τ

Figure 5: Scenario I: predictive MSE as the number of predictors p increases. The sample size is $n = 100$. All three methods use the full set of p predictors. The top row has homogeneous factor strength $\tau = -1/3$; the bottom row has mixed $\tau = (-1/4, -1/3, -1/2, -1/2)$. PCA and CV-Lasso minimum MSE values are reported in the legend when their curves exceed the plotted range.

4.2 Scenario II: intentionally added noise

We now consider the case where the economist knows which predictors are informative and intentionally adds noise to implement noise-regularization. The p_0 informative predictors are augmented with $p - p_0$ columns of $N(0, 1)$ noise. The noise-regularized estimator is computed for $B = 50$ times with independent noise draws, and the predictions are averaged.

We set $p_0 \in \{200, 300\}$ and compare four methods: Noise-regularized with the number of added noise variables selected via leave-one-out cross-validation[2]; CV-Ridge which applies Ridge regression on the p_0 informative predictors and tuned by 10-fold cross-validation; CV-Lasso and PCA. Except for noise-regularization, all three competing methods use only the informative predictors: they are oracle methods with perfect variable selection.



(a) $p_0 = 200$, $\tau = -1/3$

(b) $p_0 = 300$, $\tau = -1/3$

(c) $p_0 = 200$, mixed τ

(d) $p_0 = 300$, mixed τ

Figure 6: Scenario II: predictive MSE as the number of predictors p increases. The sample size is $n = 100$; the noise-regularized estimator averages $B = 50$ Rao-Blackwell draws. The vertical dashed line marks p_{cv} selected by LOO cross-validation. CV-Ridge, CV-Lasso, and PCA use only the p_0 informative predictors (oracle variable selection).

Figure 6 plots the predictive MSE as p increases. The horizontal dashed lines show the MSE of the oracle methods, and the solid curve is the noise-regularized forecast. The

[2]Algorithm 2 in Section A.1 of the appendix presents a detailed algorithm for selecting the number of added noise using leave-one-out method.

vertical dashed line indicates p_{cv} selected by leave-one-out cross-validation. The cross-validation selects p_{cv} close to the oracle optimum. In panel (a), for instance, we select $p_{cv} \approx 400$ while the minimum of the MSE curve occurs at $p \approx 380$.

In all four panels, noise-regularization at p_{cv} outperforms CV-Ridge, despite the latter having access to the informative predictors only. The improvement is most pronounced in the mixed-τ panels: with $p_0 = 200$, the noise-regularized MSE is 2.85 versus 2.93 for CV-Ridge; with $p_0 = 300$, the gap widens to 3.32 versus 3.41. Even in the homogeneous case ($\tau = -1/3$), noise-regularization matches or slightly beats Ridge (2.15 vs. 2.17 for $p_0 = 200$; 2.40 vs. 2.41 for $p_0 = 300$).

The advantage over Ridge is amplified when factor strengths are heterogeneous. Ridge regression applies a single penalty λ to all coefficients, whereas the implicit regularization through noise addition adapts to the signal structure—coefficients associated with stronger factors are shrunk less. This heterogeneous shrinkage is a consequence of the randomized regularization mechanism and cannot be replicated by a single Ridge penalty. CV-Lasso and PCA both perform substantially worse. Lasso suffers from the dense factor structure, while PCA fails to detect the weakest factors, especially in the mixed-τ panels where two factors have $\tau = -1/2$.

5 Empirical Applications

We apply the noise-regularized forecast to three settings in which high-dimensional prediction is of first-order importance: U.S. inflation, cross-country GDP growth, and the U.S. equity premium. In each application the economist observes p_0 economic predictors and augments them with $p - p_0$ columns of i.i.d. $N(0, 1)$ noise. The Rao-Blackwell denoised prediction averages $B = 50$ independent noise draws. The total dimension p is selected by data-driven cross-validation, specifically leave-one-out in the cross-sectional application (growth) and rolling-window time-series CV in the two time-series applications (inflation and equity premium). We compare with CV-Lasso and CV-Ridge, both tuned by five-fold cross-validation and using only the p_0 economic predictors.

In the subsequent subsections, we study the three applications in detail. Table 1 summarizes the main results across all studies.

Table 1: Summary of empirical results

Application	p_0	p_{cv}	Noise-reg	CV-Lasso	CV-Ridge	PCA
Predictive MSE						
Inflation	102	526	**0.982**	1.103	1.102	1.234
Growth	60	396	**0.940**	1.262	1.193	13.58
Out-of-sample R^2 (%)						
Equity premium (annual)	15	1,000	**+1.81**	−14.49	−11.92	−1,276
Equity premium (quarterly)	15	4,000	**+1.08**	−4.55	−2.29	−56.91
Equity premium (monthly)	15	12,000	**+0.37**	−0.46	−1.59	−11.50

5.1 U.S. inflation forecast

Accurate inflation forecasts are essential for monetary policy, business planning, and investment decisions. We use the FRED-MD dataset of McCracken and Ng (2016) to forecast the change in the U.S. inflation rate, defined as

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

The dataset contains $p_0 = 102$ macroeconomic predictors spanning June 1959 to January 2024. We use 120-month (10-year) rolling windows to produce one-month-ahead forecasts. This dataset is widely recognized for its inherent challenge of relatively weak factors: data-driven criteria such as Bai and Ng (2002)'s information criteria typically suggests 8–10 factors, explaining only 50–62% of total variation.

Adding noise predictors inflates the eigenvalues of the design matrix and provides the implicit regularization that avoids estimating weak factors. Noise-regularization augments the p_0 macro variables with $p - p_0$ columns of $N(0, 1)$ noise, and the dimension p is selected by time-series cross-validation on a burn-in period prior to the out-of-sample evaluation. The optimal number of predictors is $p_{cv} = 526$, corresponding to roughly 420 added noise columns. CV-Lasso, CV-Ridge, and PCA use only the original macro predictors throughout.

Figure 7 plots the predictive MSE as p increases on a logarithmic grid. The noise-regularized forecast exhibits a sharp spike near the interpolation threshold ($p \approx n = 120$) but its MSE drops rapidly as more noise is added, outperforming benchmarks by $p \approx 400$. At $p_{cv} = 526$ the noise-regularized MSE is 0.982, an 11% improvement over

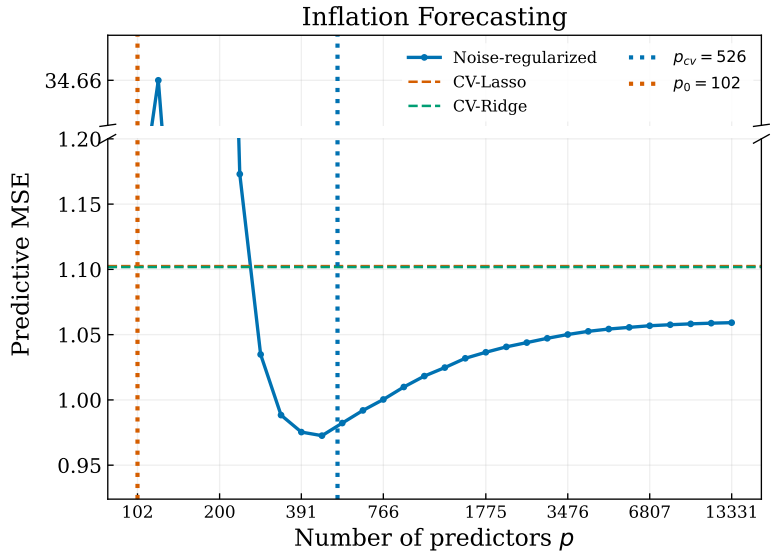


Figure 7: Inflation forecasting: predictive MSE as the number of predictors p increases. The $p_0 = 102$ FRED-MD macro predictors are augmented with $p - p_0$ noise columns. Rolling 120-month windows, one-month-ahead forecast. The vertical dashed line marks $p_{cv} = 526$ selected by time-series CV. CV-Lasso and CV-Ridge use only the original macro predictors.

CV-Lasso (1.103) and CV-Ridge (1.102). Beyond p_{cv}, the MSE curve remains nearly flat through $p > 13{,}000$, demonstrating the robustness to the number of added noise variables predicted by the theory.

5.2 Growth forecasts

Economic growth forecasting has been one of the earliest high-dimensional applications in economics (Sala-I-Martin, 1997). We use the dataset of Barro and Lee (1994), which contains $p_0 = 60$ socio-economic and geographical characteristics for 90 countries. In each replication we randomly split the countries into a training set of $n = 45$ and a test set of 45, fit the model on the training set, and evaluate predictions on the test set. Results are averaged over 50 random splits. The noise-regularized method augments the 60 predictors with $p - p_0$ noise columns. The selected total number of predictors, including noise, is $p_{cv} = 396$ using leave-one-out cross-validation.

Figure 8 plots the predictive MSE. The MSE using only the original 60 predictors via ridgeless OLS is very large (≈ 14) due to overfitting in the $p \approx n$ regime. After the interpolation threshold, the MSE drops precipitously as noise dimensions are added and stabilizes below 1.0 by $p \approx 200$. At $p_{cv} = 396$ the noise-regularized MSE is 0.940, a

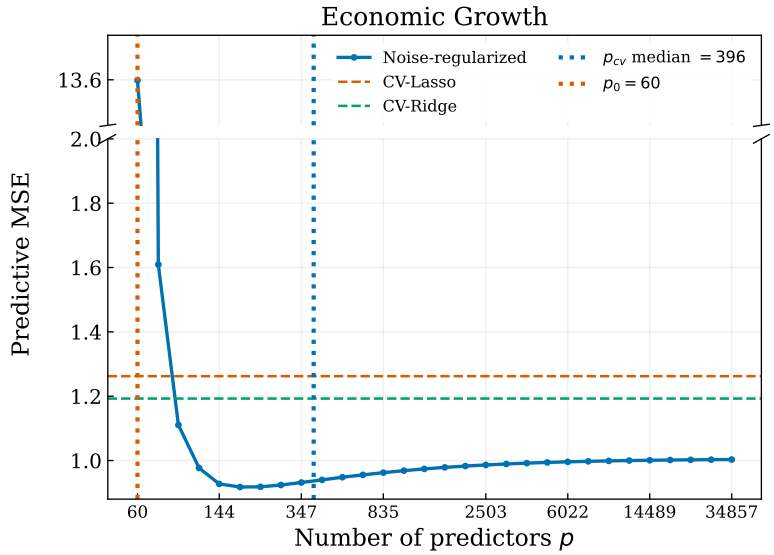


Figure 8: Growth forecasting: predictive MSE as the number of predictors p increases. The $p_0 = 60$ socio-economic variables from Barro and Lee (1994) are augmented with $p - p_0$ noise columns. Random 50/50 train/test splits, averaged over 50 replications. The vertical dashed line marks $p_{cv} = 396$. CV-Lasso and CV-Ridge use only the original 60 predictors.

25% improvement over CV-Lasso (1.262) and a 21% improvement over CV-Ridge (1.193). The MSE remains stable through $p > 34{,}000$, i.e., nearly 600 times the original predictor count, underscoring the method's robustness.

5.3 U.S. equity premium prediction

Predictability of the U.S. equity premium is a central and contentious question in asset pricing. Welch and Goyal (2008) conducted a comprehensive out-of-sample examination of prevailing predictive models, concluding that most fail to beat the historical-mean benchmark. Goyal et al. (2023) updated the analysis with new predictors and methods, arriving at qualitatively similar conclusions: the majority of models yield negative or barely positive out-of-sample R^2.

We revisit this question using the 15 predictors described by Welch and Goyal (2008): dividend-price ratio, earnings-price ratio, book-to-market, Treasury bill rate, default yield spread (AAA, BAA), long-term yield, net equity expansion, risk-free rate, inflation, long-term return, corporate bond return, stock variance, and lagged market returns. We forecast the equity premium at annual, quarterly, and monthly horizons.[3] The out-of-

[3]We drop the cross-sectional premium (csp) because it is discontinued after 2002; the remaining 15

sample evaluation covers 1960–2024 using 20-year rolling windows, harmonized across all three frequencies. The out-of-sample R^2 is defined as

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

where $\bar{y}_t$ is the expanding historical mean of returns up to time t.

At each out-of-sample date, we augment the 15 predictors with $p - 15$ columns of $N(0, 1)$ noise, compute the noise-regularization method averaged over $B = 50$ independent draws as Rao-Blackwellization. The dimension p is selected by time-series cross-validation on a burn-in period (pre-1960 data).

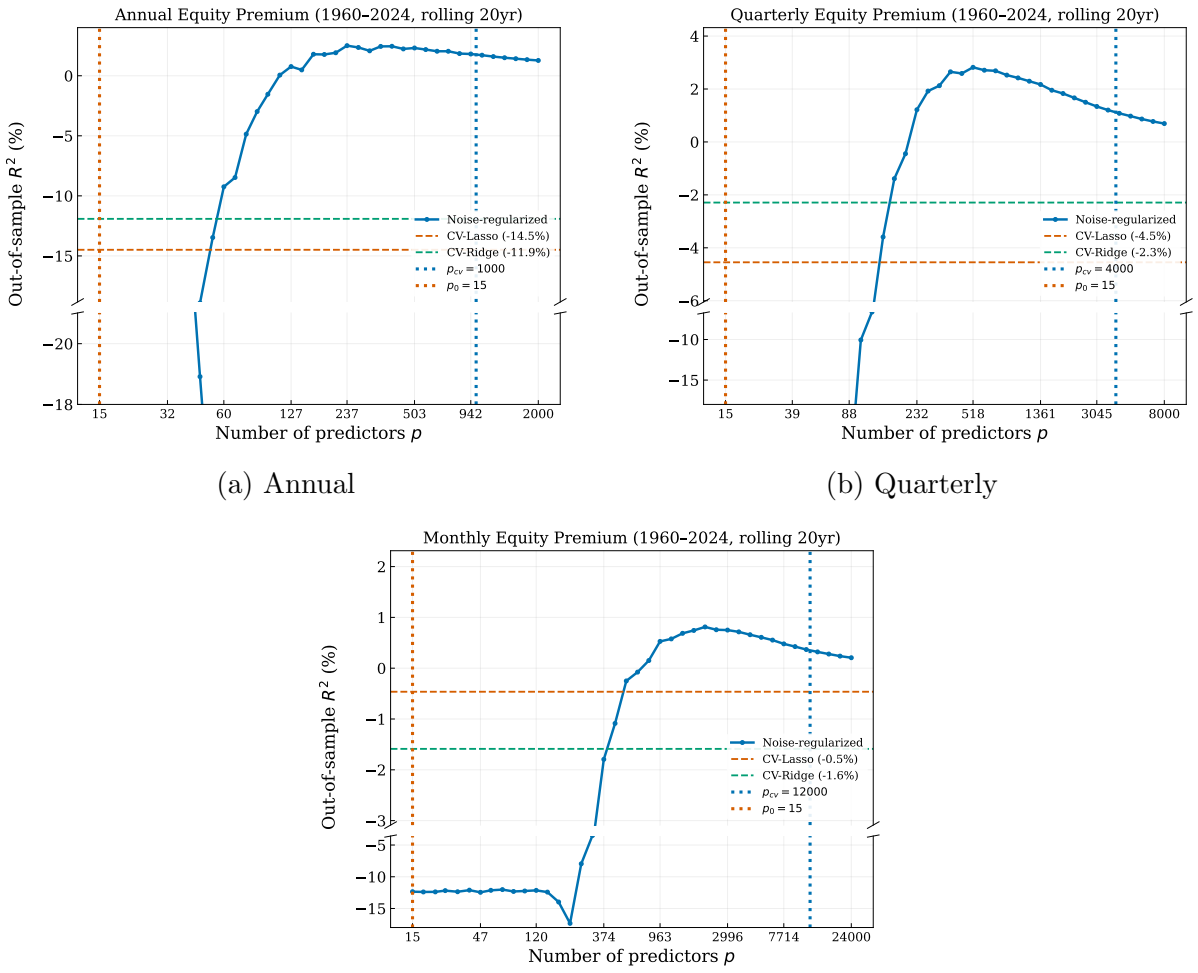
Figure 9 displays the results. The pattern is strikingly consistent across all three frequencies. For small p, the noise-regularized forecast has strongly negative R^2 because the augmented design is near or below the interpolation threshold ($p \approx n$). As p grows into the overparameterized regime, the R^2 rises sharply and stabilizes in the positive territory: precisely the second-descent behavior predicted by the theory.

At the annual frequency (panel a), the noise-regularized forecast achieves $R^2_{\text{oos}} = +1.81\%$ at $p_{cv} = 1{,}000$, compared with -14.49% for CV-Lasso and -11.92% for CV-Ridge. The annual horizon offers the strongest gains, consistent with the well-documented finding that return predictability is more pronounced at lower frequencies (Cochrane, 2011).

At the quarterly frequency (panel b), the R^2 reaches $+1.08\%$ at $p_{cv} = 4{,}000$. Both CV-Lasso (-4.55%) and CV-Ridge (-2.29%) remain negative. At the monthly frequency (panel c), the noise-regularized R^2 is $+0.37\%$ at $p_{cv} = 12{,}000$, while CV-Lasso (-0.46%) and CV-Ridge (-1.59%) again fail to beat the historical mean. Although the monthly R^2 is modest in absolute terms, a positive out-of-sample R^2 at the monthly horizon is economically meaningful: as Campbell and Thompson (2008) note, monthly R^2 values in the range of 0.5% can generate substantial utility gains for a mean-variance investor.

Three features merit emphasis. First, the cross-validated p_{cv} scales roughly in proportion to the sample size: p_{cv}/n is approximately 50 at all three frequencies (1,000/20, 4,000/80, 12,000/240), suggesting a stable optimal noise-to-signal ratio. Second, the R^2 curve is remarkably flat beyond p_{cv}; even doubling the number of noise variables has little effect, confirming the robustness documented in the simulations. Third, the improvement over CV-Lasso and CV-Ridge does not come from better variable selection but from the opposite direction: by adding pure noise, the method achieves implicit regularization that

predictors extend through 2024.



(a) Annual

(b) Quarterly

(c) Monthly

Figure 9: Out-of-sample R^2 (%) for predicting the U.S. equity premium at three horizons, 1960–2024, using 20-year rolling windows. The solid curve shows the noise-regularized forecast as p increases (15 economic predictors plus $p - 15$ noise columns, averaged over $L = 50$ Rao-Blackwell draws). Horizontal dashed lines show CV-Lasso (red) and CV-Ridge (green), both using only the 15 economic predictors. The vertical dotted lines mark $p_0 = 15$ and p_{cv} selected by time-series CV.

is better adapted to the dense, weak-factor structure of equity premium predictors than explicit penalization with a single tuning parameter.

6 Conclusion

When economic outcomes are driven by latent factors shared across a large predictor panel, the reduced-form forecasting model is dense. This paper takes denseness as a

starting point and establishes four results. The population working model converges to the oracle that uses the true factors as predictor dimension grows (Theorem 1). Perfect variable selection fails to achieve first-order forecast optimality when the number of retained predictors is comparable to n, because the design matrix lacks the dimensionality needed for eigenvalue regularization (Proposition 1). Noise regularization, i.e. retaining or deliberately adding uninformative predictors restores optimality by inflating the eigenvalue spectrum, without estimating the factors or their number (Theorems 2 and 3). The AR(p) symmetry property clarifies that overparameterization through additional lags cannot substitute for expanding the predictor set (Proposition 2).

Empirically, noise regularization improves and stabilizes forecasts of U.S. inflation, cross-country GDP growth, and the U.S. equity risk premium, outperforming Lasso, Ridge, and PCA benchmarks across all three applications. These results concern forecast accuracy in factor-structured environments; for structural interpretation or causal inference, variable selection remains appropriate. The central message is that the relevant margin for improving dense linear forecasts is the dimension of the predictor space, not the informativeness of its individual components.

Our empirical applications illustrate this principle in three areas of economic forecasting.

A Proofs

This appendix elaborates technical proofs of the theoretical statements in the main text.

A.1 Proof of Theorem 1

Proof. The factor model implies that $\Sigma_X = \Lambda\Sigma_f\Lambda' + \Sigma_u$, where $\Sigma_f = \mathbb{E}f_t f_t'$ is invertible. By the definition of linear projection, the projection coefficient is

$$\beta = [\mathbb{E}x_t x_t']^{-1}\mathbb{E}x_t y_t = \Sigma_X^{-1}\mathbb{E}\left(\Lambda f_t + u_t\right)\left(f_t'\rho + \epsilon_{y,t}\right) = \Sigma_X^{-1}\Lambda\Sigma_f\rho$$

under the setting of (2.1) and (2.2) with $\mathbb{E}\left[f_t(u_t', \epsilon_{y,t})\right] = 0$.

Part (i). By the Woodbury matrix identity, we expand β as

$$\begin{aligned}
\beta &= \Sigma_X^{-1}\Lambda\Sigma_f\rho \\
&= \left(\Sigma_u^{-1} - \Sigma_u^{-1}\Lambda(\Sigma_f^{-1} + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\right)\Lambda\Sigma_f\rho \\
&= \Sigma_u^{-1}\Lambda\left(I - (\Sigma_f^{-1} + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}\Lambda'\Sigma_u^{-1}\Lambda\right)\Sigma_f\rho \\
&= \Sigma_u^{-1}\Lambda(\Sigma_f^{-1} + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}\rho.
\end{aligned}$$

The L_2-norm of β is bounded by

$$\begin{aligned}
\|\beta\| &\leq \left\|\Sigma_u^{-1/2}\right\| \left\|\Sigma_u^{-1/2}\Lambda(\Sigma_f^{-1} + \Lambda'\Sigma_u^{-1}\Lambda)^{-1}\right\| \|\rho\| \\
&\leq \left\|\Sigma_u^{-1/2}\right\| \left\|(\Sigma_f^{-1} + \Lambda'\Sigma_u^{-1}\Lambda)^{-1/2}\right\| \|\rho\| \\
&\leq \left\|\Sigma_u^{-1/2}\right\| \left\|(\Lambda'\Sigma_u^{-1}\Lambda)^{-1/2}\right\| \|\rho\| \\
&= \frac{\|\Sigma_u^{-1/2}\|}{\sigma_{\min}(\Sigma_u^{-1/2})} \left\|(\Lambda'\Lambda)^{-1/2}\right\| \|\rho\|.
\end{aligned}$$

Under Assumption 1, we have $\|\Lambda'\Lambda\| \asymp \psi_{p,n}$, while $\sigma_{\max}(\Sigma_u)/\sigma_{\min}(\Sigma_u) = C_U/c_u \in (0, \infty)$ and $\|\rho\| = O(1)$ do not change the order. As a result, $\|\beta\| = O\left(\psi_{p,n}^{-1/2}\right)$.

Part (ii). We decompose the error term of the working model as

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

If follows

$$\mathbb{E}[(u_t'\beta)^2] = \beta'\Sigma_u\beta \leq C_U \|\beta\|^2 = O\left(\psi_{p,n}^{-1}\right)$$

Moreover, the explicit expression of β yields

$$\rho - \Lambda'\beta = \rho - \Lambda'\Sigma_X^{-1}\Lambda\Sigma_f\rho = \left(I_K - \Lambda'\left(\Lambda\Sigma_f\Lambda' + \Sigma_u\right)^{-1}\Lambda\Sigma_f\right)\rho = \Sigma_f^{-1}(\Sigma_f^{-1} + \Lambda'\Sigma_u\Lambda)^{-1}\rho$$

by the Woodbury matrix identity, so that its norm is bounded by

$$\begin{aligned}\|\rho - \Lambda'\beta\| &\leq \left\|\Sigma_f^{-1}\right\| \left\|(\Sigma_f^{-1} + \Lambda'\Sigma_u\Lambda)^{-1}\right\| \|\rho\| \\ &\leq \left\|\Sigma_f^{-1}\right\| \left\|(\Lambda'\Sigma_u\Lambda)^{-1}\right\| \|\rho\| = O\left(\psi_{p,n}^{-1}\right).\end{aligned} \tag{A.1}$$

Hence $\text{Var}(f_t'(\rho - \Lambda'\beta)) \leq \|\rho - \Lambda'\beta\|^2 \text{ tr}(\Sigma_f) = O(\psi_{p,n}^{-2})$.

The conclusion follows given that $\text{Var}\left(f_t'(\rho - \Lambda'\beta) - u_t'\beta\right) = O\left(\psi_{p,n}^{-1}\right)$. □

A.2 Auxiliary lemmas

The proof of Theorem 1 is calculated in the population. To proceed to the sample, we introduce a few auxiliary lemmas in this standalone section. These lemmas will be used repeatedly in the remaining theoretical results.

Lemma 4. *Let A be positive semidefinite matrix, and B be a generic compatible matrix. Then*

(i). $\text{tr}(AB) \leq \|B\| \text{tr}(A)$.

(ii). $\text{tr}(B'AB) \geq \lambda_{\min}(A) \text{tr}(BB')$

Proof. **Part (i)**. von Neumann (1937)'s trace inequality gives

$$\text{tr}(AB) \leq \sum_i \sigma_i(A)\sigma_i(B)$$

Since A is positive semidefinite, we continue the inequality:

$$\sum_i \sigma_i(A)\sigma_i(B) \leq \sum_i \sigma_i(A)|\sigma_i(B)| \leq \|B\| \sum_i \sigma_i(A) = \|B\| \text{tr}(A).$$

Part (ii). Denote the dimension of A as $m \times m$, and thus BB' is also an $m \times m$ positive semidefinite matrix. There is another corollary of von Neumann's trace inequality:

$$\text{tr}(B'AB) = \text{tr}(ABB') \geq \sum_{i=1}^{m} \sigma_i(A)\sigma_{m-i+1}(BB')$$

$$\geq \sigma_{\min}(A) \sum_{i=1}^{m} \sigma_{m-i+1}(BB') = \sigma_{\min}(A) \operatorname{tr}(BB'). \qquad \square$$

The following lemma helps clarify the relationship between conditional expectation and stochastic orders.

Lemma 5. *For four generic random variables $a, b, d \geq 0$ and X, if $\mathbb{E}[a|X] \leq b\mathbb{E}[d|X]$ with $b = O_P(r_n)$ and $\mathbb{E}[d] = \mu_d \in (0, \infty)$, then $\mathbb{E}[a|X] = O_P(r_n)$.*

Proof. This is an elementary and straightforward result. We lay out the proof for completeness. For any constants $C_b, C_d > 0$ and $C_a = 2C_bC_d$, we have

$$\begin{aligned}
& P(\mathbb{E}[a|X] \geq C_a r_n) \\
\leq\ & P(\mathbb{E}[a|X] \geq C_a r_n \,|\, d \leq C_d) + P(d \geq C_d) \\
\leq\ & P(b \cdot \mathbb{E}[d|X] \geq 2C_bC_d r_n \,|\, d \leq C_d, b \leq C_b r_n) + P(b \geq C_b r_n) + P(d \geq C_d) \\
=\ & P(b \geq C_b r_n) + P(d \geq C_d) \\
\leq\ & P(b \geq C_b r_n) + \frac{\mu_d}{C_d},
\end{aligned}$$

where the last inequality follows from the Markov inequality. Under the assumption $b = O_P(r_n)$ and $\mu_d \in (0, \infty)$, we choose C_b and C_d sufficiently large to make the right-hand side of the above inequality arbitrarily small. $\square$

The above two lemmas use generic notations. In contrast, the following lemma will involve random matrices in the context of our paper. Recall $B_X = (X'X)^+X'$.

Lemma 6. *Suppose Assumptions 1 and 2 hold. When $n, p \to \infty$ and $n/p \to 0$, we have*

(i). $\frac{9}{16}c_u \leq \sigma_{\min}(UU'/p) \leq \sigma_{\max}(UU'/p) \leq \frac{25}{16}C_u$ *w.p.a.1.*

(ii). $\sigma_{\min}(XX'/p) \geq \frac{1}{4}c_u$ *w.p.a.1.*

(iii). $\|B_X\|^2 = \|(X'X)^+\| = O_P(1/p)$ *and* $\|B_X'\Lambda\|^2 = O_P(1/n)$.

Proof. **Part (i)** For a constant $\gamma \in (0, 1/16)$, define $n_\gamma = \lfloor \gamma p \rfloor$. Since $n/p \to 0$, without loss of generality, we can assume p is sufficiently large so that $n_\gamma > n$. We define an $n_\gamma \times p$ matrix $W^* = (W', \tilde{W}')'$ where $\tilde{W}$ is an $(n_\gamma - n) \times p$ matrix with i.i.d. entries that follow the same distribution as those in W. Bai and Yin (1993, Theorem 2) yields

$$\sigma_{\max}(W^*W^{*\prime}/p) \xrightarrow{a.s.} (1 + \sqrt{\gamma})^2 \quad \text{and} \quad \sigma_{\min}(W^*W^{*\prime}/p) \xrightarrow{a.s.} (1 - \sqrt{\gamma})^2$$

as $n \to \infty$ and $n_\gamma/p \to \gamma$. Since WW' is the top-left submatrix block of $W^*W^{*\prime}$, w.p.a.1 we have

$$\frac{9}{16} \leq \sigma_{\min}\left(\frac{W^*W^{*\prime}}{p}\right) \leq \sigma_{\min}\left(\frac{WW'}{p}\right) \leq \sigma_{\max}\left(\frac{WW'}{p}\right) \leq \sigma_{\max}\left(\frac{W^*W^{*\prime}}{p}\right) \leq \frac{25}{16}.$$

It follows that w.p.a.1 the minimum eigenvalue

$$\sigma_{\min}\left(UU'/p\right) \geq \sigma_{\min}(\Sigma_u) \cdot \sigma_{\min}(WW'/p) \geq \frac{9}{16}c_u$$

and the maximum eigenvalue

$$\sigma_{\max}(UU'/p) \leq \|\Sigma_u\| \cdot \sigma_{\max}(WW'/p) \leq \frac{25}{16}C_u$$

as stated.

Part (ii). For any $v \in \mathbb{R}^n$ such that $\|v\| = 1$, we decompose

$$\begin{aligned} v'\frac{XX'}{p}v &= v'F\frac{\Lambda'\Lambda}{p}F'v + 2v'F\frac{\Lambda'U'}{p}v + v'\frac{UU'}{p}v = \alpha'\alpha + 2\alpha\eta + v'\frac{UU'}{p}v \\ &\geq \|\alpha\|^2 - 2\|\alpha\|\|\eta\| + v'\frac{UU'}{p}v, \end{aligned}$$

where $\alpha = (\Lambda'\Lambda/p)^{1/2} F'v$ and $\eta = (\Lambda'\Lambda/p)^{-1/2} \frac{\Lambda'}{p}U'v$ are two r-dimensional vectors. The latter is bounded by

$$\begin{aligned} \|\eta\|^2 &= v'U\frac{\Lambda}{p}\left(\frac{\Lambda'\Lambda}{p}\right)^{-1}\frac{\Lambda'}{p}U'v \\ &\leq \frac{1}{p\sigma_{\min}(\Sigma_u)} \cdot v'W\Sigma_u^{1/2}\frac{\Lambda}{\sqrt{p}}\left(\frac{\Lambda'\Sigma_u\Lambda}{p}\right)^{-1}\frac{\Lambda'}{\sqrt{p}}\Sigma_u^{1/2}W'v \\ &\leq (c_u p)^{-1}\sigma_{\max}(\tilde{W}'\tilde{W}) \end{aligned}$$

where $\tilde{W} = \left(\frac{\Lambda'\Sigma_u\Lambda}{p}\right)^{-1/2}\frac{\Lambda'}{\sqrt{p}}\Sigma_u^{1/2}W'$ is a $K \times n$ matrix with i.i.d. entries of zero mean and unit variance. Since K is finite, We have $\frac{n}{p}\sigma_{\max}\left(\frac{\tilde{W}'\tilde{W}}{n}\right) \xrightarrow{p} \lim\frac{n}{p} = 0$, and therefore $\|\eta\| = o_p(1)$; in other words, w.p.a.1 the event $\{\|\eta\| \leq \sqrt{c}\}$ occurs. Under this event, and together with Part (i), we have

$$\sigma_{\min}\left(XX'/p\right) \geq \|\alpha\|^2 - \|\alpha\|\sqrt{c_u} + \frac{9}{16}c_u = (\|\alpha\| - \frac{1}{2}\sqrt{c})^2 + \frac{5}{16}c_u > \frac{1}{4}c_u.$$

Part (iii). Note that $B_X B_X' = (X'X)^+$, and then Part (ii) immediately yields

$$\|B_X\|^2 = \|(X'X)^+\| \leq \sigma_{\min}^{-1}(X'X) \leq 4c_u^{-1}p^{-1} = O_P(1/p).$$

For the other expression, given $\Lambda = (X-U)'F(F'F)^{-1}$ and $(F'F/n)^{-1} = O_P(1)$, we have

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

where the last line follows by the fact that $XB_X = X\left(X'X\right)^+X'$ is idempotent, thereby $\|XB_X\| = 1$. On the other hand, w.p.a.1 we have

$$\|UB_X\|^2 = \left\|U\left(X'X\right)^+U'\right\|^2 \leq \left\|\left(\frac{X'X}{p}\right)^+\right\|\left\|\frac{UU'}{p}\right\| = \frac{\sigma_{\max}\left(UU'/p\right)}{\sigma_{\min}\left(XX'/p\right)} \leq 3\frac{C_u}{c_u}$$

as shown in Parts (i) and (ii). We conclude

$$\|B_X'\Lambda\|^2 \leq \frac{2}{n}\left(1+3\frac{C_u}{c_u}\right)O_P(1) = O_P\left(1/n\right)$$

as stated in Part (iii). □

A.3 Proof of Proposition 1

Proof. Since we work with the informative regressors only, the closed-form solution of the Ridgeless with the regressors in the set "I" is $\widetilde{\beta}_I = (X_I'X_I)^+X_I'Y$. Given that this is a negative message about the Ridgeless regression with p_0 regressors, it suffices to consider a special case where $\epsilon = (\epsilon_t)_{t=1}^n$ is independent over t and independent of U and F. The independence leads to the decomposition

$$\mathbb{E}[(y_{\text{new}} - \widetilde{y}_{\text{new}}^I)^2|X_I] = \sigma_\epsilon^2 + \mathbb{E}[(\rho' f_{\text{new}} - \widetilde{y}_{\text{new}}^I)^2|X_I].$$

Notice

$$\mathbb{E}[x_{I,\text{new}}x'_{I,\text{new}}|X_I] = \Lambda_I \mathbb{E}(f_{\text{new}}f'_{\text{new}}|X_I)\Lambda'_I + \Sigma_{u,I}$$

under Assumption 1 (ii) we have

$$\begin{aligned}\mathbb{E}[(\rho' f_{\text{new}} - \widetilde{y}^I_{\text{new}})^2|X_I] &= \sigma_\epsilon^2 \text{ tr}\left[\mathbb{E}[x_{I,\text{new}}x'_{I,\text{new}}|X_I](X'_IX_I)^+\right] \geq \sigma_\epsilon^2 \text{ tr}\left[\Sigma_{u,I}(X'_IX_I)^+\right] \\ &\geq \sigma_{\min}(\Sigma_{u,I})\sigma_\epsilon^2 \text{ tr}\left[(X'_IX_I)^+\right] \geq c_u\sigma_\epsilon^2 \text{ tr}\left[(X'_IX_I)^+\right]\end{aligned}$$

by Lemma 4 (ii). We focus on the trace

$$\text{tr}\left[(X'_IX_I)^+\right] = \sum_{j=1}^{p_0\wedge n} \sigma_j^{-1}(X'_IX_I) > \sum_{j=K+1}^{p_0\wedge n} \sigma_j^{-1}(X'_IX_I) \geq \frac{p_0\wedge n}{n}\sigma_{K+1}^{-1}\left(\frac{X'_IX_I}{n}\right),$$

where the strict inequality holds when we remove the top K eigenvalues from the summation. When $p_0/n \to \gamma \in (0,1)$, Cai et al. (2020, Theorem 2.5) shows the $(K+1)$-th sample eigenvalue — the largest non-spike eigenvalue $\sigma_{K+1}(X'_IX_I/n)$ — is a consistent estimator for the corresponding population eigenvalue $\sigma_{K+1}(\Sigma_{X,I}) > 0$. There exists a constant c such that

$$\mathbb{E}[(\rho' f_{\text{new}} - \widetilde{y}^I_{\text{new}})^2|X_I] > c(\gamma \wedge 1)$$

bounded away from 0 when the sample size is sufficiently large. Therefore, we conclude that there is an asymptotically non-vanishing gap between $\mathbb{E}[(y_{\text{new}} - \widetilde{y}^I_{\text{new}})^2|X_I]$ and σ_ϵ^2. The conclusion follows. □

A.4 Proof of Proposition 2

Proof. For an AR(p) without an intercept, the $(T-p)\times p$ regressor matrix X, the $T-p$ outcome vector Y, and the p dimensional "new" regressor x_{new} are respectively given as

$$\underset{(T-p)\times p}{X} = \left(Y_{1:T-p}, Y_{2:T-p+1}, \cdots Y_{p:T-1}\right), \quad Y = Y_{p+1:T}, \text{ and } x_{\text{new}} = Y_{T-p+1:T}.$$

Meanwhile, for the AR($T-p$) model, the regressor matrix $\widetilde{X}$ is of dimension $p\times(T-p)$, the outcome vector $\widetilde{Y}$ is of dimension p, and the "new" regressor $\widetilde{x}_{\text{new}}$ is $T-p$ dimensional, given respectively as:

$$\widetilde{X} = X', \quad \widetilde{Y} = x_{\text{new}}, \text{ and } \widetilde{x}_{\text{new}} = Y.$$

If $p > T/2$, the predicted value is

$$\begin{aligned}\widehat{y}_{T+1}^{AR}(p) &= x'_{\text{new}}(X'X)^{+}X'Y = x'_{\text{new}}X'(XX')^{-1}Y = \widetilde{Y}'\widetilde{X}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{x}_{\text{new}} \\ &= \widetilde{x}'_{\text{new}}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{Y} = \widehat{y}_{T+1}^{AR}(T-p)\end{aligned}$$

By symmetry, the case of $p \leq T/2$ maintains the equivalence. □

A.5 Proof of Theorem 2

Proof. Conditioning on X, the predictive MSE is

$$\mathbb{E}[(y_{new} - \widehat{y}_{\text{new}})^2|X] = \text{Var}(\epsilon_{y,\text{new}}|X) + \mathbb{E}\left[(\rho' f_{\text{new}} - \widehat{y}_{\text{new}})^2|X\right].$$

The first term is the optimal forecast MSE using the true latent factors, which cannot be reduced by any estimation method. The second term, which depends on the estimator, can be decomposed as

$$\rho' f_{\text{new}} - \widehat{y}_{\text{new}} = (\rho' f_{\text{new}} - x'_{\text{new}}\beta) - x'_{\text{new}}(\widehat{\beta} - \beta), \tag{A.2}$$

where $\rho' f_{\text{new}} - x'_{\text{new}}\beta$ is the model approximation error due to the working linear model (2.4), which is exactly the gap between $\epsilon_{y,\text{new}}$ and e_{new}. The second term $x'_{\text{new}}(\widehat{\beta} - \beta)$ raises from the estimation error of $\hat{\beta}$, which can be further decomposed into

$$x'_{\text{new}}(\widehat{\beta} - \beta) = \{x'_{\text{new}}A_X\beta + x'_{\text{new}}B_X\left(F(\rho - \Lambda'\beta) - U\beta\right)\} + x'_{\text{new}}B_X\epsilon_y, \tag{A.3}$$

where we recall $A_X = (X'X)^{+}X'X - I_p$ and $B_X = (X'X)^{+}X'$.

We focus on the decomposition of the prediction error (A.3) due to the estimation error $(\widehat{\beta} - \beta)$, where the three terms on the right-hand side are:

1. $x'_{\text{new}}B_X\epsilon_y$ associated with the variance, since under Assumption 2 (ii) it is uncorrelated with the other two terms once it is conditional on (x_{new}, X);

2. $x'_{\text{new}}A_X\beta$ associated with the bias arising from the $p > n$ regime. When $p < n$, this term is zero as $A_X = 0$;

3. $x'_{\text{new}}B_X\left(F(\rho - \Lambda'\beta) - U\beta\right)$ associated with the approximation error of the working model. Under a conditional linear model assumption $\mathbb{E}[y|X] = X'\beta$, we do not need

to consider this approximation error. However, we prefer avoiding such a strong assumption and thus must take the approximation error into consideration.

We start with the **variance** $\mathbb{E}\left\{x'_{\text{new}}B_X\epsilon_y\epsilon'_yB'_Xx_{\text{new}}|X\right\}$. By the law of iterated expectations,

$$\begin{aligned}\mathbb{E}\left\{x'_{\text{new}}B_X\epsilon_y\epsilon'_yB'_Xx_{\text{new}}|X\right\} &= \operatorname{tr}\left(\mathbb{E}\left\{\mathbb{E}\left[x'_{\text{new}}B_X\epsilon_y\epsilon'_yB'_Xx_{\text{new}}|X,x_{\text{new}}\right]|X\right\}\right)\\ &\leq \operatorname{tr}\left(\mathbb{E}\left\{B'_X\mathbb{E}\left[x_{\text{new}}x'_{\text{new}}\,|\,X\right]B_X\,|\,X\right\}\right)\left\|\mathbb{E}\left[\epsilon_y\epsilon'_y|X,x_{\text{new}}\right]\right\|\\ &\leq \sigma^2_\epsilon\operatorname{tr}\left((X'X)^+\Sigma_{X_{\text{new}}}\right)\end{aligned}$$

where $\Sigma_{X_{\text{new}}} = \mathbb{E}(x_{\text{new}}x'_{\text{new}}\,|\,X)$. Using Lemma 4 (i),

$$\begin{aligned}\operatorname{tr}\left((X'X)^+\Sigma_{X_{\text{new}}}\right) &= \operatorname{tr}\left((X'X)^+\Lambda\mathbb{E}\left[f_{\text{new}}f'_{\text{new}}|X\right]\Lambda'\right)+\operatorname{tr}\left((X'X)^+\mathbb{E}\left[u_{\text{new}}u'_{\text{new}}|X\right]\right)\\ &\leq C_f\operatorname{tr}\left((X'X)^+\Lambda\Lambda'\right)+C_U\operatorname{tr}\left((X'X)^+\right)\end{aligned}$$

because by the Jensen's inequality

$$\left\|\mathbb{E}\left[f_{\text{new}}f'_{\text{new}}|X\right]\right\| \leq \mathbb{E}\left[\left\|\left[f_{\text{new}}f'_{\text{new}}|F,U\right]\right\||X\right] = \mathbb{E}\left[\left\|\left[f_{\text{new}}f'_{\text{new}}|F\right]\right\||X\right] \leq C_f \tag{A.4}$$

where the equality holds due to the independence between f_{new} and U implied by Assumption 2 (i) and the bound by Assumption 2 (iii), and

$$\left\|\mathbb{E}\left[u_{\text{new}}u'_{\text{new}}|X\right]\right\| = \left\|\mathbb{E}\left[u_{\text{new}}u'_{\text{new}}\right]\right\| = \|\Sigma_u\| \leq C_U$$

under Assumption 2 (i) and the bound by Assumption 1 (ii). Lemma 6 (ii) and (iii) give

$$\begin{aligned}\operatorname{tr}\left((X'X)^+\Lambda\Lambda'\right) &= \operatorname{tr}(\Lambda'B_XB'_X\Lambda) \leq K\left\|B'_X\Lambda\right\|^2 = O_P(1/n);\\ \operatorname{tr}\left((X'X)^+\right) &= \sum_{j=1}^{n}\sigma_j((XX')^{-1}) \leq \frac{n}{\sigma_{\min}(XX')} = O_P(n/p).\end{aligned}$$

The order of the variance is

$$\mathbb{E}\left\{x'_{\text{new}}B_X\epsilon_y\epsilon'_yB'_Xx_{\text{new}}|X\right\} = O_P\left(\frac{1}{n}+\frac{n}{p}\right). \tag{A.5}$$

Next, we move on to the other two terms in the decomposition (A.3). We first work with the **squared bias** $\mathbb{E}\left\{\left[x'_{\text{new}}A_X\beta\right]^2|X\right\}$. Using the notations laid out in the main text, we apply a singular decomposition $X = U_nS_nV'_n$ (notice V_n is a $p\times n$ matrix, whose

columns are the singular vectors of X corresponding to the first n singular values), and thus

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

where V_{-n} is a $p \times (p-n)$ matrix, columns being eigenvectors of the $p \times p$ matrix $X'X$ corresponding to the $p-n$ eigenvalues. Obviously, $V_{-n}' V_n = 0$. We rewrite

$$\begin{aligned}\mathbb{E}\left\{[x_{\text{new}}' A_X \beta]^2 | X\right\} &= \beta' A_X \Sigma_{X_{\text{new}}} A_X \beta = \beta' V_{-n} V_{-n}' \Sigma_{X_{\text{new}}} V_{-n} V_{-n}' \beta \\ &\leq \left\|V_{-n}' \Sigma_{X_{\text{new}}} V_{-n}\right\| \|V_{-n}\|^2 \|\beta\|^2 \leq \left\|V_{-n}' \Sigma_{X_{\text{new}}} V_{-n}\right\| O(\psi_{p,n}^{-1})\end{aligned}$$

where and $A_X = V_n V_n' - I_p = -V_{-n} V_{-n}'$. Again by (A.4):

$$\begin{aligned}\left\|V_{-n}' \Sigma_{X_{\text{new}}} V_{-n}\right\| &\leq \left\|V_{-n}' \Lambda \mathbb{E}(f_{\text{new}} f_{\text{new}}' | X) \Lambda' V_{-n}\right\| + \left\|V_{-n}' \mathbb{E}(u_{\text{new}} u_{\text{new}}' | X) V_{-n}\right\| \\ &\leq \left\|V_{-n}' \Lambda\right\|^2 \|\mathbb{E}(f_{\text{new}} f_{\text{new}}' | X)\| + \|\mathbb{E}(u_{\text{new}} u_{\text{new}}' | X)\| \|V_{-n}\|^2 \\ &\leq C_f \left\|V_{-n}' \Lambda\right\|^2 + C_U.\end{aligned}$$

To deal with $V_{-n}' \Lambda$, let V_K denote a $p \times K$ matrix whose columns are the eigenvectors of $X'X$ corresponding to the first K eigenvalues. Two useful facts follow from standard theories of factor models:

Fact 1 The top K eigenvalues of $X'X/n$ grow at rate $\psi_{p,n}$ as long as $\psi_{p,n} \gg p/n$;

Fact 2 There exists an $K \times K$ matrix H such that $\|H^{-1}\| = O_P(\psi_{p,n}^{1/2})$ and

$$\|V_K - \Lambda H\| = O_P\left(\sqrt{\frac{p}{\psi_{p,n} n}}\right).$$

We decompose

$$V_{-n}' \Lambda = V_{-n}' \left(\Lambda H - V_K + V_K\right) H^{-1} = V_{-n}' \left(\Lambda H - V_K\right) H^{-1}$$

where the second equality is due to the orthogonality $V_{-n}' V_K = 0$. Therefore,

$$\|V_{-n}' \Lambda\| \leq \|V_{-n}' (\Lambda H - V_K)\| \|H^{-1}\| \leq O_P\left(\sqrt{\frac{p}{\psi_{p,n} n}}\right) O_P(\sqrt{\psi_{p,n}}) = O_P\left(\sqrt{\frac{p}{n}}\right). \quad \text{(A.6)}$$

We conclude

$$\beta' A_X \Sigma_{X_{\text{new}}} A_X \beta \leq \left(C_f \left\| V'_{-n} \Lambda \right\|^2 + C_U \right) O(\psi_{p,n}^{-1}) = O_P \left(\frac{p}{n \psi_{p,n}} \right).$$

Next, we deal with the two terms in the **approximation error**. We first work with

$$\begin{aligned} \mathbb{E}\{[x'_{\text{new}} B_X F(\rho - \Lambda'\beta)]^2 | X\} &= \mathbb{E}\left\{(\rho - \Lambda'\beta)' F' B'_X \Sigma_{X_{\text{new}}} B_X F (\rho - \Lambda'\beta) | X\right\} \\ &\leq \|\Sigma_{X_{\text{new}}}\| \, \mathbb{E}\left\{ \|(\rho - \Lambda'\beta)' F' B'_X\|^2 \,|\, X \right\} \end{aligned} \tag{A.7}$$

by the law of iterated expectations. Assumption 2 (i) and (iii) imply

$$\begin{aligned} \|\Sigma_{X_{\text{new}}}\| &= \|\mathbb{E}\left[\Lambda f_{\text{new}} f'_{\text{new}} \Lambda' + u_{\text{new}} u'_{\text{new}} | F\right]\| \\ &\leq \|\Lambda \mathbb{E}\left[f_{\text{new}} f'_{\text{new}} | F\right] \Lambda'\| + \|\mathbb{E}\left[u_{\text{new}} u'_{\text{new}} | F\right]\| \\ &\leq \text{tr}\left(\Lambda'\Lambda\right) C_f + C_U \leq O_P(\psi_{p,n}). \end{aligned} \tag{A.8}$$

Notice

$$\begin{aligned} \mathbb{E}\left\{ \|(\rho - \Lambda'\beta)' F' B'_X\|^2 | X \right\} &\leq \|\rho - \Lambda'\beta\|^2 \, \|B'_X B_X\| \, \mathbb{E}\left\{ \text{tr}\left(FF'\right) | X \right\} \\ &= O_P\left(\frac{n}{\psi_{p,n}^2 p} \right) \mathbb{E}\left\{ \text{tr}\left(\frac{F'F}{n} \right) | X \right\} \end{aligned}$$

given $\|\rho - \Lambda'\beta\| = O\left(\psi_{p,n}^{-1}\right)$ in (A.1) and $\|B'_X B_X\| = O_P\left(1/p\right)$ by Lemma 6 (iii). Because the unconditional mean $\mathbb{E}\left[F'F/n\right] = \Sigma_f$ has all eigenvalues bounded away from 0 and ∞, we invoke Lemma 5 to obtain

$$\mathbb{E}\left\{(\rho - \Lambda'\beta)' F' B'_X B_X F (\rho - \Lambda'\beta) | X\right\} = \left(\frac{n}{\psi_{p,n}^2 p} \right). \tag{A.9}$$

Insert (A.8) and (A.9) into (A.7), we have

$$\mathbb{E}\left\{ \left[x'_{\text{new}} B_X F (\rho - \Lambda'\beta) \right]^2 \,|\, X \right\} = O_P\left(\frac{n}{\psi_{p,n} p} \right).$$

The second term in the approximation error is $x'_{\text{new}} B_X U \beta$. By parallel arguments as for (A.7), we bound its expected square by the law of iterated expectations:

$$\begin{aligned} \mathbb{E}\left[\beta' U' B'_X x_{\text{new}} x'_{\text{new}} B_X U \beta | X\right] &\leq \|\Sigma_{X_{\text{new}}}\| \|\mathbb{E}\left[\beta' U' B'_X B_X U \beta \,|\, X\right] \\ &\leq O_P(\psi_{p,n}) \mathbb{E}\left[\beta' U' B'_X B_X U \beta \,|\, X\right], \end{aligned}$$

where the last inequality follows by (A.8). We use Lemma 4 (i) to bound

$$\begin{aligned}\mathbb{E}\left[\beta' U' B_X' B_X U \beta | X\right] &\leq \left\|(X'X)^+\right\| \mathbb{E}\left[\operatorname{tr}\left(U\beta\beta' U'\right)|X\right] \\ &= \frac{n}{p}\left\|\left(\frac{X'X}{p}\right)^+\right\| \mathbb{E}\left[\beta'(U'U/n)\beta | X\right] \\ &= O_P\left(\frac{n}{p}\right) \cdot O_P(\|\beta\|^2) = O_P\left(\frac{n}{p\psi_{p,n}}\right)\end{aligned}$$

by Assumption 2 (i) about U, $\|(X'X/p)^+)\| = O_P(1)$ by Lemma 6 (iii), and $\|\beta\| = O(\psi_{p,n}^{-1/2})$ from Theorem 1. The above two display expressions give

$$\mathbb{E}\left[\beta' U' B_X' x_{\text{new}} x_{\text{new}}' B_X U \beta | X\right] \leq O_P(\psi_{p,n}) O_P\left(\frac{n}{p}\psi_{p,n}^{-1}\right) = O_P\left(\frac{n}{p}\right).$$

So far, we have handled all terms in the expected square of $x_{\text{new}}'(\widehat{\beta} - \beta)$ as decomposed in (A.3). In (A.2) there is an additional term $(\rho' f_{\text{new}} - x_{\text{new}}'\beta)$, due to the population approximation error of the linear working model (2.4), which is irrelevant to the method of estimation. The expected square this term is bounded by

$$\begin{aligned}\mathbb{E}[(\rho' f_{\text{new}} - x_{\text{new}}'\beta)^2 | X] &= \mathbb{E}[(f_{\text{new}}'(\rho - \Lambda'\beta) - u_{\text{new}}'\beta)^2 | X] \\ &\leq 2\left\|\rho - \Lambda'\beta\right\|^2 \left\|\mathbb{E}[(f_{\text{new}} f_{\text{new}}' | X]\right\| + 2\left\|\beta\right\|^2 \left\|\mathbb{E}[(u_{\text{new}} u_{\text{new}}' | X]\right\| \\ &\leq 2\left\{C_f\left\|\rho - \Lambda'\beta\right\|^2 + C_U\left\|\beta\right\|^2\right\} = O_P(\psi_{p,n}^{-1})\end{aligned}$$

given (A.4) and the order provided in Theorem 1.

To summarize, in this proof we have shown that the order of the variance is $O_P(1/n + n/p)$ and the bias from A_X is $O_P(p/(n\psi_{p,n}))$. These two are the leading terms, and the approximation error of the working model to the true factor model is no larger than them. We therefore conclude the stochastic order. Given the relative magnitude of p, n and $\psi_{p,n}$, both the leading terms are $o_P(1)$ asymptotically, and thus the consistency follows, as stated in **Part (ii)** of the statement. □

A.6 Proof of Theorem 3

Proof. Write $X = (X_I, X_N)$. Here X_N is noise so that $X = F\Lambda' + (U_I, X_N)$ with $\Lambda' = (\Lambda_I', 0)$ and $\|X_N\| = O_P(p)$. Correspondingly, the true coefficient in the working model is

$\beta = (\beta_I', 0)'$. The estimated coefficient can be written as

$$\widehat{\beta}_I = X_I'(XX')^{-1}Y = B_{I,X}Y, \quad \text{where } B_{I,X} = X_I'(X_IX_I' + X_NX_N')^{-1}.$$

Define a $p_0 \times p$ selection matrix $Q_I = (I_{p_0}, 0)$, and then the error in prediction is

$$x_{\text{new},I}'(\widehat{\beta}_I - \beta_I) = x_{\text{new},I}'Q_IA_X\beta + x_{\text{new},I}'B_{I,X}\left(F(\rho - \Lambda_I'\beta_I) - U_I\beta_I\right) + x_{\text{new},I}'B_{I,X}\epsilon_y,$$

where we recall $A_X = X'(XX')^{-1}X - I_p$. The first term is the bias, the second term is the approximation error, and the last term is associated with the variance.

The proofs of the squared expectations of the first two terms resemble those in Theorem 2, and thus omitted for brevity. They share the same rate of convergence $O_P\left(\frac{p}{n\psi_{p,n}} + \frac{n}{p}\right)$ as their counterparts, respectively, in Theorem 2.

The variance, on the other hand, enjoys an improved rate of convergence, as detailed below. We work with the case $p_0 > n$; parallel argument follows for $p_0 \leq n$. Under a finite K, the spiked eigenvalues of $X_I'X_I$ is of order $O_P(n\psi_{p,n})$ whereas the remaining $(n-K)$ non-spiked eigenvalues are $O_P(p_0)$ (Cai et al., 2020), which give

$$\text{tr}\left(X_I'X_I\right) = \sum_{j=1}^{n}\sigma_j\left(X_I'X_I\right) = O_P\left(Kn\psi_{p,n} + (n-K)p_0\right) = O_P\left(np_0\right) \tag{A.10}$$

under the condition $\psi_{p,n} = O(p_0)$ in Assumption 1 (i).

Given Assumption 2 (ii), the varaince is bounded by

$$\begin{aligned}
& \mathbb{E}\left[x_{\text{new},I}'B_{I,X}\epsilon_y\epsilon_y'B_{I,X}'x_{\text{new},I}|X\right] \\
\leq\ & \sigma_\epsilon^2\ \text{tr}\left\{\Sigma_{X_{\text{new}}}B_{I,X}B_{I,X}'\right\} \\
\leq\ & \sigma_\epsilon^2\left\|\mathbb{E}\left[f_{\text{new}}f_{\text{new}}|X\right]\right\|\text{tr}\left\{\Lambda_I'B_{I,X}B_{I,X}'\Lambda_I\right\} + \sigma_\epsilon^2\left\|\mathbb{E}\left[u_{\text{new},I}u_{\text{new},I}'|X\right]\right\|\text{tr}\left\{B_{I,X}B_{I,X}'\right\} \\
\leq\ & \sigma_\epsilon^2C_f\,\text{tr}\left\{\Lambda_I'B_{I,X}B_{I,X}'\Lambda_I\right\} + \sigma_\epsilon^2C_U\,\text{tr}\left\{B_{I,X}B_{I,X}'\right\},
\end{aligned} \tag{A.11}$$

where the last inequality follows by (A.4). The second term of the above expression is bounded, using Lemma 6 (ii) and (A.10), by

$$\begin{aligned}
\text{tr}\left\{B_{I,X}B_{I,X}'\right\} = \text{tr}\left(\left(XX'\right)^{-2}X_I'X_I\right) &\leq \left\|\left(XX'\right)^{-2}\right\|\text{tr}\left(X_I'X_I\right) \\
&= O_P(p^{-2})\cdot O_P\left(np_0\right) = O_P\left(np_0/p^2\right).
\end{aligned} \tag{A.12}$$

The derivation of the order of $\text{tr}\left\{\Lambda_I'B_{I,X}B_{I,X}'\Lambda_I\right\}$ is similar to the proof of Lemma 6

(iii). Notice $\Lambda = (\Lambda_I', 0)'$ be $p \times K$, and $B_{X,I}$ is the first p_0 rows of $B_X = X'(XX')^{-1}$. Given $U = (U_I, X_N)$ which is $n \times p$. Define $\|\cdot\|_F$ as the Frobenius norm of a matrix. In view of $X = F\Lambda' + U$, we can write

$$\begin{aligned} \text{tr}\left\{\Lambda_I' B_{I,X} B_{I,X}' \Lambda_I\right\} &= \|\Lambda_I' B_{I,X}\|_F^2 = \|\Lambda' B_X\|_F^2 = \|(F'F)^{-1}F'(X-U)B_X\|_F^2 \\ &\leq O_P(n^{-1})\|XB_X| + O_P(n^{-1})\|UB_X\| = O_P(n^{-1}) \end{aligned} \tag{A.13}$$

where the last equality follows from the same proof of Lemma 6 (iii). Putting together these pieces, we conclude that the variance is of order $O_P\left(\frac{1}{n} + \frac{np_0}{p^2}\right)$ for forecast based on the denoised estimator. □

B Simulation Comparisons with VAR Forecast

We now conduct simulations to compare with the vector autoregressive model (VAR) with lags up to $L < T$. We generate an N-dimensional dynamic factor model:

$$Y_{t+1} = \bar{\Lambda} f_{t+1} + E_{t+1}, \quad \dim(Y_{t+1}) = p_0,$$

where the objective to forecast is the first time series $y_t = Y_{1,t+1}$. The dynamic factors satisfy a VAR(1) model:

$$f_{t+1} = \Theta_f f_t + e_{f,t+1}.$$

Let $\bar{\Lambda}_1'$ denote the first row of $\bar{\Lambda}$, and $E_{1,t+1}$ denote the first element of E_{t+1}. Then $y_t = \rho' f_t + \epsilon_{y,t}$, which takes the form of (2.1) with $\rho = \Theta_f' \bar{\Lambda}_1$ and $\epsilon_{y,t} = \bar{\Lambda}_1' e_{f,t+1} + E_{1,t+1}$. In addition, we use the one-period lagged variables as the "informative predictors", which also satisfy a factor model: $X_{I,t} = Y_t$, and merge them with additional added pure noise (standard normal variables) to implement the Noise-regularization. We implement the VAR(L) model using Ridgeless regression, which is overfitting when $L \times p_0 > T - L$. Both methods can be compared in the same framework of Ridgeless regression: using the first lag of Y_t as the main informative predictors, and additionally add $(L-1)p_0$ predictors. The only difference is that the Noise-regularization adds i.i.d. independent noise variables whereas the VAR(L) adds additional lagged vectors.

Analogous to the symmetry of the AR model, the VAR model also exhibits a symmetry property on its aggregated eigenvalues of the predictor matrix. To see this, let X_L denote the $(T-L) \times NT$ predictor matrix of VAR(L), and $EigS(L)$ denote the sum of eigenvalues

of $X_L'X_L$. We have the aggregated eigenvalue symmetry for VAR(p) as follows:

$$EigS(L) = EigS(T-L). \quad [4]$$

Thus the aggregated eigenvalues of X_L do not monotonically increase as more lags are added. In contrast, the aggregated eigenvalues of the Noise-regularized regressors linearly increase as more noise are added. The sharp contrast on the aggregated eigenvalues yields a critical implication on the following quantity:

$$\text{tr}(B_I B_I'), \quad B_I = X_I(X_I X_I' + X_N X_N')^{-1}.$$

For both models X_I is the matrix of the first lag Y_{t-1}; but X_N corresponds to additional lags in the VAR model and to the added noise in the Noise-regularized model. The above trace quantity plays a pivotal role in the forecast variance: smaller $\text{tr}(B_I B_I')$ yields a smaller variance.

Figure 10 plots the MSE of the compared methods. In all cases, the MSE of the two methods converge as the number of added lags/noise increase to over 80. Yet, the cross-validation method chooses the optimal number of added noise p_{cv}, at which point the e Noise-regularized forecast outperforms the VAR. The comparison is particularly interesting that the comparison is qualitatively the same under both low serial dependence ($\rho_f = 0.2$) and strong serial dependence ($\rho_f = 0.7$). This seems counterintuitive as the VAR model should outperform the Noise-regularized forecast under stronger serial dependence because the lagged variables carry more predictive informative than pure noise. This however is not the case as shown in Figure 10. We find two interpretations for this: First, for the VAR model the sample size $n = T - L$ decreases as more lags are added, whereas the Noise-regularization keeps the same sample size $n = T - 1$.

Secondly and more fundamentally, the symmetry of the aggregated eigenvalues makes the VAR not benefit from regularizing eigenvalues as the Noise-regularization does. Figure 11 contrasts $EigS$ and $\text{tr}(B_I B_I')$ for the two models in the case $\rho_f = 0.7$. It clearly shows that EigS is symmetric for VAR and is monotonically increasing for Noise-regularization, and that $\text{tr}(B_I B_I')$ of VAR decreases slower in the over-parametrization regime.

[4] To see this, let $\mathbb{Y}_{a:b} = [Y_a, ..., Y_b]$ whose columns are Y_t ranging from $t = a, ..., b$. Then $X_L = [\mathbb{Y}_{L:T-1}', ..., \mathbb{Y}_{1:T-L}']$, and X_{T-L} is rearranging elements of X_L, thus they have the same Frobenius norm: $\|X_L\|_F = \|X_{T-L}\|_F$. So $EigS(L) = \text{tr}(X_L X_L') = \|X_L\|_F^2 = \|X_{T-L}\|_F^2 = Eig(T-L)$.

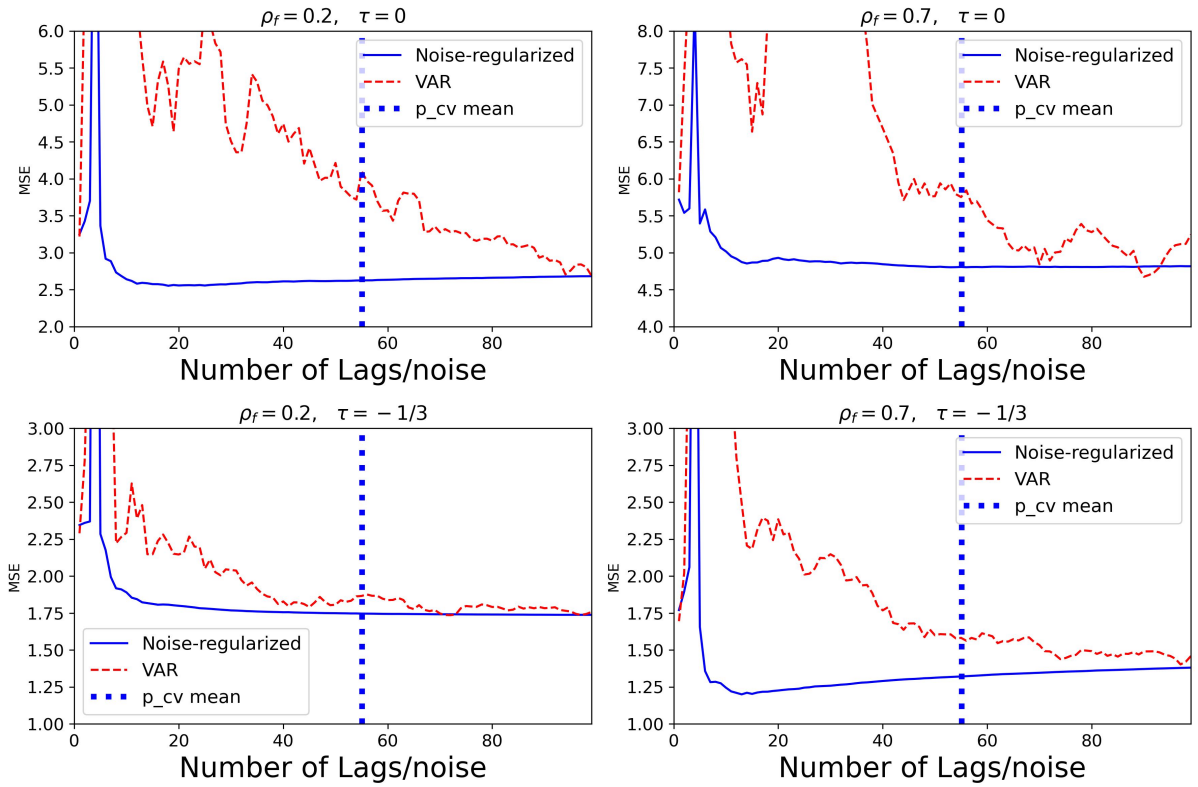


Figure 10: Comparison with VAR

Predictive MSE averaged from 50 replications as the number of predictors p increases. The sample size is fixed $n = 100$. The Noise-regularized method uses the first p_0 dimensional lagged vector as the informative predictors, and additionally adds $(L-1) \times p_0$ noise, whereas VAR(L) adds lags up to Y_{t-L}. The vertical blue dashed line p_{cv} in the last panel indicates the averaged L chosen by the leave-one-out cross validation.

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.

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

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

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.

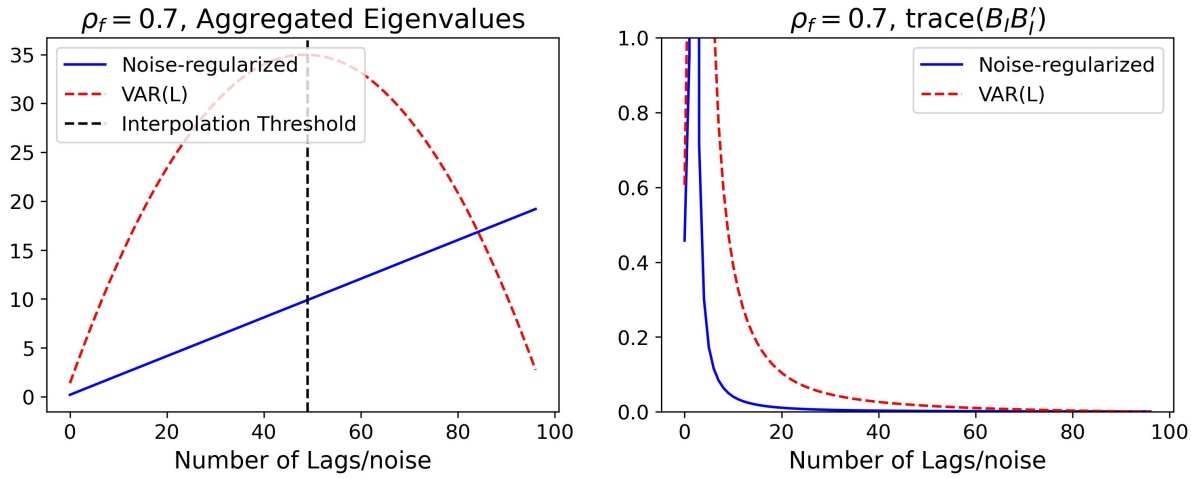


Figure 11: Aggregated Eigenvalues Comparison with VAR

Notes: the left panel plots the sum of eigenvalues of $X_N'X_N$ where X_N is the predictor matrix for VAR(L) and is pure noise for the Noise-regularization. The right panel plots tr($B_I B_I'$) for the two models.

Bartlett, P. L., P. M. Long, G. Lugosi, and A. Tsigler (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences 117*(48), 30063–30070.

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. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

Cai, T. T., X. Han, and G. Pan (2020). Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *The Annals of Statistics 48*(3), 1255–1280.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies 21*(4), 1509–1531.

Carrasco, M. and B. Rossi (2016). In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics 34*(3), 313–338.

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

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.

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

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.

Doz, C., D. Giannone, and L. Reichlin (2012). A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of economics and statistics 94*(4), 1014–1024.

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).

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.

Hirano, K. and J. H. Wright (2017). Forecasting with model uncertainty: Representations and risk reduction. *Econometrica 85*(2), 617–643.

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

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

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*.

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

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*.

Meng, X., Y. Cao, and W. Wang (2025). Estimation of out-of-sample sharpe ratio for high dimensional portfolio optimization. *Journal of the American Statistical Association*, 1–13.

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

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*(2), 244–258.

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*.

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

Stock, J. and M. Watson (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*(2), 147–162.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning 8*(1-2), 1–230.

von Neumann, J. (1937). Some matrix inequalities and metrization of matrix space. *Tomsk Univ. Rev 1*, 286–300.

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.

Supplemental Appendix

A.1 Implementation

Algorithm 2 (Rao-Blackwell Leave-One-Out (LOO) for selecting p)**.** *For each candidate total dimension p on a grid $p_1 < \cdots < p_J$:*

1. *For $\ell = 1, \ldots, L$, draw $Z^{(\ell)} \sim N(0, I_{p-p_0})$ independently and form $X^{(\ell)}_{\text{aug}} = [\mathbf{1}_n,\ X_I,\ Z^{(\ell)}]$. Compute*
$$e^{(\ell)}(p) = \text{Diag}\big(G^{(\ell)}\big)^{-1} G^{(\ell)} Y, \qquad G^{(\ell)} = \big(X^{(\ell)}_{\text{aug}} X^{(\ell)\prime}_{\text{aug}}\big)^{+}.$$
2. *Average the LOO residual* vectors *across draws*:
$$e^{RB}(p) = \frac{1}{L} \sum_{\ell=1}^{L} e^{(\ell)}(p).$$
3. *Set* $\text{LOO}(p) = \|e^{RB}(p)\|^2$.

Select $p_{cv} = \arg\min_p \text{LOO}(p)$.