Adding Noise to Reduce Noise: A Counter-Intuitive Approach to Stock Return Prediction

Shingo Goto
University of Rhode Island

Toru Yamada
Nomura Asset Management

May 2025

Abstract

This paper presents an unconventional and counter-intuitive approach to improving out-of-sample cross-sectional return prediction: deliberately introducing random noise into linear regression models. Traditional forecasting methods, such as Ridge, Lasso, and Partial Least Squares (PLS), often fail in high-dimensional settings where the number of predictors exceeds the sample size, frequently yielding negative out-of-sample R^2. We propose two complementary techniques—noise injection and noise augmentation—that exploit implicit regularization to stabilize coefficient estimates and enhance predictive performance. Grounded in machine learning insights on double descent, our framework shows that adding noise acts as an implicit form of Ridge regularization. Empirical results demonstrate that these noise-based methods consistently outperform conventional regularization techniques, reinforcing the emerging preference for dense over sparse modeling. Our findings reveal a paradoxical yet powerful insight: strategically adding noise improves out-of-sample prediction, offering a new tool for high-dimensional financial forecasting.

Adding Noise to Reduce Noise: A Counter-Intuitive Approach to Stock Return Prediction

Abstract

This paper presents an unconventional and counter-intuitive approach to improving out-of-sample cross-sectional return prediction: deliberately introducing random noise into linear regression models. Traditional forecasting methods, such as Ridge, Lasso, and Partial Least Squares (PLS), often fail in high-dimensional settings where the number of predictors exceeds the sample size, frequently yielding negative out-of-sample R^2. We propose two complementary techniques—noise injection and noise augmentation—that exploit implicit regularization to stabilize coefficient estimates and enhance predictive performance. Grounded in machine learning insights on double descent, our framework shows that adding noise acts as an implicit form of Ridge regularization. Empirical results demonstrate that these noise-based methods consistently outperform conventional regularization techniques, reinforcing the emerging preference for dense over sparse modeling. Our findings reveal a paradoxical yet powerful insight: strategically adding noise improves out-of-sample prediction, offering a new tool for high-dimensional financial forecasting.

Keywords: Cross-section of stock returns; Return predictability; Out-of-sample; High-dimensional; Regularization; Ridge; Overfittig; Sparse vs. dense modeling; Minimum norm least squares; Machine learning

JEL Classification: C13, C31, G11, G17

1 Introduction

Cross-sectional prediction of individual stock returns out-of-sample remains a major challenge in empirical finance.[1] Despite researchers identifying hundreds of firm characteristics ("signals") that predict returns (e.g., Green, Hand, and Zhang, 2017; Hou, Xue, and Zhang, 2020; Gu, Kelly, and Xiu, 2020)—many validated through rigorous replication tests (Chen and Zimmermann, 2022)—achieving robust out-of-sample predictability remains elusive. This difficulty is particularly pronounced in linear cross-sectional regression models when the number of signals ("model size" p) is large relative to the number of stocks ("sample size," n), or when the in-sample training period (T) is short. Conventional ordinary least squares (OLS) estimation often yields negative out-of-sample R^2 values, even when using well-established predictors, underscoring the persistent challenge.

Traditional solutions rely on shrinkage and regularization methods—such as Ridge, Lasso, Partial Least Squares (PLS), and related techniques—to mitigate overfitting by introducing bias to reduce prediction variance. However, even these techniques frequently fail to deliver positive out-of-sample R^2.

We propose an unconventional and counter-intuitive strategy: deliberately introducing random noise during model training. Specifically, we investigate two complementary techniques:

1. **Noise Injection**: Adding random noise directly to existing signals.

2. **Noise Augmentation**: Expanding the set of signals with purely random noise variables.

While introducing noise may seem counterproductive, prior research suggests that strategically applied noise can enhance model generalization. For example, Matsuoka (1992), Webb (1993), and Bishop (1995) demonstrated that training neural networks with noise-corrupted inputs improves

[1] As in Goyal and Welch (2008), "out-of-sample" prediction refers to forecasting returns at time $t + h$ ($h > 0$; "testing sample") using only information available at time t ("training sample"), as the training window moves forward with t.

generalization. More recently, Liao, Ma, Neuhierl, and Shi (2024) found that augmenting signals with random noise can enhance out-of-sample predictability across various economic variables, including S&P firms' earnings, the US equity premium, employment rates, and GDP growth—a phenomenon they termed the "blessing of overfitting." They attribute this effect to the "double descent" or "benign overfitting" phenomenon identified in the machine learning and statistics literature (Belkin, Hsu, Ma, and Mandal,2019; Bartlett, Long, Lugosi, and Tsigler, 2020; Hastie, Montanari, Rosset, and Tibshirani, 2022).

This study builds upon prior work in two key ways. First, we apply both noise injection and noise augmentation within linear cross-sectional regressions using the minimum norm OLS estimator (also known as the "Ridgeless" estimator) to handle high-dimensional, over-parameterized ($p > n$) settings where conventional OLS is infeasible. Second, we focus specifically on the cross-sectional predictability of US individual stock returns—a notoriously difficult forecasting problem not addressed by Liao et al. (2024).

We interpret the benefits of noise through the lens of "implicit Ridge regularization," as explored in recent literature.[2] This mechanism may help mitigate overfitting and stabilize regression coefficients. Specifically, we identify two ways in which random noise enhances predictive accuracy:

1. **Implicit Ridge Regularization via Noise Injection:** Injecting noise into existing signals approximates explicit Ridge regularization in both under-parameterized ($p < n$) and over-parameterized ($p > n$) settings.

2. **Implicit Ridge Regularization via Noise Augmentation:** Adding random noise variables creates an over-parameterized setting ($p > n$), allowing the minimum norm OLS estimator to trigger a "double descent" effect.

[2] See Bartlett et al. (2020), Derezinski, Liang, and Mahoney (2020), Hastie et al. (2022), Kobak, Lomond, and Sanchez (2020), Wu and Xu (2020), and Richards, Mourtada, and Rosasco (2022) for discussions on implicit regularization effects in over-parameterized models. An earlier version of Zhang, Bengio, Hardt, Recht, and Vinyals (2021) was among the first to propose implicit regularization as a key factor in the success of deep learning models.

Empirical results show that in scenarios where conventional methods (OLS, Ridge, Lasso, and PLS) yield negative out-of-sample R^2, our noise-based approaches significantly improve performance. Both noise injection and noise augmentation consistently achieve positive out-of-sample R^2 on average. More intensive noise injection or extensive noise augmentation (e.g., adding $5,000$ random noise variables) leads to greater improvements in out-of-sample R^2.

Across various settings, noise-based methods outperform traditional regularization techniques like Lasso, Ridge, and PLS. The paradoxical gain arises because noise induces implicit regularization, particularly when applying minimum norm OLS to over-parameterized models. Injecting or augmenting noise pushes the estimator toward a low-norm solution, reducing sensitivity to idiosyncratic training sample features and thereby improving out-of-sample prediction.

This study makes four contributions. First, we introduce noise-based regularization—via noise injection and noise augmentation—as a simple, flexible, and effective method for improving out-of-sample predictability in cross-sectional return forecasting.

Second, we provide an intuitive theoretical interpretation of how noise functions as an implicit form of Ridge regularization in minimum norm OLS, shrinking coefficients toward zero and enhancing model stability.

Third, we demonstrate practical benefits of noise-based regularization, showing that adding noise—despite being uninformative by construction—systematically enhances out-of-sample predictive accuracy.

Finally, our study contributes to the ongoing "dense" versus "sparse" modeling debate by providing evidence that retaining all signals in cross-sectional regressions, alongside implicit regularization via noise, achieves superior predictive accuracy compared to reducing predictors though methods like Lasso and PLS. While traditional finance research has favored sparse modeling approaches focusing on selecting a small number of factors from many possible signals,[3] emerging

[3]Methods such as Lasso, PLS, PCR (Principal Component Regression), IPCA (Instrumented Principal Component

evidence supports dense models, which often outperform sparse approaches in economic forecasting (e.g., Giannione, Lenza, and Primiceri, 2021; Kelly, Malamund, and Zhou, 2024; Shen and Xiu, 2024).

Our findings align with this emerging consensus: retaining all available signals and harnessing implicit regularization through random noise yields higher predictive accuracy than conventional dimension-reduction methods.

2 Setups and Motivations

2.1 Linear Regression Model for Cross-Sectional Return Prediction

This study focuses on the following linear regression model for cross-sectional return prediction:

$$y_{i,t} = x_{i,t-1}^{\intercal}\beta + \varepsilon_{i,t}, \qquad i = 1, ..., n,$$

where $y_{i,t}$ is the firm i's stock return in period t, and $x_{i,t-1}$ is the vector of p_x known predictors (signals) observed at $t-1$. $\varepsilon_{i,t}$ represents the residual with $E[\varepsilon_{i,t}] = 0$. Our discussion and empirical implementation demean individual stock returns and all signal values before each cross-sectional regression, eliminating the need for an intercept term. We stack regressions for individual firms $i = 1, ..., n$ as:

$$Y_t = X_{t-1}\beta + \epsilon_t, \tag{1}$$

where $Y_t := [y_{1,t}, ..., y_{n,t}]^{\intercal}$ and $\epsilon_t := [\varepsilon_{1,t}, ..., \varepsilon_{n,t}]^{\intercal}$ are n-vectors and $X_{t-1} := [x_{1,t}, ..., x_{n,t}]^{\intercal}$ is an $n \times p_x$ signal matrix. We use the regression model (1) as the foundation for the following discussion.

Analysis) (Kelly, Pruitt, and Su., 2020) exemplify this sparse modeling paradigm.

2.2 Minimum Norm (and Ridgeless) Ordinary Least Squares (OLS) Estimator

The classical regression estimation problem requires the sample size to be much larger than the model size, i.e., $p_x < n$. Most existing empirical finance research focuses on this classical under-parameterized regime. However, modern prediction models must consider the high-dimensional over-parameterized ($p_x < n$) setting, as the number of potential predictors can easily exceed the number of available samples. Fortunately, we can generalize the conventional OLS estimator to the over-parameterized ($p_x < n$) regime using a generalized inverse, such as the Moore-Penrose universe (e.g., Ben-Israel and Greville, 2003; Campbell and Meyer, 1979).

Setting statistics aside and focusing on linear algebra for a moment, let us consider the problem of finding solutions b for β (a p_x-vector) that make $Y_t - X_{t-1}\beta$ as small as possible. Specifically, we seek to minimizes the l_2 norm,

$$||Y_t - X_{t-1}\beta||_2 := \sqrt{(Y_t - X_{t-1}\beta)^\intercal (Y_t - X_{t-1}\beta)}.$$

A p_x-vector $\hat{b}$ is called an ordinary least squares (OLS) solution if it minimizes $||Y_t - X_{t-1}\beta||_2$. In the classical under-parameterized regime, we can find a unique OLS solution, which is the conventional OLS estimator. In the over-parameterized regime, however, infinitely many OLS solutions $\hat{b}$ achieve $Y_t = X_{t-1}\hat{b}$, that is, $\hat{b}$ interpolates the data perfectly.

Now, suppose we want to find a solution that minimizes $||\beta||_2 := \sqrt{\beta^\intercal \beta}$ among all OLS solutions that minimize $||Y_t - X_{t-1}\beta||_2$. We can then uniquely define $\hat{\beta}$ (another p_x-vector) as the "minimum norm OLS" estimator if $||\hat{\beta}||_2 < ||\hat{b}||_2$ for all other OLS solutions $\hat{b} \neq \hat{\beta}$. The unique minimum norm OLS estimator (interpolator), or the "pseudo-OLS" estimator (Liao et al. 2024), is

$$\hat{\beta} = X_{t-1}^{+} Y_t,$$

where X_{t-1}^{+} is the Moore-Penrose inverse of X_{t-1} (e.g., Campbell and Meyer, 1979, Theorem 2.1.1). We can view the minimum norm OLS estimator $\hat{\beta} = X_{t-1}^{\dagger} Y_t$ as a generalization of the conventional OLS estimator. It is well-defined for both classical under-parameterized and modern over-parameterized regimes.

- When X_{t-1} has full column rank ($p_x \leq n$), $X_{t-1}^{+} = (X_{t-1}^{\intercal} X_{t-1})^{-1} X_{t-1}^{\intercal}$ and $\hat{\beta} = (X_{t-1}^{\intercal} X_{t-1})^{-1} X_{t-1}^{\intercal} Y_t$. This is the conventional OLS estimator.
- When X_{t-1} has full row rank ($p_x \geq n$), $X_{t-1}^{+} = X_{t-1}^{\intercal}(X_{t-1} X_{t-1}^{\intercal})^{-1}$ and $\hat{\beta} = X_{t-1}^{\intercal}(X_{t-1} X_{t-1}^{\intercal})^{-1} Y_t$. This minimum norm OLS estimator interpolates the training data perfectly as $||Y_t - X_{t-1}\hat{\beta}|| = 0$. The literature on double descent (benign overfitting) focuses on the generalization capability (out-of-sample predictive accuracy) of this particular minimum norm solution.

$\hat{\beta} = X_{t-1}^{+} Y$ is also called the "Ridgeless" estimator (e.g., Hastie et al. 2022). This follows from a fundamental property of the Moore-Penrose inverse (e.g., Bernstein, 2005, Fact 6.3.10):

$$X_{t-1}^{+} = \lim_{\lambda \downarrow 0} X_{t-1}^{\intercal}(X_{t-1} X_{t-1}^{\intercal} + \lambda I_n)^{-1} = \lim_{\lambda \downarrow 0} (X_{t-1}^{\intercal} X_{t-1} + \lambda I_{p_x})^{-1} X_{t-1}^{\intercal},$$

where I_n and I_{p_x} are $n \times n$ and $p_x \times p_x$ identity matrices.

It follows that the minimum norm OLS estimator provides the Ridgeless (zero regularization) limit of the Ridge estimator, $\hat{\beta} = X_{t-1}^{+} Y = \lim_{\lambda \downarrow 0} \hat{\beta}_\lambda$, where $\hat{\beta}_\lambda = (X_{t-1}^{\intercal} X_{t-1} + \lambda I)^{-1} X_{t-1}^{\intercal}$ is the Ridge estimator (Hoerl and Kennard, 1970) that minimizes $||Y_t - X_{t-1}\beta||_2^2 + \lambda ||\beta||_2^2$, a special case of Tikhonov regularization. The Ridge penalty $\lambda > 0$ dictates the degree of regularization (shrinkage). In the bias-variance tradeoff, bias increases with λ while variance decreases with λ.

2.3 Double Descent (Benign Overfitting)

Classical textbooks on statistical learning taught us that "though it is possible to perfectly fit the training data in the high-dimensional setting, the resulting linear model will perform extremely poorly on an independent test set, and therefore does not constitute a useful model" (James, Witten, Hastie, Tibshirani, 2013, p.240). However, the recent success of highly over-parameterized machine learning models, such as deep neural networks, has drastically changed our view of over-parameterized models in high dimensional-settings. Over-parameterized models that perfectly fit (interpolate) the training data can actually offer superior out-of-sample predictive performance despite the overfitting concern. This surprising phenomenon is known as "benign overfitting" or "double descent." The double descent phenomenon is also prevalent in classical linear regression models implemented with minimum norm OLS estimation.

Adapted from Belkin et al. (2019), Figure 1 depicts the double descent pattern. The prediction risk (the vertical axis) refers to the mean-squared prediction error.

Figure 1

2.3.1 "Classical" Under-parametrized Regime ($p_x < n$)

Figure 1(a) illustrates the classical under-parameterized regime, where the model size is less than the sample size, $p_x < n$. The dotted curve shows that the in-sample prediction risk decreases as the model size increases. However, the solid curve shows that the out-of-sample prediction risk initially diaereses until it reaches a "sweet spot" but then increases as the model size increases—this is the classical overfitting problem.

Figure 1(a) illustrates the bias-variance tradeoff—a fundamental concept in classical statistical learning (e.g., Hastie, Tibshirani, and Friedman, 2009). Larger model size decreases bias but increases error variance. Once the model size passes a certain threshold (sweet spot), its prediction risk increases with the model size. When the model size reaches the sample size, the regression model fits the training data perfectly in-sample (in the training sample) but performs poorly out-of-sample. Hastie et al., (2009, p.221) note: “a model with zero training error is overfit to the training data and will typically generalize poorly.”

2.3.2 “Modern” Over-parameterized Regime ($p_x > n$)

Figure 1(b) shows what happens when the model size exceeds the sample size, i.e., $p_x > n$. The model perfectly interpolates the training data ($y_t = X_{t-1}\beta$), resulting in zero in-sample error. Out-of-sample prediction risk is initially high when the model size is close to the sample size. Surprisingly, as the model size increases even further, out-of-sample prediction risk starts decreasing again—this is the “double descent” curve proposed by Belkin et al. (2019), underpinning the success of modern over-parameterized deep learning models.

In proposing the double descent risk curve, Belkin et al. (2019) assumes the minimum norm OLS solution in the training sample, among many other solutions that interpolate the data perfectly. The use of minimum norm OLS solutions is crucial for exploiting the double descent effect. Other interpolation solutions would not yield similar double descent effects. For instance, in describing Belkin et al.’s (2019) double descent risk curve, Strang, Hiranabe, and Fernandes (2024, p. 12) note: “among many solutions, a good one is chosen: the ‘minimum norm least squares solution.’ The last fact justified the success of ‘deep learning,’ with thousands or even millions of free parameters. Ordinary Lagrange interpolation would be a disaster.”

2.4 Noise and Implicit Ridge Regularization

2.4.1 Noise Injection and Implicit Ridge Regularization

Injecting noise into existing signals is equivalent to introducing implicit Ridge regularization. Researchers have long recognized that injecting noise into inputs is equivalent to Tikhonov (Ridge) regularization (e.g., Matsuoka, 1992; Webb, 1994; Bishop, 1995). Consider the following noise injection: $X_{I,t-1} = X_{t-1} + aZ_{I,t-1}$, where $Z_{I,t-1}$ is an $n \times p_x$ matrix of independent random noise and $a > 0$ controls the intensity of noise injection. Since our empirical implementation standardizes signals to zero mean and unit variance, $a > 1$ indicates that the inject noise has higher variance than the signals. Replacing X_{t-1} with $X_{I,t-1}$, the regression (1) model becomes

$$Y_t = X_{I,t-1}\beta_I + \epsilon_{I,t}.$$

Now consider the minimum norm OLS estimator $\hat{\beta}_I = X^+_{I,t-1}Y_t$. In the classical under-parameterized regime, $\hat{\beta}_I$ equals a conventional OLS estimator but approximates a Ridge estimator:

$$\begin{aligned}\hat{\beta}_I &= (X^\intercal_{I,t-1}X_{I,t-1})^{-1}X^\intercal_{I,t-1}Y_t \\ &\xrightarrow{p} (X^\intercal_{t-1}X_{t-1} + a^2nI_{p_x})^{-1}X^\intercal_{t-1}Y_t.\end{aligned}$$

In the high-dimensional over-parameterized regime, we obtain an estimator that involves Ridge regularization:

$$\begin{aligned}\hat{\beta}_I &= X^\intercal_{I,t-1}(X_{I,t-1}X^\intercal_{I,t-1})^{-1}Y_t \\ &\xrightarrow{p} X^\intercal_{t-1}(X_{t-1}X^\intercal_{t-1} + a^2p_xI_n)^{-1}Y_t.\end{aligned}$$

In both regimes, we can view $\hat{\beta}_I$ as an implicit Ridge estimator approximating an explicit Ridge estimator where the squared noise intensity a^2 controls the degree of regularization.

Hoerl and Kennard (1970) proved in the classical $n > p_x$ regime that, in explicit Ridge regression, there always exists a positive regularization parameter that yields lower prediction risk than OLS estimation. By invoking an implicit Ridge (Tikhonov) regularization, noise injection can enhance the out-of-sample predictability of a linear regression (1) by approximating explicit Ridge regression.

2.4.2 Noise Augmentation and Implicit Ridge Regularization

Augmenting existing signals with noise variables is equivalent to implicit Ridge regularization in the over-parameterized regime. Even when the number of signals is less than the number of observations, $p_x < n$, we can augment the signal matrix with additional noise variables to transform the regression model into an over-parameterized regime. Let $X_{A,t-1} := [X_{t-1}, Z_{t-1}]$ be a noise-augmented signal matrix, where Z_{t-1} is an $n \times p_z$ matrix of random noise, ensuring $(p_x + p_z) > n$. In the over-parameterization regime, the regression becomes and interpolation problem:

$$Y_t = X_{t-1}\beta_A + Z_{t-1}\gamma_A. \tag{2}$$

To obtain the minimum norm OLS estimator for (2), we solve:

$$\min \left\| \begin{matrix} \beta_A \\ \gamma_A \end{matrix} \right\|_2 \text{ subject to } Y_t = X_{t-1}\beta_A + Z_{t-1}\gamma_A.$$

The solution is

$$\begin{bmatrix} \hat{\beta}_A \\ \hat{\gamma}_A \end{bmatrix} = X^{+}_{A,t-1} Y_t.$$

When $n < (p_x + p_z)$, the minimum norm OLS estimator $\hat{\beta}_A$ is

$$\begin{aligned} \hat{\beta}_A &= X_{t-1}^{\intercal} \left(X_{t-1} X_{t-1}^{\intercal} + Z_{t-1} Z_{t-1}^{\intercal} \right)^{-1} Y_t \\ &\xrightarrow{p} X_{t-1}^{\intercal} \left(X_{t-1} X_{t-1}^{\intercal} + p_z I_n \right)^{-1} Y_t. \end{aligned} \tag{3}$$

With noise augmentation, the minimum norm OLS estimator $\hat{\beta}_A$ approximates a Ridge estimator, where including more noise variables (increasing p_z) strengthens the regularization effect. However, estimator $\hat{\beta}_A$ differs from the conventional Ridge estimator because it operates within a model containing a much larger number of signals—including numerous additional noise variables.

2.5 Relationship between Two Noise-Based Regularization Strategies: Noise Injection and Noise Augmentation

Practical applications of noise injection and noise augmentation often involve a situation where the number of stocks exceeds the number of signals ($p_x < n$) but we augment the signals with numerous noise variables to create an over-parameterized setting, $p_x + p_z > n$.

This section discusses how the OLS estimator with noise injection ($\hat{\beta}_I$) and the minimum norm OLS estimator with noise augmentation ($\hat{\beta}_A$) are related. To examine the approximate relationship between $\hat{\beta}_I$ and $\hat{\beta}_A$, let us temporarily drop the time subscript $t-1$ and write

$$\hat{\beta}_I \approx (X^{\intercal} X + a^2 n I_{p_x})^{-1} X^{\intercal} Y.$$

$$\hat{\beta}_A \approx X^{\intercal} \left(X X^{\intercal} + p_z I_n \right)^{-1} Y.$$

We assume X has a full column rank, $rank\,(X) = p_{x.}$and $p_x + p_z > n$.

Consider a singular value decomposition of X,

$$X = [U_x, U_0] \begin{bmatrix} S \\ 0_{(n-p_x)\times p_x} \end{bmatrix} V^\intercal$$

where U_x is an $n \times p_x$ orthogonal matrix satisfying $U_x^\intercal U_x = I_{p_x}$. U_0 is an $n \times (n - p_x)$ orthogonal matrix with $U_0^\intercal U_0 = I_{n-p_x}$. U_x and U_0 are orthogonal: $U_x^\intercal U_0 = 0_{p_x\times(n-p_x)}$ and $U_x^\intercal U_0 = 0_{(n-p_x)\times p_x}$. V is a $p_x \times p_x$ orthogonal matrix with $V^\intercal V = I_{p_x}$. $S := diag(s_1, ..., s_{p_x})$ is an $p_x \times p_x$ diagonal matrix of singular values, $s_1 \geq ... \geq s_{p_x} > 0$. Since none of the signals are redundant, all p_x singular values are positive.

Then, we can express the fitted values $X\hat{\beta}_I$ and $X\hat{\beta}_A$ as follows:

$$\begin{aligned}
X\hat{\beta}_I &\approx U_x S V^\intercal (V S^2 V^\intercal + a^2 n I_{p_x})^{-1} V S U_x^\intercal Y \\
&= U_x S \left(S^2 + a^2 n I_{p_x}\right)^{-1} S U_x^\intercal Y \\
&= U_x \cdot diag\left(\frac{s_1^2}{s_1^2 + a^2 n}, ..., \frac{s_{p_x}^2}{s_{p_x}^2 + a^2 n}\right) \cdot U_x^\intercal Y
\end{aligned}$$

$$\begin{aligned}
X\hat{\beta}_A &\approx U_x S V^\intercal V S U_x^\intercal \cdot [U_x, U_0] \begin{bmatrix} \left(S^2 + p_z I_{p_x}\right)^{-1} & 0 \\ 0 & (p_z I_{n-p_x})^{-1} \end{bmatrix} \begin{bmatrix} U_x^\intercal \\ U_0^\intercal \end{bmatrix} Y \\
&= U_x S^2 \left(S^2 + p_z I_{p_x}\right)^{-1} U_x^\intercal Y \\
&= U_x \cdot diag\left(\frac{s_1^2}{s_1^2 + p_z}, ..., \frac{s_{p_x}^2}{s_{p_x}^2 + p_z}\right) \cdot U_x^\intercal Y
\end{aligned}$$

This shows that the two model fitted-values, $X\hat{\beta}_I$ and $X\hat{\beta}_A$, are approximately equal when $a^2 n = p_z$ or $a = \sqrt{p_z/n}$. The intensity parameter a is approximately equivalent to the ratio of the number of additional random variables to the sample size.

For example, when we include $p_z = 5,000$ (or $10,000$) noise variables in the noise augmentation strategy, $\hat{\beta}_A$ becomes approximately equal to $\hat{\beta}_I$ from a noise injection strategy when we set the noise integration intensity at $a = \sqrt{5,000/1000} \fallingdotseq 2.24$ (or $a = \sqrt{10,000/1000} \fallingdotseq 3.16$). Alternatively, choosing a noise injection intensity of $a = 2$ (or 3) is approximately equivalent to including $p_z = 2^2 \times 1,000 = 4,000$ (or $3^2 \times 1,000 = 9,000$) noise variables in the data augmentation strategy

2.6 Minimum Norm OLS, Implicit Ridge Regularization, and Double Descent: A Discussion

The double descent (or benign overfitting) phenomenon has attracted significant attention in the machine learning and statistics community. Many studies have theoretically identified conditions under which extreme overfitting can improve a model's generalization (out-of-sample prediction) performance using random matrix theories. One key discussion topic is whether explicit Ridge regularization helps improve out-of-sample prediction of over-parameterized models beyond the minimum norm (and Ridgeless) OLS estimation.

Notably, Kobak et al. (2020), Wu and Xu (2020), and Richards et al. (2022), among others, emphasize the crucial role of implicit Ridge regularization in driving double descent. While this discussion does not directly involve noise injection or noise augmentation, the topic remains relevant because noise augmentation may effectively exploit descent through implicit Ridge regularization. This view aligns with Kobak et al. (2020, Abstract), who show that "augmenting any linear model with random covariates and using minimum norm estimator is asymptotically equivalent to adding the ridge penalty." Wu and Xu (2020) suggest that over-parameterization may not result in overfitting due to the implicit regularization of the minimum norm OLS estimator.

In high-dimensional setting ($p > n$), the signal matrix may contain weak or redundant signals that act similarly to random noise variables. For example, suppose we partition a signal matrix X_{t-1} into two sub-matrices, $X_{t-1} = [X_{s,t-1}, X_{w,t-1}]$, where $X_{s,t-1}$ contains p_s strong signals while $X_{w,t-1}$ contains p_w weak (or redundant) signals that provide no incremental information about Y_t beyond what's in $X_{s,t-1}$. The number of both strong and weak signals exceeds the number of samples, $p_s + p_w > n$, so we are considering an interpolation problem: $Y_t = X_{s,t-1}\beta_s + X_{w,t-1}\beta_w$.

For simplicity, suppose $X_{w,t-1}$ are orthogonal to the strong signals $X_{1,t-1}$. Then, the weak (or redundant) signal matrix $X_{w,t-1}$ resembles the noise matrix Z_{t-1} in the interpolation model with random noise (2). It follows that the minimum-norm (and Ridgeless) OLS solution for the coefficients on $X_{s,t-1}$, $\hat{\beta}_s$, takes the following form similar to $\hat{\beta}_A$ in (3):

$$\begin{aligned}\hat{\beta}_s &= X_{s,t-1}^{\intercal}\left(X_{s,t-1}X_{s,t-1}^{\intercal} + X_{w,t-1}X_{w,t-1}^{\intercal}\right)^{-1} Y_t \\ &\xrightarrow{p} X_{s,t-1}^{\intercal}\left(X_{s,t-1}X_{s,t-1}^{\intercal} + p_w\Sigma_w\right)^{-1} Y_t\end{aligned} \tag{4}$$

where $\Sigma_w := E\left[X_{w,t-1}X_{w,t-1}^{\intercal}\right]$ is an $n \times n$ weight matrix. If Σ_w is approximately diagonal (i.e., when weak signals in $X_{w,t-1}$ are not strongly correlated), the Ridgeless OLS estimator $\hat{\beta}_s$ in (4) involves implicit regularization approximating an weighted Ridge (Tikhonov) regularization with weight matrix Σ_w. The number of weak signals p_w affects the strength of regularization.

Despite being "Ridgeless," a minimum norm OLS estimator may embody implicit Ridge regularization when the number of weak (or redundant) signals is large and weak signals are not strongly correlated. When the implicit regularization effect in the minimum norm OLS estimator is already strong, explicit Ridge (Tikhonov) regularization may not help—or may even harm—the model's out-of-sample predictive accuracy.[4]

3 Methodology

3.1 Data and Sample

We draw on well-established return-predicting signals compiled by Chen and Zimmermann (2022), who replicated 319 cross-sectional return predictors (signals) from the academic literature. Their

[4]Even negative regularization may become optimal in explicit Ridge (Tikhonov) regularization. See Kobak et al. (2020), Wu and Xu (2020), and Richards et al. (2022) for more rigorous discussions.

replication generally confirms the original studies' reported t-values. Of these, Andrew Chen publicly released firm-level signal values for 209 signals via openassetpricing.com,[5] which we access from the October 2024 release. To these, we add three additional signals constructed directly from CRSP, bringing the total number of signals to 212.

We collect signal values annually at the end of June from 1963 to 2022 and corresponding one–year ahead stock returns from June 1964 to June 2023. Because our rolling training windows span up to 10 years, our out-of-sample testing begins in July 1973 and ends in June 2023—covering a 50 year period.

Not all firms have valid data for every signal in each year. In any given year, we exclude signals with fewer than 25% valid observations across firms. This results in between 140 and 182 usable signals, with an average of 168.7, per year. Except for binary indicators, all signals are winsorized at the 2nd and 98th percentiles and standardized to have zero mean and unit variance in each annual cross section.

Our sample consists of US publicly-traded firms excluding financials (SIC codes 6000-6999). As predictive power tends to deteriorate in samples dominated by large-cap stocks, we exclude the 300 largest firms by market capitalization at each June. We then construct sample of size $n = 100$, 250 or $1,000$ by selecting the next largest firms.[6] The smallest sample ($n = 100$) reflects an over-parameterized regime even without noise. Our noise-based strategies also create over-para+meterization intentionally.

We work with log returns rather than arithmetic returns to mitigate the influence of extreme values. Our focus is not on portfolio performance, but on cross-sectional out-of-sample predictive accuracy, as measured by out-of-sample R^2.

[5]We thank Andrew Chen for making the data publicly available.

[6]Section 4.6 will discuss the case when we retain the largest 300 firms in our samples.

3.2 Data Preparation for Rolling Training and Out-of-Sample Testing

We adopt a rolling-window framework for in-sample training and out-of-sample testing from June 1974 to June 2023. In each year t, we collect log annual stock returns Y_t for n firms, and construct an $n \times p_x$ signal matrix X_{t-1} (firm signals from the previous June).. We then demean Y_t and each column of X_{t-1} and standardize all signals to mean zero and unit variance before running cross-sectional regressions.

The number of stocks (n) is fixed at 100, 250, or $1,000$, while the number of available signals (p_x) varies over time between 140 and 182 (mean$=168.7$).

3.3 Estimation Strategies

3.3.1 Noise Injection Strategies

We create 10 random noise matrices with dimension $n \times p_x$, $Z_{t-1}^{(j)}$ ($j = 1, ..., 10$), with each element drawn independently from $N(0, 1)$. We then inject this noise into the signal matrix as:

$$X_{I,t-1}^{(j)} = X_{t-1} + aZ_{I,t-1}^{(j)}; \qquad j = 1, ..., 10,$$

where $a = \{1, 2, 3, 4, 5\}$ controls noise intensity.

For each noise-injected matrix, we compute OLS estimates, $\hat{\beta}_{I,t}^{(j)} = (X_{I,t-1}^{(j)\intercal} X_{I,t-1}^{(j)})^{-1} X_{I,t-1}^{(j)\intercal} Y_t$. ($j = 1, ..., 10$).and average them:

$$\hat{\beta}_{I,t} = \frac{1}{10} \sum_{j=1}^{10} \hat{\beta}_{I,t}^{(j)}.$$

3.3.2 Noise Augmentation Strategies

We simulate 10 random noise matrices $Z_{t-1}^{(j)}$ $(j = 1, ..., 10)$.with dimension $n \times p_z$, again drawing elements independently from $N(0, 1)$. We then augment the signal matrix via:

$$X_{A,t-1}^{(j)} = [X_{t-1}, Z_{t-1}^{(j)}]$$

for $j = 1, ..., 10$, $p_z \in \{100,\ 500,\ 1,000,\ 5,000, 10,000\}$.

For each noise-augmented signal matrix $X_{A,t-1}^{(j)}$, we solve for the minimum norm coefficients $\hat{\beta}_{A,t}^{(j)}$ using

$$\hat{\beta}_{A,t}^{(j)} = X_{t-1}^{\intercal} \left(X_{t-1} X_{t-1}^{\intercal} + Z_{t-1}^{(j)} Z_{t-1}^{(j)\intercal} \right)^{-1} Y_t.$$

The final coefficients are averaged over simulations:

$$\hat{\beta}_{A,t} = \frac{1}{10} \sum_{j=1}^{10} \hat{\beta}_{A,t}^{(j)}.$$

3.3.3 Conventional Regularization Methods Without Noise

We implement traditional regularization techniques for comparison.

Lasso and Ridge: We estimate Lasso and Ridge regressions each year by solving:

$$\hat{\beta}_{Lasso,t} = \arg\min_{\beta} |||Y_t - X_{t-1}\beta||_2^2 \text{ subject to } ||\beta||_1 \leq c_{Lasso},$$

$$\hat{\beta}_{Ridge,t} = \arg\min_{\beta} |||Y_t - X_{t-1}\beta||_2^2 \text{ subject to } ||\beta||_2^2 \leq c_{Ridge},$$

where $||\beta||_1$ and $||\beta||_2$ are l_1 and l_2 norms of β. Optimal constraint values c_{Ridge} and c_{Lasso} are chosen via cross-validation.

Partial Least Squares (PLS): We also apply PLS as a 3-factor model through iterative marginal regressions. We select 3 factors as including more factors decreased out-of-sample R^2.:

1. First factor

   - Run marginal regressions of Y_t on $\xi_{k,t-1}$, the k-th signal (i.e., the k-th column of X_{t-1}), $k = 1, ..., p_x$, to obtain the OLS estimate $\hat{\gamma}_{k,t}$. Form the first factor: $\hat{\varphi}_{1,t} = \sum_{k=1}^{p_x} \xi_{k,t-1} \hat{\gamma}_{k,t}$
   - Regress Y_t on $\hat{\varphi}_{1,t}$ to get coefficient $\hat{\eta}_{1,t}$. $\hat{Y}_t(\hat{\varphi}_{1,t}) := \hat{\varphi}_{1,t}\hat{\eta}_{1,t}$ is the predicted Y_t using the first factor.

2. Second factor:

   - Regress residuals $Y_t - \hat{\varphi}_{1,t}\hat{\eta}_{1,t}$ on $\xi_{k,t-1}, k = 1, ..., p_x$, to obtain the OLS estimate $\hat{\gamma}_{k,t}$ to form the second factor $\hat{\varphi}_{2,t}$ and coefficient $\hat{\eta}_{2,t}$.

3. Third factor:

   - Repeat with residual $Y_t - \hat{\varphi}_{1,t}\hat{\eta}_{1,t} - \hat{\varphi}_{2,t}\hat{\eta}_{2,t}$.

Through these iterative steps—known as the boosted marginal regression approach—PLS constructs a lower-dimensional representation of the predictors that optimally captures their covariance with Y_t. The final regression model remains linear in X_{t-1}. We use $\hat{\beta}_{PLS,t}$ to denote estimated PLS coefficients.

We can obtain $\hat{\beta}_{Lasso,t}$, $\hat{\beta}_{Ridge,t}$, and $\hat{\beta}_{PLS,t}$ both in under-parameterized ($p_x < n$) and over-parameterized ($p_x > n$) regimes. When $p_x < n$, we also obtain a conventional OLS estimate, $\hat{\beta}_{OLS,t}$. In over-parameterized ($p_x > n$) settings, we use the minimum norm OLS estimate for $\hat{\beta}_{OLS,t}$.

3.4 In-Sample Training

For each year t, we consider a T-year training window: $(Y_t, X_{t-1}), (Y_{t-1}, X_{t-2}), \ldots, (Y_{t+1-T}R^2_{t+1}, X_{t-T})$. Let $\hat{\beta}_{t+1-\tau}$ be the p_x-vector of cross-sectional regression coefficients estimated from $(Y_{t+1-\tau}, X_{t-\tau})$ for $\tau = 1, ..., T$. Using Fama and MacBeth's ("FMB") (1973) approach, we average past T years of regression coefficients:[7]

$$\hat{\beta}_t^{(T)} := \frac{1}{T}\sum_{\tau=1}^{T}\hat{\beta}_{t+1-\tau}.$$

We apply this FMB-style averaging to all estimation strategies: noise injection ($\hat{\beta}_{I,t}$), noise augmentation ($\hat{\beta}_{A,t}$), Lasso, Ridge, PLS, and OLS. We explore training period lengths: $T \in \{1,\ 3,\ 5,\ 10\}$. We then form the out-of-sample predicted return as

$$\hat{Y}_{t+1}^{(T)} = X_t\hat{\beta}_t^{(T)}.$$

3.5 Out-of-Sample Testing

For each strategy, we calculate the out-of-sample R^2 in period $t+1$ as:

$$R^2_{t+1} = 1 - \frac{||Y_{t+1} - X_t\hat{\beta}_t^{(T)}||_2^2}{||Y_{t+1}||_2^2}, \tag{5}$$

where $||Y_{t+1} - X_t\hat{\beta}_t^{(T)}||_2^2 := (Y_{t+1} - X_t\hat{\beta}_t^{(T)})^\intercal(Y_{t+1} - X_t\hat{\beta}_t^{(T)})$ and $||Y_{t+1}||_2^2 := Y_{t+1}^\intercal Y_{t+1}$. Note that all variables are demeaned in each cross-section. For each strategy, we report time-series averages of yearly out-of-sample R^2, as defined in (5), calculated over the full sample period (R^2_{t+1} from 1974 to 2023). Higher positive values indicate greater predictive accuracy.

[7] We require at least $T/2$ years of data for the time average calculation.

4 Evidence

We present and discuss out-of-sample R^2 of different estimation strategies using the full sample period spanning 50 years. Empirical implementation of our noise-based regularization strategies—noise injection and noise-augmentation—use simulations to generate noise and average over simulated coefficients. Therefore, our results may vary slightly in each empirical run. Our Lasso and Ridge regression results may also vary in each test, because we choose regularization parameters with cross-validation.

4.1 Out-of-Sample R^2: Conventional Estimation Without Noise

Table 1 summarizes out-of-sample R^2 values from conventional estimation methods (OLS, Minimum-norm OLS, Lasso, Ridge, and PLS) across $n = 100$, 250, 1,000 and training lengths $T = 1$, 3, 5, 10. Shading indicates positive out-of-sample R^2, with darker shades for higher values.

Results indicate that OLS is infeasible for $n = 100$ and unstable for $n = 250$, yielding large negative out-of-sample R^2 values. Minimum norm OLS also produces negative out-of-sample R^2. Without explicit regularizations, (minimum) OLS estimation is damaging rather than helpful.

However, regularization methods, such as Ridge, and PLS, reduce predictive harm and can yield positive out-of-sample R^2 with larger n and longer T. Among the three conventional regularization methods, Ridge generally outperforms Lasso and PLS. PLS shows the least effectiveness among the three conventional regularization methods. Even with regularization, positive out-of-sample R^2 values only appear with large samples ($n = 1,000$) and longer training periods ($T = 10$). The highest out-of-sample R^2 (2.15%) comes from Ridge with $n = 1,000$ and $T = 10$.

Table 1

4.2 Out-of-Sample R^2: Noise Injection Strategies

Table 2 presents results from noise injection strategies at noise intensities $a \in \{1,\ 2,\ 3,\ 4,\ 5\}$.

We highlight the following results. First of all, noise injection significantly improves out-of-sample R^2.Noise injection performance is comparable or superior to Ridge regression. For small n or short T, higher noise intensity (a) works better. High-intensity noise injections (e.g., $a = 3, 4, 5$) create positive out-of-sample R^2 even when conventional regularization methods fail. With $T = 10$, noise injection with $a = 2$ achieves the highest out-of-sample R^2 of 2.42% ($n = 1,000$).

Table 2

4.3 Out-of-Sample R^2: Noise Augmentation Strategies

Table 3 shows the performance of noise augmentation with varying $p_z \in \{100,\ 500,\ 1,000,\ 5,000,\ 10,000\}$.

Our results reveal that adding thousands of random noise variables (e.g., $p_z = 5,000,\ 10,000$) significantly improves out-of-sample R^2, even when n is small or T is short. Performance resembles the double descent phenomenon: more over-parameterization with including more noise variables generally helps improve out-of-sample forecasting performance. In our analysis, augmenting existing signals with $p_z = 5,000$ noise variables achieves strong results. With $T = 10$, augmenting with $1,000$ or more noise variables achieves out-of-sample R^2 matches or exceeds Ridge, especially at $p_z = 5,000$ or $10,000$. In particular, the highest out-of-sample R^2 of 2.42% is achieved with $p_z = 5,000$ noise variables when $n = 1,000$.

When the sample size is $n = 1,000$, we can achieve the highest out-of-sample R^2 when $p_z = 5,000$ (when $T = 10$) and $10,000$ (when $T = 5$) in Table 3. As the section 2.5 discusses, $p_z = 5,000$ and $10,000$ approximately correspond to $a \fallingdotseq 2.24$ and 3.16, respectively. In Table 2, $a = 2$ and $a = 3$ achieves the highest out-of-sample R^2 when $T = 10$ and $T = 5$, respectively. This confirms that the relationship between the two noise-based regularization strategies—noise injection and noise augmentation—can be summarized by a simple approximate relationship: $a \approx \sqrt{p_z/n}$.

Table 3

4.4 Out-of-Sample R^2: Practical Implications

Our results demonstrate that noise injection and noise augmentation significantly improve out-of-sample predictive performance when applied to well-established return predicting signals. These benefits are especially pronounced in settings with smaller cross-sectional sample sizes (n) or shorter training periods (T).

Importantly, the frequent occurrence of negative out-of-sample R^2 values does not imply that the 212 well-established signals fail to predict stock returns. On the contrary, each signal is a robust and statistically significant cross-sectional return predictor (Chen and Zimmermann, 2022). Rather, these negative out-of-sample R^2 values highlight the difficulty of dynamically combining these signals into effective predictive models.

That said, we find that the maximum out-of-sample R^2 of 2.42% is encouraging. This value is attained with both a noise injection strategy ($a = 2$, $n = 1,000$, $T = 10$) and a noise augmentation strategy ($p_z = 5,000$, $n = 1,000$, $T = 10$).

From a practitioner's perspective, out-of-sample R^2 corresponds to the squared information coefficient (IC^2) (Grinold and Kahn, 2000). Since $IC = \sqrt{R^2}$, and out-of-sample R^2 of 2.42% translates to $IC = 0.156$. Grinold and Kahn (2000, p.272) provide a helpful benchmark:

> "A good forecaster has $IC = 0.05$, a great forecaster has $IC = 0.10$, and a world-class forecaster has $IC = 0.15$. An IC higher than 0.20 usually signals a faulty backtest or imminent investigation for insider dealing."

Thus, achieving $IC = 0.156$ suggests world-class forecasting performance—a striking result given it is obtained by injecting or augmenting established signal with random noise variables.

4.5 Noise Augmentation and Stability of Regression Coefficients

Because noise injection and noise augmentation introduce implicit Ridge regularization, they help reduce overfitting and stabilize regression coefficients. To evaluate this regularization (shrinkage) effect, we examine the time-series variability of estimated coefficients, focusing on noise augmentation strategies with $n = 1,000$ stocks.

We concentration signals from the valuation and profitability categories—those with the strongest historical predictive performance. According to Chen and Zimmermann (2022), these categories contain up to 17 and 8 signals, respectively.

Table 4 reports out-of-sample R^2 of regression models using either the valuation or profitability signals alone. Adding 100, 500, or even $1,000$ random noise variables does not improve out-of-sample prediction. Out-of-sample R^2 values remain largely negative. While Lasso. Ridge, and PLS help reduce harm from overfitting, only extensive noise augmentation—with $5,000$ or $10,000$ noise variables—makes out-of-sample R^2 values positive, even with short training periods (e.g., $T \leq 5$).

Table 4

4.5.1 Coefficient Stability Analysis

We then analyze how noise augmentation affects the stability of signal coefficients over time. Specifically, we examine regression coefficients on the B/M (book-to-market) and cash-based operating profitability (Ball, Gerakos, Linnainmaa, and Nikolaev, 2016) signals.

We compare Ridge, PLS, and minimum norm OLS estimation under noise augmentation with

$p_z = 5,000$ noise variables.

Panels A-B of Figure 2 show the time-series of B/M coefficients for $T = 1$ and $T = 5$. At $T = 1$ (Panel A), Ridge and PLS coefficients vary wildly, frequently switching signs, suggesting unstable forecasts and likely negative out-of-ample R^2. With noise augmentation, the coefficients are more tightly shrunk toward zero and substantially more stable. At $T = 5$ (Panel B), Ridge and PLS coefficients are smoothed over time, yet still exhibit significant fluctuations. Noise augmentation continues to deliver more stable coefficient estimates, reflecting stronger regularization.

Panels C-D show similar patterns for the cash-based operating profitability signal. Noise-augmented coefficients are consistently more stable than those from Ridge and PLS regularizations.

Figure 2

4.5.2 Double Descent and Over-parameterization

Figure 3 illustrates how the number of noise variables (p_z) in noise augmentation strategies affects the standard deviation of the B/M coefficient (Panel A), and the out-of-sample R^2 (Panel B) using $n = 1,000$ and $T = 1$.

As the number of noise variables increases: Standard deviation spikes when $n \approx p_x + p_z$, i.e., when the model is at or near the interpolation threshold. Beyond this threshold, as we enter the over-parameterization regime, further increases in noise variables lead to reduced coefficient variability and improved out-of-sample R^2, albeit still negative or close to zero at $T = 1$.

These patterns exemplify the double descent phenomenon, where increasing model size eventually improves generalization (out-of-sample predictive accuracy) though implicit regularization.

Figure 3

4.6 Out-of-Sample Stock Return Predictability of the Largest US Firms

So far, our empirical analysis has excluded the 300 largest firms, constructing samples of size n = 100, 250 or 1,000 by selecting the next largest firms in the US. This exclusion was intentional, as our focus is on proposing a new return prediction approach and highlighting the benefits of noise-based regularization methods. Predicting returns for the largest 300 stocks is particularly challenging in the US.

In Table 5, we report out-of-sample R^2 values for Lasso, Ridge, PLS, and noise augmentation with varying $p_z \in \{500,\ 1,000,\ 5,000\}$ for samples including the largest 300 stocks. When these 300 largest stocks are included, most out-of-sample R^2 values are negative, particularly in smaller samples (n = 100 and 250). However, positive out-of-sample R^2 is observed only when existing signals are augmented with $p_z = 5,000$ random noise variables and the training period is $T = 10$ years.

For $n = 1,000$, both Lasso and Ridge regularization achieve positive out-of-sample R^2 when the training period is $T = 10$ years, with Ridge outperforming lasso (1.21% vs. 0.71%). However, noise augmentation with $p_z = 5,000$ yields an even higher out-of-sample R^2 of 1.47% when $T = 10$. Furthermore, noise augmentation with $p_z = 5,000$ achieves positive out-of-sample R^2 even with a shorter training period ($T = 5$), where other methods fail to deliver positive out-of-sample R^2.

These findings underscore the difficulty of out-of-sample return prediction for the largest US stocks. Nevertheless, extensive noise augmentation strategies—such as augmenting with $p_z = 5,000$ additional noise variables—can significantly enhance out-of-sample return predictability, outperforming conventional regularization methods.

Table 5

5 Conclusion

This study demonstrates that deliberately introducing random noise into prediction models can significantly improve the out-of-sample cross-sectional predictability of individual stock returns. By employing noise injection and noise augmentation techniques, we achieve positive out-of-sample R^2 values in scenarios where traditional regularization methods—such as Ridge, Lasso, PLS—consistently fail.

Grounded in the machine learning literature on benign overfitting and double descent, our theoretical framework reveals that noise-based approaches serve as implicit Ridge regularization mechanisms. When combined with minimum-norm OLS estimation, random noise shrinks coefficient vectors toward zero, stabilizing estimates and improving out-of-sample forecasting performance, even in over-parameterized settings.

Empirical findings strongly support three key conclusions. First, noise-based regularization offers a simple yet powerful alternative to conventional shrinkage methods, particularly in low signal-to-noise ratio environments or when training samples are limited. Second, the counter-intuitive success of adding uninformative variables highlights the crucial role of implicit regularization effects in high-dimensional financial modeling. Third, dense modeling approaches outperform traditional sparse methods in out-of-sample prediction tasks, reinforcing the emerging consensus that retaining all signals—rather than focusing on a strict signal or factor selection—enhances predictive accuracy.

These insights carry important implications for both practitioners and researchers. Instead of prioritizing signal (or factor) selection or model parsimony, our findings suggest that retaining a comprehensive set of signals while leveraging implicit regularization through strategic noise addition can yield superior predictive performance. This shift from sparse to dense modeling, coupled with noise-based regularization, opens new avenues for advancing high-dimensional financial forecasting.

References

[1] Ball, Ray, Joseph Gerakos, Juhani T. Linnainmaa, and Valeri Nikolaev. 2016. "Accruals, Cash Flows, and Operating Profitability in the Cross Section of Stock Returns." Journal of Financial Economics 121 (1): 28–45. https://doi.org/10.1016/j.jfineco.2016.03.002.

[2] Bartlett, Peter L., Philip M. Long, Gábor Lugosi, and Alexander Tsigler. 2020. "Benign Overfitting in Linear Regression." Proceedings of the National Academy of Sciences 117 (48): 30063–70. https://doi.org/10.1073/pnas.1907378117.

[3] Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. "Reconciling Modern Machine-Learning Practice and the Classical Bias–Variance Trade-Off." Proceedings of the National Academy of Sciences 116 (32): 15849–54. https://doi.org/10.1073/pnas.1903070116.

[4] Ben-Israel, Adi and Thomas N. E. Greville. 2003. Generalized Inverses: Theory and Applications, 2nd ed. Springer.

[5] Bernstein , Dennis S. 2005. Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory. Princeton University Press.

[6] Bishop, Chris M. 1995. "Training with Noise Is Equivalent to Tikhonov Regularization." Neural Computation 7 (1): 108–16. https://doi.org/10.1162/neco.1995.7.1.108.

[7] Campbell, S. L. and C. D. Meyer. 1979. Generalized Inverses of Linear Transformations. Dover Publications.

[8] Chen, Andrew Y. and Tom Zimmermann. 2022. "Open Source Cross-Sectional Asset Pricing." Critical Finance Review 11 (2): 207-64. http://dx.doi.org/10.1561/104.00000112

[9] Dereziński, Michał, Feynman Liang, and Michael W. Mahoney. 2020. "Exact Expressions for Double Descent and Implicit Regularization via Surrogate Random Design." arXiv. https://doi.org/10.48550/arXiv.1912.04533.

[10] Fama, Eugene F. and James D. MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." The Journal of Political Economy 81 (3): 607–36. https://doi.org/10.1086/260061

[11] Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri. 2021. "Economic Predictions With Big Data: The Illusion of Sparsity." Econometrica 89 (5): 2409–37. https://doi.org/10.3982/ecta17842.

[12] Goyal, Amit and Ivo Welch. 2008. "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction." Review of Financial Studies 21 (4): 1455–1508. https://doi.org/10.1093/rfs/hhm014.

[13] Green, Jeremiah, John R. M. Hand, and X. Frank Zhang. 2017. "The Characteristics That Provide Independent Information about Average U.S. Monthly Stock Returns." The Review of Financial Studies 30 (12): 4389–4436. https://doi.org/10.1093/rfs/hhx019.

[14] Grinold, Richard C. and Ronald N. Kahn. 2000. Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk. McGraw-Hill.

[15] Gu, Shihao, Bryan Kelly, and Dacheng Xiu. 2020. "Empirical Asset Pricing via Machine Learning." The Review of Financial Studies 33 (5): 2223–73. https://doi.org/10.1093/rfs/hhaa009.

[16] Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. 2022. "Surprises in High-Dimensional Ridgeless Least Squares Interpolation." The Annals of Statistics 50 (2). https://doi.org/10.1214/21-aos2133.

[17] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer.

[18] Hoerl, Arthur E., and Robert W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics 12 (1): 55. https://doi.org/10.2307/1267351.

[19] Hou, Kewei, Chen Xue, and Lu Zhang. 2020. "Replicating Anomalies." The Review of Financial Studies 33 (5): 2019–2133. https://doi.org/10.1093/rfs/hhy131.

[20] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. An Introduction to Statistical Learning with Applications in R. Springer.

[21] Kelly, Bryan, Semyon Malamud, and Kangying Zhou. 2024. "The Virtue of Complexity in Return Prediction." The Journal of Finance 79 (1): 459–503. https://doi.org/10.1111/jofi.13298.

[22] Kelly, Bryan T., Seth Pruitt, and Yinan Su. 2019. "Characteristics Are Covariances: A Unified Model of Risk and Return." Journal of Financial Economics 134 (3): 501–24. https://doi.org/10.1016/j.jfineco.2019.05.001.

[23] Kobak, Dmitry, Jonathan Lomond, and Benoit Sanchez. 2020. "The Optimal Ridge Penalty for Real-World High-Dimensional Data Can Be Zero or Negative Due to the Implicit Ridge Regularization." Journal of Machine Learning Research 21 (1): 6863–78.

[24] Liao, Yuan, Xinjie Ma, Andreas Neuhierl, and Zhentao Shi. 2024. "Does Noise Hurt Economic Forecasts?" Working Paper available at SSRN: https://ssrn.com/abstract=4659309 or http://dx.doi.org/10.2139/ssrn.4659309

[25] Matsuoka, Kiyotoshi. 1992. "Noise Injection into Inputs in Back-Propagation Learning." IEEE Transactions on Systems, Man, and Cybernetics 22 (3): 436–40. https://doi.org/10.1109/21.155944.

[26] Richards, Dominic, Jaouad Mourtada, and Lorenzo Rosasco. 2021. "Asymptotics of Ridge (Less) Regression under General Source Condition." arXiv. https://doi.org/10.48550/arXiv.2006.06386.

[27] Shen, Zhouyu, and Dacheng Xiu. 2024. "Can Machines Learn Weak Signals?" Chicago Booth Research Paper No. 24-03, available at SSRN: https://ssrn.com/abstract=4722678 or http://dx.doi.org/10.2139/ssrn.4722678working paper.

[28] Strang, Gilbert, Kenji Hiranabe, and Ashley Fernandes. 2024. "The Art of Linear Algebra." PRIMUS, March, 1–14. https://doi.org/10.1080/10511970.2024.2321349.

[29] Webb, Andrew R. 1994. "Functional Approximation by Feed-Forward Networks: A Least-Squares Approach to Generalization." IEEE Transactions on Neural Networks 5 (3): 363–71. https://doi.org/10.1109/72.286908.

[30] Wu, Denny, and Ji Xu. 2020. "On the Optimal Weighted l_2 Regularization in Overparameterized Linear Regression." Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020).

[31] Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. "Understanding Deep Learning (Still) Requires Rethinking Generalization." Communications of the ACM 64 (3): 107–15. https://doi.org/10.1145/3446776.

Tables and Figures

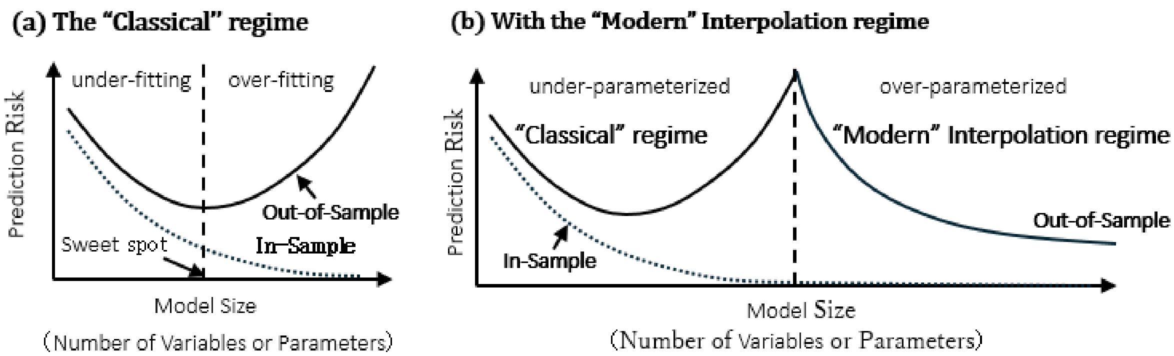


Figure 1: Prediction risk (mean squared prediction error) vs model size

Notes: This figure illustrates the relationship between prediction risk (measured as mean squared prediction error) and model size (i.e., the number of predictors or parameters). It is adapted from Belkin et al.'s (2019) Figure 1, with a few modifications.

Table 1: Out-of-Sample R^2 for Conventional Estimation Methods Without Noise

Number of stocks (n)	Training period Length (T)	OLS	Min-norm OLS	Lasso	Ridge	PLS
100	1	N/A	-251.81%	-13.81%	-8.07%	-49.45%
	3	N/A	-97.82%	-4.24%	-2.12%	-17.17%
	5	N/A	-61.47%	-2.19%	-0.43%	-9.10%
	10	N/A	-35.55%	-0.35%	0.63%	-3.81%
250	1	-292.68%	-292.68%	-11.77%	-7.77%	-36.87%
	3	-136.02%	-136.02%	-3.08%	-1.27%	-10.58%
	5	-92.37%	-92.37%	-0.85%	-0.25%	-7.02%
	10	-48.09%	-48.09%	0.65%	1.13%	-2.26%
1,000	1	-41.40%	-41.40%	-11.15%	-8.54%	-19.09%
	3	-12.19%	-12.19%	-1.72%	-0.89%	-3.96%
	5	-6.24%	-6.24%	0.20%	0.69%	-1.44%
	10	-1.58%	-1.58%	1.85%	2.15%	1.25%

Notes: This table reports the out-of-sample R^2 of cross-sectional regressions of stock returns on approximately 200 established signals (from Chen and Zimmermann 2022), evaluated across varying model sizes (n), training period lengths (T) for conventional estimation methods: OLS, Minimum-Norm OLS, Lasso, Ridge, and PLS. It presents time-series averages of yearly out-of-sample R^2, defied as:$R^2_{t+1} = 1 - (||Y_{t+1} - X_t\hat{\beta}_t^{(T)}||_2^2 / ||Y_{t+1}||_2^2)$, calculated over the full sample period ($t+1$ spanning from 1974 and 2023). Shaded cells indicate positive average out-of-sample R^2 values.

Table 2: Out-of-Sample R^2: Noise Injection Strategies

Number of stocks (n)	Training period Length (T)	Lasso	Ridge	PLS	Noise Injection Strategies (Min-Norm OLS)				
					a = 1	2	3	4	5
100	1	-13.81%	-8.07%	-49.45%	-28.00%	-9.65%	-5.06%	-2.19%	-0.83%
	3	-4.24%	-2.12%	-17.17%	-9.35%	-1.86%	-0.64%	-0.10%	-0.09%
	5	-2.19%	-0.43%	-9.10%	-4.83%	-0.75%	0.45%	0.76%	0.39%
	10	-0.35%	0.63%	-3.81%	-1.76%	0.56%	0.75%	0.75%	0.62%
250	1	-11.77%	-7.77%	-36.87%	-26.5%	-10.72%	-4.75%	-3.17%	-1.60%
	3	-3.08%	-1.27%	-10.58%	-7.5%	-2.12%	-0.44%	0.01%	0.32%
	5	-0.85%	-0.25%	-7.02%	-4.42%	-0.89%	0.38%	0.53%	0.62%
	10	0.65%	1.13%	-2.26%	-0.77%	0.87%	1.10%	1.07%	0.98%
1,000	1	-11.15%	-8.54%	-19.09%	-9.10%	-2.86%	-0.71%	0.30%	0.69%
	3	-1.72%	-0.89%	-3.96%	-0.86%	0.94%	1.35%	1.49%	1.37%
	5	0.20%	0.69%	-1.44%	0.74%	1.73%	1.83%	1.73%	1.50%
	10	1.85%	2.15%	1.25%	2.18%	2.42%	2.14%	1.88%	1.55%

Notes: This table reports the out-of-sample R^2 of cross-sectional regressions of stock returns on approximately 200 established signals (from Chen and Zimmermann 2022), evaluated across varying model sizes (n), training period lengths (T) for minimum OLS estimation methods with noise injection strategies. Each year (at the end of June), we exclude the 300 largest firms and construct samples of size $n = 100$, 250 or $1,000$ by selecting the next largest firms. We examine noise injection intensity parameter of $a = 1, 2, 3, 4, 5$. The table presents time-series averages of yearly out-of-sample R^2, defied as:$R^2_{t+1} = 1 - (||Y_{t+1} - X_t\hat{\beta}_t^{(T)}||_2^2/||Y_{t+1}||_2^2)$, calculated over the full sample period ($t + 1$ spanning from 1974 and 2023). Shaded cells indicate positive average out-of-sample R^2 values.

Table 3: Out-of-Sample R^2: Noise Augmentation Strategies

Number of stocks (n)	Training period Length (T)	Lasso	Ridge	PLS	Noise Augmentation Strategies (Min-Norm OLS)				
					100	500	1,000	5,000	10,000
100	1	-15.36%	-7.77%	-49.45%	-41.78%	-10.64%	-5.27%	-0.32%	0.06%
	3	-4.81%	-1.89%	-17.17%	-13.45%	-2.74%	-0.95%	0.38%	0.32%
	5	-2.68%	-0.27%	-9.10%	-7.05%	-0.60%	0.28%	0.61%	0.42%
	10	-0.37%	0.74%	-3.81%	-2.93%	0.55%	0.89%	0.67%	0.43%
250	1	-11.54%	-7.80%	-36.87%	-301.6%	-17.06%	-9.12%	-1.12%	-0.10%
	3	-3.21%	-1.21%	-10.58%	-119.3%	-3.87%	-1.26%	0.58%	0.62%
	5	-1.05%	-0.16%	-7.02%	-76.33%	-2.12%	-0.26%	0.86%	0.74%
	10	0.53%	1.13%	-2.26%	-39.45%	0.12%	1.05%	1.15%	0.85%
1,000	1	-11.34%	-8.30%	-19.09%	-41.45%	-43.57%	-26.80%	-2.81%	-0.48%
	3	-1.79%	-0.85%	-3.96%	-12.30%	-13.19%	-6.19%	0.99%	1.46%
	5	0.07%	0.69%	-1.44%	-6.33%	-6.61%	-2.33%	1.75%	1.86%
	10	1.82%	2.14%	1.25%	-1.61%	-1.45%	0.93%	2.42%	2.18%

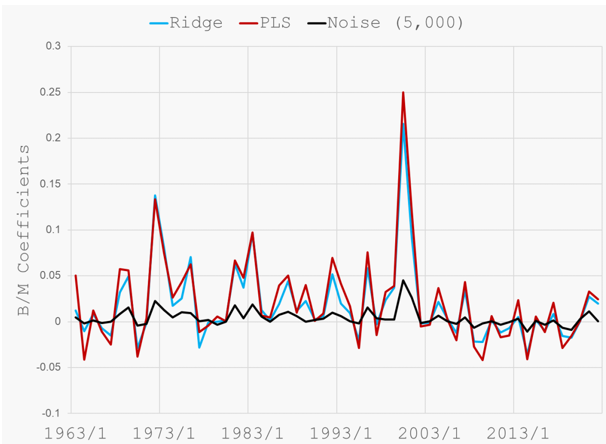
Notes: This table reports the out-of-sample R^2 of cross-sectional regressions of stock returns on approximately 200 established signals (from Chen and Zimmermann 2022), evaluated across varying model sizes (n), training period lengths (T) for minimum OLS estimation methods with noise augmentation strategies. Each year (at the end of June), we exclude the 300 largest firms and construct samples of size $n = 100$, 250 or $1,000$ by selecting the next largest firms. We examine noise augmentations with $p_z = 100$, 500, 1000, $5,000$, and $10,000$ additional noise variables. The table presents time-series averages of yearly out-of-sample R^2, defied as: $R^2_{t+1} = 1 - (||Y_{t+1} - X_t\hat{\beta}^{(T)}_t||^2_2/||Y_{t+1}||^2_2)$, calculated over the full sample period ($t+1$ spanning from 1974 and 2023). Shaded cells indicate positive average out-of-sample R^2 values.

Table 4: Out-of-Sample R^2 of Regressions with Valuation or Profitability Signals.

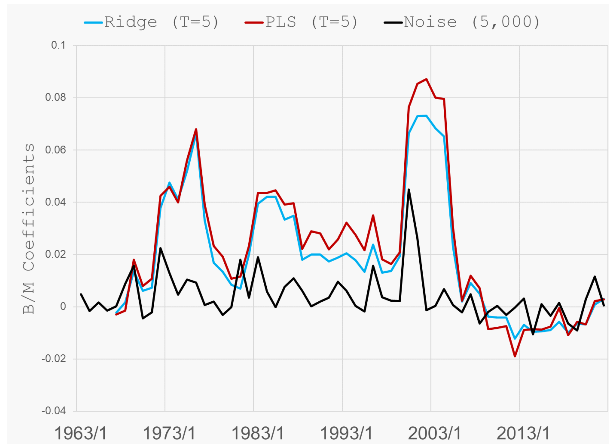
Signal Categories	*Training period Length (T)*	Lasso	Ridge	PLS	Noise Augmentation Strategies (Min-Norm OLS)				
					100	500	1,000	5,000	10,000
Valuation	1	-6.55%	-6.26%	-8.68%	-9.17%	-9.14%	-19.84%	-0.18%	0.25%
	3	-2.14%	-1.97%	-2.77%	-3.12%	-3.12%	-8.03%	0.53%	0.51%
	5	-0.83%	-0.75%	-1.41%	-1.54%	-1.54%	-4.51%	0.71%	0.59%
	10	0.34%	0.37%	0.04%	0.01%	-0.01%	-1.58%	0.78%	0.59%
Profitability	1	-3.46%	-3.01%	-4.76%	-4.8%	-5.01%	-12.41%	0.67%	0.57%
	3	-1.98%	-1.66%	-2.39%	-2.4%	-2.51%	-7.30%	0.47%	0.41%
	5	-0.92%	-0.71%	-1.22%	-1.19%	-1.28%	-4.62%	0.49%	0.38%
	10	-0.13%	0.01%	-0.25%	-0.25%	-0.27%	-2.12%	0.51%	0.36%

Notes: This table reports the out-of-sample R^2 of cross-sectional regressions of stock returns on valuation signals (17 or less during the sample period) and profitability signals (8 or less during the sample period), based on Chen and Zimmermann's (2022) signal categorization, evaluated for $n = 1,000$ stocks and across varying training period lengths (T) for minimum norm OLS estimation methods with noise augmentation strategies. Each year (at the end of June), we exclude the 300 largest firms and construct samples of size n = 100, 250 or 1,000 by selecting the next largest firms. We examine noise augmentations with p_z = 100, 500, 1000, 5,000, and 10,000 additional noise variables. The table presents time-series averages of yearly out-of-sample R^2, defied as:$R^2_{t+1} = 1 - (||Y_{t+1} - X_t\hat{\beta}_t^{(T)}||_2^2/||Y_{t+1}||_2^2)$, calculated over the full sample period ($t+1$ spanning from 1974 and 2023). Shaded cells indicate positive average out-of-sample R^2 values.

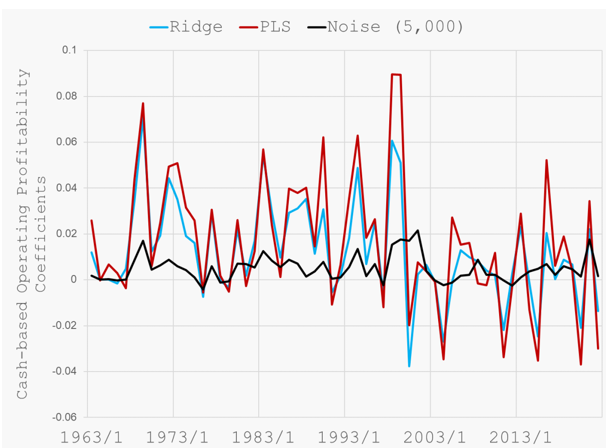
Figure 2: Time Variations of Coefficients on Selected Signals



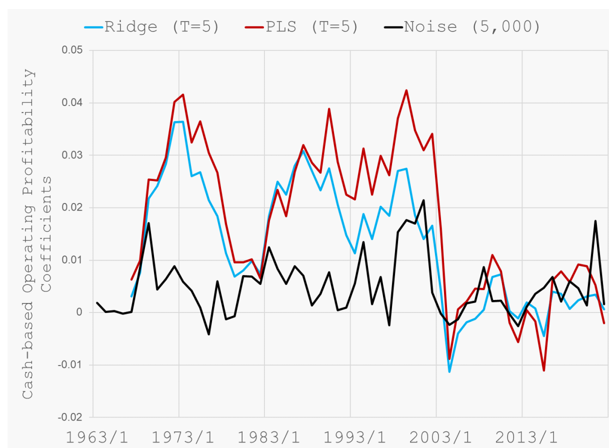
Panel A: Coefficients on the B/M Signal: $T = 1$, $n = 1,000$.Ridge, PLS, and Minimum Norm OLS with Noise Augmentation ($p_z = 5,000$).



Panel B: Coefficients on the B/M Signal: $T = 5$, $n = 1,000$.Ridge, PLS, and Minimum Norm OLS with Noise Augmentation ($p_z = 5,000$).

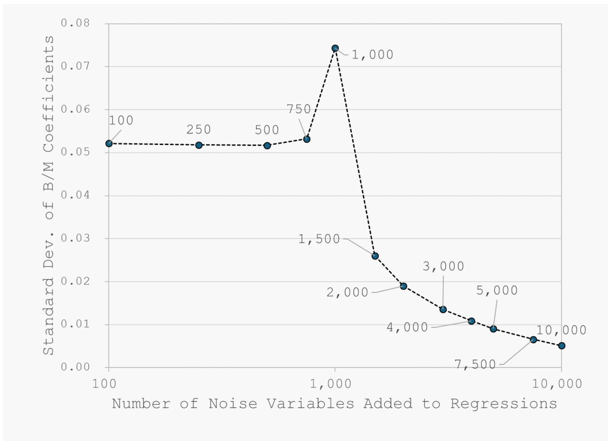


Panel C: Coefficients on the Cash-based Operating Profitability Signal: $T = 1$, $n = 1,000$.Ridge, PLS, and Minimum Norm OLS with Noisee Augmentation ($p_z = 5,000$).

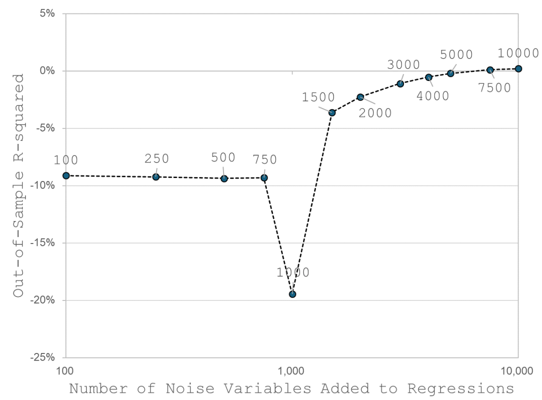


Panel D: Coefficients on the Cash-based Operating Profitability Signal: $T = 5$, $n = 1,000$.Ridge, PLS, and Minimum Norm OLS with Noisee Augmentation ($p_z = 5,000$).

Figure 3: Relationship between Time-Series Properties of B/M Coefficients and the Number of Additional Noise Variables



Panel A: Standard Deviation of the B/M Coefficient vs. Number of Noise Variables



Panel B: Average Out-of-Sample R^2vs. Number of Noise Variables

Notes to Figure 2: Panels A-D show the time-variation of regression coefficients for the B/M signal (Panels A-B) and the cash-based operating profitability signal (Panels C-D), estimated using Ridge, PLS, and minimum norm OLS with noise augmentation ($p_z = 5,000$ additional noise variables). Panels A and C present results for $T = 1$, while Panels B and D illustrate the $T = 5$ case.

Notes to Figure 3 : Panel A illustrates how the time-series standard deviation of the B/M signal coefficient varies with the number of noise variables (p_z) included in the noise augmentation strategy. Panel B presents the out-of-sample R^2 of a cross-sectional regression using established signals. For both panels, we set $n = 1,000$ and $T = 1$.

Table 5: Out-of-Sample R^2 for Samples Including the 300 Largest Stocks

Number of stocks (n)	Training period Length (T)	Lasso	Ridge	PLS	Noise Augmentation (Min-Norm OLS)		
					500	1,000	5,000
100	1	-22.21%	-11.11%	-55.53%	-15.27%	-8.39%	-1.11%
	3	-8.07%	-3.72%	-18.97%	-4.92%	-2.68%	-0.33%
	5	-4.79%	-1.80%	-12.09%	-2.65%	-1.31%	-0.03%
	10	-1.54%	-0.40%	-5.17%	-0.82%	-0.23%	0.21%
500	1	-18.46%	-14.23%	-34.92%	-40.48%	-19.18%	-4.89%
	3	-5.29%	-4.15%	-10.72%	-13.05%	-5.68%	-1.19%
	5	-2.93%	-2.59%	-6.82%	-8.35%	-3.55%	-0.43%
	10	-0.65%	-0.16%	-2.24%	-2.57%	-0.64%	0.48%
1,000	1	-14.61%	-11.86%	-24.37%	-53.92%	-34.76%	-5.61%
	3	-2.99%	-1.75%	-5.38%	-17.54%	-8.89%	-0.25%
	5	-1.29%	-0.53%	-2.97%	-10.52%	-4.86%	0.50%
	10	0.71%	1.21%	-0.04%	-4.13%	-1.16%	1.47%

Notes: This table reports the out-of-sample R^2 of cross-sectional regressions of stock returns on approximately 200 established signals (from Chen and Zimmermann 2022), evaluated across varying model sizes (n), training period lengths (T) for minimum OLS estimation methods with noise augmentation strategies. Each year (at the end of June), we construct samples of size $n = 100$, 250 or $1,000$ by selecting the largest firms, including the largest 300. We examine noise augmentations with $p_z = 500$, 1000, and $5,000$ additional noise variables. The table presents time-series averages of yearly out-of-sample R^2, defied as: $R^2_{t+1} = 1 - (||Y_{t+1} - X_t\hat{\beta}_t^{(T)}||_2^2/||Y_{t+1}||_2^2)$, calculated over the full sample period ($t + 1$ spanning from 1974 and 2023). Shaded cells indicate positive average out-of-sample R^2 values.