Confident Risk Premia: Economics and Econometrics of Machine Learning Uncertainties

Rohit Allena *
Goizueta Business School
Emory University

First draft: October 16, 2020

Current draft: March 19, 2021

Abstract

This paper derives ex-ante standard errors of risk premium predictions from neural networks (NNs). Considering standard errors, I provide improved investment strategies and ex-post out-of-sample (OOS) statistical inferences relative to existing literature. The equal-weighted (value-weighted) confident high-low strategy that takes long-short positions exclusively on stocks that have precise risk premia earns an OOS average monthly return of 3.61% (2.21%). In contrast, the conventional high-low portfolio yields 2.52% (1.48%). Existing OOS inferences do not account for ex-ante estimation uncertainty and thus are not adequate to statistically compare the OOS returns, Sharpe ratios and mean squared errors of competing trading strategies and return prediction models (e.g., linear, NN and random forest). I develop a bootstrap procedure that delivers robust OOS inferences. The bootstrap tests reveal that large OOS return and Sharpe ratio differences between NN and benchmark linear models' traditional high-low portfolios are statistically insignificant. However, the NN-based confident high-low portfolios significantly outperform all competing strategies. Economically, standard errors reflect time-varying market uncertainty and spike after financial shocks. In the cross-section, the level and precision of risk premia are correlated, thus NN-based investments deliver more gains in the long positions.

Keywords: Machine Learning, Neural Networks, Standard Errors, Risk Premium, Novel Investment Strategies, Robust Out-of-Sample Inferences, Average Return Comparisons, Sharpe Ratio Comparisons, Machine Learning Uncertainties

*I am grateful to Jay Shanken (committee chair) and Tarun Chordia (committee co-chair) for their mentorship and invaluable suggestions. I benefited from discussions with my other committee members and faculty at Emory University, including William Mann, Jegadeesh Narasimhan, Gonzalo Maturano, Christoph Herpfer, and Donald Lee, as well as seminar participants at the business schools of Boston College, Boston University, Copenhagen, HEC Paris, HKUST, Indian School of Business, National University of Singapore, Tulane University, Universities of Florida, Georgia, Houston, and Yale University. Special thanks to Bryan Kelly and Jonathan Lewellen for their insightful feedback. All errors are mine.
Address: 1300 Clifton Road, Atlanta, 30322, E-mail Address: rohit.allena@emory.edu

I. Introduction

Modern empirical asset pricing literature applies machine learning (ML) models to estimate asset risk premia (i.e., expected returns in excess of the risk free rate), as these models can accommodate non-linear relations amongst a high-dimensional set of predictors. In an influential work, Gu, Kelly, and Xiu (2020) (GKX) examine various ML models, such as neural networks (NNs) and random forests, to predict individual stock's monthly risk premia. They argue that NNs statistically outperform the benchmark linear models examined by Lewellen (2015) (henceforth Lewellen) in predicting stock risk premia.[1]

However, the burgeoning ML literature has not ascertained the ex-ante precision (i.e., standard errors and confidence intervals) of risk premium predictions from NNs. Fama and French (1997) and Pástor and Stambaugh (1999) show that expected return estimates from traditional factor-based models are unavoidably imprecise due to uncertainty about unknown parameters, including asset exposures to factors (betas) and factor premia (gammas). Consequently, they argue that factor-based risk premium measurements are not suitable for making cost-of-equity capital decisions. Given that NNs entail a massive number of parameters, determining the precision of NN-based risk premia is important.

This paper develops a novel and easy-to-implement procedure to estimate predictive standard errors of NN-based risk premium predictions at both the stock-level and portfolio-level (e.g., industry portfolios). These *ex-ante* measures capture estimation uncertainty related to risk premium predictions. Whereas standard errors of traditional, linear, factor-based and characteristics-based risk premium estimates are available in the literature, those of highly complex, NN-based risk premia are not. I tackle this challenge by adapting the NNs of GKX to simultaneously deliver risk premium predictions and their standard errors every period. The predictive standard errors resemble classical bootstrap-based estimators but are available in real-time with no additional computation costs. The obtained standard errors are then theoretically justified, and empirically validated using Monte-Carlo simulations.

Importantly, I present novel insights demonstrating why and how ex-ante standard errors must be explicitly considered to address two core asset pricing problems that appear in virtually every study in the burgeoning ML literature: (i) forming long-short trading portfolios using NN-based or any ML-based risk premium predictions and (ii) statistically evaluating the *ex-post* out-of-sample (OOS) performance of any model-based risk premia and corresponding trading strategies.[2] Considering ex-ante standard errors in answering both of these questions is of fundamental importance and has not been established in the literature.

[1] Bianchi, Büchner, and Tamoni (2020) and Bali, Goyal, Huang, Jiang, and Wen (2020) employ NNs to estimate bond and corporate bond risk premia, respectively.

[2] The standard errors also impact the cost of capital decision-making with NN-based risk premia. In the spirit of Fama and French (1997) and Pástor and Stambaugh (1999), Allena (2020b) separately addresses this question.

Ex-ante standard errors provide investment gains. Many researchers (e.g. GKX and Avramov, Cheng, and Metzker (2020)) sort stocks into deciles based solely on their return predictions, and they take long-short positions on the extreme predicted-return deciles. This paper provides substantial enhancements to these conventional high-low (HL) investment strategies by exploiting the cross-sectional variation in the ex-ante precision of risk premia. I introduce novel "Confident-HL" trading portfolios that exclusively take long-short positions on a subset of stocks in the extreme predicted-return deciles that have more confident risk premia (i.e., high absolute ratios of risk premium predictions and their standard errors, or absolute t-ratios).[3] These strategies deliberately exclude stocks with relatively imprecise risk premium estimates and thus deliver large OOS average return and Sharpe ratio improvements.

Ex-ante standard errors impact ex-post OOS statistical inferences. To compare the ex-post OOS performance of these HL trading strategies or, any competing return prediction models or associated investment portfolios, researchers use two approaches: (i) reporting *point estimates* of models' OOS R^2s (OOS-R^2s) and investment portfolios' OOS average returns and Sharpe ratios (e.g., Chen, Pelger, and Zhu (2020)) or (ii) conducting simple t-tests motivated by Diebold and Mariano (2002) (henceforth DM) (e.g., GKX, Bianchi et al. (2020), Avramov et al. (2020), and Bali et al. (2020)).[4] I show that these ex-post OOS inferences are inadequate because they do not account for ex-ante standard errors (i.e., estimation uncertainty).[5]

This paper presents a bootstrap procedure, robust to ex-ante estimation uncertainty, for valid statistical comparisons of any two portfolios' ex-post OOS returns and Sharpe ratios. Likewise, the method also compares the predictive performance of any two competing return prediction models (e.g., linear, random forests and NNs). Simulations suggest that whereas the 5%-level bootstrap tests yield accurate sizes close to 5%, the DM tests deliver distorted sizes between 13% and 42%, depending on the degree of estimation uncertainty.

Importantly, the bootstrap tests reveal that existing inferences with the DM tests over-reject the benchmark Lewellen model in favor of NNs. I find that the difference between both models' conventional HL portfolios' OOS returns and Sharpe ratios are either moderately significant or statistically insignificant. However, NNs exceptionally outperform on subsamples of stocks that have *confident* NN-based risk premia. Likewise, NN-based Confident-HL portfolios, which exclude stocks with relatively imprecise risk premia, statistically outperform all other competing strategies. Thus, considering ex-ante standard errors of NN-based risk premia is necessary for both real-time trading strategies and ex-post OOS inferences. Although this paper focuses primarily on NNs

[3] I measure the precision of risk premium predictions using their confidence-levels (i.e., absolute t-stats). See section C.C3 for an analytical motivation. Alternatively, I also present results using the inverse standard errors as proxies for the precision, and my conclusions are the same.

[4] Using simulations, Allena (2020a) shows that inferences based only on OOS point estimates are highly misleading.

[5] Diebold (2015) and GKX emphasize that the DM tests are not suitable for comparing model-based forecasts with estimation uncertainty. GKX acknowledge this limitation and conduct the DM tests. I illustrate why and how to account for parameter uncertainty to obtain accurately sized tests.

because of their predominance, I emphasize that the arguments hold for all ML-based risk premia.

I begin by showing that ex-ante standard errors of NN-based, or any ML-based, risk premium predictions predict their (future) squared forecast errors and thus yield large economic gains.[6] For example, when the standard errors of specific stock risk premium predictions are large, so are their squared forecast errors. This result is due to the "bias-variance" tradeoff. Expected squared forecast errors equal the sum of ex-ante "variances" and squared "biases". Whereas bias represents model misspecification, variance quantifies estimation uncertainty. Because predictions from ML models entail flexible functions involving many parameters, variances rather than biases predominantly determine their squared forecast errors. As a consequence, I establish that the Confident-HL portfolios that deliberately drop stocks with imprecise risk premia earn superior expected returns.

A simple example provides the central intuition. Consider two stocks A and B with risk premia μ_A and μ_B, respectively. Let $\hat{\mu}_A$ and $\hat{\mu}_B$ be their risk premium predictions, which are normal, uncorrelated and unbiased, with the measurement error variance σ^2. The unbiased assumption suits ML-based predictions. Then the expected OOS return of the HL strategy that takes a long (short) position on the stock with the highest (lowest) risk premium prediction equals

$$E(HL) = (\mu_A - \mu_B)P(\hat{\mu}_A > \hat{\mu}_B) + (\mu_B - \mu_A)P(\hat{\mu}_B > \hat{\mu}_A) = (\mu_A - \mu_B)\left[2\Phi\left(\frac{\mu_A - \mu_B}{\sqrt{2}\sigma}\right) - 1\right], \quad (1)$$

where $P(.)$, $\Phi(.)$ denote the probability and standard normal distribution measures, respectively. (1) indicates that the expected HL return monotonically decreases with the variance of risk premium predictions. In other words, between any two sets of stocks with the same levels of risk premia, the HL strategy formed from more precise predictions yields higher OOS expected returns. Intuitively, besides the level of risk premium predictions, the precision helps better determine the cross-sectional ranking among stocks and thus generates higher HL expected returns.[7]

Consistent with this intuition, the empirical section documents enormous economic gains from the Confident-HL portfolios. In particular, I consider a 3-layer NN (NN-3) examined by GKX to predict a large sample of U.S stock returns between 1987 and 2016. The conventional equal-weighted (EW) and value-weighted (VW) HL portfolios formed using NN-3-based risk premia earn ex-post OOS average monthly returns of 2.52% and 1.48%, with annualized Sharpe ratios of 1.5 and 0.9, respectively. However, the EW (VW) Confident-HL portfolio formed from a small subset of stocks confidently predicted by NN-3 delivers corresponding measures of 3.61% (2.21%) and 1.75 (1.09), respectively. Thus, dropping imprecise predictions enhances the OOS average returns by 43% (49%) and Sharpe ratio by 16% (21%). In contrast, measures of the EW (VW) "Low-Confident" portfolio that instead takes long-short positions on the subset of stocks with the most

[6] Forecast errors equal the differences between true and predicted risk premia.

[7] Mathematically, the prediction uncertainty induces downward bias to the maximum possible expected HL return that can be obtained when true risk premia are known. This result follows from Jensen's inequality (see section II).

imprecise risk premia are relatively much lower, 2.35% (1.31%) and 1.18 (0.55), respectively.

The Confident-HL portfolio's impressive performance hinges on the theoretical result showing that NN-based predictions' ex-ante standard errors predict their ex-post squared forecast errors. Consistent with this result, I find that the ex-ante confidence and ex-post OOS-R^2 of NN-based predictions are monotonically related. The bottom decile containing the stocks with the most imprecise ex-ante return predictions attain an OOS-R^2 of 0.81%. In contrast, the top decile of stocks confidently predicted by NN-3 delivers a dramatic 2.21% OOS-R^2, an increase of 170%.

Notably, Confident-HL portfolios based on simple models involving a few parameters (e.g., Lewellen) are less likely to deliver impressive gains. Biases rather than variances predominantly determine expected forecast errors of simple models. Consistent with this result, I find that the Confident-HL portfolios formed using the Lewellen model's risk premium predictions and standard errors do not yield economic gains. Unfortunately, it is not possible to construct "Low-Bias-HL" portfolios (analogous to "Confident-HL" portfolios) for simple models using ex-ante biases (rather than standard errors) because true risk premia are unknown.

To assess whether the documented NN-3-based Confident HL portfolios' OOS gains *statistically* outperform other strategies, I first show that the existing DM tests are inadequate because they do account for ex-ante standard errors. Although ex-ante estimation uncertainty impacting ex-post OOS inferences seems instinctively puzzling, a simple example demonstrates the main intuition.

Consider comparing OOS returns of any two model-based HL portfolios. These portfolios could be expressed as different weighted sums of excess returns, depending on which stocks comprise the portfolios' long and short legs. Every period, the weights are estimated using all past data. The DM t-test thus equals the ratio of the HL return differentials' time-series average to its standard error estimate. DM show that this test yields valid asymptotic inferences only under the assumption that the return differential series is covariance stationary. However, the precision of the portfolios' estimated weights increases over time as more data are available. Thus, the HL return differentials exhibit time-varying second moments, breaking down the DM assumption.

Consistent with this intuition, I empirically establish that all model-based HL returns violate the DM assumption. The covariance-stationarity tests of Pagan and Schwert (1990) lends support to non-stationarities in the HL returns, suggesting that the DM tests are inadequate. To conduct valid OOS inferences, I develop a bootstrap procedure that is robust to non-stationarities induced by estimation uncertainty. The method builds on the block bootstrap procedure of Kunsch (1989), which provides asymptotically valid inferences in the presence of non-stationarities (Gonçalves and White (2002, 2005)).

The bootstrap tests suggest that the differences between NN-3 and Lewellen-based conventional HL strategies' OOS returns and Sharpe ratios are either statistically insignificant or moderately significant. For example, a seemingly large 0.72% (0.37%) difference between the EW (VW) NN-3-

based and Lewellen-based HL portfolios' average monthly OOS returns are statistically insignificant at the 1% (10%) level.[8]

However, the NN-3-based Confident-HL strategy statistically outperforms all other competing strategies, including NN-3-based conventional HL portfolios, as well as Lewellen-based HL and Confident-HL portfolios. Moreover, the relative performance of NN-3 over Lewellen increases monotonically with the precision of NN-3-based risk premia. For example, the average monthly return difference between NN-3 and Lewellen VW HL portfolios formed using the stocks most confidently predicted by NN-3 is a highly significant 0.82%. In contrast, the difference is a significantly negative -1.2% on the subset of stocks most imprecisely predicted by NN-3. These results demonstrate that besides risk premium predictions, ex-ante standard errors are crucial for constructing desirable NN-based investment portfolios.

Avramov et al. (2020) argue that investments based on NN-3 predictions primarily extract gains from microcaps (i.e., stocks with market capital smaller than the 20^{th} NYSE size percentile) and deliver insignificant OOS returns on non-microcaps. However, I find that the Confident-HL portfolios yield significant economic gains even on non-microcaps. For example, the EW (VW) Confident-HL portfolio yields an average OOS monthly return of 2.25% (2.07%), whereas the HL strategy delivers 1.66% (1.42%). The Confident-HL portfolios' performance is robust to transaction costs, traditional factor model risk exposures and higher-moment risks that penalize losses more than rewarding gains.

To ensure that the Confident-HL strategies' superior performance is not driven by inadvertently taking long (short) positions on the stocks that have higher (lower) risk premium predictions, I construct several matching strategies. These portfolios resemble the conventional HL strategies but are matched to have the same "predicted-return" averages as those of the Confident-HL portfolios. Whereas the EW-Confident HL portfolio yields a 3.61% monthly OOS return, the matching HL strategy makes 3.07%. This result, consistent with the previously described example, reiterates that for the same levels of risk premia, trading strategies formed from stocks with more confident risk premia earn higher expected returns. The significant 0.55% monthly return difference between the two portfolios precisely captures the economic value of incorporating standard error information into trading strategies.

In the final exploration, I document interesting time-series and cross-sectional variations in the ex-ante standard errors that have important economic relevance. In the time-series, aggregate monthly standard errors (i.e., cross-sectional averages of ex-ante standard errors) reflect time-varying financial market uncertainty. Bloom (2009) and Baker, Bloom, and Davis (2016) docu-

[8] My results do not directly compare with GKX for one main reason, among others. Lewellen (2015) advocates three benchmark linear models with either three, seven, or fifteen characteristics. Whereas GKX use the model with three predictors, I examine the model with fifteen that Lewellen showed to exhibit superior return forecasting ability. Nevertheless, the conclusion that the DM tests over-reject any of Lewellen's models in favor of NNs remains valid.

ment that market uncertainty jumps up after major shocks (e.g., Black Monday, Lehman Brothers bankruptcy). Consistent with these studies, the aggregate standard errors spike an average of at least twice the value of other periods. Because many individual predictors (e.g., size, price trends, and stock market volatility) in the NN-3 model substantially deviate from their usual distributions when markets are uncertain, risk premium predictions based on these unusual predictors would be hugely imprecise. Thus, the aggregate standard errors capture market uncertainty.

In the cross-section, the NN-3 model (*ex-ante*) confidently predicts risk premia of stocks associated with small market capital, high book-to-market ratios, high 1-year momentum returns, and high risk premium predictions. Thus, the NN-3-based investment strategies deliver more gains in the long-leg rather than the short-leg. This result contrasts with the "arbitrage asymmetry" studies, which argue that anomaly-based investment portfolios yield relatively more profits in the short-leg (e.g., Stambaugh, Yu, and Yuan (2012) and Avramov, Chordia, Jostova, and Philipov (2013)). Thus, possible mechanisms that lead to the association between the level and precision of (NN-based) risk premium predictions still need to be explored.

To summarize, this paper quantifies the *ex-ante* precision of the NN-based risk premium predictions and exploits this information to construct desirable Confident-HL investment portfolios. To statistically assess these portfolios' OOS performance, the paper shows that the existing DM tests are inadequate because they do not take into account ex-ante estimation uncertainty. I propose a bootstrap test that permits valid OOS inferences. The tests suggest that the NN-3-based Confident-HL portfolios significantly outperform the traditional NN3-HL and Lewellen-HL portfolios in terms of their OOS returns and Sharpe ratios, whereas the reported dominance of the conventional NN3-HL over the Lewellen-HL portfolio is statistically insignificant.

A. Contribution

The paper makes three crucial methodological and investment-related contributions.

Ex-ante standard errors. This paper generalizes the "dropout" procedure developed by Gal and Ghahramani (2016) to obtain standard errors of NN-based risk premium predictions. They show that an NN that employs dropout regularization is a Bayesian NN with a similar structure, and they estimate standard errors of NN-based predictions using the comparable Bayesian models' instantly available posterior variances. However, these are standard errors of individual "raw" predictions (equivalent to excess return predictions), not of "prediction means" (comparable to risk premium predictions). Moreover, they do not discuss how to obtain "joint densities" of different predictions from Bayesian NNs, which are necessary to compute portfolio-level standard errors. Nor do they show whether these Bayesian standard errors satisfy frequentist properties.

To my knowledge, this is the first paper to compute stock-level and portfolio-level standard errors of NN-based *risk premium* estimates by explicitly deriving the marginal and joint densities of

expected return predictions from Bayesian NNs. I draw an equivalence between the frequentist and Bayesian standard errors and use simulations to show that the computed standard errors satisfy frequentist properties with accurate coverage probabilities. For example, simulations indicate that 95% (or any x% with $0 < x < 100$) confidence intervals constructed from risk premium predictions and their standard errors cover the true simulated risk premia with nearly 95% (x%) probability.

Out-of-Sample Comparisons. The paper relates to studies that compare competing return forecast models, including Goyal and Welch (2003, 2008), GKX, Bianchi et al. (2020), Bali et al. (2020), and Chen et al. (2020). These studies use either the OOS DM tests or assess the point estimates of OOS Sharpe ratios and OOS-R^2s, without accounting for estimation uncertainty. In contrast, this paper's block bootstrap method generalizes the DM tests by automatically accounting for non-stationarities induced by estimation uncertainty. This method can be employed to assess OOS performance of any model-based return predictions.

Investment Portfolios. The paper relates to studies, including GKX, Chinco, Clark-Joseph, and Ye (2019), and Avramov et al. (2020), that construct traditional HL portfolios based on various model-based return predictions. Alternatively, this paper shows how Confident-HL strategies could deliver superior expected returns. These strategies generally apply to all model-based return predictions, as long as their predictive standard errors are informative about their squared forecast errors.

B. Paper Overview

I organize the rest of the paper as follows. Section II provides the basics of model-based risk premium predictions and shows why the Confident-HL portfolios yield superior expected returns. Section III presents the statistical framework of NN-based risk premia and derives their standard errors. Section IV shows how to conduct valid OOS inferences. Section V presents the empirical results. Section VI concludes. Appendix includes proofs of propositions and simulations. Internet Appendix contains additional robustness checks and simulations.

II. Risk Premium Predictions and Predictive Standard Errors

This section presents the fundamental premise of measuring risk premia based on general econometric models, including the traditional linear and more advanced ML models (e.g., NN). It builds on the bias-variance tradeoff to explain why ML models' predictive standard errors are informative about their squared forecast errors, thus yielding large economic gains in terms of appropriate investment portfolios.

A. Basics of model-based risk premium predictions

In the spirit of GKX, consider a general additive prediction error model for realized stock returns in excess of the risk-free rate, given by

$$r_{i,t+1} = E_t(r_{i.t+1}) + \epsilon_{i,t+1},\ E_t(\epsilon_{i,t+1}) = 0,\ V_t(\epsilon_{i,t+1}) = \sigma^2 \tag{2}$$

where $r_{i,t+1}$ is the excess return of stock i at period $t+1$; $E_t(r_{i,t+1})$ is the stock i's unobserved conditional risk premium at period t; and $\epsilon_{i,t+1}$ is the unexpected component of returns due to new information at $t+1$, which is unpredictable at t. $E_t(.)$ and $V_t(.)$ denote the conditional expectation and variance operations, respectively. $\epsilon_{i,t+1}$ are iid over time and across stocks.

Let a flexible model $f(z_{it};\beta)$, involving stock-level predictors $\{z_{it}\}_{(it)}$ and parameters β, estimates unobserved risk premia. The set of predictors could be potentially large, containing many characteristics (e.g., size and book-to-market) and macroeconomic variables (e.g., earnings-to-price, stock market volatility). Like GKX, the parametric form of the model, $f(.)$, remains the same across different stocks and over time, thereby exploiting information from the entire panel of data to yield stable risk premium measurements. Because the true parameters, β, are unknown, the risk premia are estimated by

$$E_t(r_{i,t+1}) \approx f(z_{it};\hat{\beta}),\ \forall \text{ stocks } i, \tag{3}$$

where $\hat{\beta}$ are estimated parameters from the past data. The expected squared forecast errors of the model-based risk premium predictions are given by

$$E_t\left[\left(E_t(r_{i,t+1}) - f(z_{i,t};\hat{\beta})\right)^2\right] = E_t\left[\left(r_{i,t+1} - f(z_{i,t};\hat{\beta})\right)^2\right] - V_t(\epsilon_{i,t+1}),\ \forall i. \tag{4}$$

Because $\epsilon_{i,t+1}$ and $\{z_{it}\}_{(i,t)}$ are independent, minimizing the risk-premium squared forecast errors is equivalent to minimizing the realized return squared forecast errors. Thus, the best risk premium measurements are those that accurately predict subsequent returns. Consequently, the literature uses the following specification to estimate the true risk premia:

$$r_{i,t+1} = f(z_{it};\beta) + \eta_{i,t+1},\ E_t(\eta_{i,t+1}) = 0, \tag{5}$$

where risk premium and next period return ($\hat{r}_{i,t+1}$) predictions are given by

$$E_t(r_{i,t+1}) \approx \hat{r}_{i,t+1} = f(z_{it};\hat{\beta}) \tag{6}$$

Importantly, the expected squared forecast errors of return predictions based on (5) could be

decomposed as the sum of three terms, given by

$$E_t\left[(r_{i,t+1} - f(z_{i,t};\hat{\beta}))^2\right] = \underbrace{\Big(E_t(r_{i,t+1}) - E_t(f(z_{i,t};\hat{\beta}))\Big)^2}_{Bias^2} + \underbrace{E_t\left(f(z_{i,t};\hat{\beta}) - E_t(f(z_{i,t};\hat{\beta}))\right)^2}_{Variance} + V_t(\epsilon_{i,t+1}). \tag{7}$$

The first term in the right hand side of (7), popularly known as "squared-bias", measures the model misspecification of $f(.)$ in estimating the true risk premia. The second, known as "variance", quantifies parameter uncertainty. The ex-ante predictive standard errors, which are the main focus of this paper, exactly equal the square root of the variance component. The final term, known as "irreducible-variance", captures the realized return variation due to unpredictable new information. Under the assumption that $V_t(\epsilon_{i,t+1})$ is constant across the stocks, the squared-bias and variance components wholly determine the cross-sectional variation in squared forecast errors. These components also explain the squared forecast errors' time-series variation.

Remark-1: Ex-post squared forecast errors of risk premium predictions based on simple linear models are challenging to predict ex-ante. Such models comprise few parameters and thus yield small predictive standard errors. However, they are grossly misspecified when the true risk premia are non-linear functions of many predictors. Hence, squared-bias rather than variance largely governs their forecast-squared errors. Because true risk premia are unobserved, ex-ante measurement of squared-bias is not possible, rendering simple models' forecast-squared errors unpredictable ex-ante.

Remark-2: In contrast, ex-post forecast errors of ML-based predictions are ex-ante predictable. These predictions use many predictors and parameters and thus are less likely to be misspecified. However, their massive predictive standard errors, which reflect parameter uncertainty, predominantly determine their forecast-squared errors. These standard errors, unlike biases, are readily obtainable, rendering ML models' forecast-squared errors predictable ex-ante. For instance, in the cross-section, stocks whose ML-based risk premium predictions have large ex-ante standard errors also have large ex-post squared forecast errors.

Consistent with these remarks, the empirical section documents that the ex-ante predictive standard errors of the NN-based risk premium predictions strikingly predict their ex-post squared forecast errors, whereas those of the Lewellen-based predictions do not. The following subsection illustrates how these ex-ante standard errors could be used in real-time to form desirable investment portfolios that yield large economic gains.

B. Risk Premium Predictions, Standard Errors and Investment Portfolios

This subsection introduces the Confident-HL portfolios that deliberately exclude or downweight stocks with large predictive standard errors from the extreme predicted-return decile stocks. I

restate the example provided in the introduction to illustrate why these portfolios yield superior expected returns relative to the conventional HL strategies.

Example-1. Consider two stocks A and B with true risk premia μ_A and μ_B ($< \mu_A$), respectively. Let $\hat{\mu}_A$ and $\hat{\mu}_B$ be the predicted risk premia based on an econometric model, satisfying

$$\hat{\mu}_A = \mu_A + \epsilon_A,\ \hat{\mu}_B = \mu_B + \epsilon_B,\ \epsilon_A, \epsilon_B \sim N(0, \sigma^2),\ \epsilon_A \perp \epsilon_B. \tag{8}$$

Recall that the assumption of unbiased predictions ($E(\epsilon_A), E(\epsilon_B) = 0$) is more likely to hold for ML-based rather than traditional linear models. For simplicity, (8) assumes uncorrelated predictions with the same predictive standard error, σ. Proposition-1 relaxes this assumption and generalizes for heteroskedastic standard errors.

The expected return of the traditional HL portfolio that goes long (short) on the stock with the highest predicted risk premium is then given by

$$E(HL) = (\mu_A - \mu_B)P(\hat{\mu}_A > \hat{\mu}_B) + (\mu_B - \mu_A)P(\hat{\mu}_B > \hat{\mu}_A) = (\mu_A - \mu_B)\left[2\Phi\left(\frac{\mu_A - \mu_B}{\sqrt{2}\sigma}\right) - 1\right], \tag{9}$$

where $P(.)$, $\Phi(.)$ denote the probability and standard normal distribution measures, respectively.

Thus, (9) indicates that the expected HL return monotonically increases (decreases) with the precision of risk premium predictions (σ). Mathematically, the prediction uncertainty induces bias to the maximum possible expected HL return that can be obtained when true risk premia are known. For example, the HL strategy formed from the zero standard error predictions delivers the maximum possible expected return of $(\mu_A - \mu_B)$, as the strategy always takes the long (short) position on A (B) by perfectly ranking the stocks. In contrast, the HL portfolio formed from grossly imprecise predictions ($\sigma = \infty$) earns zero expected returns, with a bias of $(\mu_A - \mu_B)$. This result follows from Jensen's inequality: " The expectations of the maximum (minimum) of a given set of risk premium predictions are lower (higher) than the maximum (minimum) of the expectations of predicted risk predicted risk premia". The lower the variance of risk premium predictions, lower will be the difference between both.

The following proposition builds on this intuition and formally establishes the Confident-HL strategies' superiority over the conventional HL portfolios.

Consider four stocks A_1, A_2, B_1, and B_2 with true risk premia μ_A, μ_A, μ_B ($< \mu_A$), and μ_B, respectively. Predictions are unbiased, independent, and normal, but could have different predictive standard errors. To form trading strategies, stocks are sorted into two quantiles, denoted by Q_S and Q_L. Q_L (Q_S) comprises the two stocks with the highest (lowest) risk premium predictions. Now, consider the following three long-short investment strategies:

1. **HL**: The traditional HL strategy takes the EW long (short) positions on the 2 Q_L (Q_S) stocks.

2. **PW-HL**: The "precision-weighted" (PW) HL portfolio also takes the long (short) positions on the two Q_L (Q_S) stocks, but overweights ($> 50\%$) the precisely predicted stock in each quantile.

3. **Confident-HL**: This strategy takes the long (short) position only on the stock with the lowest predictive standard error in each quantile, deliberately excluding the stock with imprecise risk premium.

Then, the expected returns of these portfolios are in the order of

Proposition 1:

$$E(\text{HL}) \leq E(\text{PW-HL}) \leq E(\text{Confident-HL}). \tag{10}$$

Proof. See Appendix (A.1). □

The proof is similar to the previous example. Thus, proposition-1 indicates that the Confident-HL portfolios dominate the traditional HL portfolios in terms of earning higher expected returns. Proposition-1 makes the stylized assumption of uncorrelated predictions for mathematical tractability, as it is not possible to generalize this result with correlated predictions. However, Internet Appendix C.C1 (table A) presents an extensive simulation study to validate proposition-1 for general cases with many stocks, correlated return predictions and Confident-HL portfolios formed from various other quantile portfolios (e.g., decile).

Consistent with these results, the empirical section documents large economic gains emanating from the Confident-HL portfolios based on the NN-3 risk premium predictions and their standard errors. Such large gains would not be realized from the Lewellen-based Confident-HL portfolios, as their predictive standard errors do not predict their squared forecast errors.

Before deriving NN-based risk premia's predictive standard errors to form the Confident-HL portfolios, it is worth emphasizing a couple of important points. First, dropping stocks with imprecise risk premia improves the expected returns of HL strategies, not necessarily their variance, as it may reduce the diversification benefit. Determining the trade-off between expected HL returns and their variances is ultimately an empirical question. The empirical section shows that the Confident-HL portfolios formed using the standard decile-sorted rules deliver superior Sharpe ratios, suggesting that the expected return improvements are relatively larger. Second, the Confident-HL strategies exploit information only from the variance of risk premium predictions and not predicted return variances nor covariances. Forming optimal portfolios using all stock returns' joint predictive density requires a Bayesian framework, thus left for a future study.

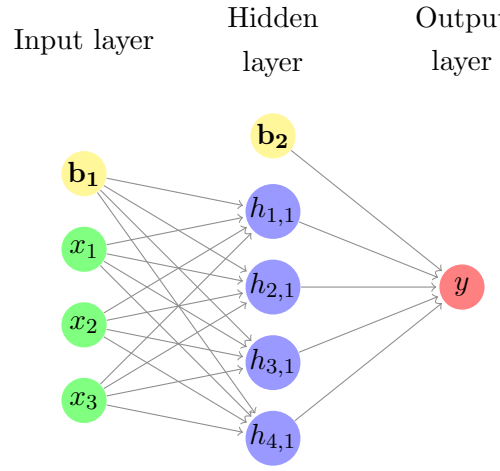
III. NN-based Risk Premia and Standard Errors

This section presents the statistical framework to predict individual stock- and portfolio-level risk premia using NN. It then theoretically derives their standard errors, shown to be easily obtainable with no additional computation cost. In particular, an NN that employs a specific regularization known as "dropout" is identical to a Bayesian NN with a similar structure (Gal and Ghahramani (2016)). A simple analogy to this identity is the equivalence between linear regressions with L_2 regularization (i.e., Ridge regressions) and Bayesian linear regressions. Thus, NN-based predictive standard errors are estimated using the comparable Bayesian models' instantly available posterior variances.

Although Bayesian posterior variances and frequentist standard errors philosophically represent different entities, the section justifies why and how the obtained standard errors satisfy critical frequentist properties with accurate coverage probabilities. This is important, because no frequentist alternative currently exists (to my knowledge) to provide standard errors.

A. Neural Networks

Figure 1. Example of a 1-layer Neural Network



Note: An example of a 1-layer, feed-forward neural network.

Like GKX, this paper considers conventional "feed-forward" NNs, which consist of an "input layer" of raw predictors, one or more "hidden layers" and an "output layer" of a final prediction, in that order. Each layer is composed of neurons that aggregate information from the neurons of (immediately) preceding layer. Thus, information hierarchically flows from the raw predictors of

the input layer to the neurons in the hidden layers and finally to the final prediction in the output layer. To understand how NNs systematically conduct this prediction exercise, figure (1) shows a simple example of a 1-layer NN (NN-1) with 3 and 4 neurons in the input and hidden layers, respectively.

In figure (1), $\{x_1, x_2, x_3\}$, $\{h_{k,1}\}_{k=1}^{4}$, and y are the sets of neurons in the input, hidden, and output layers, respectively. Furthermore, $\{x_i\}_{i=1}^{3}$ are raw individual predictors, and y is the final output prediction. Each neuron in the hidden layer applies a nonlinear function (ϕ) to an aggregate signal received from the preceding (input) layer. The aggregate signal is a weighted sum of the preceding layer's neurons plus an intercept, known as "bias". Thus,

$$h_{k,1} = \phi\left(b_{1k} + \sum_{j=1}^{3} w_{1jk} x_j\right), \text{ for } k = 1, 2, 3, 4, \tag{11}$$

where b_{1k} is the intercept associated with the input (first) layer and k^{th} neuron in the (next) hidden layer, and w_{1jk} is the weight associated with the j^{th} predictor (neuron) in the input layer and the k^{th} neuron in the hidden layer. The linear sum, $(b_{1k} + \sum_{j=1}^{3} w_{1jk} x_j)$, is the aggregated signal received by the hidden layer's $h_{j,1}$ neuron from the input layer. In the spirit of GKX, the nonlinear function ϕ takes the rectified linear unit functional form (ReLU). However, the theory developed in this section holds for any general function. The ReLU is given by

$$\phi(x) = ReLU(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ otherwise.} \end{cases} \tag{12}$$

Likewise, the final output is given by

$$y_{output} = b_2 + \sum_{j=1}^{4} w_{2j} h_{j,1}, \tag{13}$$

where w_{2j} is the weight associated with the j^{th} neuron in the hidden layer and the output. Thus, given an input of Q individual predictors, x, the final prediction, y_{output}, based on a general NN-1 model with K hidden neurons can be expressed in the parametric form

$$y_{output} = b_2 + \phi(b_1 + xW_1)W_2, \tag{14}$$

where $\{W_1, W_2, b_1, b_2\}$ are the unknown parameters. W_1 and W_2 are the weight matrices connecting the imput layer to the hidden layer and hidden layer to the output layer, respectively. Intercepts b_1 and b_2 are added to the hidden and output layers, respectively. W_1 is a $Q \times K$ matrix, W_2 is a $K \times 1$ vector, b_1 is a $K \times 1$ vector, and b_2 is a scalar.

B. Parameter Estimation, Regularization, and Dropout

For simplicity, the rest of the section focuses on NN-1 models. However, the theory that follows holds in general for any feed-forward NN with an arbitrary number of hidden layers and neurons. Consider the return prediction specification in (5),

$$r_{it+1} = f(z_{it}; \beta) + \eta_{i,t+1}, \tag{15}$$

where $r_{i,t+1}$ is stock i's excess return at period $t+1$, and z_{it} is the set of stock i's raw predictors at time t. When f is an NN-1, it takes the parametric form in (14), with $\beta = \{W_1, W_2, b_1, b_2\}$.

Because the parameters are unknown, risk premia are measured as $E_t(r_{i,t+1}) \approx f(z_{it}; \hat{\beta})$, where $\hat{\beta}$ are estimated parameters of β. Given a panel of "training data", the literature typically minimizes the mean of squared forecast errors to estimate the parameters, i.e

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{N_{Tr} N_S} \sum_{t \in Tr} \sum_{i \in S} \left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it} W_1) W_2)\right)^2, \tag{16}$$

where Tr is the training sample over N_{Tr} periods, and S is the total set of N_S stocks. The estimated parameters from (16) often overfit the data by taking extreme values. To alleviate this concern, the literature adds various penalties such as L_2 regularization to the usual squared forecast error loss function. Under L_2 regularization, the estimated parameters are given by

$$\begin{aligned} \hat{\beta}_\lambda = \arg\min_{\beta} \frac{1}{N_{Tr} N_S} \sum_{t \in Tr} \sum_{i \in S} & \left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it} W_1) W_2)\right)^2 \\ & + \lambda \left[||W_1||^2 + ||W_2||^2 + ||b_1||^2 + ||b_2||^2\right], \end{aligned} \tag{17}$$

where $||.||$ represents the L_2 norm operator, and λ is known as the "hyperparameter". Note that the estimated parameters depend on the hyperparameter λ. From a given set of hyperparameters, the standard practice chooses the λ that minimizes the forecast-squared error mean in a panel of "validation data" that do not overlap with the training data. In particular,

$$\lambda = \arg\min_{\lambda \in \Lambda} \frac{1}{N_V N_S} \sum_{t \in V} \sum_{i \in S} \left(r_{i,t+1} - f(z_{it}, \hat{\beta}_\lambda)\right)^2, \tag{18}$$

where V is the validation sample over N_V periods, and Λ is a given set of hyperparameters.

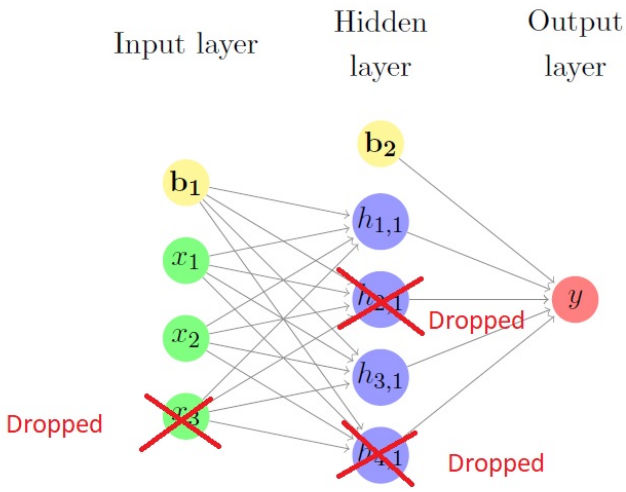
Thus, (17) and (18) together determine the estimated parameters and hyperparameters. Because the optimal parameters that minimize (17) are not available in closed-forms, numerical algorithms start with an initial estimate (guess), and then iteratively update the parameters by feeding each observation into the training data one-by-one. This procedure could be computationally intensive. Thus, a popular algorithm known as stochastic gradient descent (SGD) considers random

samples (rather than the full sample) from the training data to iteratively update the parameters until they converge.[9]

Besides L_2, GKX use several other regularizations, such as L_1, to minimize overfitting. This subsection introduces another popular regularization known as dropout that can be employed either exclusively or simultaneously with other penalties. Dropout stands out among others because it boosts the performance of NN models and helps determine predictive standard errors. GKX do not discuss the dropout procedure. In a recent working paper, Chen et al. (2020) use dropout to fit various NNs for predicting stock returns. However, they do not address how such a regularization could be exploited to obtain predictive standard errors.

Dropout. Dropout is a simple but powerful regularizations proposed by Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014).[10]

Figure 2. NN-1 with Dropout Regularization



Note: The figure shows an NN-1 with dropout regularization. At each training iteration, a random subset of all neurons in one or more layers, including the input layer, but always excluding the output layer, is dropped. Each iteration's dropped out neurons temporarily output 0 (during that iteration), but might become active in the next iteration.

At each training iteration during parameter estimation, every neuron, including the input neurons, but always excluding the output neurons, has a probability $(1-p)$ of being temporarily dropped. These dropped out neurons are deliberately set to output 0 (equivalently, discarded) during that iteration but are allowed to become active in the next iteration. Like λ for L_2, $(1-p)$

[9] See GKX for a detailed review of parameter estimation using SGD.
[10] See Géron (2019) for an excellent non-technical summary on dropout regularization.

(p) is a hyperparameter known as "dropout rate" ("retention rate"), and thus chosen (typically between 10% and 50%) to minimize the validation forecast-squared error. After training and obtaining estimated parameters, neurons are no longer dropped (i.e., to make a new prediction). Figure (2) shows an example of an NN-1 with dropout regularization.

To summarize, during parameter estimation, dropout randomly disconnects a few neurons at each iteration to avoid overfitting and improves performance. Consider a random sample of 1000 observations from training data for parameter estimation. The SGD algorithm takes 1000 iterations to estimate the parameters. Employing dropout would imply 1000 different NNs are trained, yielding 1000 distinct estimated weights. These weights are not independent but are nevertheless all different. The final estimated weights could be interpreted as an average of these distinct weights, thereby alleviating parameter uncertainty.

Estimated parameters of an NN-1 that employ dropout and L_2 regularizations satisfy

$$\hat{\beta}_{\lambda,p} = \arg\min_{\beta} \frac{1}{N_{Tr}N_S} \sum_{t\in Tr}\sum_{i\in S} (r_{i,t+1} - (b_2 + \phi(b_1 + z_{it}(p_{1it}W_1))(p_{2it}W_2)))^2 + \lambda \left[||W_1||^2 + ||W_2||^2 + ||b_1||^2 + ||b_2||^2\right], \tag{19}$$

where each element in p_{1it} and p_{2it} is an independent draw from a *Bernoulli* distribution with parameter (p) ((1-dropout rate)). p_{1it} and p_{2it} are $(Q \times Q)$ and $(K \times K)$ diagonal matrices, respectively. Thus, unknown parameters could be estimated by solving (19).[11] Hereafter, an NN that employs L_2 and dropout regularizations will be called a "dropout NN".

Stock-level risk premia. Given newly observed "test data" (Te) of raw predictors that do not overlap with the training and validation data sets, a dropout NN-1-based risk premium prediction is given by

$$E_t(r^*_{i,t+1}) \approx E^*_{it,Dropout} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}W_{1,\{\lambda,p\}})W_{2,\{\lambda,p\}}),\ r^*_{i,t+1}, z^*_{it} \in Te, \tag{20}$$

where the parameters, $\{b_{2,\{\lambda,p\}}, b_{1,\{\lambda,p\}}, W_{1,\{\lambda,p\}}, W_{2,\{\lambda,p\}}\}$, are given in (19). $E^*_{it,Dropout}$ represents the dropout NN-1-based risk premium prediction of stock i at period t. Note that no neurons are dropped out while making predictions on the test data. However, these predictions rely on estimated parameters that employ dropout regularization. In fact, Srivastava et al. (2014) establish that the predictions given in (20) are approximately equal to the sample averages of corresponding predictions that employ dropout at the test time as well. In particular,

$$E^*_{it,Dropout} \approx \frac{1}{D}\sum_{d=1}^{D}(b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1id}W_{1,\{\lambda,p\}}))(p_{2id}W_{2,\{\lambda,p\}})),\ r^*_{i,t+1}, z^*_{it} \in Te, \tag{21}$$

[11] The most commonly used software programs, including Python and Matlab, readily solve (19).

where each element in $\{p_{1i,d}, p_{2i,d}\}_{i=1}^{D}$ is an independent draw from $\sim Bernoulli(p)$, and D is the total number of distinct predictions drawn at the test time with dropout applied.

Portfolio-level risk premia. The risk premium prediction, $E^*_{Pt,Dropout}$, of portfolio P formed using a set of stock-level weights $\{\omega_{P,i,t}\}_{i=1}^{S}$ at the beginning of period $t+1$ is given by

$$E_t(r^*_{P,t+1}) = \sum_{i=1}^{S} \omega_{P,i,t} r^*_{i,t+1} \approx E^*_{Pt,Dropout} \approx \sum_{i=1}^{S} \omega_{P,i,t} E^*_{it,Dropout},\ r^*_{i,t+1} \in Te, \tag{22}$$

where $E^*_{it,Dropout}$ is given in (21).

Importantly, it turns out that the risk premium estimates in (20) (or (21)) and (22) are approximately equal to the respective risk premia's posterior density means under an equivalent Bayesian NN with a similar structure. Using this approximation but before formally discussing Bayesian NNs, the following subsection illustrates how to instantly obtain standard errors of general dropout NN-based risk premium predictions.

C. Standard Errors of Risk Premium Predictions based on Neural Networks

Stock-level standard errors. Given a new observation of a stock's raw predictors z^*_{it} in the test data, consider its risk premium prediction based on a dropout NN-1

$$E_t(r^*_{i,t+1}) \approx E^*_{it,Dropout} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it} W_{1,\{\lambda,p\}}) W_{2,\{\lambda,p\}}),\ r_{i,t+1}, z^*_{it} \in Te. \tag{23}$$

Then the predictive standard error of $E^*_{it,Dropout}$ is estimated by the sample standard deviation of distinct predictions that are obtained by randomly dropping out neurons (with probability $(1-p)$) at the test (prediction) time. In particular,

$$SE_t(E^*_{it,Dropout}) = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t+1} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t+1}\right)^2}, \tag{24}$$

where D is the total number of distinct predictions ($\hat{E}_{i,d,t}$) drawn, with each $\hat{E}_{i,d,t}$ given by

$$\hat{E}_{i,d,t} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1d} W_{1,\{\lambda,p\}}))(p_{2d} W_{2,\{\lambda,p\}})),\ z^*_{it} \in Te. \tag{25}$$

Every element in $p_{1,d}$, $p_{2,d}$ is an *iid* draw from the $Bernoulli(p)$ distribution. The empirical section considers $D = 100$ to estimate the standard errors, as simulations confirm that it yields well-calibrated estimates.[12]

To summarize, after estimating an NN-1 model's weights using the training and validation

[12] The higher D is, the more accurate uncertainty estimates will be. However, inference time also increases with D. Thus, an ideal D trades-off between latency and accuracy.

data sets, standard errors of risk premium predictions on the test data are quickly available by collecting predictions that deliberately assign 0 to randomly selected weights. Intuitively, as the following subsection shows, this procedure is equivalent to drawing samples from the risk premium's predictive distribution based on a comparable Bayesian NN having the same number of neurons and hidden layers as the considered NN-1.

Portfolio-level standard errors. Likewise, the predictive standard error of a portfolio-level prediction is given by

$$SE_t(E^*_{Pt,Dropout}) = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2}, \tag{26}$$

where

$$\hat{E}_{P,d,t} = \sum_{i=1}^{S}\omega_{P,i,t}\left(b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1d}W_{1,\{\lambda,p\}}))(p_{2d}W_{2,\{\lambda,p\}})\right),\ z^*_{it} \in Te, \tag{27}$$

and $p_{1,d}$, $p_{2,d}$ are iid draws from $Bernoulli(p)$.

The procedure for computing portfolio-level standard errors deserves emphasis. Note that the dropped weights (i.e., p_{1d}, p_{2d} draws) are the *same* across the stocks that composite P, thereby preserving correlations among stock-level risk premium predictions to yield unbiased standard error estimates, as shown in the following subsection.

The outlined procedure for obtaining standard errors in (24) and (26) generally applies to all predictions based on NNs with an arbitrary number of layers and neurons as long as their weights are estimated using dropout and L_2 regularizations (Gal and Ghahramani (2016)). The procedure is also robust to adding more regularizations, such as implementing the SGD algorithm with an arbitrary learning rate.

It is worth emphasizing that (24) and (26) yield standard errors of risk premium predictions and not excess return predictions. Fama and French (1997) and Pástor and Stambaugh (1999) also compute risk premium estimates' standard errors. Recall that realized excess returns equal the sum of risk premia and unexpected returns due to unpredictable new information. Thus, their predictive variances equal the sum of predictive variances of risk premium predictions and "irreducible-variance" due to unexpected returns (see (7)). The validation data's mean squared error is an asymptotically unbiased estimate of irreducible-variance (Zhu and Laptev (2017)). Thus, predictive variances of return predictions are easily obtainable as well.

D. Dropout Neural Networks and Bayesian Interpretation

This subsection illustrates a profound connection between dropout NNs and Bayesian NNs to formally validate the previously presented standard errors under a Bayesian framework.

In an influential work, Gal and Ghahramani (2016) first prove that dropout NNs have a Bayesian interpretation. In doing so, they draw upon the probability theory of Gaussian processes, thereby limiting the potential audience for their work. Moreover, they show how to estimate standard errors of *individual* NN-based "raw" predictions (analogous to return predictions) but not those of "prediction means" (equivalent to risk premium predictions). They also do not discuss how to obtain "joint densities" of different NN-based predictions, which are necessary to compute portfolio-level standard errors.

I use a simple Bayesian model to provide a straightforward but rigorous discussion of their central conclusions. In a significant contribution, I (Bayesian) theoretically derive the standard errors (24, 26) of stock and portfolio-level *risk premia*.

Bayesian Neural Network. Consider the Bayesian NN analogous to the previously considered NN-1, with the parametric form given by

$$r_{i,t+1} = b_2 + \phi(b_1 + z_{it}W_1)W_2 + \eta_{i,t+1}, \;\; E_t(\eta^2_{i,t+1}) = \sigma^2_\eta \tag{28}$$

where the parameters $\{W_1, W_2\}$ are random. σ^2_η and $b = (\{b_1, b_2\})$ are assumed to be known for simplicity.[13] Denote the risk premia by μ_{it}, where

$$\mu_{i,t} = E_t(r_{it+1}) = b_2 + \phi(b_1 + z_{it}W_1)W_2. \tag{29}$$

Specify the unknown weight matrices with the standard Gaussian priors,

$$[W_1, W_2] = \mathcal{N}(0, l^{-2}I),$$

where I is the $(NK + K) \times (NK + K)$ identity matrix, and l is a hyperparameter. Then the predictive density of stock i's risk premium given a set of its raw predictors, z^*_{it}, from the test data, and the past training and validation data sets, denoted by $\{R, Z\}$, is given by

$$P(\mu^*_{i,t}|z^*_{it}, R, Z) = \int P(\mu^*_{i,t}|z^*_{it}, R, Z, W_1, W_2, b, \sigma^2_\eta)P(W_1, W_2|R, Z, b, \sigma^2_\eta)dW_1dW_2, \tag{30}$$

where $P(W_1, W_2|R, Z, b, \sigma^2_\eta)$ is the posterior density of the weight matrices given past data. Because this density is not available in a closed-form, the literature often uses one of the powerful methods known as variational inference (VI) to directly approximate the intractable posterior.

[13] The theory generalizes when $\{b_1, b_2\}$ are allowed to be random as well, in which case these parameters could be specified with Gaussian priors.

The following discussion introduces VI and shows that approximating the posterior of the weight matrices using VI and frequentist estimation the weights with dropout and L_2 regularizations, as in (17), are equivalent. Thus, dropout NNs are approximations to Bayesian NNs.

Variational Inference (VI). To approximate a given posterior density $P(W|data)$, VI first considers a family of some known densities, $\{q_\theta(W)\}$, parameterized by θ, and then finds the optimal parameter, θ^*, such that the Kullback-Leibler divergence between $q_{\theta^*}(W)$ and the true posterior density is minimized. Thus, VI approximates the true posterior with $q_{\theta^*}(W)$, where the optimal parameter θ^* would be a function of data. The key is to consider a "good" family of densities that guarantee the (almost surely) convergence of $q_{\theta^*}(W)$ to the true posterior.[14] As a reference, in the finance literature, Allena and Chordia (2020) develop the first VI method to approximate the intractable posterior density of true stock liquidity and equilibrium prices.

Variational Inference for Bayesian Neural Networks. Gal and Ghahramani (2016) consider the following family of independent Gaussian mixture densities to approximate the posterior of the NN weight matrices

$$q_{\{M_1,M_2\}}(W_1,W_2) = q_{M_1}(W_1)q_{M_2}(W_2), \text{ with } q_{M_i}(W_i) = \prod_{k=1}^{K_i} q_{m_{iq}}(w_{iq}), \text{ for } i=1,2, \text{ where}$$
$$q_{m_{iq}}(w_{iq}) = p\mathcal{N}(m_{iq},\sigma^2 I) + (1-p)\mathcal{N}(0,\sigma^2 I) \text{ for } i=1,2, \tag{31}$$

with $M_1 = [(m_{1q})]$ and $M_2 = [(m_{2q})]$. These are the "variational" parameters to be optimized. Also, $W_1 = [(w_{1q})]$ and $W_2 = [(w_{2q})]$. σ^2 and p are known scalars. K_i is the number of neurons in the i^{th} layer. Thus, $K_1 = Q$ and $K_2 = K$. M_1 and M_2 are matrices with the same dimensions as W_1 and W_2, respectively.

The optimal set of parameters $\{M_1^*, M_2^*\}$ that best approximate the true posterior is given by

$$\{M_1^*, M_2^*\} = \arg\min_{\{M_1,M_2\}} KL\left(q_{M_1}(W_1)q_{M_2}(W_2)||P(W_1,W_2|R,Z_b,\sigma_\eta^2)\right), \tag{32}$$

where $KL(x||y)$ represents the Kullback-Leibler divergence between the two random variables, x and y.

Bayesian and Dropout Neural Network Equivalence. Interestingly, given the sample of training data, it turns out that the optimal parameters in (32) minimize the loss function that

[14] See Blei, Kucukelbir, and McAuliffe (2017) for an excellent review of VI. They address two fundamental questions: i) what family of densities to consider? ii) how to obtain the optimal density in the family that best approximates the true posterior?

resembles a dropout NN's loss function, as in (19). In particular,

$$\{M_1^*, M_2^*\} = \arg \min_{\{M_1, M_2\}} \frac{1}{N_{Tr} N_S} \sum_{t \in Tr} \sum_{i \in S} (r_{i,t+1} - (b_2 + \phi(b_1 + z_{it}(p_{1it} M_1))(p_{2it} M_2)))^2 + \mu_1 ||M_1||^2 + \mu_2 ||M_2||^2 + \mu_3 ||b_1||^2 + \mu_4 ||b_2||^2, \quad (33)$$

where each element in p_{1it} and p_{2it} is an independent draw from a *Bernoulli* distribution with parameter (p). $\{\mu_1, \dots \mu_4\}$ are different scalars that are distinct functions of $\{l_1, \sigma_\eta^2, \sigma^2\}$.

Thus, for an appropriate choice of l_1, the variational parameters, $\{M_1^*, M_2^*\}$, that best approximate the (Bayesian) NN weight matrices' posterior density are identical to the comparable (frequentist) dropout NN's estimated weights. This implies

$$M_1^* = W_{1,\{\lambda,p\}}, \text{ and } M_2^* = W_{2,\{\lambda,p\}}. \quad (34)$$

The predictive density of a risk premium given in (29) can be approximated by

$$P(\mu_{i,t}^* | z_{it}^*, R, Z) \approx Q(\mu_{i,t}^* | z_{it}^*, R, Z) = \int P(\mu_{i,t}^* | z_{it}^*, R, Z, W_1, W_2, b, \sigma_\eta^2) q_{M_1^*, M_2^*}(W_1, W_2) dW_1 dW_2, \quad (35)$$

where $\{M_1^*, M_2^*\}$ are given in (34), and $q(.)$ in (31).

As an immediate corollary, (35) implies that the mean of a risk premium's (approximated) Bayesian predictive density is

$$E\left[Q(\mu_{i,t}^* | z_{it}^*, R, Z)\right] \approx E_{it,Dropout}^* \approx \frac{1}{D} \sum_{d=1}^{D} (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z_{it}^*(p_{1id} W_{1,\{\lambda,p\}}))(p_{2id} W_{2,\{\lambda,p\}})), \; z_{it}^*, \in Te, \quad (36)$$

where each element in $p_{1id}, p_{2id} \sim Bernoulli(p)$.

Thus, the mean of a risk premium's Bayesian predictive density (36) precisely matches with the comparable dropout NN-based risk premium prediction, as in (21). In simpler words, predicting risk premia using dropout NNs and Bayesian NNs are equivalent.

Bayesian Justification for Stock-level Standard Errors. Due to (36), under usual regularity conditions (e.g., prior mass is not concentrated at a single point), and for large data, the standard deviation of a risk premium's Bayesian predictive density should proxy for the standard

error of its frequentist counterpart ($E^*_{it,Dropout}$).[15] This implies

$$SE_t(E^*_{it,Dropout}) = SD\left[Q(\mu^*_{i,t}|z^*_{it}, R, Z)\right], \tag{37}$$

where $SD\left[Q(\mu^*_{i,t}|z^*_{it}, R, Z)\right]$ represents the standard deviation of $\mu^*_{i,t}$'s Bayesian predictive density. This is given by the following proposition.

Proposition 2:

$$SD\left[Q(\mu^*_{i,t}|z^*_{it}, R, Z)\right] \approx \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2}, \tag{38}$$

where $\hat{E}_{i,d,t}$ *is given in* (25).

Proof. See Appendix (A.2). □

Thus, the standard errors of dropout NN-based stock-level risk premium predictions, as in (24), are justified from a Bayesian standpoint.

Bayesian Justification for Portfolio-level Standard Errors. Likewise, the standard error of a portfolio P's risk premium prediction should satisfy

$$SE_t(E^*_{Pt,Dropout}) = SD\left[Q(\mu^*_{P,t}|\{z^*_{it}\}_{i=1}^S, R, Z)\right], \tag{39}$$

where $\mu^*_{P,t} = \sum_{i=1}^S \omega_{P,i,t}\mu^*_{i,t+1}$, and $Q(\mu^*_{P,t}|\{z^*_{it}\}_{i=1}^S, R, Z)$ is the Bayesian predictive density of P's risk premium, given a set of stock-level characteristics.

Obtaining this density is not straightforward, as it involves computing the *joint* predictive density of risk premia of all stocks that compose P, $Q\left(\mu^*_{1,t}, \mu^*_{2,t}, \ldots, \mu^*_{S,t}|\{z^*_{it}\}_{i=1}^S, R, Z\right)$. The following proposition formally derives the joint density to compute the standard deviation of P's posterior risk premium density.

Proposition 3:

$$SD\left[Q(\mu^*_{P,t}|\{z^*_{it}\}_{i=1}^S, R, Z)\right] \approx \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2}, \tag{40}$$

where $\hat{E}_{P,d,t}$ *are given in* (27).

[15] This property, known as "frequentist consistency" of posteriors, is due to the Bernstein-von Mises theorem. Whereas literature often demonstrates this result for true posteriors, Wang and Blei (2019) establish that, under standard regularity conditions, approximated posteriors using VI are consistent as well. In any case, the following subsection empirically validates this result.

Proof. See Appendix (A.3). □

Thus, the standard errors of dropout NN-based portfolio-level risk premium predictions, as in (26), are theoretically justified as well.

E. Frequentist Justification for Standard Errors

Recall that the paper trades the Bayesian standard errors for the frequentist standard errors, as the former are instantly available but no valid method exists to compute the latter directly (to my knowledge). This subsection justifies the obtained standard errors from a frequentist standpoint, by drawing the equivalence between both standard errors and conducting extensive Monte-Carlo simulations.

Under a large sample, observed data dominates prior, rendering Bayesian and frequentist standard errors nearly identical (see result 8 in section 4.7 of Berger (1985)). However, NN-based predictions generally employ substantial regularization, which is equivalent to starting with proper priors. In such cases, data may not always dominate prior, resulting in differences between the Bayesian and frequentist approaches under specific parameters. However, such issues typically occur at the atypical values of the parameters, such as when they approach infinity (Kyung, Gill, Ghosh, and Casella (2010)).[16] Thus, for a wide range of parameters, Bayesian and frequentist standard errors should be equivalent.

Consistent with this result, an extensive simulation study in appendix B.1 (table (I)) confirms that the proposed standard errors are well-calibrated in the frequentist sense. Using a high dimensional predictor set, risk premia are simulated from four different data generating processes. Whereas the first two model returns as a linear function of predictors with homoscedastic and correlated residuals, respectively, the last two entertain non-linear functions. Across all models, 95% (or any x% with $0 < x < 100$) confidence intervals constructed from risk premium predictions and their standard errors cover the true simulated risk premia with nearly 95% (x%) probability.

IV. Ex-ante Estimation Uncertainty and Ex-post OOS Inferences

Recall that sections II and III showed how to estimate valid standard errors of NN-based risk premium predictions and exploit them to construct desirable Confident-HL portfolios. This section derives a formal method to assess the Confident-HL or any model-based investment portfolios' ex-post OOS performance.

In doing so, the section first documents that existing tests violate the central assumption

[16] In fact, Kyung et al. (2010) motivate the same to compute the otherwise intractable standard errors of LASSO based predictions using their Bayesian counterparts.

required for the DM tests' asymptotic validity. The section then presents a bootstrap methodology to deliver valid OOS comparisons in the presence of estimation uncertainty. The section concludes by showing that the method yields well-sized tests, whereas the DM tests lead to significant size distortions using simulated data.

A. Out-of-Sample Comparisons with the Diebold and Mariano (2002) Tests

OOS returns of HL strategies and DM tests. Consider any two competing model-based HL strategies, HL_1 and HL_2. These portfolio returns could be expressed as different weighted sums of excess returns, depending on which stocks comprise their long and short legs. Thus,

$$HL_{1t} = \sum_{i \in S} \hat{w}_{1,i,t-1} r_{i,t}, \;\; HL_{2t} = \sum_{i \in S} \hat{w}_{2,i,t-1} r_{i,t}, \tag{41}$$

where $r_{i,t}$ denotes the excess return of stock i at period t, and $\{\hat{w}_{1,i,t-1}\}_{i \in S}$ and $\{\hat{w}_{2,i,t-1}\}_{i \in S}$ represent the weights with which individual stocks compose the HL_1 and HL_2 portfolios, respectively. The weights are estimated using all data until $t-1$. This specification is consistent with the "recursive estimation scheme" typically employed by researchers (e.g., GKX, Bianchi et al. (2020)).

Consider the return differentials over the OOS period,

$$d_{12,t} = HL_{1t} - HL_{2t}, \;\; t \in Te, \tag{42}$$

where Te denotes the OOS test period. Then the DM statistic to test the null of equal return means, $H_0 : E(d_{12,t} = 0) \; \forall t$, is a simple t-ratio given by

$$DM_{HL} = \frac{\bar{d}_{12}}{\hat{\sigma}_{d_{12}}} \sim \mathcal{N}(0,1), \tag{43}$$

where $\bar{d}_{12} = \frac{1}{N_{Te}} \sum_{t \in Te} d_{12,t}$ is the sample average of return differentials over N_{Te} OOS periods and $\hat{\sigma}_{d_{12}}$ is a heteroskedastic and autocorrelation robust standard error estimate for $\bar{d}_{12}$. Whereas Avramov et al. (2020) use Newey-West standard errors of return differentials as a proxy for $\hat{\sigma}_{d_{12}}$, most studies use the standard OLS standard errors.

OOS MSEs and DM tests. Likewise, existing studies employ the DM tests to compare OOS mean squared errors (MSEs) of any two competing models as well. Given two models M_1 and M_2, let $f_1(Z_{i,t-1}; \hat{\beta}_{1,t-1})$, $f_2(Z_{i,t-1}; \hat{\beta}_{2,t-1})$ be the return predictions for period t based on M_1 and M_2, respectively. Then the forecast-squared error differentials over the OOS period are given by

$$D_{12,t} = e^2_{1,t} - e^2_{2,t}, \;\; \text{where } e^2_{k,t} = \frac{1}{N_s} \sum_{i \in S} \Big(r_{i,t} - f_k(Z_{i,t-1}; \hat{\beta}_k) \Big)^2, \;\; k = 1,2, \;\; t \in Te, \tag{44}$$

with each $e_{k,t}^2$ representing the cross-sectional average of forecast-squared errors at period t under M_k, $k = 1, 2$. Like in the previous case, the model parameters $\hat{\beta}_{k,t-1}$ are estimated using all data until $t-1$. Then the DM statistic to test the null of equal predictive ability is given by

$$DM = \frac{\bar{D}_{12}}{\hat{\sigma}_{D_{12}}} \sim \mathcal{N}(0,1), \tag{45}$$

where $\bar{D}_{12} = \frac{1}{N_{Te}} \sum_{t \in Te} D_{12,t}$ is the sample mean of squared forecast error differentials and $\hat{\sigma}_{D_{12}}$ is a heteroskedastic and autocorrelation robust standard error estimate of $\bar{D}_{12}$. GKX use Newey-West standard errors of squared forecast error differentials as a proxy for $\bar{D}_{12}$.[17]

Asymptotic validity of DM tests. DM emphasize that their tests (43, 45) yield asymptotically valid inferences only under the assumption that the loss differentials, $\{d_{12,t}\}\{D_{12,t}\}$, are covariance stationary. Equivalently,

$$E(d_{12,t}) = \mu_1, \ cov(d_{12,t}, d_{12,t-\tau}) = \gamma_1(\tau), \ \forall t, \tau \geq 0, \text{ and} \tag{46}$$

$$E(D_{12,t}) = \mu_2, \ cov(D_{12,t}, D_{12,t-\tau}) = \gamma_2(\tau), \ \forall t, \tau \geq 0. \tag{47}$$

However, this assumption is violated when the parameters, such as $\{\hat{w}_{k,i,t-1}\}_{i \in S}$ and $\hat{\beta}_{k,t-1}$, are estimated from econometric models. Their estimation uncertainties introduce time-varying temporal dependencies between the loss differentials, thereby breaking down the covariance stationarity assumption. A simple intuition demonstrates the central idea.

Recall that $\{\hat{w}_{1,k,t-1}\}_{i \in S}$ are estimated using all data until $t-1$. Thus, their precision (variance) increases (decreases) as time proceeds and more data are available. Consequently, the HL return differentials exhibit time-varying moments and temporal dependencies, rendering the covariance stationarity assumption infeasible.

B. Violation of Covariance Stationarity: Empirical Evidence

Consistent with the previous intuition, appendix B.2 (table II) empirically documents that the loss differentials computed using NN-3 and Lewellen-based return predictions significantly violate the covariance stationarity assumption. This result reaffirms that the existing DM-based conclusions are misleading.

In particular, B.2 conducts covariance stationarity tests proposed by Pagan and Schwert (1990) on three different loss differentials over the 360 OOS months. The first comprises the forecast-squared error differences between the NN-3 and Lewellen-based return predictions. The second

[17] To be precise, the DM tests were original designed for time-series data. GKX adapted these tests on panel data by cross-sectionally averaging the forecast-squared errors at each period, as in (44). In a recent working paper, Timmermann and Zhu (2019) show that this adapted statistic yields asymptotically valid inferences, of course, only under the assumption that there is no parameter uncertainty.

(third) contains the return differences between the EW (VW) HL portfolios based on the NN-3 and Lewellen models.

If these loss differentials were covariance stationary, then each of their sample standard deviations over the first 180 periods should be close to those over the last 180 periods. However, the initial period standard deviations are significantly (5, 1.85, and 1.75 times) higher than those of the final period. Thus, the null of covariance stationarity is rejected across the loss differentials. Also, relatively large beginning period standard deviations may reflect a "recursive estimation scheme", in which case parameter uncertainty decreases as time progresses, when true model parameters are time-invariant.

C. Bootstrap Tests for Out-of-Sample Comparisons

This subsection presents a bootstrap test that accommodates non-stationary loss differentials. The method builds on the moving block bootstrap procedure of Kunsch (1989). Although it was initially designed for stationary processes, Gonçalves and White (2002, 2004) establish their asymptotic validity for non-stationary processes under certain assumptions that govern the degree of non-stationarity.

First, they assume that the mean heterogeneity of the given series is not too strong. The return differentials in (42) satisfy this condition, as their unconditional means are the same.[18] Second, they assume that the series is a near epoch dependent on an underlying mixing process (Billingsley (1999)). This condition is less stringent than "mixing conditions" that researchers, including DM, typically assume to derive limiting distributions. Importantly, near epoch dependent processes allow for considerable heterogeneity (of (co)variances) and also for dependence. Thus, their assumptions suit this paper's framework.

Why bootstrap works. Recall that the DM tests make two parametric approximations. The tests use heteroskedastic and autocorrelation standard errors and draw critical values from the standard normal. Such approximations likely fail under complex scenarios (e.g., when the series is not stationary). However, bootstrap-based tests do not make such parametric simplifications and thus likely yield valid asymptotic inferences even in challenging situations. Of course, even bootstrap could fail under certain circumstances (see section 4.5 from Horowitz (2001)). Thus, the literature recommends complementary simulation checks, as described in the next subsection.

I now explicitly discuss how to conduct bootstrap-based OOS inferences.

[18] It is less clear whether forecast-squared-error differentials theoretically have the same unconditional means. However, empirical tests suggest that the null of equal means over different periods do not get rejected. This result supports the assumption laid out by Gonçalves and White (2002).

C.1. Tests of equal return means or forecast-squared errors.

Consider a series of loss differentials $\{\Delta_t\}_{t=1}^T$. These could be either HL return ($d_{12,t}$) or squared forecast error differentials ($D_{12,t}$). Then the procedure for obtaining critical values, or p-values, under the null hypothesis $H_0 : E(\frac{1}{T}\sum_{t=1}^T \Delta_t) = 0$ is as follows.

1. Choose a block-size l. For each iteration i,
   (a) draw $n = (T/l)$ random numbers, $\{b_i\}_{i=1}^n$, from the set $\{1, 2, \ldots, T-l\}$ with replacement,
   (b) draw a block bootstrap sample D_i=$\{\Delta_{b_1}, \Delta_{b_1+1}, \ldots \Delta_{b_1+l-1};\ \Delta_{b_2}, \Delta_{b_2+1}, \ldots \Delta_{b_2+l-1};$ $\ldots; \Delta_{b_n}, \Delta_{b_n+1}, \ldots \Delta_{b_n+l-1}\}$, where D_i contains a total number of T differentials, and
   (c) impose the null and compute the bootstrap-based t-ratio, $t_i = \left(\bar{D}_i - \bar{\Delta}\right)/std(D_i)$, where $\bar{D}_i$ and $std(D_i)$ are the sample mean and standard deviation of D_i, respectively. $\bar{\Delta}$ is the sample mean of the original loss differentials.
2. Repeat step (1) many times. The generalized p-value equals the proportion of times the absolute value of t_i is greater than the original sample's realized absolute t-ratio, which equals $t = \left(\bar{\Delta}\right)/std(\Delta)$, where $std(\Delta)$ is the sample standard deviation of the loss differentials $\{\Delta_j\}_{j=1}^T$.

The optimal block-size l, shown in the literature to be $O(T^{1/2})$, is close to 2 years of data on a sample over 30 years. Thus, the empirical section uses a block size of 24. However, the results are qualitatively similar across other block lengths of 6, 12, 18, and 36.

C.2. Tests of equal Sharpe ratios.

I further generalize the procedure to compare OOS Sharpe ratios of any two model-based investment strategies. Let $\{HL_{1t}\}$ and $\{HL_{2t}\}$ be two such series, with squared Sharpe ratios

$$Sh_i^2 = \frac{(\frac{1}{T}\sum_{t=1}^T HL_{it})^2}{\frac{1}{T}\sum_{t=1}^T (HL_{it} - \frac{1}{T}\sum_{t=1}^T HL_{it})^2}, \text{ for } i = 1, 2. \tag{48}$$

The p-value for testing the null of equal squared Sharpe ratios, $H_0 : E(Sh_1^2) = E(Sh_2^2)$, can be computed as follows.

1. Choose a block-size l. For each iteration i.
   (a) draw $n = (T/l)$ random numbers, $\{b_i\}_{i=1}^n$, from the set $\{1, 2, \ldots, T-l\}$ with replacement,
   (b) normalize the returns to impose the null,

$$HL_{it}^* = \sqrt{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^T HL_{it})/\sqrt{\sum_{t=1}^T (HL_{it} - \frac{1}{T}\sum_{t=1}^T HL_{it})^2}, \tag{49}$$

(c) draw a block bootstrap sample $\{H_{ki}\}$ from the normalized returns;

$$H_{ki} = \{HL^*_{k,b_1}, HL^*_{k,b_1+1}, \dots HL^*_{k,b_1+l-1}; HL^*_{k,b_2}, HL^*_{k,b_2+1}, \dots HL^*_{k,b_2+l-1}; \dots; HL^*_{k,b_n}, HL^*_{k,b_n+1}, \dots HL^*_{k,b_n+l-1}\} \text{ for } k = 1, 2, \text{ and}$$

(d) compute the bootstrap-based squared Sharpe ratio difference, $Sh^2_{1i} - Sh^2_{2i}$.

$$Sh^2_{ki} = \frac{(\frac{1}{T}\sum_{t=1}^{T} H_{kit})^2}{\frac{1}{T}\sum_{t=1}^{T}(H_{kit} - \frac{1}{T}\sum_{t=1}^{T} H_{kit})^2}, \text{ for } k = 1, 2, \text{ where } H_{kit} = t^{th}\text{element of } H_{ki}.$$

2. Repeat step (1) many times. The p-value equals the proportion of times the absolute value of $(Sh^2_{1i} - Sh^2_{2i})$ is greater than the absolute value of $Sh^2_1 - Sh^2_2$.

D. Performance of the Methodology: Monte Carlo Evidence

Extensive simulations in Appendix B.3 reveal that this paper's bootstrap-based tests are well-sized. In contrast, DM-based tests lead to massive size distortions.

In particular, (B.3) (figure (3)) simulates return time series with zero means under three distinct models, each allowing for a different degree of time-varying temporal dependency. It then conducts the zero return mean tests on the simulated data using three methods that include the DM-test with OLS standard errors, the DM-test with Newey-West standard errors, and this paper's bootstrap method with a block size of 24. Across all simulations, bootstrap-based 5% level tests yield accurate sizes close to 5%. However, DM-based 5% level tests deliver hugely distorted sizes between 13% and 42%, depending on how strong the temporal dependencies are.

Figure (4) shows the power curves for the three methods and confirms that bootstrap-based test "size" refinements come at the expense of only small power losses.

V. Empirical Results

This section presents the main empirical results of the paper. Recall that the theoretical sections imply two central predictions. (1) Ex-ante precision of NN-based risk premium predictions proxy for their ex-post forecast-squared errors, and thus (2) the Confident-HL investment portfolios that deliberately exclude stocks with imprecise risk premium estimates should yield huge OOS economic gains. I empirically demonstrate both of these predictions.

A. Data, Definitions, and Replication Study

A.1. Data

The sample contains monthly excess stock returns of all individual firms listed in the NYSE, AMEX, and NASDAQ exchanges between March of 1957 and December of 2016. The data include 26667 total stocks, with an average of more than 6000 stocks per month. The data also comprise a high-dimensional set of 176 raw predictors examined by GKX and Avramov et al. (2020), including 94 individual stock characteristics analyzed by Green, Hand, and Zhang (2017) (e.g., size, book-to-market, 1-year momentum returns). Another 74 are industry-sector dummy variables based on the first two digits of the Standard Industrial Classification codes. The final eight are aggregate macroeconomic variables used by Goyal and Welch (2008).[19] The Treasury-bill rate proxies for the risk-free rate.

A.2. Models

Neural Network. The paper primarily focuses on a feed-forward NN with three hidden layers (NN-3) and 32, 16, and 18 neurons per layer. This model was examined by GKX and Avramov et al. (2020). I precisely mimic their "recursive scheme" to estimate the model parameters. The scheme first divides the data into 18 years of training (1957-1974), 12 years of validation (1975-1986), and 30 years (1987-2016) of OOS test samples. It then estimates the parameters and hyperparameters using objective functions to minimize the training sample's regularized MSE (17) and the validation sample's MSE (18), respectively. At the end of each year, it re-estimates the model parameters, increasing the training sample by one year. The validation sample rolls forward every year to include the most recent year's data, maintaining the same size (12 years).

I implement this estimation framework to obtain risk premium predictions, as well as their standard errors, over the OOS test sample. Whereas GKX and Avramov et al. (2020) mainly apply L_1 regularization to estimate the parameters, I use dropout and L_2. As discussed in section III, this approach enhances the model's predictive performance and delivers standard errors of predictions. I retain the other hyperparameters (e.g., SGD learning rate, Adam optimization, early-stopping) used by GKX. The Internet appendix tabulates all regularizations with their values.[20]

Lewellen. To compare the economic gains from NN-3-based risk premium predictions and their standard errors with those of simple benchmark models, I also examine one of Lewellen (2015)'s linear models. This Lewellen model predicts stock returns using a pooled regression on

[19] Besides these 176 predictors, GKX and Avramov et al. (2020) also consider (94×8) interactions between the stock characteristics and macroeconomic variables. They do so as they examine several linear models (e.g., Lasso, Instrumented Principal Components) that do not explicitly account for variable interactions. Because NNs automatically capture such interactions, this paper excludes those additional variables.

[20] See GKX for a detailed review of these regularizations.

15 firm-level characteristics (e.g., size, book-to-market, accruals, asset growth ratio). The Internet appendix describes the exact model. This model, unlike NN-3, does not entail regularization. Thus, to make a fair assessment, I estimate the regression parameters using both training and validation data-sets. The OOS test data remain the same.

A.3. Definitions of Performance Metrics

I lay out the definitions of ex-ante and ex-post precision measures that I use repeatedly throughout the rest of the paper.

Ex-ante Confidence. I compute ex-ante confidence of stock-level risk premium predictions using their absolute t-ratios

$$EC_{it} = \frac{|\hat{r}_{i,t+1}|}{se_t(\hat{r}_{i,t+1})}, \tag{50}$$

where EC is ex-ante confidence, $\hat{r}_{i,t+1}$ is the risk premium prediction of stock i at period t (for $t+1$) and $se_t(\hat{r}_{i,t+1})$ is its ex-ante predictive standard error. $|.|$ denotes the absolute value. Ex-ante confidence proxying for a prediction's precision is consistent with the notion that an estimate's standard error must always be understood in the context of the estimate's mean. See section (C.C3) in the internet appendix for a formal discussion using a simple linear model in the spirit of the capital asset pricing model.[21] However, my central conclusions are the same when I use inverse standard errors as proxies for precision. Table B in Appendix C.C1 presents the results.

Whereas I calculate the ex-ante confidence of NN-3-based risk premium predictions using the theory derived in section III, those of Lewellen-based predictions are available in the closed-form expressions. For example, consider a linear regression model $R = Z\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, where R and Z are panels of stock-level returns and characteristics, respectively. Given a stock i's risk premium prediction $z_i\hat{\beta}$, its standard error equals $z_i'(Z'Z)^{-1}z_i\hat{\sigma}^2$, where $\{\hat{\beta}, \hat{\sigma}^2\}$ are the ordinary least squares (OLS) estimates of β and σ^2, respectively. The OLS standard errors are consistent with the model specification of GKX, given in (2).[22]

Ex-post Out-of-Sample-R^2. Given a set of risk premium predictions $\mathcal{S}$, I compute their ex-post OOS R^2 using the following measure motivated by GKX

$$\text{OOS-}R^2 = 1 - \frac{\sum_{(i,t)\in\mathcal{S}}(r_{i,+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{S}} r_{i,t+1}^2}, \tag{51}$$

where $r_{i,t+1}$ is the realized excess return of stock i at period $t+1$.

[21] Recall that proposition-1 makes a highly stylized assumption of invariant risk premia across the stocks in the top (bottom) decile. In more realistic scenarios, this assumption does not hold, in which case I argue that considering the absolute t-ratios as proxies for the precision leads to superior performance relative to the inverse standard errors. See section (C.C3).

[22] Alternatively, I also consider Fama-Macbeth standard errors for Lewellen-based risk premia to account for cross-sectional correlations of residuals. The conclusions are the same.

A.4. Replication of Gu, Kelly, and Xiu (2020)

To ensure that this paper's NN-3-based risk premium measurements are comparable with GKX and Avramov et al. (2020), I replicate their studies. For every period in the OOS test sample, I sort stocks into deciles, decile-1 to decile-10, according to their NN-3-based return predictions for the next month. Decile-1 (decile-10) comprises the bottom (top) 10% of stocks with the lowest (highest) return predictions. Figure 5 (6) presents the EW (VW) average OOS returns and Sharpe ratios of the decile portfolios. All of these monotonically increase from decile-1 through decile-10, thereby confirming that the realized OOS returns align with their predictions. Furthermore, the EW (VW) HL portfolio that takes long-short positions on the extreme decile portfolios (i.e., decile-10 minus decile-1) earns an enormous OOS return of 2.51% (1.47%) and an annualized Sharpe ratio of 1.56 (0.96). These results reflect the success of NN-3 in terms of impressive economic gains. They also qualitatively and quantitatively match with GKX and Avramov et al. (2020), respectively.

Having outlined the data and showing that this paper's NN-3-based return predictions match those of the previous studies, I move on to test the theoretical predictions.

B. Ex-ante Confidence and Ex-post Out-of-Sample-R^2

I first validate remarks 1 and 2 of section II asserting that the ex-ante precision of NN-based risk premium predictions significantly predict their ex-post precision, whereas those of Lewellen-based predictions do not.

Figure 7 confirms this result for NN-3. For every month, I sort stocks into deciles according to their NN-3-based ex-ante confidence. I then calculate the OOS-R^2 attained by these decile subsamples over the 30-year OOS period. Figure 7 reveals that the ex-post OOS-R^2 monotonically increases with the level of ex-ante confidence. For example, the bottom decile, containing stocks most imprecisely predicted by NN-3, attains an OOS-R^2 of 0.81%. In contrast, the top decile with the most confident predictions delivers a much improved OOS-R^2 of 2.21%. This result reinforces that the ex-post precision of NN-based predictions is ex-ante predictable.

Table III further shows that these OOS-R^2 refinements translate into large economic gains. In particular, I construct EW (VW) HL portfolios on each of these confident-decile subsamples, further sorting stocks into deciles according to their next period's (NN-3-based) return predictions. Table III demonstrates that the EW (VW) HL portfolios formed on precise deciles earn remarkably higher OOS returns and Sharpe ratios than those formed on imprecise deciles. For example, the extremely imprecise decile's HL portfolio yields a modest 0.88% (0.34) and 0.71 (0.23) average monthly return and annualized Sharpe ratios, respectively. However, those of the most confident decile's HL portfolio are 3.10% (1.59%) and 1.44 (0.80), respectively, nearly 250% (300%) and 100% (300%) larger than the imprecise decile's counterparts.

Interestingly, the HL portfolios constructed on deciles 9 and 1 have nearly the same average return predictions. However, the average realized OOS return of the relatively more precise decile's EW (VW) HL portfolio, 2.03% (1.16%), is at least twice (thrice) more than that of the imprecise decile, 0.88% (1.16). This result is in the spirit of example-1 in section 2, which shows that between any two sets of stocks with the same risk premium levels, the HL strategy formed on the relatively precisely predicted set has higher expected returns.

Figure 8 repeats the analysis for Lewellen-based predictions and supports the theory as well. Their ex-post OOS-R^2s, unlike NN-based OOS-R^2s, do not monotonically increase with the ex-ante precision. For example, decile 10, containing the stocks with the highest ex-ante precision, has a markedly lower ex-post OOS-R^2 (0.41%) than the OOS-R^2 (0.93%) of decile 7 with relatively lower ex-ante precision. This result is consistent with remark 1, which posits that "bias" rather than "variance" predominantly determines the ex-post precision of a "simple" model-based prediction, rendering it unpredictable ex-ante. Interestingly, though, predictions from the lowest ex-ante precision decile (1) also registers awful ex-post OOS-R^2. The result perhaps reflects the decile's drastically large ex-ante "variances", which dominate average "biases" across other predictions to yield cross-sectionally higher ex-post squared forecast errors.

Overall, ex-post OOS-R^2s of Lewellen-based predictions are not as conspicuously predictable as NN-3-based OOS-R^2s. Consequently, table III indicates that Lewellen-based HL portfolios formed on (Lewellen-based) precise deciles do not earn significantly higher OOS returns than those on imprecise deciles. This result contrasts with the massive economic gains realized by the NN-3 HL portfolios formed on precise rather than imprecise deciles.

To summarize, this subsection demonstrated that the ex-post squared forecast errors of NN-3-based predictions are ex-ante predictable. Before moving on to show how the Confident-HL portfolios exploit this result to yield spectacular economic gains, I first describe the procedure for forming various HL portfolios.

C. Portfolio Construction

1. EW(VW)-HL. These are the conventional HL portfolios. For every month, I sort stocks into deciles according to their next month's return predictions. Let L and H represent the lowest and highest prediction deciles, respectively. Then the EW(VW)-HL portfolios take EW (VW) long and short positions on H and L, respectively.

2. EW(VW)-Confident-HL. These portfolios deliberately drop stocks with imprecise risk premium predictions from the conventional HL portfolios. In particular, both L and H are further partitioned into deciles, $\{L_1, L_2, \ldots, L_{10}\}$, and $\{H_1, H_2, \ldots, H_{10}\}$, based on their ex-ante confidence. Let L_{10} (L_1) and H_{10} (H_1) denote the subsets with the highest (lowest) ex-ante confidence from L and H, respectively. Then the EW(VW)-Confident-HL portfolios take EW (VW) long and

short positions only on the highest ex-ante confident subsets, H_{10} and L_{10}, respectively.

3. EW(VW)-Low-Confident-HL. In contrast, these portfolios take EW (VW) long and short positions on the lowest ex-ante confident subsets, L_1 and H_1, respectively.

4. PW-HL. Rather than completely ignoring low ex-ante confident subsets, the "precision-weighted" strategies disproportionately downweight them while forming portfolios. In particular, the portfolios take long (short) positions on each subset H_j (L_j) with the weights proportional to $1/(11-j)$, for $j = 1, 2, \ldots, 10$. Thus, the higher a subset's precision, the more weight it has.

5. LPW-HL. In contrast, the "low-precision-weighted" portfolios take long (short) positions on each subset H_j (L_j) with the weights proportional to $1/j$.

6. Matching portfolios. To fairly assess the Confident-HL portfolios' performance, I also construct several matching strategies. These portfolios, represented by "HL_{CM}", resemble conventional HL portfolios but are matched to have the same "predicted-return" averages as those of the Confident-HL portfolios. For example, based on NN-3, the EW-Confident-HL portfolio's monthly return predictions average 1.97%. It turns out that a traditional HL strategy that takes EW long (short) positions on the top (bottom) 5% of stocks with the highest (lowest) return forecasts also has an average predicted-return of 1.97%. Thus, this strategy serves as an apt benchmark for EW-Confident-HL. The difference between the two portfolios' ex-post OOS performance precisely captures the economic value of dropping stocks with low ex-ante precision.

In general, I construct the matching portfolios as follows. Every month, EW(VW)-HL_{CM} takes long (short) positions on the top (bottom) x% of the stocks with the highest (lowest) predicted returns for the next month. I choose x so that the time-series average of EW(VW)-HL_{CM} portfolio's predicted return precisely matches that of the EW(VW)-Confident-HL portfolio.[23] Likewise, I construct the "EW(VW)-HL_{LCM}", "LPW-HL_{M}", and "PW-HL_{M}" portfolios to match the average predicted-returns of the EW(VW)-Low-Confident-HL, LPW-HL, and PW-HL, respectively.

7. Double-Sorted portfolios. As an additional robustness check, I consider various double-sorted predicted-return strategies matched to contain the same number of stocks as the Confident-HL portfolios. In particular, I partition the extreme predicted return deciles, L and H, into deciles, $\{L_{1,d}, L_{2,d}, \ldots, L_{10,d}\}$, and $\{H_{1,d}, H_{2,d}, \ldots, H_{10,d}\}$, based on their predicted-returns, respectively. Let $H_{10,d}$ ($L_{10,d}$) and $H_{1,d}$ ($L_{1,d}$) denote the subsets with the highest and lowest predicted returns from H (L), respectively. Then the EW(VW)-double-sorted-HL portfolios take EW (VW) long and short positions on the highest and lowest predicted return subsets, $H_{10,d}$ and $L_{1,d}$, respectively.[24] Despite containing the same number of stocks as the Confident-HL portfolios, these strategies (unlike matching portfolios) do not serve as apt benchmarks for assessing the Confident-HL portfolios' performance because they have higher predicted-returns by construction. Nevertheless, table C

[23] Because x is determined ex-post, the matching portfolios could be interpreted as counterfactual strategies.

[24] Simply, the double-sorted portfolios take long (short) positions on stocks that have predicted risk premia higher (lower) than the top (bottom) 1% of all stocks.

in Internet Appendix C.C1 reports that the Confident-HL portfolios significantly dominate these double-sorted portfolios in terms of Sharpe ratios and information ratios.

D. Economic Gains from Confident-HL Portfolios

I now establish the dominance of the Confident-HL over the conventional HL portfolios.

D.1. OOS Average Returns and Sharpe ratios of Confident-HL Portfolios

Table IV presents the main results. The Confident-HL and precision-weighted (PW-HL) portfolios remarkably outperform the conventional HL portfolios in terms of extensive economic measures. These measures include the OOS average realized returns, Sharpe ratios, as well as abnormal returns (α) and information ratios relative to Fama and French (2015) augmented to the momentum factor (FF-5+UMD) and Stambaugh and Yuan (2017) (SY) models. For example, the traditional EW(VW)-HL portfolio earns an impressive OOS average monthly return of 2.52% (1.48%) and an annualized Sharpe ratio of 1.5 (0.9). However, the EW(VW)-Confident-HL portfolio outperforms this strategy with the same measures of 3.61%(2.21%) and 1.75 (1.09). These are massive 43% (49%) and 17% (21%) increases, respectively. Likewise, the PW-HL also outperforms the EW-HL with an average return and Sharpe ratio of 2.87% and 1.67, respectively.

Note that the matching EW(VW)-HL_{CM} and the EW(VW)-Confident-HL portfolios have the same average NN-3-based predicted-returns. However, the former yields a considerably lower average return and Sharpe ratio than the latter. The 0.54% (0.48%) monthly return difference between the two signifies the economic value of incorporating the ex-ante precision information into forming NN-3-based HL portfolios. In contrast, the Low-Confident-HL and low-precision-weighted (LPW-HL) portfolios containing stocks with imprecise risk premium predictions underperform the traditional HL and Confident-HL portfolios. For example, although the EW(VW)-Low-Confident-HL portfolio has higher average predicted-returns than that of the EW(VW)-HL, it earns a drastically lower average-return and Sharpe ratio. Particularly, the VW-Low-Confident-HL strategy's annualized Sharpe ratio and the FF-6-adjusted and SY-adjusted information ratios are almost or even less than half the corresponding measures of the VW-HL portfolio. This result demonstrates the enormous imprecision of Low-Confident-HL portfolios.

Table IV reveals that the expected returns of the EW-HL, PW-HL, and EW-Confident-HL portfolios are in increasing order, thus validating proposition-1 of section II. Of course, all inferences drawn so far are based on the OOS point-estimates of various economic measures. To establish their statistical significance, I conduct pairwise comparisons using the moving block bootstrap tests developed in section IV.

Table V presents the bootstrap results, and the central conclusions are the same. The OOS

annualized squared-Sharpe and squared-information ratio differences between the Confident-HL and conventional HL portfolios and between the Confident-HL portfolios and their matching HL strategies are significant at the 1% level. Likewise, the corresponding differences between the PW-HL and conventional EW-HL portfolios and between the precision-weighted HL portfolio and its matching HL strategy are also significant at 1%. Even the OOS average return and alpha differences between the Confident-HL and conventional-HL are significant at 1%. Thus, these results statistically validate the superiority of the Confident-HL portfolios.

Similarly, squared Sharpe and squared information ratios of the low-Confident-HL and low-precision-weighted-HL (LPW-HL) portfolios are significantly lower than those of their matching portfolios and the conventional HL portfolios at the 1% level. Interestingly, though, a seemingly large 0.17% monthly average return difference between the EW(VW)-HL and EW(VW)-Low-Confident-HL is statistically insignificant. Because the Low-Confident-HL portfolio returns are excessively imprecise (volatile), zero-mean comparison tests with them perhaps have less "power" to reject the null. However, Sharpe ratio tests vividly indicate the underwhelming performance of the Low-Confident-HL portfolios.

To summarize, the statistical tests distinctly reject the conventional HL portfolios in favor of the Confident-HL portfolios. As mentioned in section IV, the bootstrap tests use a block-size of 24. However, the conclusions are the same for block-lengths of 6, 12, 18, and 36.

D.2. Robustness of Confident-HL Portfolios on Non-Microcaps

In a recent working paper, Avramov et al. (2020) document that NN-3-based HL strategies primarily extract economic gains from microcap stocks. Thus, to investigate the extent to which these stocks drive the Confident-HL portfolio results, I retrain NN-3 on non-microcaps by excluding microcaps.

Table VI presents the portfolios' OOS performance. Table VII shows their statistical significance. Even on the non-microcap subsample, the EW Confident-HL portfolio significantly outperforms comparable alternative HL strategies. For example, the VW-Confident-HL and its matching VW-HL_{CM} have the same average predicted-returns. However, the difference between the former and latter portfolio's average monthly return is a large 0.48% (5.76% at the annual level), which is statistically significant at 5%. Likewise, the former portfolio yields a 15% higher annualized Sharpe ratio (1.00) compared with the latter (0.87), statistically distinct at the 1% level.

D.3. Robustness to Higher-Moment Risks and Transaction-Costs

Higher-Moment Risks. Because NN-3-based HL portfolios are known to display positive skewness and excess kurtosis (Avramov et al. (2020)), I also examine several higher-moment-

adjusted performance measures that reflect the portfolios' downside risk. I consider Omega, Sortio, and upside-potential ratio measures that asymmetrically penalize portfolio losses more than rewarding gains, typically examined by practitioner-researchers as alternatives for Sharpe ratios.[25]

Table VIII presents the results. The Confident-HL and PW-HL handily outperform the conventional HL and equivalent matching portfolios across the higher-order measures. Thus, dropping or downweighing stocks with lower ex-ante precision from an investment portfolio also mitigates its downside risk.

Transaction-Costs. To evaluate whether the economic gains from the Confident-HL portfolios come at the expense of high transaction-costs, I calculate their portfolio turnovers. I find that the Confident HL-portfolios deliver impressive transaction-adjusted returns as well. The "Turnover" column of table VIII shows the portfolio turnovers, representing their average monthly percentage change in holdings. The higher the turnover, the larger the transaction costs. In fact, Avramov et al. (2020) extrapolate that a deduction of (0.005× turnover) from a portfolio's realized return roughly approximates the portfolio's transaction-cost adjusted returns.

The Confident-HL portfolio turnovers, thereby transaction costs, are significantly higher relative to the conventional HL portfolios. This result is expected, as they predominantly take long-short positions on a much smaller subset of stocks, thereby requiring more rebalancing. However, the Confident-HL portfolios' trading-cost adjusted returns are substantially larger than the conventional HL and corresponding matching portfolios. For example, the adjusted returns of the EW(VW)-Confident-HL are 2.68% (1.89%), whereas those of the EW(VW)-HL are much lower, 1.26% (0.79%), respectively.

In summary, I demonstrate that the NN-3-based Confident-HL portfolios statistically outperform the traditional HL counterparts across various economic measures. Plus, these results are robust on non-microcaps and to transaction-costs and higher-moment risks. Now, I compare these portfolios with the benchmark Lewellen-based HL portfolios.

E. Reassessing NN-3 and Lewellen Model Comparisons Using Bootstrap Tests

Recall from section IV that the OOS model comparisons conducted by the existing studies (GKX) using the DM tests are inadequate, as they do not account for estimation uncertainty. This section reevaluates the predictive performance of NN-3 relative to the benchmark Lewellen model using the bootstrap tests. I assess the models' performance in terms of their OOS MSEs and the HL portfolios' average returns and Sharpe ratios.

[25] See the following Wikipedia pages for the definitions of these measures: Omega, Sortino, and up-side potential.

E.1. NN-3 versus Lewellen: Out-of-Sample Mean Squared Error Comparisons

First, I test the null hypothesis that the MSEs of the NN-3 and Lewellen models are equal. Figure 9 presents the p-values computed using the bootstrap tests and the DM tests on various subsamples. In particular, every month, I sort stocks into deciles according to their NN-3-based risk premium predictions' ex-ante confidence, NN-3-EC. The blue line (yellow dotted-line) displays the bootstrap (DM) p-values on the subsamples that dropout 10%, 20%, ..., and 90% of the stocks with the lowest NN-3-EC, respectively. These subsamples contain the forecasts that NN-3 confidently predicts. In contrast, the red line (purple dotted-line) represents the p-values on subsamples comprising the forecasts that NN-3 imprecisely predicts.

Figure 9 reveals that the DM-based p-value is less than 0.01 on the entire OOS data comprising all stocks. Thus, consistent with GKX, the DM test rejects the Lewellen model in favor of NN-3 at the 1% level. However, with a p-value of 3.03%, the bootstrap test does not reject the null at 1% significance. Although the null of equal predictive abilitiy is rejected at the 5% significance in favor of NN-3, the difference between both p-values suggest that the DM-based tests over reject the null.

Interestingly, figure 9 illustrates that the predictive dominance of NN-3 monotonically increases with the level of ex-ante confidence. For example, dropping out 10%, 50%, and 90% of stocks with the lowest NN-3-EC significantly decreases the p-value to 2.86%, 2.24%, and 1.01%, respectively. Thus, the likelihood in favour of NN-3 increases considerably on the subsamples containing forecasts confidently predicted by NN-3. In contrast, excluding 10%, 50%, and 90% of the stocks with the highest NN-3-EC substantially increases the p-values to 4.11%, 5.72%, and 7.91%.

Of course, p-value comparisons may not provide adequate information about the models' performance on different subsamples. For example, consider the effect of changing the sample size, holding the model MSEs constant. The smaller samples would yield larger standard errors and larger p-values, although the true MSEs remain the same. Thus, to draw more informative inferences, the following subsection compares the two models in terms of their HL portfolios' OOS returns and Sharpe ratios

E.2. NN-3 versus Lewellen: High-Low Portfolio Comparisons

Fig 10 plots the OOS return and Sharpe ratio differences between both models' VW HL portfolios on various subsamples. Like in the previous figure, the economic gains from the NN-3 monotonically increase with the NN-3-EC. For example, on the entire sample containing all stocks, the difference between NN-3 and Lewellen HL portfolios' average returns (squared Sharpe ratios) is 0.38% (0.02), and statistically insignificant (at 10%). However, the difference soars to a highly significant 0.82% (0.52) on the subsample comprising the top 10% stocks with the highest NN-3-

EC. In contrast, for the bottom 10% of stocks with the lowest NN-3-*EC*, Lewellen statistically outperforms NN-3. The average return (square-Sharpe ratio) difference between NN-3 and Lewellen HL portfolios is significantly negative -1.2% (-0.58).

Finally, I compare the conventional and Confident-HL portfolios formed from the NN-3 and Lewellen models. The portfolio definitions and notations remain the same as in section V.C. In addition, I denote all Lewellen-based HL portfolios by attaching the subscript "$_{\text{L}}$" to HL. For example, the conventional EW-HL portfolio based on the Lewellen model is represented by EW-HL_{L}.

Table IX presents the results. It reveals that the difference between the conventional EW (VW) NN-3-HL and Lewellen-HL portfolios' squared-Sharpe ratios is statistically insignificant at 1% (10%). Moreover, the analogous difference between the NN-3-Low-Confident-HL and Lewellen-HL is significantly negative, suggesting the Lewellen model's dominance on the subsample of forecasts imprecisely predicted by NN-3. In contrast, the corresponding difference between the NN-3-Confident-HL and Lewellen-HL portfolios is highly positive and significant at 1%. These results confirm the superiority of NN-3-based Confident-HL portfolios.

To make a fair assessment, I also compare the NN-3-Confident-HL portfolios with Lewellen-Confident-HL portfolios. The conclusions are the same. The NN-3-Confident-HL portfolios remarkably outperform in terms of squared-Sharpe ratios. This result is expected, as Confident-HL portfolios' performance hinges on ex-ante precision predicting ex-post squared forecast errors. Because it is less likely to hold for the benchmark Lewellen model (as shown in II and V.B), the Lewellen-Confident-HL portfolios do not deliver superior performance.

In sum, this section shows that existing studies significantly overestimate the overall predictive performance of NN-3 relative to the Lewellen model. The difference between the performance of both models' conventional HL portfolios' is moderately significant or insignificant. However, NN-3 exceptionally outperforms on subsamples of forecasts that it confidently predicts. Likewise, the NN-3-based Confident-HL portfolios statistically dominate the comparable Lewellen model's portfolios.

In the following two sections, I explore the time-series and cross-sectional properties of NN-3-based ex-ante precision.

F. Time-Series Variation in Ex-ante Standard Errors

To understand the time-series variation in the estimation uncertainty of NN-3-based risk premia, I compute the cross-sectional average of their ex-ante standard errors and call these "aggregate standard errors". Figure 11 plots the time-series of the aggregate standard errors. The series clearly reflects time-varying financial market uncertainty. For example, Bloom (2009) and Baker et al. (2016) document that market uncertainty appears to jump up after major shocks, such as Black Monday, the Dotcom Bubble, the Russian default, the failure of Lehman Brothers, and the

2011 debt ceiling dispute. Consistent with these studies, the aggregate standard errors spike after such shocks.

Table X presents the time-series average of aggregate standard errors over the OOS period and periods of shocks. Whereas the average monthly standard error across all periods is 1.06%, it is 2.31% during crisis periods. Because many individual predictors (e.g., size, price trends, and stock market volatility) in the NN-3 model substantially deviate from their usual distributions during these crisis periods, resulting risk premium predictions would also be hugely imprecise. Thus, the aggregate standard errors proxy for market uncertainty. For example, the standard errors are 38% correlated with the widely-used uncertainty proxy, the monthly market return standard deviation computed using daily data.

G. Cross-sectional Variation in Ex-ante Confidence

Table XI presents the cross-sectional properties of various ex-ante confidence sorted deciles. It reveals that NN-3 confidently predicts stocks with small market capital, high book-to-market ratios and high 1-year momentum returns. Because these characteristics associate with higher expected returns, NN-3-based HL portfolios deliver more gains in the long-leg rather than the short-leg. This result contrasts with the "arbitrage asymmetry" studies that argue, under trading frictions, anomaly-based investment portfolios yield relatively more profits in the short-leg (e.g., Stambaugh et al. (2012)). Avramov et al. (2020) note similar observations, albeit examining *ex-post* OOS long-leg and short-leg returns of investment portfolios based on various ML models, including NN-3. Possible reasons for understanding the association between the level and precision of NN-based risk premium predictions warrant a future study.

Moreover, NN-3 confidently predicting risk premia of small-sized stocks lends support to Avramov et al. (2020), who argue that NN-3-based HL portfolios derive more economic gains from microcaps. Table XI shows why. Because such stock risk premia are more confidently predicted, HL portfolios containing microcaps yield relatively larger economic gains.

Interestingly, I find that a significant proportion of non-microcaps have confidently risk premium predictions. Table XII presents the results. It shows that 34% of the stocks with the most precise risk premium predictions have market caps greater than the median size across all individual stocks. Thus, NN-3-based Confident-HL portfolios yield impressive gains even on sub-samples containing large-sized stocks.

VI. Conclusions

I develop an easy-to-implement method to estimate ex-ante standard errors of risk premium predictions from neural networks. To my knowledge, this is the first paper to explicitly derive the precision of NN-based risk premia at the stock-level and portfolio-level. I show that considering ex-ante standard errors leads to enhanced investment portfolios and out-of-sample statistical inferences.

The neural-network-based confident high low trading strategies that take long-short positions on stocks that have more risk premium estimates yield at least 40% higher returns and 15% higher Sharpe ratios than the neural-network-based conventional high-low portfolios. In evaluating whether these improvements are statistically significant, this paper shows that existing out-of-sample inferences that do not account for ex-ante standard errors are inadequate. I develop a bootstrap method, robust to estimation uncertainty, to compare OOS returns and Sharpe ratios of any two model-based investment strategies. The method also can be employed to compare mean squared errors of any two competing return predictions.

The bootstrap tests suggest that the neural-network-based confident high-low portfolios significantly outperform the neural-network-based conventional high-low portfolios, as well as the traditional high-low and confident high-low portfolios formed using the benchmark Lewellen model. However, the difference between the conventional neural-network-based and Lewellen-based high-low portfolios' out-of-sample returns and Sharpe ratios are either statistically insignificant or moderately significant. Thus, considering ex-ante standard errors is necessary for both real-time trading strategies and ex-post out-of-sample inferences.

A. Appendix: Proofs

1. Proof of Proposition-1:

Let the risk premium predictions of A_1, A_2, A_3, and A_4 be $\hat{a}_1$, $\hat{a}_1$, $\hat{b}_1$, and $\hat{b}_2$, respectively. Let pse_{a1}, pse_{a2}, pse_{b1}, and pse_{b2} be the predictive standard errors of A_1, A_2, A_3, and A_4, respectively.

The expected HL return equals the sum of the following measures

$$\begin{aligned} E(HL) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+ (\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) \\ &+ 0 \times p_3, \end{aligned} \tag{52}$$

where $p_3 = 1 - P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) - P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right)$.

Case1: When $pse_{a1} \geq \{pse_{b1}, pse_{b2}\}$ and $pse_{a2} \geq \{pse_{b1}, pse_{b2}\}$, the expected Confident-HL return equals

$$\begin{aligned} E(\text{Confident-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+ (\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) \\ &+ 0 \times p_3 \\ &= E(HL). \end{aligned} \tag{53}$$

Case2: Similarly, when $pse_{b1} \geq \{pse_{a1}, pse_{a2}\}$ and $pse_{b2} \geq \{pse_{a1}, pse_{a2}\}$

$$\begin{aligned} E(\text{Confident-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+ (\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) \\ &+ 0 \times p_3 \\ &= E(HL). \end{aligned} \tag{54}$$

Case3: When predictive standard errors do not align with either case1 or case2, without loss of generality, let $pse_{a1} \leq pse_{b1} \leq pse_{a2} \leq pse_{b2}$. Then,

$$\begin{aligned} E(\text{Confident-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+(\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) + (\mu_a - \mu_b) \times p_4 + (\mu_b - \mu_a) \times p_5, \end{aligned} \tag{55}$$

where $p_4 = P\left(\{\hat{a}_1, \hat{b}_2\} \in Q_L\right)$, $p_5 = P\left(\{\hat{a}_2, \hat{b}_1\} \in Q_L\right)$, and $P(.)$ is the probability measure. Because $\hat{a}_1$ and $\hat{b}_1$ are (relatively) precisely measured, $\hat{a}_1$ and $\hat{b}_2$ are more likely to be in Q_L and Q_S, respectively. Consistent with this intuition, it turns out that $p_4 > p_5$. Thus,

$$\begin{aligned} E(\text{Confident-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+(\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) + (\mu_a - \mu_b) \times (p_4 - p_5) \\ &> E(HL) \end{aligned} \tag{56}$$

Similarly, the expected return of PW-HL is given by

$$\begin{aligned} E(\text{PW-HL}) =&(\mu_a - \mu_b) \times P\left(\left[\hat{a}_1 > \{\hat{b}_1, \hat{b}_2\}, \hat{a}_2 > \{\hat{b}_1, \hat{b}_2\}\right]\right) \\ &+(\mu_b - \mu_a) \times P\left(\left[\hat{b}_1 > \{\hat{a}_1, \hat{a}_2\}, \hat{b}_2 > \{\hat{a}_1, \hat{a}_2\}\right]\right) + (2w - 1) \times (\mu_a - \mu_b) \times (p_4 - p_5), \end{aligned} \tag{57}$$

where w (> 0.5) is the weight assigned to the precise stock in each quantile. When $w = 1$, PW-HL reduces to Confident-HL, as it takes long (short) position only on the stock with the precise risk premium prediction. Thus,

$$E(\text{HL}) \leq E(\text{PW-HL}) \leq E(\text{Confident-HL}) \tag{58}$$

2. Proof of Proposition-2

Proof. Using Gal and Ghahramani (2016), the following expressions are directly obtained for the (approximated) Bayesian marginal predictive distribution of returns and their variances, respectively.

$$\begin{aligned} &Q(r^*_{i,t+1}|z^*_{it}, R, Z) = P(r_{i,t+1}|z^*_{it}, R, Z, \Omega)q(\Omega) \\ &q(\Omega) = \prod_{k=1}^{K} p_{i,k}, \text{ where each } p_{i,k} \sim Bern(p), \\ &P(r_{i,t+1}|z^*_{it}, R, Z, \Omega) = \mathcal{N}(\hat{E}_{i,\Omega,t}, \sigma^2_\eta I), \end{aligned} \tag{59}$$

where $Bern()$ represents Bernoulli distribution. $\hat{E}_{i,\Omega,t}$ is given by (25), with d replaced by Ω. And

$$Var\left[Q(r^*_{i,t+1}|z^*_{it}, R, Z)\right] \approx \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2 + \sigma^2_\eta \tag{60}$$

Denote $Var\left[Q(r^*_{i,t+1}|z^*_{it},R,Z)\right]$ by $V_Q(r^*_{i,t+1})$, where V_Q represents the variance operation under the probability distribution $Q(r^*_{i,t+1}|z^*_{it},R,Z)$. Note that by the law of total variance

$$V_Q(r^*_{i,t+1}) = V_Q(E(r^*_{i,t+1}|W_1,W_2)) + E_Q(V(r^*_{i,t+1}|W_1,W_2)), \tag{61}$$

where W_1, W_2 are the unknown weight matrices of the NN-1, and E_Q represents the expectation operation under the probability distribution $Q(r^*_{i,t+1}|z^*_{it},R,Z)$.

(61) further implies that

$$V_Q(r^*_{i,t+1}) = V_Q(\mu^*_{i,t}) + \sigma^2_\eta, \tag{62}$$

because $E(r^*_{i,t+1}|W_1,W_2) = \mu^*_{i,t}$, and $V(r^*_{i,t+1}|W_1,W_2) = \sigma^2_\eta$, which is assumed to be known.

Thus, (60) and (61) implies

$$V_Q(\mu^*_{i,t}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2. \tag{63}$$

□

3. Proof of Proposition-3

Proof. To compute portfolio-level standard errors, joint (approximated) density of return predictions are required. Straightforward algebra implies that it is given by

$$Q(r^*_{1,t+1}, r^*_{2,t+1}, \dots r^*_{S,t+1}|\{z^*_{it}\}_{i=1}^S, R, Z) = P(r^*_{1,t+1}, r^*_{2,t+1}, \dots r^*_{S,t+1}|\{z^*_{it}\}_{i=1}^S, R, Z, \Omega)q(\Omega)$$
$$q(\Omega) = \prod_{k=1}^{K} p_{i,k}, \text{ where each } p_{i,k} \sim Bern(p),$$
$$P(r^*_{1,t+1}, r^*_{2,t+1}, \dots r^*_{S,t+1}|\{z^*_{it}\}_{i=1}^S, R, Z, \Omega) = \mathcal{N}(\hat{E}_{S,\Omega,t}, \sigma^2_\eta I), \text{ where } \hat{E}_{S,\Omega,t} = \begin{bmatrix} \hat{E}_{1,\Omega,t} \\ \hat{E}_{2,\Omega,t} \\ \vdots \\ \hat{E}_{S,\Omega,t} \end{bmatrix}, \tag{64}$$

with each $\hat{E}_{i,\Omega,t}$ given by (25). The key is to use the same Ω across the stocks, as discussed in the main section of the paper. Then, the predictive variance of the portfolio P is given by

$$V_Q(r^*_{P,t+1}) = E_Q\left(V(r^*_{P,t+1}|\Omega)\right) + V_Q\left(E(r^*_{P,t+1}|\Omega)\right), \tag{65}$$

where $r^*_{P,t+1} = \sum_{i\in S} \omega_{P,i,t} r^*_{i,t+1}$. Moreover, $V(r^*_{P,t+1}|\Omega) = \sum_{i\in S} \omega^2_{P,i,t}\sigma^2_\eta$. And due to (64), $V_Q\left(E(r^*_{P,t+1}|\Omega)\right)$ can be approximated by

$$V_Q\left(E(r^*_{P,t+1}|\Omega)\right) \approx \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2, \tag{66}$$

with $\hat{E}_{P,d,t}$, and $p_{1,d}$, $p_{2,d}$ given in (27).

Thus, (65) further implies that

$$V_Q(r^*_{P,t+1}) = \sum_{i\in S}\omega^2_{P,i,t}\sigma^2_\eta + \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2. \tag{67}$$

Now, to compute the predictive variance of P's risk premium, note that

$$V_Q(r^*_{P,t+1}) = E_Q\left(V(r^*_{P,t+1}|W_1,W_2)\right) + V_Q\left(E(r^*_{P,t+1}|W_1,W_2)\right) = \sum_{i\in S}\omega^2_{P,i,t}\sigma^2_\eta + V_Q(\mu^*_{P,t}). \tag{68}$$

Thus, from (67) and (68),

$$V_Q(\mu^*_{P,t}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2 \tag{69}$$

□

B. Appendix: Simulations and Testing the Diebold and Mariano (2002) Assumption

1. Validity of Standard Errors: Monte Carlo Evidence

Table I
Calibration of the Confidence Intervals: Monte Carlo Evidence
This table validates the proposed standard errors using Monte Carlo simulations. The data comprise monthly stock risk premia and their raw predictors simulated under four different models 1-4. On the simulated data, confidence intervals (CIs) of various levels are constructed using NN-based risk premium predictions and their standard errors. Each row presents the confidence level and probabilities with which the corresponding level's confidence intervals cover the true simulated risk premia under the four models.

	Probability that CI contains true risk premium			
Confidence level	Model 1	Model 2	Model 3	Model 4
1%	1.26%	1.49%	1.08%	0.91%
5%	6.23%	6.65%	4.64%	3.63%
10%	11.81%	13.16%	8.98%	7.57%
20%	23.83%	26.26%	17.78%	16.17%
50%	48.72%	61.62%	46.85%	43.64%
60%	57.73%	73.10%	59.38%	55.52%
80%	78.94%	90.73%	83.60%	79.66%
90%	90.24%	96.48%	93.72%	90.36%
95%	96.03%	98.56%	97.39%	95.20%
99%	99.33%	99.74%	99.36%	98.75%

2. Tests of Covariance Stationarity

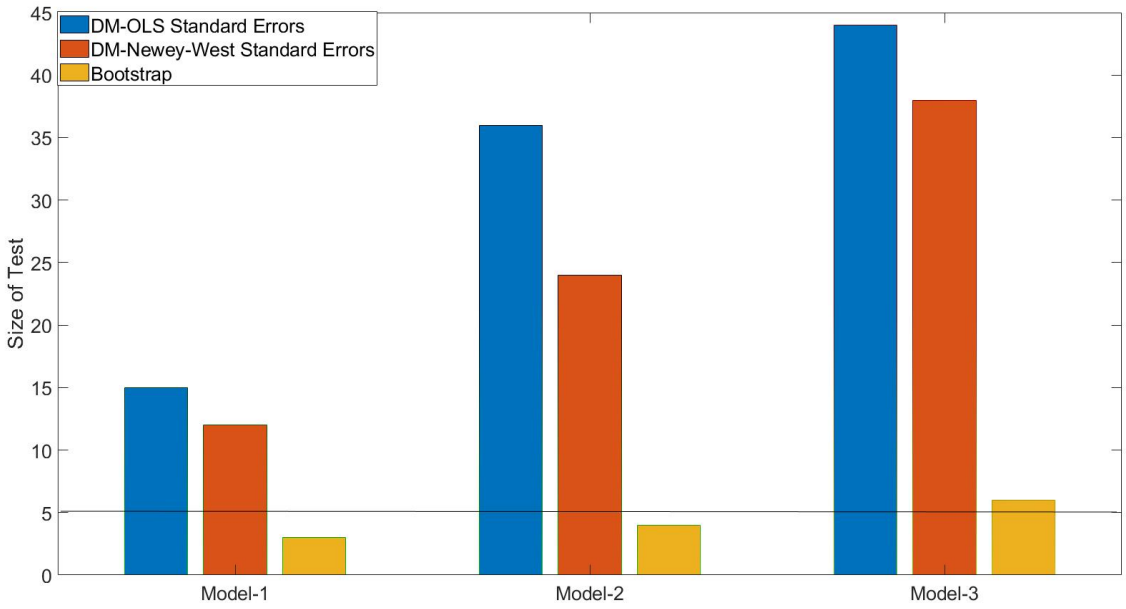
Table II
Violation of Diebold and Mariano (2002) conditions : Non-Stationarities due to Estimation Uncertainty

This table shows that the model-based loss differentials violate the covariance stationarity assumption required for the Diebold and Mariano (2002) tests' asymptotic validity. The table presents three loss differential series over the 360 out-of-sample periods. The first comprises the forecast-squared error differences between the NN-3 and Lewellen-based return predictions. The second contains the return differences between the equal-weighted high-low portfolios based on the NN-3 and Lewellen-based models. The third includes the return differences between the value-weighted high-low portfolios based on the NN-3 and Lewellen-based models. The First 180 Months column presents the loss differentials' sample standard deviations over the first 180 OOS periods, whereas the Last 180 Months column shows those over the last 180 periods. The Ratio column presents the ratio of the first and last 180 month standard deviations. The p-value column presents the p-value under the hypothesis that the ratio equals one, with critical values based on Pagan and Schwert (1990).

(NN-3 − Lewellen) Differentials	Standard Deviation of Loss Function			
Loss Function	First 180 Months	Last 180 Months	Ratio	p-value
Mean Squared Forecast Errors	0.12%	0.02%	5.03	< 0.001
Equal-weighted High-low Returns	0.35%	0.19%	1.85	< 0.001
Value-weighted High-low Returns	0.47%	0.27%	1.75	< 0.001

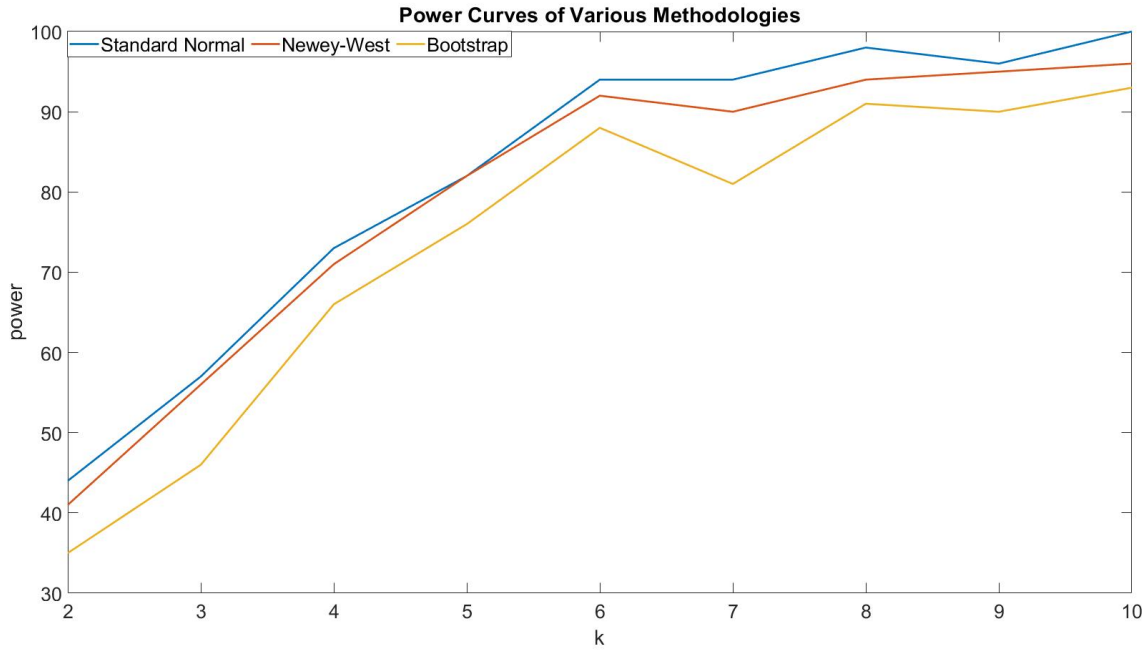
3. Performance of this paper's OOS Comparison Method: Monte Carlo Evidence

Figure 3. Test Sizes of OOS Comparison Methodologies



Note: This figure presents the "test sizes" of various methodologies at the 5% level. Test size represents the probability of *incorrectly rejecting* the null when it is true. Return time series with zero means are simulated under three distinct models, each imposing a different degree of time-varying temporal dependency. On the simulated data, tests of zero return means are conducted using three methods. The first (in blue) performs DM tests with the OLS standard errors. The second (in red) executes DM tests with Newey-West standard errors. The third (in orange) implements this paper's bootstrap method.

Figure 4. Power Curves of OOS Comparison Methodologies



Note: This figure presents the "power curves" of various methodologies at the 5% level. Power represents the probability of *correctly rejecting* the null when it is not true. Return time series are simulated under nine models, denoted by k, allowing for time-varying temporal dependencies. The mean return under model k equals k$\times\sigma$, where σ is a known scalar calibrated to match the standard deviation of the market risk premium. On the simulated data, tests of zero return means are conducted using three methods. The first (in blue) performs DM tests with the OLS standard errors. The second (in red) executes DM tests with Newey-West standard errors. The third (in orange) implements this paper's bootstrap method.

References

Allena, Rohit, 2020a, Comparing Asset Pricing Models with Non-Traded Factors and Principal Components, *SSRN Electronic Journal* .

Allena, Rohit, 2020b, Industry Costs of Equity with Machine Learning, *Working Draft, Goizueta Business School* .

Allena, Rohit, and Tarun Chordia, 2020, True Liquidity and Equilibrium Prices: US Tick Pilot, *Working Paper, Goizueta Business School* .

Avramov, Doron, Si Cheng, and Lior Metzker, 2020, Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability, SSRN Scholarly Paper ID 3450322, Social Science Research Network, Rochester, NY.

Avramov, Doron, Tarun Chordia, Gergana Jostova, and Alexander Philipov, 2013, Anomalies and financial distress, *Journal of Financial Economics* 108, 139–159.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, 2016, Measuring Economic Policy Uncertainty*, *The Quarterly Journal of Economics* 131, 1593–1636.

Bali, Turan G., Amit Goyal, Dashan Huang, Fuwei Jiang, and Quan Wen, 2020, The Cross-Sectional Pricing of Corporate Bonds Using Big Data and Machine Learning, SSRN Scholarly Paper ID 3686164, Social Science Research Network, Rochester, NY.

Berger, James O., 1985, *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics, second edition (Springer-Verlag, New York).

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni, 2020, Bond Risk Premiums with Machine Learning, *The Review of Financial Studies* .

Billingsley, Patrick, 1999, *Convergence of Probability Measures*, second edition (Wiley).

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe, 2017, Variational Inference: A Review for Statisticians, *Journal of the American Statistical Association* 112, 859–877.

Bloom, Nicholas, 2009, The Impact of Uncertainty Shocks, *Econometrica* 77, 623–685.

Chen, Luyang, Markus Pelger, and Jason Zhu, 2020, Deep Learning in Asset Pricing, SSRN Scholarly Paper ID 3350138, Social Science Research Network, Rochester, NY.

Chinco, Alex, Adam D. Clark-Joseph, and Mao Ye, 2019, Sparse Signals in the Cross-Section of Returns, *The Journal of Finance* 74, 449–492.

Diebold, Francis X., 2015, Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests, *Journal of Business & Economic Statistics* 33, 1–1.

Diebold, Francis X, and Roberto S Mariano, 2002, Comparing Predictive Accuracy, *Journal of Business & Economic Statistics* Vol.20(1), p.134-144.

Fama, Eugene F., and Kenneth R. French, 1997, Industry costs of equity, *Journal of Financial Economics* 43, 153–193.

Fama, Eugene F., and Kenneth R. French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Gal, Yarin, and Zoubin Ghahramani, 2016, Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, *Proceedings of the 33 rd International Conference on Machine Learning, New York, NY, USA, 2016* 10.

Gonçalves, Sílvia, and Halbert White, 2002, The bootstrap of the mean for dependent heterogeneous arrays, *Econometric Theory* 18, 1367–1384.

Gonçalves, Sílvia, and Halbert White, 2004, Maximum likelihood and the bootstrap for nonlinear dynamic models, *Journal of Econometrics* 119, 199–219.

Gonçalves, Sílvia, and Halbert White, 2005, Bootstrap Standard Error Estimates for Linear Regression, *Journal of the American Statistical Association* 100, 970–979.

Goyal, Amit, and Ivo Welch, 2003, Predicting the Equity Premium with Dividend Ratios, *Management Science* 49, 639–654, Publisher: INFORMS.

Goyal, Amit, and Ivo Welch, 2008, A Comprehensive Look at the Empirical Performance of Equity Premium Prediction, *The Review of Financial Studies* 21, 1455–1508.

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, 4389–4436.

Géron, Aurélien, 2019, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, second edition (O'Reilly Media, Inc.).

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical Asset Pricing via Machine Learning, *The Review of Financial Studies* 33, 2223–2273.

Horowitz, Joel L., 2001, Chapter 52 - The Bootstrap, in James J. Heckman, and Edward Leamer, eds., *Handbook of Econometrics*, volume 5, 3159–3228 (Elsevier).

Kunsch, Hans R., 1989, The Jackknife and the Bootstrap for General Stationary Observations, *The Annals of Statistics* 17, 1217–1241.

Kyung, Minjung, Jeff Gill, Malay Ghosh, and George Casella, 2010, Penalized regression, standard errors, and Bayesian lassos, *Bayesian Analysis* 5, 369–411.

Lewellen, Jonathan, 2015, The Cross-section of Expected Stock Returns, *Critical Finance Review* 4, 1–44.

Pagan, Adrian R., and G. William Schwert, 1990, Testing for covariance stationarity in stock market data, *Economics Letters* 33, 165–170.

Pástor, Ľuboš, and Robert F. Stambaugh, 1999, Costs of Equity Capital and Model Mispricing, *The Journal of Finance* 54, 67–121.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, 2014, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research* Vol.15, 1929–1958.

Stambaugh, Robert F., Jianfeng Yu, and Yu Yuan, 2012, The short of it: Investor sentiment and anomalies, *Journal of Financial Economics* 104, 288–302.

Stambaugh, Robert F., and Yu Yuan, 2017, Mispricing Factors, *The Review of Financial Studies* 30, 1270–1315.

Timmermann, Allan, and Yinchu Zhu, 2019, Comparing Forecasting Performance with Panel Data, *SSRN Electronic Journal* .

Wang, Yixin, and David M. Blei, 2019, Frequentist Consistency of Variational Bayes, *Journal of the American Statistical Association* 114, 1147–1161.

Zhu, Lingxue, and Nikolay Laptev, 2017, Deep and Confident Prediction for Time Series at Uber, in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, 103–110, ISSN: 2375-9259.

Figure 5. Out-of-Sample (OOS) Performance of Equal-weighted Deciles Based on NN-3 Predictions.

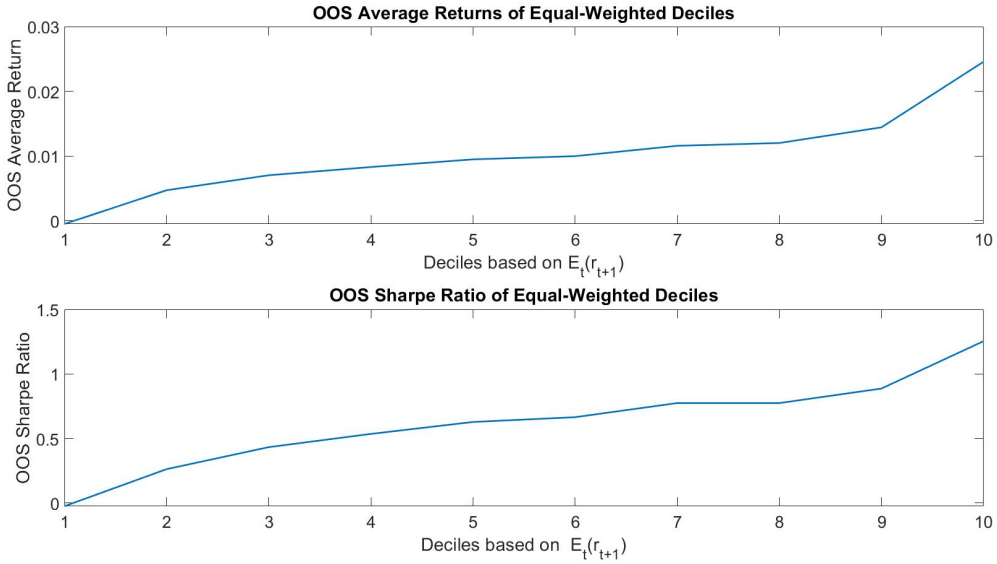
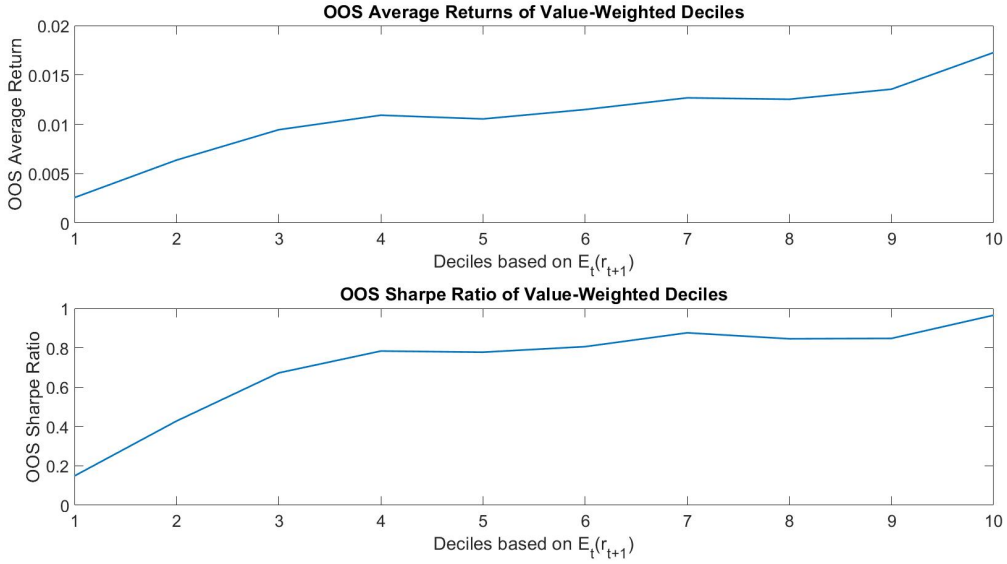
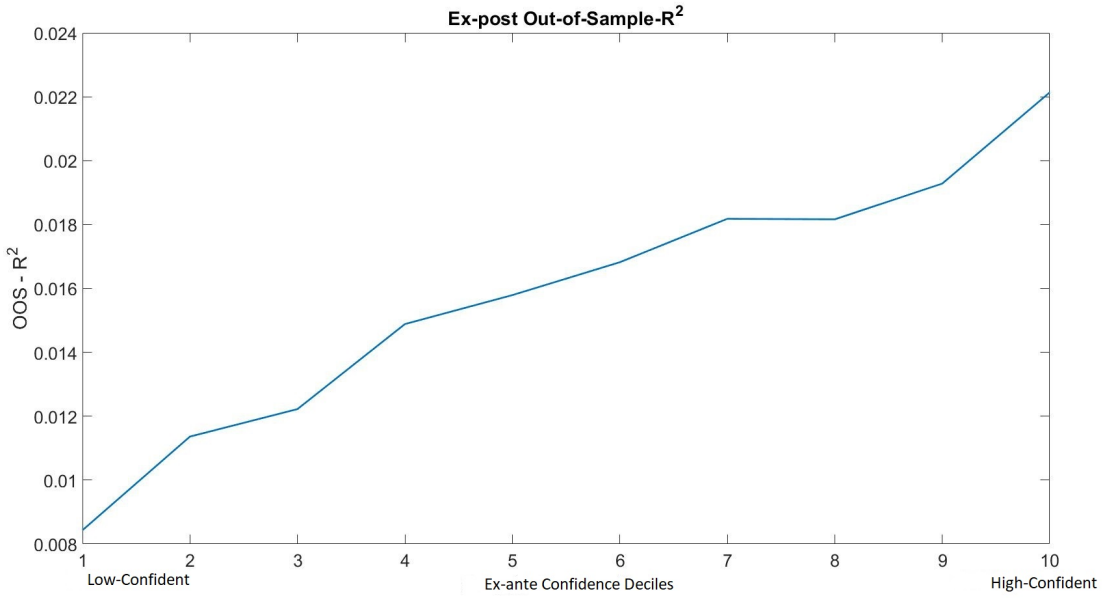


Figure 6. Out-of-Sample (OOS) Performance of Value-weighted Deciles Based on NN-3 Predictions.



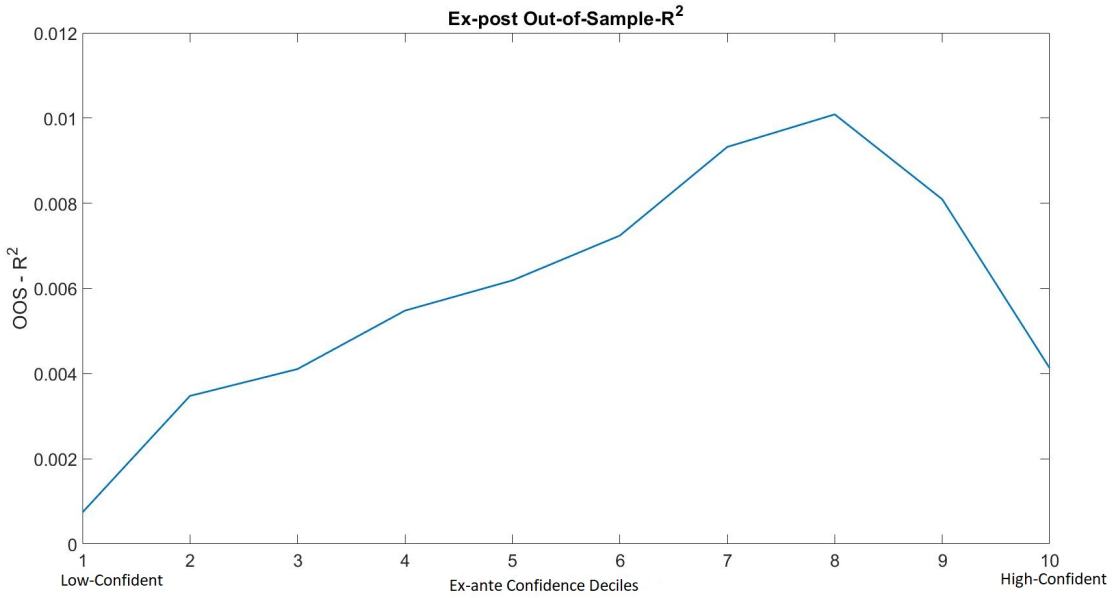
Note: Figure 5 (6) presents the performance of equal-weighted (value-weighted) prediction-sorted portfolios over the 30-year out-of-sample. At each period, stocks are sorted into deciles according to their NN-3-based risk premium predictions. Decile-10 (decile-1) comprises the top (bottom) 10% stocks with the lowest (highest) return predictions. The top figure shows the average monthly returns of each decile, whereas the bottom represents their annualized Sharpe ratios.

Figure 7. Ex-ante Confidence and Ex-post OOS-R^2: NN-3-based Predictions and Standard Errors



Note: This figure presents the out-of-sample (OOS) R^2s of various ex-ante confidence-sorted subsamples over the 30-year test sample. At each period, stocks are sorted into deciles according to their NN-3-based risk premium predictions' ex-ante confidence (EC). Decile-10 (decile-1) comprises the top (bottom) 10% stocks with the lowest (highest) precision. The y-axis represents the ex-post OOS R^2s attained by the decile subsamples.

Figure 8. Ex-ante Confidence and Ex-post OOS-R^2: Lewellen-based Predictions and Standard Errors



Note: This figure presents the out-of-sample (OOS) R^2s of various ex-ante confidence-sorted subsamples over the 30-year test sample. At each period, stocks are sorted into deciles according to their Lewellen-based risk premium predictions' ex-ante confidence. Decile-10 (decile-1) comprises the top (bottom) 10% of stocks with the lowest (highest) precision. The y-axis represents the ex-post OOS R^2s attained by the decile subsamples.

Table III

Long-short Portfolios' Performance on Subsamples with Different Levels of Ex-ante Confidence

This table reports the performance of model-based high-low (HL) portfolios over the 30-year out-of-sample (OOS) period on various subsamples. Each period, stocks are first sorted into deciles according to their ex-ante confidence levels of model-based risk premium predictions. On each decile, equal-weighted (value-weighted) HL portfolios are formed by further sorting stocks into deciles according to their next month's model-based return predictions and taking long-short positions on the extreme deciles. The NN-3-HL and Lewellen-HL columns present each precision-decile's HL portfolio's performance under the NN-3 and Lewellen models, respectively. The Pred Ret column reports the HL portfolio's average return predictions. The Avg Ret, Std, Sharpe columns respectively represent the average, standard deviation, and Sharpe ratio of the HL portfolio's realized returns. Panels A and B present the equal-weighted and value-weighted strategies, respectively.

Panel A: Performance of equal-weighted-HL on various precision-sorted subsamples

	NN-3-HL				Lewellen-HL			
Precision decile	Pred Ret	Avg Ret	Std	Sharpe	Pred Ret	Avg Ret	Std	Sharpe
1 (Low-Confident)	0.72%	0.88%	4.29%	0.71	1.83%	0.81%	6.85%	0.41
2	0.52%	1.14%	4.80%	0.83	2.97%	1.89%	6.53%	1.00
3	0.54%	0.75%	4.62%	0.56	2.30%	1.53%	6.88%	0.77
4	0.58%	1.31%	4.74%	0.96	1.83%	1.80%	7.60%	0.82
5	0.62%	1.31%	5.15%	0.88	1.62%	1.70%	7.02%	0.84
6	0.64%	1.77%	5.42%	1.13	1.49%	1.44%	5.99%	0.83
7	0.66%	1.40%	5.44%	0.89	1.49%	1.93%	6.12%	1.09
8	0.68%	1.78%	5.59%	1.10	1.50%	1.53%	5.27%	1.01
9	0.71%	2.03%	7.43%	0.95	1.43%	2.01%	4.99%	1.40
10 (High-Confident)	0.88%	3.10%	7.48%	1.44	1.07%	1.42%	4.90%	1.00

Panel B: Performance of value-weighted-HL on various precision-sorted subsamples

	NN-3-HL				Lewellen-HL			
Precision decile	Pred Ret	Avg Ret	Std	Sharpe	Pred Ret	Avg Ret	Std	Sharpe
1 (Low-Confident)	0.70%	0.34%	5.12%	0.23	1.79%	1.00%	5.46%	0.64
2	0.49%	0.65%	5.82%	0.39	2.89%	1.27%	8.64%	0.51
3	0.52%	0.86%	5.60%	0.53	2.16%	1.57%	7.39%	0.74
4	0.56%	0.65%	5.21%	0.43	1.76%	1.07%	6.39%	0.58
5	0.60%	0.80%	5.55%	0.50	1.45%	1.06%	6.30%	0.58
6	0.62%	0.68%	5.59%	0.42	1.32%	1.01%	5.43%	0.64
7	0.62%	0.43%	6.02%	0.25	1.29%	1.27%	5.40%	0.82
8	0.67%	0.67%	6.52%	0.36	1.35%	1.13%	5.11%	0.77
9	0.70%	1.16%	7.68%	0.52	1.33%	1.33%	5.98%	0.77
10 (High-Confident)	0.89%	1.59%	6.86%	0.80	0.99%	0.66%	5.93%	0.39

Table IV
Performance of Confident and Low-Confident Long-Short Portfolios: All Stocks

This table reports the performance of various NN-3-based long-short portfolios over the 30-year out-of-sample (OOS) period. EW(VW)-HL represents the traditional equal(value)-weighted long-short portfolio. EW(VW)-Confident-HL and EW(VW)-Low-Confident-HL denote the equal(value)-weighted Confident and Low-Confident long-short portfolios that only include stocks with the most *confident* and *imprecise* risk premium predictions, respectively. LPW-HL and PW-HL are the "imprecision" and "precision" weighted portfolios that overweight stocks with imprecise and precise return predictions, respectively. EW(VW, LPW)-HL_{LCM} is the conventional EW(VW, LPW) HL portfolio matched to have the same average predicted returns as that of the EW-Low-Confident-HL (EW-Low-Confident-HL, LPW-HL) portfolio. EW(VW)-HL_{CM} is a traditional EW(VW)-HL portfolio matched to have the same average predicted returns as that of the EW-Confident-HL (VW-Confident-HL) portfolio. Likewise, LPW(PW)-HL_{M} is a traditional EW-HL portfolio matched with LPW(PW)-HL. See section V.C for a detailed description of the portfolios. All portfolio returns are also adjusted for Fama-French 5-factors plus momentum (FF-5+UMD) and Stambaugh-Yuan 4-factor (SY) models. The "pred ret" column represents the average predicted returns. The "avg ret" column shows the average realized returns. The "α" columns indicate abnormal returns. The "t" columns denote the t-stats of "average returns" and "α". The "SR" and "IR" columns represent the annualized Sharpe and Information ratios, respectively.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL_{LCM}, HL_{CM} and HL_{M} are matching high-low portfolios.

Panel A: Equal-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
EW-HL	1.69%	2.52%	8.21	1.50	2.20%	7.63	1.39	2.18%	7.15	1.31
EW-HL_{LCM}	1.77%	2.64%	8.20	1.50	2.34%	7.7	1.41	2.33%	7.25	1.32
EW-Low-Confident-HL	1.79%	2.35%	6.46	1.18	1.97%	5.65	1.03	1.96%	5.28	0.96
EW-HL_{CM}	1.97%	3.07%	8.65	1.58	2.77%	8.26	1.51	2.75%	7.8	1.42
EW-Confident-HL	1.97%	3.61%	9.58	1.75	3.29%	9.02	1.65	3.27%	8.6	1.57

Panel B: Value-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
VW-HL	1.62%	1.48%	4.95	0.90	0.90%	3.26	0.59	0.77%	2.68	0.49
VW-HL_{LCM}	1.77%	1.50%	4.61	0.84	0.87%	2.87	0.52	0.76%	2.38	0.44
VW-Low-Confident-HL	1.78%	1.31%	3.02	0.55	0.48%	1.15	0.21	0.39%	0.88	0.16
VW-HL_{CM}	1.90%	1.73%	4.92	0.90	1.12%	3.39	0.62	1.02%	2.95	0.54
VW-Confident-HL	1.90%	2.21%	5.95	1.09	1.79%	4.77	0.87	1.43%	3.82	0.70

Panel C: Precision-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
EW-HL	1.69%	2.52%	8.21	1.50	2.20%	7.63	1.39	2.18%	7.15	1.31
LPW-HL_{M}	1.69%	2.52%	8.21	1.50	2.20%	7.63	1.39	2.18%	7.15	1.31
LPW-HL	1.70%	2.36%	7.63	1.39	2.02%	6.95	1.27	2.00%	6.48	1.18
PW-HL_{M}	1.77%	2.64%	8.20	1.50	2.34%	7.7	1.41	2.33%	7.25	1.32
PW-HL	1.77%	2.87%	9.14	1.67	2.57%	8.68	1.59	2.55%	8.16	1.49

Table V
Statistical Comparison of Long-Short Portfolios: All Stocks

This table conducts pairwise statistical comparisons of the out-of-sample (OOS) performance of various NN-3-based long-short portfolios. The tests are based on the moving block bootstrap procedure developed in section IV, with a block-length of 24. The Investment Strategy column shows the comparing pair of portfolios. The avg ret column presents the average return differences between the pair of investment strategies. The α column shows the average abnormal return differences. The $Sharpe^2$ and IR^2 columns show the annualized squared-Sharpe and squared-information ratio differences between the investment portfolios, respectively. The numbers in parenthesis are p-values. *, ** and *** denote significance at the 1%, 5% and 10% levels, respectively. See table IV and section V.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL_{LCM}, HL_{CM} and HL_{M} are matching high-low portfolios.

Panel A : OOS Performance Differences of Equal-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − EW-Low-Confident-HL	0.17% (0.373)	0.859^{***} (0)	0.23% (0.207)	1.008^{***} (0)	0.22% (0.267)	0.941^{***} (0)
EW-HL_{LCM} − EW-Low-Confident-HL	0.30% (0.142)	0.853^{***} (0)	$0.36\%^{*}$ (0.06)	1.049^{***} (0)	$0.36\%^{*}$ (0.083)	0.998^{***} (0)
EW-Confident-HL − EW-HL	$1.10\%^{***}$ (0)	0.808^{***} (0)	$1.09\%^{***}$ (0)	0.884^{***} (0)	$1.09\%^{***}$ (0)	0.92^{***} (0)
EW-Confident-HL − EW-Low-Confident-HL	$1.27\%^{***}$ (0.001)	1.666^{***} (0)	$1.32\%^{***}$ (0)	1.892^{***} (0)	$1.31\%^{***}$ (0.001)	1.861^{***} (0)
EW-Confident-HL − EW-HL_{CM}	$0.55\%^{**}$ (0.03)	0.563^{***} (0)	$0.52\%^{**}$ (0.039)	0.502^{***} (0)	$0.52\%^{**}$ (0.043)	0.527^{***} (0)

Panel B : OOS Performance Differences of Value-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
VW-HL − VW-Low-Confident-HL	0.17% (0.542)	0.511^{***} (0.001)	$0.42\%^{*}$ (0.094)	0.356^{***} (0.001)	0.38% (0.136)	0.258^{***} (0.002)
VW-HL_{LCM} − VW-Low-Confident-HL	0.19% (0.503)	0.404^{***} (0.002)	0.39% (0.144)	0.266^{***} (0.003)	0.37% (0.173)	0.198^{***} (0.008)
VW-Confident-HL − VW-HL	$0.73\%^{***}$ (0.003)	0.364^{***} (0.003)	$0.89\%^{***}$ (0)	0.467^{***} (0)	$0.66\%^{***}$ (0.007)	0.3^{***} (0.001)
VW-Confident-HL − VW-Low-Confident-HL	$0.90\%^{**}$ (0.032)	0.875^{***} (0)	$1.31\%^{***}$ (0)	0.823^{***} (0)	$1.04\%^{***}$ (0.009)	0.558^{***} (0)
VW-Confident-HL − VW-HL_{CM}	$0.48\%^{*}$ (0.086)	0.374^{***} (0.004)	$0.67\%^{***}$ (0.003)	0.433^{***} (0.001)	0.41% (0.128)	0.238^{***} (0.008)

Panel C : OOS Performance Differences of Precision-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − LPW-HL	$0.15\%^{**}$ (0.031)	0.307^{***} (0)	$0.18\%^{***}$ (0.007)	0.38^{***} (0)	0.38% (0.136)	0.258^{***} (0.002)
LPW-HL_{M} − LPW-HL	$0.15\%^{**}$ (0.031)	0.307^{***} (0)	$0.18\%^{***}$ (0.007)	0.38^{***} (0)	0.37% (0.173)	0.198^{***} (0.008)
PW-HL − EW-HL	$0.36\%^{***}$ (0)	0.535^{***} (0)	$0.37\%^{***}$ (0)	0.658^{***} (0)	$0.66\%^{***}$ (0.007)	0.3^{***} (0.001)
PW-HL − LPW-HL	$0.51\%^{***}$ (0.001)	0.842^{***} (0)	$0.55\%^{***}$ (0)	1.038^{***} (0)	$1.04\%^{***}$ (0.009)	0.558^{***} (0)
PW-HL − PW-HL_{M}	$0.23\%^{**}$ (0.014)	0.541^{***} (0)	$0.23\%^{***}$ (0.007)	0.617^{***} (0)	0.41% (0.128)	0.238^{***} (0.008)

Table VI
Performance of Confident and Low-Confident Long-Short Portfolios: Non-Microcap Stocks
This table reports the performance of various NN-3-based long-short portfolios over the 30-year out-of-sample (OOS) period. Every period, the sample excludes microcap stocks with market capital smaller than the 20^{th} NYSE size percentile. See table IV and section V.C for a description of the portfolios. All portfolio returns are also adjusted for Fama-French 5-factors plus momentum (FF-5+UMD) and Stambaugh-Yuan 4-factor (SY) models. The pred ret column represents the average predicted returns. The avg ret column shows the average realized returns. The α columns indicate abnormal returns. The t columns denote the t-stats of average returns and α. The SR and IR columns represent the annualized Sharpe and Information ratios, respectively.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL_{LCM}, HL_{CM} and HL_{M} are matching high-low portfolios.

Panel A: Equal-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
EW-HL	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.84	1.24%	3.99	0.73
EW-HL_{LCM}	0.74%	1.83%	5.57	1.02	1.51%	4.76	0.87	1.37%	4.13	0.75
EW-Low-Confident-HL	0.74%	1.50%	3.98	0.73	1.10%	2.96	0.54	0.89%	2.32	0.42
EW-HL_{CM}	0.74%	1.83%	5.57	1.02	1.51%	4.76	0.87	1.37%	4.13	0.75
EW-Confident-HL	0.74%	2.25%	6.68	1.22	2.04%	6.03	1.10	1.93%	5.49	1.00

Panel B: Value-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
VW-HL	0.66%	1.42%	4.64	0.85	1.09%	3.58	0.65	0.98%	3.1	0.57
VW-HL_{LCM}	0.73%	1.58%	4.76	0.87	1.25%	3.76	0.69	1.10%	3.2	0.59
VW-Low-Confident-HL	0.74%	1.25%	3.13	0.57	0.88%	2.26	0.41	0.74%	1.83	0.33
VW-HL_{CM}	0.73%	1.58%	4.76	0.87	1.25%	3.76	0.69	1.10%	3.2	0.59
VW-Confident-HL	0.72%	2.07%	5.48	1.00	1.84%	4.78	0.87	1.64%	4.14	0.76

Panel C: Precision-Weighted Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
EW-HL	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.84	1.24%	3.99	0.73
LPW-HL_M	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.84	1.24%	3.99	0.73
LPW-HL	0.69%	1.60%	4.99	0.91	1.26%	4.06	0.74	1.13%	3.47	0.63
PW-HL_{M}	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.84	1.24%	3.99	0.73
PW-HL	0.69%	1.80%	5.93	1.08	1.52%	5.17	0.94	1.41%	4.57	0.83

Table VII
Statistical Comparison of Long-Short Portfolios: Non-Microcap Stocks

This table conducts pairwise statistical comparisons of the OOS performance of various NN-3-based long-short portfolios. Every period, the sample excludes microcap stocks with market capital smaller than the 20^{th} NYSE size percentile. The tests are based on the moving block bootstrap procedure developed in section IV, with a block-length of 24. The Investment Strategy column shows the comparing pair of portfolios. The avg ret column presents the average return differences between the pair of investment strategies. The α column shows the average abnormal return differences. The $Sharpe^2$ and IR^2 columns show the annualized squared-Sharpe and squared-information ratio differences between the investment portfolios. The numbers in parenthesis are p-values. *, ** and *** denote significance at the 1%, 5% and 10% levels, respectively. See table IV and section V.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL_{LCM}, HL_{CM} and HL_{M} are matching high-low portfolios.

Panel A : Performance Differences of Equal-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − EW-Low-Confident-HL	0.16% (0.393)	0.454^{***} (0.000)	0.25% (0.183)	0.469 (0.000)	$0.35\%^{*}$ (0.064)	0.427^{***} (0.000)
EW-HL_{LCM} − EW-Low-Confident-HL	$0.33\%^{*}$ (0.076)	0.505^{***} (0.000)	$0.41\%^{**}$ (0.023)	0.535^{***} (0.000)	$0.48\%^{***}$ (0.008)	0.471^{**} (0.000)
EW-Confident-HL − EW-HL	$0.59\%^{***}$ (0.000)	0.505^{***} (0.000)	$0.69\%^{***}$ (0.000)	0.588^{***} (0.000)	$0.69\%^{***}$ (0.000)	0.572^{***} (0.000)
EW-Confident-HL − EW-Low-Confident-HL	$0.75\%^{**}$ (0.016)	0.959^{***} (0.000)	$0.94\%^{***}$ (0.002)	1.058^{***} (0.001)	$1.03\%^{***}$ (0.001)	0.999^{***} (0.000)
EW-Confident-HL − EW-HL_{CM}	$0.42\%^{**}$ (0.015)	0.454^{***} (0.000)	$0.53\%^{***}$ (0.001)	0.523^{***} (0.000)	$0.56\%^{***}$ (0.001)	0.528^{***} (0.000)

Panel B : Performance Differences of Value-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
VW-HL − VW-Low-Confident-HL	0.17% (0.509)	0.391^{***} (0.000)	0.20% (0.438)	0.296^{***} (0.000)	0.24% (0.341)	0.253^{***} (0.001)
VW-HL_{LCM} − VW-Low-Confident-HL	0.33% (0.214)	0.428^{***} (0.000)	$0.37\%^{**}$ (0.166)	0.348^{***} (0.000)	$0.36\%^{*}$ (0.168)	0.280^{**} (0.001)
VW-Confident-HL − VW-HL	$0.65\%^{***}$ (0.005)	0.285^{***} (0.000)	$0.75\%^{***}$ (0.001)	0.382^{***} (0.000)	$0.66\%^{***}$ (0.005)	0.304^{***} (0.000)
VW-Confident-HL − VW-Low-Confident-HL	$0.82\%^{**}$ (0.029)	0.676^{***} (0.000)	$0.95\%^{***}$ (0.009)	0.679^{***} (0.000)	$0.90\%^{**}$ (0.012)	0.557^{*} (0.000)
VW-Confident-HL − VW-HL_{CM}	$0.48\%^{**}$ (0.041)	0.248^{***} (0.001)	$0.59\%^{**}$ (0.011)	0.331^{***} (0.000)	$0.54\%^{**}$ (0.024)	0.277^{***} (0.000)

Panel C : Performance Differences of Precision-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − LPW-HL	0.06% (0.348)	0.152^{***} (0.000)	0.09% (0.146)	0.172^{***} (0.000)	$0.11\%^{*}$ (0.082)	0.157^{***} (0.000)
LPW-HL_{M} − LPW-HL	0.06% (0.348)	0.152^{***} (0.000)	0.09% (0.146)	0.172^{***} (0.000)	$0.11\%^{*}$ (0.082)	0.157^{***} (0.000)
PW-HL − EW-HL	$0.14\%^{**}$ (0.014)	0.192^{***} (0.000)	$0.17\%^{***}$ (0.002)	0.222^{***} (0.000)	$0.17\%^{***}$ (0.001)	0.198^{***} (0.000)
PW-HL − LPW-HL	$0.20\%^{*}$ (0.088)	0.343^{***} (0.000)	$0.27\%^{**}$ (0.015)	0.394^{***} (0.000)	$0.28\%^{**}$ (0.011)	0.355^{***} (0.000)
PW-HL − PW-HL_{M}	$0.14\%^{**}$ (0.014)	0.192^{***} (0.000)	$0.17\%^{***}$ (0.002)	0.222^{***} (0.000)	$0.17\%^{***}$ (0.001)	0.198^{***} (0.000)

Table VIII
Transaction Costs and Higher-Moment Adjusted Performance of Confident-HL Portfolios
This table reports the transaction costs and higher-moment-risk-adjusted performance of various NN-3-based long-short portfolios over the 30-year out-of-sample period. The Turnover column presents a portfolio's average monthly percentage change in holdings (i.e., turnover). A deduction of (0.005×Turnover) from a portfolio's realized return roughly approximates its transaction-cost-adjusted returns. The Omega, Sortino and Upside columns respectively represent the Omega, Sortino and Upside potential ratios. These ratios measure the higher-moment-risk-adjusted performance of portfolios, explicitly penalizing losses more than realizing gains. See table IV and section V.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low; HL_{LCM}, HL_{CM} and HL_{M} are matching high-low portfolios.

Equal-Weighted Portfolios: Higher-Moment Adjusted Performance

	All Stocks				Non-Microcaps			
Investment Strategy	Turnover	Omega	Sortino	Upside	Turnover	Omega	Sortino	Upside
EW-HL	1.27	4.22	0.98	1.28	1.12	2.46	0.51	0.86
EW-HL_{LCM}	1.37	4.18	0.96	1.27	1.23	2.49	0.54	0.89
EW-Low-Confident-HL	1.88	2.83	0.71	1.10	1.89	1.89	0.37	0.80
EW-HL_{CM}	1.53	4.44	1.05	1.36	1.45	2.49	0.54	0.89
EW-Confident-HL	1.85	4.70	1.28	1.62	1.84	2.84	0.66	1.01

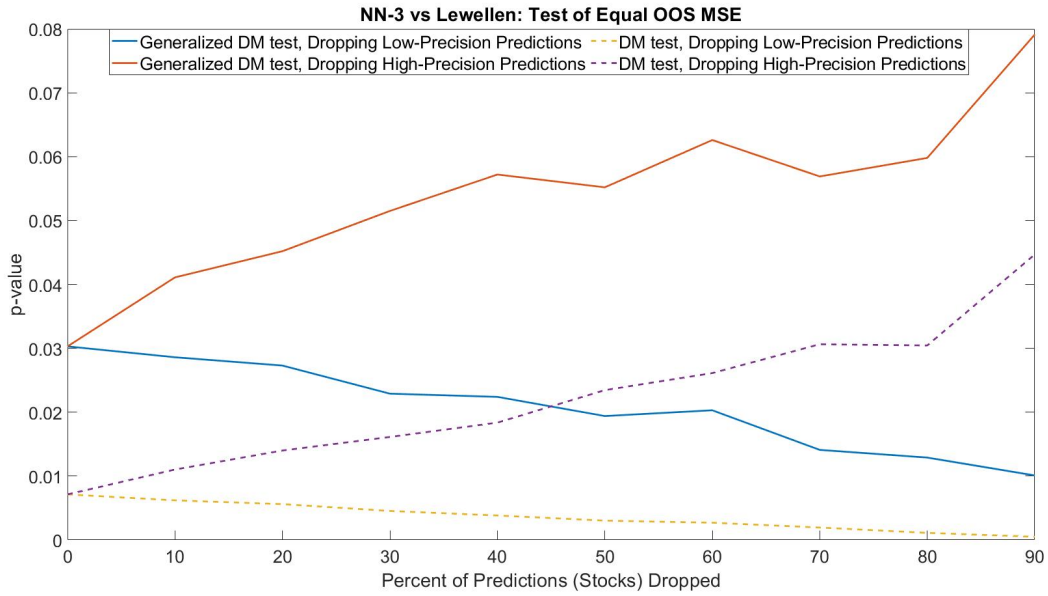
Value-Weighted Portfolios: Higher-Moment Adjusted Performance

	All Stocks				Non-Microcaps			
Investment Strategy	Turnover	Omega	Sortino	Upside	Turnover	Omega	Sortino	Upside
VW-HL	1.37	2.24	0.53	0.96	1.2	2.12	0.43	0.82
VW-HL_{LCM}	1.51	2.12	0.49	0.93	1.37	2.14	0.46	0.86
VW-Low-Confident-HL	1.90	1.58	0.26	0.71	1.86	1.59	0.26	0.71
VW-HL_{CM}	1.62	2.23	0.54	0.98	1.5	2.14	0.46	0.86
VW-Confident-HL	1.89	2.43	0.63	1.07	1.88	2.43	0.56	0.96

Precision-Weighted Portfolios: Higher-Moment Adjusted Performance

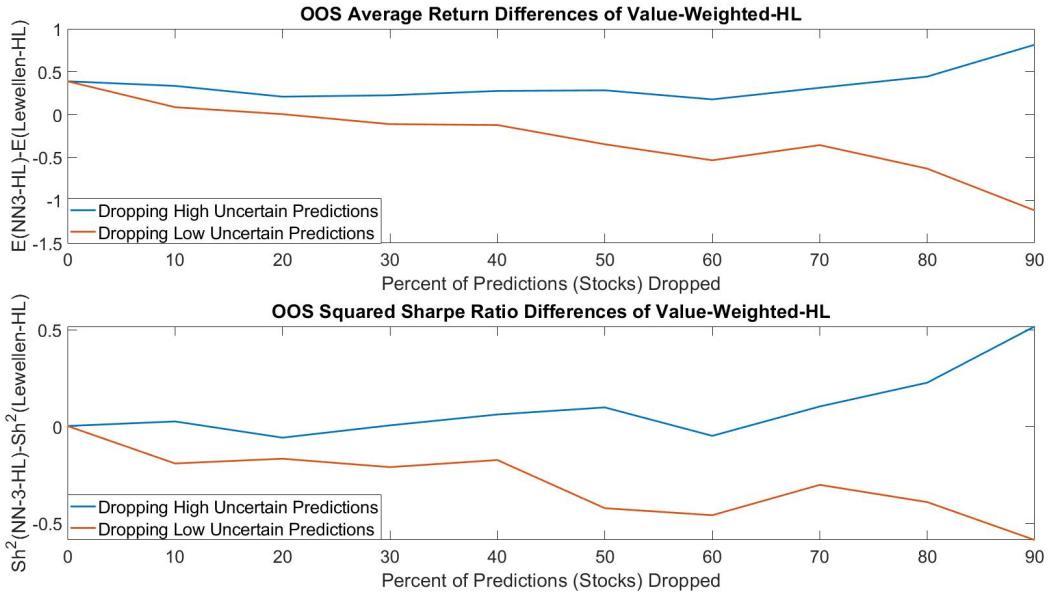
	All Stocks				Non-Microcaps			
Investment Strategy	Turnover	Omega	Sortino	Upside	Turnover	Omega	Sortino	Upside
PW-HL	1.27	4.22	0.98	1.28	1.12	2.46	0.51	0.86
PW-HL_{M}	1.27	4.22	0.98	1.28	1.12	2.46	0.51	0.86
PW-Low-Confident-HL	1.54	3.74	0.91	1.24	1.43	2.26	0.47	0.85
PW-HL_{M}	1.37	4.18	0.96	1.12	1.38	2.46	0.51	0.86
PW-Confident-HL	1.51	4.80	1.13	1.42	1.43	2.66	0.56	0.90

Figure 9. Comparing predictive performance of NN-3 with the benchmark Lewellen (2015) model



Note: This figure presents the p-values under the null hypothesis that the mean squared error of the NN-3 and Lewellen models are equal on various subsamples over the 30-year out-of-sample period. Every month, stocks are sorted into deciles according to their NN-3-based risk premium predictions' ex-ante confidence, NN-3-EC. The blue line (yellow dotted-line) displays the bootstrap (DM) p-values on the subsamples that dropout 10%, 20%, ... and 90% of the stocks with the lowest NN-3-EC, respectively. Thus, these subsamples contain the forecasts that NN-3 confidently predicts. In contrast, the red line (purple dotted-line) represents the p-values on the subsamples comprising the forecasts that NN-3 imprecisely predicts, excluding the 10%, 20%, ... and 90% stocks with the highest NN-3-EC, respectively.

Figure 10. Comparing predictive performance of NN-3 with the benchmark Lewellen (2015) model



Note: This figure presents the out-of-sample average return and squared-Sharpe-ratio differences between the value-weighted high-low (HL) portfolios formed using the NN-3 and Lewellen models on various subsamples. Every month, stocks are sorted into deciles according to their NN-3-based risk premium predictions' ex-ante confidence, NN-3-EC. The blue line in the top (bottom) of the figure displays the HL portfolios' average return (squared-Sharpe-ratio) differences on the subsamples that dropout 10%, 20%, ..., and 90% of the stocks with the lowest NN-3-EC, respectively. Thus, these subsamples contain the forecasts that NN-3 confidently predicts. In contrast, the red line at the top (bottom) of the figure corresponds to the subsamples comprising the forecasts that NN-3 imprecisely predicts, excluding the 10%, 20%, ... and 90% highest NN-3-EC stocks, respectively.

Table IX
Statistical Comparison of Long-Short Portfolios: NN-3 versus Lewellen (2015)

This table conducts pairwise statistical comparisons of the OOS performance of various long-short portfolios based on the NN-3 and Lewellen models. The tests are based on the moving block bootstrap procedure developed in section IV, with a block-length of 24. The Investment Strategy column shows the comparing pair of portfolios. The avg ret column presents the average return differences between the pair of investment strategies, the α column shows the average abnormal return differences. The $Sharpe^2$ and IR^2 columns show the annualized squared-Sharpe and squared-information ratio differences between the investment portfolios. The "HL" and "HL_L" portfolios are based on the NN-3 and Lewellen models, respectively. The numbers in parenthesis are p-values. *, **, and *** denote significance at the 1%, 5% and 10% levels, respectively. See table IV and section V.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low portfolio based on NN-3; HL_L=high-low portfolio based on Lewellen

Panel A : Performance Differences of Equal-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − EW-HL_L	$0.72\%^{**}$ (0.016)	0.247^{**} (0.036)	$0.66\%^{**}$ (0.036)	0.255^{**} (0.025)	$0.70\%^{**}$ (0.033)	0.446^{***} (0.002)
EW-Low-Confident-HL − EW-HL_L	$0.55\%^{**}$ (0.089)	-0.611^{***} (0.002)	0.44% (0.23)	−0.753 (0)	0.49% (0.21)	-0.495^{***} (0)
EW-Confident-HL − EW-HL_L	$1.82\%^{***}$ (0)	1.055^{***} (0)	$1.75\%^{***}$ (0)	3.071^{***} (0)	$1.80\%^{***}$ (0)	1.366^{***} (0)
EW-Low-Confident-HL − EW-Low-Confident-HL_L	$1.94\%^{***}$ (0)	1.33^{***} (0)	$1.64\%^{***}$ (0)	1.225^{***} (0)	$1.61\%^{***}$ (0)	1.08^{***} (0)
EW-Confident-HL − EW-Confident-HL_L	$0.99\%^{*}$ (0.059)	1.034^{***} (0.001)	$1.25\%^{**}$ (0.02)	2.511^{***} (0)	$1.42\%^{***}$ (0.009)	1.253^{***} (0)

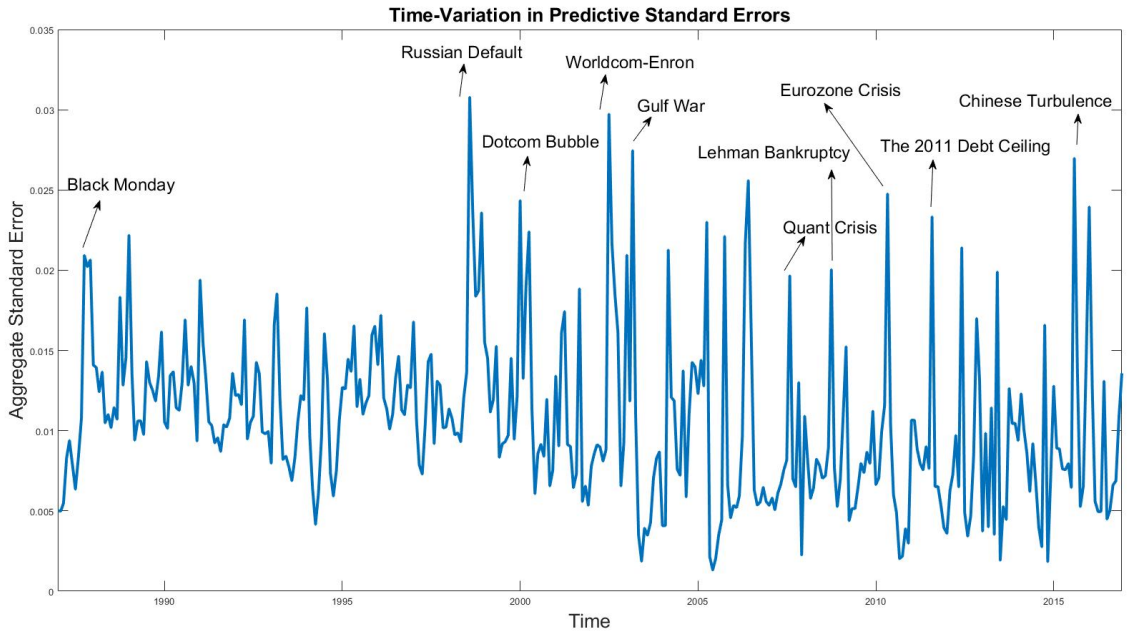
Panel B : Performance Differences of Value-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
VW-HL − VW-HL_L	0.39% (0.249)	0.002 (0.925)	0.33% (0.256)	−0.004 (0.869)	0.27% (0.36)	0.036 (0.423)
VW-Low-Confident-HL − VW-HL_L	0.22% (0.659)	-0.509^{***} (0.004)	−0.09% (0.847)	-0.36^{***} (0.005)	−0.12% (0.793)	-0.221^{**} (0.023)
VW-Confident-HL − VW-HL_L	$1.12\%^{***}$ (0.003)	0.366^{**} (0.013)	$1.22\%^{***}$ (0.001)	0.873^{***} (0)	$0.93\%^{**}$ (0.015)	0.337^{***} (0.004)
VW-Low-Confident-HL − VW-Low-Confident-HL_L	0.98% (0.109)	0.281^{**} (0.036)	0.20% (0.715)	0.051 (0.293)	0.12% (0.818)	0.013 (0.672)
VW-Confident-HL − VW-Confident-HL_L	0.44% (0.344)	0.377^{**} (0.014)	$0.86\%^{**}$ (0.03)	0.855^{***} (0)	$0.81\%^{*}$ (0.072)	0.419^{***} (0.001)

Panel C : Performance Differences of Precision-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
EW-HL − EW-HL_L	$0.72\%^{**}$ (0.016)	0.247^{**} (0.04)	$0.66\%^{**}$ (0.033)	0.255^{**} (0.023)	$0.70\%^{**}$ (0.035)	0.446^{***} (0.002)
LPW-HL − EW-HL_L	$0.57\%^{**}$ (0.049)	−0.06 (0.319)	0.49% (0.127)	−0.125 (0.11)	0.52% (0.127)	0.076 (0.225)
PW-HL − EW-HL_L	$1.08\%^{***}$ (0.002)	0.782^{***} (0.002)	$1.03\%^{***}$ (0.002)	2.796^{***} (0)	$1.07\%^{***}$ (0.002)	1.071^{***} (0)
LPW-HL − LPW-HL_L	$1.06\%^{***}$ (0.001)	0.798^{***} (0.001)	$1.05\%^{***}$ (0.001)	1.787^{***} (0)	$1.05\%^{***}$ (0.003)	0.978^{***} (0)
PW-HL − PW-HL_L	$0.60\%^{*}$ (0.099)	0.273^{**} (0.046)	$0.82\%^{***}$ (0.03)	1.977^{***} (0)	$0.90\%^{***}$ (0.023)	0.529^{***} (0.002)

Figure 11. Time-Series Variation in Standard Errors of NN-based Risk Premia



Note: This figure plots the time-series of aggregate standard errors, which are the cross-sectional averages of NN-3-based risk premium predictions' ex-ante standard errors . The labels, such as "Black Monday", "Russian Default", represent periods of major shocks.

Table X
Aggregate Standard Errors of NN-3-based Risk Premia
This table reports time-series averages of aggregate standard errors over different periods. The aggregate standard errors equal the cross-sectional averages of NN-based risk premium predictions' standard errors.

Panel A: Overall Period

Event	Standard Error	Time Period
Overall Data	1.06%	Jan 1987 to Dec 2016

Panel B: Periods of major Shocks

Event	Standard Error	Time Period
Black Monday	2.05%	Oct 1987 to Nov 1987
Russian LTCM Defualt	3.08%	Sep 1998 to Sep 1998
Dotcom Bubble	2.24%	Apr 2000 to Apr 2000
Worldcom and Enron	2.33%	Jul 2002 to Sep 2002
Gulf War	2.75%	Mar 2003 to Mar 2003
Quant Crisis	1.97%	Aug 2007 to Aug 2007
Lehman Bankruptcy	2.00%	Oct 2008 to Oct 2008
The 2011 Debt-Ceiling	2.32%	Aug 2011 to Aug 2011
Crisis Period Average	2.31%	
Non-Crisis Period Average	1.02%	

Table XI
Cross-sectional Characteristics of Confidence-sorted Deciles

This table reports average characteristics of various confidence-sorted deciles. Every month, stocks are sorted into deciles according to their ex-ante confidence of NN-3-based risk premium predictions. Each row under All Stocks Columns represents the equal-weighted average of various characteristics across all stocks in the corresponding precision-sorted decile. The table also presents the characteristics of confidence-sorted portfolios from the long and short legs, separately. Every period stocks are first sorted into deciles according to their NN-based risk premia, with H and L representing the deciles containing the highest and lowest predicted returns. Both H and L are further partitioned into deciles according to their ex-ante confidence. The Long-Leg columns represent the average characteristics of confidence-sorted deciles of H, whereas Short-Leg columns show those of L.

Ex-ante Precision Decile	All Stocks			Long-Leg			Short-Leg		
	Size	BM	mom12m	Size	BM	mom12m	Size	BM	mom12m
1	1811	1.62	0.01	816	3.45	0.23	1939	0.76	-0.11
2	1836	1.76	0.05	810	3.37	0.23	2003	0.88	-0.08
3	1838	1.97	0.07	793	3.33	0.24	2084	0.92	-0.06
4	1788	2.12	0.08	877	3.20	0.25	2043	0.99	-0.06
5	1750	2.29	0.10	846	3.58	0.26	2102	1.04	-0.06
6	1627	2.39	0.11	805	3.58	0.26	2049	1.03	-0.05
7	1521	2.54	0.12	829	3.50	0.29	2188	0.97	-0.05
8	1394	2.62	0.13	798	3.56	0.31	2206	0.99	-0.05
9	1233	2.72	0.16	706	3.74	0.34	2283	0.89	-0.05
10	988	3.16	0.22	628	4.53	0.42	2347	1.02	-0.07

Table XII
Characteristics Distributions of Stocks in the Decile Containing the Most Confident Risk Premium Predictions

This table reports various characteristic distributions of stocks in the top decile with the most confident risk premium predictions. Every month, stocks are sorted into deciles according to their ex-ante confidence. The first row of the Size column presents the proportion of stocks in the top-most confident decile that have market capital lower than the 10^{th} percentile of sizes across all stocks. Similarly, the second (third, ..., tenth) row of the Size column shows the proportion of stocks in the top-most confident decile that have market capital between the 10^{th} and 20^{th} (20^{th} and 30^{th}, ..., 90^{th} and 100^{th}) percentile of sizes across all stocks. The BM, mom12m, and illiq columns represent equivalent proportions for book-to-market, 1-year momentum and illiquidity characteristics.

Decile	Size	BM	mom12m	illiq
1 (Low-Characteristic)	18.50%	10.02%	9.58%	7.23%
2	15.05%	8.21%	8.33%	6.94%
3	12.61%	8.34%	7.98%	7.03%
4	10.38%	11.39%	8.25%	7.53%
5	8.96%	14.09%	7.89%	8.14%
6	7.92%	11.61%	7.96%	9.21%
7	7.17%	7.64%	9.47%	10.61%
8	6.62%	10.55%	10.88%	12.36%
9	6.56%	13.43%	13.07%	14.54%
10 (High-Characteristic)	6.51%	15.10%	17.04%	16.50%

C. Internet Appendix

C1. Internet Appendix: Simulation Results and Robustness Checks

Table A
Performance of High-Low and Confident High-Low Portfolios: Simulation Evidence
This table compares the performance of the confident high-low portfolios with the conventional high-low portfolios on simulated data. The data contains 200 stock-level simulated true risk premia, NN-3-based estimated risk premia and their standard errors over 60 out-of-sample periods. Every period, the "True High-Low" portfolios take long (short) positions on the stocks with the simulated true risk premia greater (lower) than the $x\%$ ($100-x\%$) percentile of the true risk premia across 200 stocks. x equals 80, 70 and 90 under rule 1, 2 and 3, respectively. The "High-Low" portfolios take long (short) positions on the stocks with NN-3-based risk premium estimates greater (lower) than the $x\%$ ($100-x\%$) percentile of the predicted risk premia in the cross-section. Extreme predicted-return deciles are further partitioned into quantiles according to their precision measures. Panel A (Panel B) presents the results using the absolute t-ratios (inverse standard errors) as proxies for the precision. The "Confident High-Low" portfolios take long-short positions on the top $y\%$ subset of stocks in the extreme predicted return deciles that have the highest precision. y equals 80, 80 and 50 under rule 1, 2 and 3, respectively. The "Matching High-Low" portfolios take (short) positions on the stocks with NN-3-based risk premium predictions greater (lower) than the $z\%$ ($100-z\%$) percentile of the predicted risk premia in the cross-section. See section (C.C2) and equation (75) for a detailed description of the simulated data.

Panel A: Confident-HL Portfolios Constructed Using Absolute t-ratios

	Rule 1		Rule 2		Rule 3	
Portfolio	pred ret	avg ret	pred ret	avg ret	pred ret	avg ret
True High-Low	2.45%	2.45%	2.16%	2.16%	2.74%	2.74%
High-Low	3.04%	1.69%	2.60%	1.45%	3.57%	1.88%
Matching High-Low	3.64%	1.90%	3.45%	1.84%	3.72%	1.92%
Confident High-Low	3.65%	2.31%	3.47%	2.23%	3.74%	2.23%

Panel B: Confident-HL Portfolios Constructed Using Standard Errors

	Rule 1		Rule 2		Rule 3	
Portfolio	pred ret	avg ret	pred ret	avg ret	pred ret	avg ret
True High-Low	2.45%	2.45%	2.16%	2.16%	2.74%	2.74%
High-Low	3.04%	1.69%	2.60%	1.45%	3.57%	1.88%
Confident High-Low	2.72%	2.18%	2.34%	1.99%	3.41%	2.18%

Table B

Performance of Various Long-Short Portfolios: Inverse Standard Errors as Precision Measures

This table reports the performance of various NN-3-based long-short portfolios over the 30-year out-of-sample (OOS) period. This table uses inverse standard errors (rather than the absolute t-ratios) of risk premium predictions as proxies for ex-ante precision (i.e., ex-ante confidence). See table IV and section V.C for a description of the portfolios. The pred ret column represents the average predicted returns. The avg ret column shows the average realized returns. The t, SR and SR^2 columns denote the t-stats of the average returns, annualized Sharpe ratios and squared Sharpe ratios, respectively. Notes: EW = equal-weighted; VW = value-weighted

All Stocks: Equal-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2
EW-HL	1.69%	2.52%	8.21	1.50	2.25
EW-Low-Confident-HL	1.92%	3.02%	7.62	1.39	1.93
EW-Confident-HL	1.69%	3.07%	8.46	1.54	2.39
EW-Confident-HL − EW-HL		0.55%** (0.013)			0.14*** (0.046)
EW-Confident-HL − EW-Low-Confident-HL		0.05% (0.916)			0.45*** (0.001)

All Stocks: Value-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2
VW-HL	1.62%	1.48%	4.95	0.90	0.82
VW-Low-Confident-HL	1.88%	1.13%	2.47	0.45	0.20
VW-Confident-HL	1.64%	1.83%	5.68	1.04	1.08
VW-Confident-HL − VW-HL		0.35%* (0.067)			0.26*** (0.022)
VW-Confident-HL − VW-Low-Confident-HL		0.70%* (0.071)			0.87*** (0.000)

Non-Microcaps: Equal-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2
EW-HL	0.68%	1.66%	5.43	0.99	0.980
EW-Low-Confident-HL	0.72%	1.30%	3.53	0.64	0.35
EW-Confident-HL	0.66%	1.87%	5.95	1.08	1.17
EW-Confident-HL − EW-HL		0.23%** (0.041)			0.19** (0.02)
EW-Confident-HL − EW-Low-Confident-HL		0.57%*** (0.000)			0.82*** (0.000)

Non-Microcaps: Value-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2
VW-HL	0.66%	1.42%	4.64	0.85	0.72
VW-Low-Confident-HL	0.71%	1.25%	2.90	0.53	0.27
VW-Confident-HL	0.65%	1.91%	5.68	1.04	1.08
VW-Confident-HL − VW-HL		0.49%** (0.041)			0.36** (0.001)
VW-Confident-HL − VW-Low-Confident-HL		0.66%* (0.0723)			0.81*** (0.000)

Table C
Comparing Confident-HL Portfolios with Double-sorted HL Portfolios

This table compares the out-of-sample performance of the Confident-HL portfolios with the HL portfolios that are double sorted on predicted-returns. EW(VW)-Confident-HL represents the equal(value)-weighted Confident long-short portfolio that only include stocks with the most confident risk premium predictions. See section V.C for a detailed description of the portfolios. Each period, stocks are sorted into quantiles according to their NN-based risk premia. EW-double-sorted-HL and VW-double-sorted-HL denote the HL portfolios that take equal-weighted and value-weighted long (short) positions on stocks that have greater (lower) predicted-returns than the predicted-return of the 99^{th} (1^{st}) quantile, respectively. The avg ret column presents the average return differences between the pair of investment strategies. The $Sharpe^2$ and IR^2 columns show the annualized squared-Sharpe and squared-information ratio differences between the investment portfolios. The numbers in parenthesis are p-values. *, ** and *** denote significance at the 1%, 5% and 10% levels, respectively.

All Stocks: Equal-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2	IR^2_{FF}	IR^2_{SY}
EW-Confident-HL	1.97%	3.61%	9.58	1.75	3.06	3.12	2.99
EW-double-sorted-HL	2.54%	3.99%	8.58	1.57	2.46	2.49	1.87
Difference		−0.37% (0.168)			0.60^{***} (0.000)	0.96^{***} (0.000)	1.12^{**} (0.000)

All Stocks: Value-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2	IR^2_{FF}	IR^2_{SY}
VW-Confident-HL	1.90%	2.21%	5.95	1.09	1.18	0.87	0.59
VW-double-sorted-HL	2.51%	2.39%	5.28	0.96	0.93	0.5	0.42
Difference		−0.18% (0.61)			0.25^{**} (0.02)	0.37^{**} (0.016)	0.17^{**} (0.03)

Non-Microcaps: Equal-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2	IR^2_{FF}	IR^2_{SY}
EW-Confident-HL	0.66%	2.25%	6.68	1.22	1.49	1.39	1.22
EW-double-sorted-HL	1.02%	2.39%	5.56	1.01	1.02	0.87	0.66
Difference		−0.13% (0.62)			0.47^{**} (0.000)	0.52^{**} (0.000)	0.56^{**} (0.000)

Non-Microcaps: Value-Weighted High-low Portfolios

Strategy	pred	avg	t	SR	SR^2	IR^2_{FF}	IR^2_{SY}
VW-Confident-HL	0.72%	2.07%	5.48	1.00	1.00	0.97	0.69
VW-double-sorted-HL	1.01%	2.20%	4.71	0.86	0.74	0.69	0.44
Difference		−0.13% (0.73)			0.26^{**} (0.000)	0.28^{**} (0.000)	0.25^{**} (0.000)

C2. Internet Appendix: Simulation Details

To assess the finite sample performance of this paper's standard errors and Confident-HL portfolios, I replicate the simulation exercise of GKX.[26] I simulate a 3-factor model for excess returns, for $t = 1, 2, \ldots, T$:

$$r_{i,t+1} = g(z_{i,t}) + e_{i,t+1},\ e_{i,t+1} = \beta_{i,t} v_{t+1} + \epsilon_{i,t+1},\ z_{i,t} = (1, x_t)' \otimes c_{i,t},\ \beta_{i,t} = (c_{i1,t}, c_{i2,t}, c_{i3,t}), \quad (70)$$

where c_t is a 200×180 matrix of characteristics, v_{t+1} is a 3×1 vector of factors, x_t is a univariate time series, and ϵ_{t+1} is a 200×1 vector of idiosyncratic errors. I choose $v_{t+1} = 0,\ \forall t$ under models 1 and 3 and $v_{t+1} \sim \mathcal{N}(0, 0.05^2 \times I)$ under models 2 and 4, respectively. I specify $\epsilon_{i,t+1} \sim \epsilon_{i,t+1} \sim \mathcal{N}(0, 0.05^2)$. These parameters are calibrated so that the average time series R^2 is 50% (40%) and annualized volatility is 24% (30%) under models 1 and 3 (2 and 4). The OOS-R^2 of NN-3-based risk premium predictions on the simulated data is 3.8% (3.2%) under models 1 and 3 (2 and 4).

I simulate the panel of characteristics by

$$c_{ij,t} = \frac{2}{N+1} CSrank(\bar{c}_{ij,t}) - 1,\ \bar{c}_{ij,t} = \rho_j \bar{c}_{ij,t-1} + \epsilon_{ij,t},\ \text{for } 1 \leq i \leq 200,\ 1 \leq j \leq 180, \quad (71)$$

where $CSrank$ denotes the cross-sectional rank.

And the time-series x_t is given by

$$x_t = \rho x_{t-1} + u_t, \quad (72)$$

where $u_t \sim \mathcal{N}(0, 1 - \rho^2)$, and $\rho = 0.95$ so that x_t is highly persistent.

Under models 1 and 2, the parametric form of $g(.)$ is linear and given by

$$g(z_{i,t}) = (c_{i1,t}, c_{i2,t}, c_{i3,t})\theta_0,\ \text{where } \theta_0 = (0.02, 0.02, 0.02)'. \quad (73)$$

In contrast, under models 3 and 4, $g(.)$ takes the following non-linear functional form

$$g(z_{i,t}) = (c_{i1,t}^2, c_{i1,t} \times c_{i2,t}, sgn(c_{i3,t} \times x_t))\theta_0,\ \text{where } \theta_0 = (0.04, 0.03, 0.012)'. \quad (74)$$

To summarize, the simulated true risk premia are linear in characteristics under models 1 and 2, whereas they are non-linear under models 3 and 4. Models 1 and 3 do not entertain cross-sectional temporal residual correlations, whereas models 2 and 4 do.

Lastly, I divide the whole time-series into three consecutive subsamples of equal length (60) for training, validation, and testing, respectively. Although this paper's standard errors are derived

[26] I thank GKX for making their code publicly available.

under the assumption that the residual errors are uncorrelated in the time-series and cross-section, table (I) of the main section indicates that the standard errors are well-calibrated even under models 2 and 4.

Simulations for table (A) of the Internet Appendix use the non-linear specification of model 3, given by

$$r_{i,t+1} = g(z_{i,t}) + e_{i,t+1},\ e_{i,t+1} = \epsilon_{i,t+1},\ z_{i,t} = (1, x_t)' \otimes c_{i,t}, \tag{75}$$

where $\epsilon_{i,t+1} \sim \epsilon_{i,t+1} \sim \mathcal{N}(0, 0.05^2)$, $g(z_{i,t})$ is given by (74) and $c_{i,t}$ is given by (71).

C3. Why Confidence-levels are Better Measures of Precision Relative to Inverse Standard Errors

In this section, I present a simple example showing why the absolute t-stat is a better measure relative to the inverse standard error for constructing Confident-HL portfolios. Consider regressing a given cross-section of excess stock returns on one of stock characteristics (e.g., betas)

$$r_i = \lambda\beta_i + \epsilon_i,\ \epsilon_i \sim MVN(0, \sigma^2 I),\ i = 1, 2, \ldots N \tag{76}$$

where r_i, β_i are assumed to be given. λ, which can be interpreted as the market premium, is an unknown parameter. Assume $\lambda > 0$ without loss of generality. Let $\hat{\lambda}$ be the OLS estimate of λ obtained from the cross-sectional regression in (76).

Now, consider four stocks in the out-of-sample that have betas β_1^*, β_2^*, β_3^* and β_4^*, respectively. Let $0 < \beta_1^* < \beta_2^* < \beta_3^* < \beta_4^*$. Their predicted excess returns are then given by $\beta_1^*\hat{\lambda}$, $\beta_2^*\hat{\lambda}$, $\beta_3^*\hat{\lambda}$ and $\beta_4^*\hat{\lambda}$, respectively. Straightforward algebra implies that these predictions' standard errors equal $\frac{\beta_1^*\hat{\sigma}}{\sum\beta_i^2}$, $\frac{\beta_2^*\hat{\sigma}}{\sum\beta_i^2}$, $\frac{\beta_3^*\hat{\sigma}}{\sum\beta_i^2}$ and $\frac{\beta_4^*\hat{\sigma}}{\sum\beta_i^2}$, respectively. $\hat{\sigma}$ is the OLS estimate of σ in (76).

Thus, the standard errors are proportional to the stock betas. In contrast, the absolute t-ratios are invariant across stocks. In other words, the "confidence-level" of predicting returns is the same across all stocks. In the following paragraph, I show that Confident-HL-se portfolios formed using the standard errors yield sub-optimal returns relative to the traditional HL portfolios. In contrast, Confident-HL-t portfolios formed using the absolute "t-ratios" do not.

Consider the following trading strategies using these four stocks' predicted returns and their precision measures.

1. Conventional-HL: Takes equal-weighted long (short) positions on the top (bottom) stocks with the highest (lowest) predicted returns.

2. Confident-HL-t: Sort stocks into two quantiles based on their predicted returns. Take the long (short) position on the stock in the top (bottom) quantile that has the highest absolute t-ratio. If two stocks have the same absolute t-ratios, take the equal-weighted average.

3. Confident-HL-se: Sort stocks into two quantiles based on their predicted returns. Take the long (short) position on the stock in the top (bottom) quantile that has the lowest standard error. If two stocks have the same absolute standard errors, take the equal-weighted average.

Then the expected return of these three strategies are given by

$$E(\text{Conventional-HL}) = E(\text{Conventional-HL-t}) = \left[\frac{(\beta_3^* + \beta_4^*)}{2} - \frac{(\beta_1^* + \beta_2^*)}{2}\right]\Big(P(\hat{\lambda} > 0) - P(\hat{\lambda} < 0)\Big) \tag{77}$$

$$E(\text{Confident-HL-se}) = (\beta_3^* - \beta_1^*)\left(P(\hat{\lambda} > 0) - P(\hat{\lambda} < 0)\right) \tag{78}$$

For sufficiently large β_4^*, $E(\text{Conventional-HL}) > E(\text{Confident-HL-se})$. Thus, standard errors must always be evaluated relative to the "level" of predictions to obtain better measures of precision.