Confident Risk Premiums and Investments using Machine Learning Uncertainties

Rohit Allena [*]
C.T. Bauer College of Business
University of Houston

First draft: October 16, 2020

Current draft: February 10, 2022

[*] This article extends a few sections of my first dissertation essay at Emory University, which was circulated as "Confident Risk Premia: Economics and Econometrics of Machine Learning Uncertainties". This research was previously presented in seminars at the business schools of Boston College, Boston University, Copenhagen, HEC Paris, HKUST, Indian School of Business, National University of Singapore, Universities of Florida, Georgia, Houston, Rice, Tulane and Yale. It was also presented at the NFA (2021) and FMA (2021) conferences. 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 seminar participants, including Michael Halling (discussant), Praveen Kumar, William Mann (committee), Jegadeesh Narasimhan (committee), Gonzalo Maturano, Christoph Herpfer, and Donald Lee (committee), and Ariel Viale (discussant). Special thanks to Bryan Kelly and Jonathan Lewellen for their insightful comments. All errors are mine.
Address: 4750 Calhoun Road, Houston, TX 77204, E-mail Address: rallena@central.uh.edu

Confident Risk Premiums and Investments using Machine Learning Uncertainties

Abstract

This paper derives ex-ante (co)variances of stock-level and portfolio-level risk premium predictions from neural networks (NNs). Based on the precision of risk premium forecasts, I provide improved investment strategies. The confident high-low strategies that take long-short positions exclusively on stocks with precise risk premium forecasts deliver superior out-of-sample returns and Sharpe ratios than traditional high-low strategies. Mean-variance strategies that incorporate covariances of return predictions also outperform existing strategies. Risk premium variances reflect time-varying market uncertainty and spike after financial shocks. Cross-sectionally, the level and precision of risk premiums are correlated, thus NN-based investments deliver more gains in the long positions.

Keywords: Neural Networks, Return Predictions, Risk Premiums, Standard Errors, Confidence Intervals, Investment Strategies, Machine Learning Uncertainties

I. Introduction

Modern empirical asset pricing literature applies machine learning (ML) models to estimate expected stock excess returns (i.e., risk premiums), as these models can accommodate non-linear relations amongst a high-dimensional set of predictors. In an influential paper, Gu, Kelly, and Xiu (2020) (GKX) document that ML models, particularly neural networks (NNs), outperform linear characteristic-based models examined by Lewellen (2015) (henceforth Lewellen) in predicting stock risk premiums out-of-sample (OOS). However, not much is known about the ex-ante precision (i.e., standard errors and confidence intervals) of risk premium predictions from NNs. For example, 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 premiums (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 premiums is important.

This paper develops an easy-to-implement procedure to estimate predictive standard errors and covariances of NN-based *expected return* predictions at both the stock-level and portfolio-level (e.g., 48 industry portfolios of Fama and French (1997)). 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 premiums are not. I tackle this challenge by adapting the NNs of GKX to simultaneously deliver risk premium predictions and their (co)variances every period. These (co)variances resemble classical bootstrap-based estimators but are available in real-time with no additional computation costs. The obtained (co)variances are then statistically justified using a Bayesian framework, and empirically validated using Monte-Carlo simulations.

Determining the precision of risk premium measurements has several applications, such as making cost-of-capital decisions (e.g., Fama and French (1997)) and conducting out-of-sample inferences (e.g., Allena (2021)). As a novel application, this paper demonstrates why and how incorporating ex-ante precision into trading strategies is important, as it leads to sizable OOS return and Sharpe ratio improvements. In particular, 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 significant enhancements to these HL strategies by exploiting the cross-sectional variation in the ex-ante precision of risk premium measurements. I introduce "Confident-HL" trading strategies that exclusively take long-short positions on the subset of stocks in the extreme predicted-return deciles that have relatively more confident risk premium forecasts (i.e., high absolute ratios of risk premium predictions and their standard errors). Thus, Confident-HL portfolios deliberately exclude stocks in the extreme predicted-return deciles with relatively imprecise risk premium forecasts.

Confident-HL strategies formed using NN-based, or any ML-based, risk premium predictions deliver superior OOS returns and Sharpe ratios. The reason is that ex-ante standard errors of NN-based, or any ML-based, risk premium predictions also predict their ex-post squared forecast errors.[1] 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" decomposition. 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 non-linear functions involving many parameters, variances rather than biases predominantly determine their squared forecast errors. As a consequence, Confident-HL strategies that deliberately drop stocks with imprecise risk premium forecasts earn superior OOS *expected returns*. A simple example provides the central intuition.

Example-1: Consider two stocks A and B with risk premiums μ_A and μ_B, respectively, and $\mu_A > \mu_B$. Let $\hat{\mu}_A$ and $\hat{\mu}_B$ be their risk premium predictions that are normal, uncorrelated and unbiased, with the measurement error variance σ^2. Note that the unbiased assumption suits ML-based predictions, as their variance component dominates the bias component. Then the expected OOS return of the HL strategy that takes a long (short) position on the stock with the highest

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

(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 OOS 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 premiums, the HL strategy formed from more precise predictions yields higher OOS expected returns. For example, when the predictions are precisely measured with $\sigma = 0$, the HL strategy always, and correctly, assigns A (B) in the long (short) leg, yielding the maximum possible expected spread return of $\mu_A - \mu_B$. In contrast, when the predictions are grossly imprecise with $\sigma \to \infty$, the HL strategy wrongly assigns A (B) in the short (long) leg with 50% probability, thus delivering zero expected return. Intuitively, besides the level of risk premium predictions, the precision helps better determine the cross-sectional ranking among stocks. Thus, strategies exclusively containing stocks with precise risk premiums generate higher HL expected returns.

Consistent with this intuition, the empirical section documents large economic gains from the Confident-HL portfolios. 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 premiums 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 in the extreme predicted-return deciles with confident risk premiums 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%). For perspective, this Sharpe ratio improvement translates to 6.8% (4.7%) *annualized holding period return* difference between the EW (VW) Confident-HL and and EW (VW) strategies.

In contrast, OOS returns and Sharpe ratios of the EW (VW) "Low-Confident" portfolio that instead takes long-short positions on the subset of stocks in the extreme predicted-return deciles with the most imprecise risk premiums are relatively much lower, 2.35% (1.31%) and 1.18 (0.55), respectively. Thus, the strategy involving imprecise risk premium predictions reduces the OOS return by 7% (12%) and Sharpe ratio by 22% (39%), compared with the EW (VW) HL strategy.

Recall that the Confident-HL portfolios' impressive performance hinges on the result 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. For example, 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 large 2.21% OOS-R^2, an increase of 170%.

Avramov et al. (2020) argue that NN-based HL strategies 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 strategy yields OOS monthly return and annualized Sharpe ratios of 2.25% (2.07%) and 1.22 (1.00), respectively, whereas the EW (VW) HL yields 1.66% (1.42%) and 0.99 (0.85). 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. The Confident- HL portfolios also outperform two other benchmark strategies that are specifically constructed to have the same predicted return averages and number of stocks, respectively, as the Confident-HL portfolios.

Economically, I show that the Confident-HL strategies could be interpreted as *robust* trading strategies of ambiguity-averse investors who maximize their expected OOS returns, by explicitly considering the confidence-intervals around risk premium predictions. In contrast, the traditional HL portfolios correspond to non-robust strategies that ignore confidence intervals. Thus, this paper's results are consistent with Garlappi, Uppal, and Wang (2007), who documented the superiority of confidence-interval-based *robust* strategies OOS.

Besides the Confident HL portfolios, I also construct mean-variance strategies that incorporate the entire covariance structure (not just individual standard errors) of expected return predictions to maximize the OOS mean-variance utility. These strategies are shown to be equivalent to robust stochastic discount factors (SDFs) proposed by Kozak, Nagel, and Santosh (2019) that take into account estimation uncertainty of return means. I find that even the mean-variance strategies yield significant Sharpe ratio enhancements relative to the existing HL portfolios. For example, the mean-variance strategy formed using non-microcaps earns an annualized Sharpe ratio of 1.26, which is 25% (46%) larger than the EW (VW) HL's Sharpe ratio of 0.99 (0.85). And the mean-variance strategy constructed using 48 industry portfolios of Fama and French (1997) yields an annualized Sharpe ratio of 1.23, which is nearly double the Sharpe ratio (0.66) of the EW strategy.

One may wonder whether analogous Confident HL and mean-variance strategies constructed using predictions from simple models with a few parameters (e.g., Lewellen), rather than NNs, also deliver economic gains OOS. However, I argue and document that such strategies formed using Lewellen's risk premium predictions and their (co)variances do not yield significant gains. For example, the OOS Sharpe ratio of Lewellen's mean-variance strategy is one-third the Sharpe ratio of Lewellen' HL strategy. The reason is biases (or model misspecification), rather than variances, more likely determine expected squared forecast errors of simple models. Thus, the ex-ante-precision-based strategies such as Confident-HL and mean-variance strategies deliver gains predominantly in the context of NN-based (or any ML-based) predictions that are relatively less biased.

Finally, to *statistically* compare the OOS returns and Sharpe ratios of all the considered trading strategies, I conduct moving block bootstrap tests of Allena (2021).[2] These tests are more conservative than Diebold and Mariano (2002) (DM, henceforth) inferences because they take into account the ex-ante parameter uncertainty.[3] The bootstrap tests suggest that the NN-3-based Confident-HL and mean-variance strategies statistically outperform all other competing strategies, including NN-3-based conventional HL strategies, as well as Lewellen-based HL, Confident-HL and mean variance strategies. More interestingly, I find that the relative performance of NN-3 over Lewellen increases monotonically with the precision of NN-3-based risk premiums. For example, the average monthly return difference between NN-3 and Lewellen VW HL portfolios formed using the stock returns most confidently predicted by NN-3 is a highly significant 0.82%. In contrast, the corresponding difference is a significantly negative -1.2% on the subset of stock returns most impre-

[2] Allena and Robotti (2021) study the asymptotic properties of these moving block bootstrap tests.

[3] Thus, if the bootstrap tests imply that results are significant, then the DM tests imply significance too.

cisely predicted by NN-3. Thus, these results highlight the paper's central point that incorporating ex-ante (co)variances of ML-based risk premium predictions into trading strategies is important.

In the end, to shed some light on the dynamics of the precision-based strategies, I document several time-series and cross-sectional properties of the ex-ante precision. The time-series of standard errors reflect stock market uncertainty, with standard errors increasing by a factor of two, on average, after major shocks such as Black Monday and Lehman Bankruptcy. Because many individual predictors (e.g., price trends) in the NN-3 model substantially deviate from their usual distributions when markets are uncertain, risk premium predictions based on these unusual predictors will also be imprecise. As result, standard errors capture time-varying market uncertainty. In the cross-section, the NN-3 model (*ex-ante*) confidently predicts risk premiums of stocks with small marketcaps, with high book-to-market ratios, with high 1-year momentum returns, and with high risk premium predictions. Thus, NN-3-based investment strategies will deliver more gains in the long positions rather than in the short positions. Possible mechanisms that lead to the cross-sectional differences in precision of risk premium predictions warrants a future study.

Overall, this paper estimates the ex-ante precision of the NN-based risk premium predictions and demonstrates that incorporating the precision into NN-based trading strategies is important.

A. Contribution

The paper makes three methodological and investment-related contributions.

Methodological. This is the first paper to estimate the stock-level and portfolio-level (co)variances of NN-based *risk premium* predictions. I make methodological advancements relative to Gal and Ghahramani (2016), who estimated standard errors of NN-based predictions, in *three dimensions*. First, Gal and Ghahramani (2016) do not apply their method on financial data. Moreover, they compute the variances of individual "raw" predictions (equivalent to excess return predictions), not of "prediction means" (comparable to risk premium predictions), which are more relevant in the finance literature. Second, while they estimate "raw prediction" variances using a Bayesian framework, they do not provide "joint densities" of different prediction means nor their covariances, which are necessary for computing mean-variance strategies and portfolio-level variances. Last, they do not show whether their Bayesian standard errors satisfy frequentist properties. In this paper, I show how to compute the marginal and joint densities of NN-based risk premium predictions, and thereby I show how to estimate the covariances of risk premium forecasts. In addition, I prove the frequentist consistency of the Bayesian (co)variance estimators.

The paper also relates to several methodological papers outside the finance literature that conduct inferences based on various ML-based predictions. For example, Farrell, Liang, and Misra (2021) provide nonasymptotic high-probability bounds for neural network predictions using a semi-parametric framework. However, they do not explicitly provide confidence intervals nor joint densities of NN-based predictions, which are the main focus of this paper. Wager, Hastie, and Efron (2014), and Wager and Athey (2018) provide methods to estimate standard errors of predictions based on random forests. Likewise, Kyung, Gill, Ghosh, and Casella (2010) provide a Bayesian framework to estimate otherwise intractable standard errors of LASSO-based predictions.

Investment Portfolios. The paper relates to GKX and Avramov et al. (2020), who construct HL portfolios based on various ML-based return predictions. Alternatively, this paper shows how Confident-HL and mean-variance strategies could deliver superior performance OOS. In related studies, Ahn, Conrad, and Dittmar (2009) and Bryzgalova, Pelger, and Zhu (2020) construct a set of statistically-motivated test assets to test asset pricing models. Instead, I focus on obtaining a single optimal portfolio (rather than many test assets) by explicitly deriving the (co)variances of risk premium predictions, which both studies do not consider.

Finally, the Confident-HL strategies developed in this paper fundamentally differ from the idiosyncratic volatility (IVOL) strategies (Ang, Hodrick, Xing, and Zhang (2006)). Whereas IVOL strategies take short positions on stocks with relatively *large* idiosyncratic return volatilities, the Confident-HLs totally exclude stocks with *relatively large* risk premium variances. In addition, this paper's ex-ante risk premium variances are conditional measures that depend on the information from the predictor set, whereas the IVOL measures are not. In fact, an analogous IVOL-based-Confident-HL strategy that is constructed using IVOLs (rather than ex-ante risk premium variances of this paper) delivers significantly lower OOS returns and Sharpe ratios than the HL portfolios.

II. NN-based Risk Premiums and their Ex-ante Precision

This section presents the statistical framework to estimate NN-based stock-level and portfolio-level risk premium predictions and their (co)variances. Noting that an NN that employs *dropout* regularization is identical to a Bayesian NN with a similar structure, I estimate NN-based risk premium variances using the comparable Bayesian models' *instantly* available posterior variances.[4] Although Bayesian posterior variances and frequentist variances philosophically represent different entities, the section proves the *frequentist consistency* of the estimated variances.

Notations: $\widehat{(.)}$ represents an estimator for the underlying parameter $(.)$; Var represents the variance operation; $\xrightarrow{p}$ denotes convergence in probability; $||f - g||_{\text{TV}}$ denotes the total-variation distance between the two densities f and g, which equals $\frac{1}{2}\int |f(x) - g(x)|dx$. Internet appendices C and D provide all proofs and conditions required for establishing the frequentist consistency.

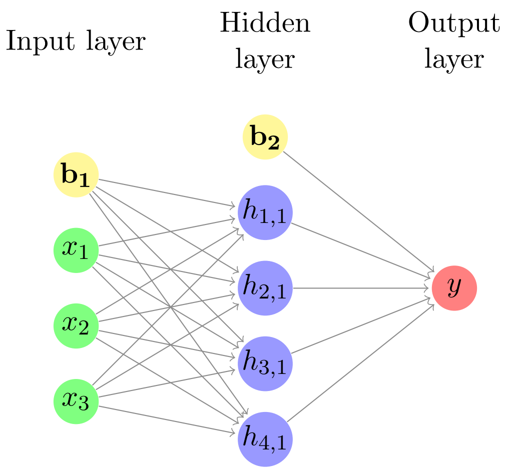
A. Neural Networks

Like GKX, this paper considers "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 the 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.

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. $\{x_1, x_2, x_3\}$, $\{h_{k,1}\}_{k=1}^4$, 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

[4] "NNs with dropout regularization being identical to Bayesian NNs" is analogous to "linear regressions that employ L_2 regularization (i.e., Ridge regressions) being identical to Bayesian linear regressions".

Figure 1. Example of a 1-layer Neural Network



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

sum of the preceding layer's neurons plus an intercept, known as "bias". Thus,

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

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. Like 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

$$\phi(x) = ReLU(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ otherwise.} \end{cases} \tag{3}$$

Likewise, the final output is given by

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

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

$$y_{output} = b_2 + \phi(b_1 + xW_1)W_2, \tag{5}$$

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 a general additive prediction error model for realized excess returns,

$$r_{it+1} = f(z_{it}; \beta) + \eta_{i,t+1}, \;\; \eta_{i,t+1} \sim N(0, \sigma_\eta^2 I) \tag{6}$$

where $r_{i,t+1}$ is stock i's excess return at period $t+1$; z_{it} is a large set of stock i's raw predictors, such as size, book-to-market, 1-month momentum returns, at time t; f is a flexible non-linear model, which takes the parametric form in (5), with $\beta = \{W_1, W_2, b_1, b_2\}$, $x = z_{it}$, when the model is NN-1.

Because the parameters are unknown, risk premiums are measured using $f(z_{it}; \hat{\beta})$, where $\hat{\beta}$ are estimated parameters of β. Given a panel of "training data", the literature typically minimizes the regulaized mean of squared forecast errors to estimate the parameters, i.e

$$\hat{\beta}_{\lambda} = \arg\min_{\beta} \frac{1}{N_{Tr}N_S} \sum_{t\in Tr} \sum_{i\in S} \left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it}W_1)W_2)\right)^2 \\ + \lambda\left[||W_1||^2 + ||W_2||^2 + ||b_1||^2 + ||b_2||^2\right], \tag{7}$$

where Tr is the training sample over N_{Tr} periods; S is the total set of N_S stocks; $||.||$ represents the L_2 norm operator; and λ is the L_2 "hyperparameter", which prevents overfitting.

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 mean forecast squared error on a panel of "validation data" that do not overlap with the training data. In particular,

$$\lambda = \arg\min_{\lambda\in\Lambda} \frac{1}{N_V N_S} \sum_{t\in V} \sum_{i\in S} \left(r_{i,t+1} - f(z_{it}, \hat{\beta}_{\lambda})\right)^2, \tag{8}$$

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

Thus, (7) and (8) together determine the estimated parameters and hyperparameters. Because minimizing (7) is not possible 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. Since this procedure could be computationally intensive, literature uses stochastic gradient descent (SGD) algorithm that considers random samples (rather than the full sample) from the training data to iteratively update the parameters until they converge.[5]

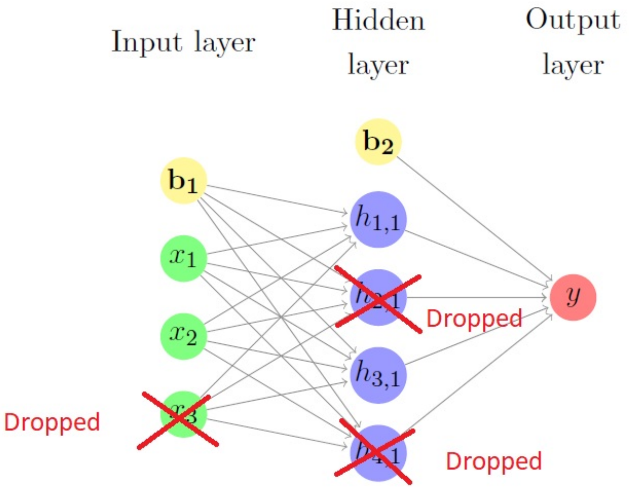
Besides L_2, I discuss another regularization known as *dropout* that can be employed either exclusively or simultaneously with other penalties, such as L_2 or L_1.[6] Dropout is useful because it boosts the performance of NN models and simultaneously delivers prediction variances.

Dropout. 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,

[5] See GKX for a detailed review of parameter estimation using SGD and other regularizations such as L_1.

[6] Dropout is proposed by Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014).

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.

discarded) during that iteration but are allowed to become active in the next iteration. Like λ for L_2, $(1-p)$ (p) is a hyperparameter for Dropout. Thus, the optimal "dropout rate" ("retention rate") $1-p$ (p) is chosen to minimize the validation mean squared error. After training and obtaining estimated parameters, neurons are no longer dropped 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.

Thus, estimated parameters of an NN-1 that employs dropout and L_2 regularizations satisfy

$$\hat{\beta}_{\lambda,p} = \arg\min_{\beta} \frac{1}{N_{Tr}N_S} \sum_{t\in Tr} \sum_{i\in S} \left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it}(p_{1it}W_1))(p_{2it}W_2))\right)^2 \\ + \lambda\left[||W_1||^2 + ||W_2||^2 + ||b_1||^2 + ||b_2||^2\right], \tag{9}$$

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 (9). Hereafter, an NN that employs L_2 and dropout regularizations will be called a "dropout NN".

Stock-level risk premiums. 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

$$\hat{E}_t(r^*_{i,t+1}) = E^*_{it,Dropout} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it} W_{1,\{\lambda,p\}}) W_{2,\{\lambda,p\}}), \;\; r^*_{i,t+1}, z^*_{it} \in Te, \tag{10}$$

where the parameters, $\{b_{2,\{\lambda,p\}}, b_{1,\{\lambda,p\}}, W_{1,\{\lambda,p\}}, W_{2,\{\lambda,p\}}\}$, are given in (9). $E^*_{it,Dropout}$ 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) (see section 7.5) show that the predictions given in (10) approximately equal the sample averages of corresponding predictions that employ dropout at the test time as well. In particular,

$$E^*_{it,Dropout} \approx \frac{1}{D} \sum_{d=1}^{D} (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1id} W_{1,\{\lambda,p\}}))(p_{2id} W_{2,\{\lambda,p\}})), \;\; r^*_{i,t+1}, z^*_{it} \in Te, \tag{11}$$

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 premiums. 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

$$\hat{E}_t(r^*_{P,t+1}) = E^*_{Pt,Dropout} = \sum_{i=1}^{S} \omega_{P,i,t} E^*_{it,Dropout}, \;\; r^*_{i,t+1} \in Te, \tag{12}$$

where $r^*_{P,t+1} = \sum_{i=1}^{S} \omega_{P,i,t} r^*_{i,t+1}$, and $E^*_{it,Dropout}$ is given in (10).

Before formally discussing Bayesian NNs, I now discuss how to instantly obtain (co)variances of dropout NN-based risk premium predictions.

C. (Co)Variances of Risk Premium Predictions based on Neural Networks

Stock-level variances. 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

$$E^*_{it,Dropout} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}W_{1,\{\lambda,p\}})W_{2,\{\lambda,p\}}), \; r_{i,t+1}, z^*_{it} \in Te. \tag{13}$$

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

$$\widehat{Var}_t(E^*_{it,Dropout}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t+1} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t+1}\right)^2, \tag{14}$$

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

$$\hat{E}_{i,d,t} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1d}W_{1,\{\lambda,p\}}))(p_{2d}W_{2,\{\lambda,p\}})), \; z^*_{it} \in Te. \tag{15}$$

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.[7]

Intuition. To summarize, after estimating an NN-1 model's weights using the training and validation data sets, variances 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 next subsection shows, this procedure is equivalent to drawing samples from the risk premium's predictive distribution under a comparable Bayesian NN having the same number of neurons and hidden layers as the considered NN-1.

Stock-level covariances. The predictive covariance between any two estimated stock risk

[7] An ideal D trades-off between latency and accuracy because the former (latter) decreases (increases) with D.

premiums $E^*_{it,Dropout}$ and $E^*_{jt,Dropout}$ is estimated by

$$\widehat{Covar}_t(E^*_{it,Dropout}, E^*_{jt,Dropout}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t+1} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t+1}\right)\left(\hat{E}_{j,d,t+1} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{j,d,t+1}\right), \tag{16}$$

where $\hat{E}_{i,d,t}$ and $\hat{E}_{i,d,t}$ are given in (15).

Portfolio-level variances. The predictive variance of a portfolio-level risk premium prediction is estimated by

$$\widehat{Var}_t(E^*_{Pt,Dropout}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2, \tag{17}$$

where

$$\hat{E}_{P,d,t} = \sum_{i=1}^{S}\omega_{P,i,t}\left(b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1d}W_{1,\{\lambda,p\}}))(p_{2d}W_{2,\{\lambda,p\}})\right), \; z^*_{it} \in Te, \tag{18}$$

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

The procedures for computing stock-level covariances and portfolio-level variances deserve emphasis. Note that the dropped weights (i.e., p_{1d}, p_{2d} draws) are the *same* across stocks i and j, and across all stocks that compose P, respectively. I prove that this preserves cross-sectional correlations among stock-level risk premium predictions, delivering consistent estimators.

The outlined procedure for obtaining standard errors in (14) and (17) 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 also worth emphasizing that (14) and (17) yield variances of risk premium predictions and not excess return predictions. Because realized excess returns equal the sum of risk premiums and unexpected returns due to unpredictable new information, their predictive variances equal the sum of predictive variances of risk premium predictions and "irreducible-variance" due to unexpected returns. 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 could be easily estimated as well.

D. Dropout Neural Networks and Bayesian Interpretation

This subsection statistically validates the previously presented (co)variance estimators by showing that dropout NNs and Bayesian NNs are identical. Gal and Ghahramani (2016) proved the dropout NN and Bayesian NN equivalence by drawing upon the probability theory of Gaussian processes, thereby limiting the potential audience for their work. So, I use a simple Bayesian model to provide a straightforward but rigorous discussion of their central conclusions. In addition, I derive stock-level and portfolio-level *risk premium* (co)variances and prove their frequentist consistency, which Gal and Ghahramani (2016) do not discuss.

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

$$r_{i,t+1} = b_2 + \phi(b_1 + z_{it}W_1)W_2 + \eta_{i,t+1},\ E_t(\eta_{i,t+1}^2) = \sigma_\eta^2 \tag{19}$$

where the parameters $\{W_1, W_2\}$ are random. σ_η^2 and $b = (\{b_1, b_2\})$ are assumed to be known for simplicity.[8] Denote the risk premiums by μ_{it}, where

$$\mu_{i,t} = E_t(r_{it+1}) = b_2 + \phi(b_1 + z_{it}W_1)W_2. \tag{20}$$

Specify the unknown weight matrices with the standard Gaussian priors,

$$[W_1, W_2] = \mathcal{N}(0, l^{-2}I), \tag{21}$$

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,

[8] $\{b_1, b_2\}$ could be treated random as well, in which case these parameters must be specified with Gaussian priors.

and the past training and validation data sets, denoted by $\{R, Z\}$, is given by

$$P(\mu^*_{i,t}|z^*_{it}, R, Z) = \int P(\mu^*_{i,t}|z^*_{it}, R, Z, W_1, W_2, b, \sigma^2_\eta)P(W_1, W_2|R, Z, b, \sigma^2_\eta)dW_1 dW_2, \tag{22}$$

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 uses one of the powerful methods known as variational inference (VI) to directly approximate the intractable posterior.

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 (7), 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 convergence (in total-variation) of $q_{\theta^*}(W)$ to the true posterior.[9] For reference in the finance literature, see Allena and Chordia (2020), who develop a specialized VI method to approximate the intractable posterior density of true stock liquidity and prices, accounting for tick-size induced rounding biases.

Variational Inference for Bayesian Neural Networks. To approximate the posterior of the NN weight matrices, Gal and Ghahramani (2016) consider the following family of Gaussian mixture densities containing two components:

$$q_{\{M_1,M_2\}}(W_1, W_2) = q_{M_1}(W_1)q_{M_2}(W_2), \text{ with } q_{M_1}(W_1) = \prod_{k=1}^{Q} q_1(w_{1q}),\ q_{M_2}(W_2) = \prod_{k=1}^{K} q_2(w_{2q}),$$

$$\text{where } q_i(w_{iq}) = p\mathcal{N}(m_{iq}, \sigma^2 I_i) + (1-p)\mathcal{N}(0, \sigma^2 I_i) \text{ for } i = 1, 2, \tag{23}$$

with $M_1 = [(m_{1q})]$ and $M_2 = [(m_{2q})]$ being the "variational" parameters to be optimized. σ^2

[9] See Blei, Kucukelbir, and McAuliffe (2017) for an excellent review of VI, where they discuss: i) what family of densities to consider? ii) how to obtain the optimal density in the family that best approximates the true posterior?

and p are known scalars. I_1 (I_2) is the identity matrix of dimension K (1); M_1 and M_2 are matrices with the same dimensions as W_1 and W_2, respectively. Note that the variational density $q_{M_1,M_2}(W_1, W_2)$ induces strong joint correlations over the rows of matrices W_i, which will help capture the correlations among different risk premium predictions.

The optimal variational parameters $\{M_1^*, M_2^*\}$ that best approximate the true posterior are

$$\{M_1^*, M_2^*\} = \arg \min_{\{M_1,M_2\}} KL \left(q_{M_1}(W_1) q_{M_2}(W_2) || P(W_1, W_2 | R, Z_b, \sigma_\eta^2) \right), \tag{24}$$

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

Bayesian and Dropout Neural Network Equivalence. It turns out that, given the sample of training data, and as the number of neurons $K \to \infty$, the optimal parameters in (24) minimize the loss function that resembles a dropout NN's frequentist-based loss function (9).

$$\{M_1^*, M_2^*\} = \arg \min_{\{M_1,M_2\}} \frac{1}{N_{Tr} N_S} \sum_{t \in Tr} \sum_{i \in S} \left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it}(p_{1it} M_1))(p_{2it} M_2)) \right)^2$$
$$+ \mu_1 ||M_1||^2 + \mu_2 ||M_2||^2 + \mu_3 ||b_1||^2 + \mu_4 ||b_2||^2, \tag{25}$$

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

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

$$M_1^* = W_{1,\{\lambda,p\}}, \text{ and } M_2^* = W_{2,\{\lambda,p\}}. \tag{26}$$

Thus, predicting risk premiums using dropout NNs and Bayesian NNs are equivalent. As a consequence, the following results follow.

Denote the VI-based approximated posterior densities of risk premiums by

$$P_{VI}(\mu^*_{i,t}|z^*_{it}, R, Z) = \int P(\mu^*_{i,t}|z^*_{it}, R, Z, W_1, W_2, b, \sigma^2_\eta) q_{M^*_1, M^*_2}(W_1, W_2) dW_1 dW_2, \tag{27}$$

where the VI-based density $P_{VI}(\mu^*_{i,t}|z^*_{it}, R, Z)$ approximates the true posterior $P(\mu^*_{i,t}|z^*_{it}, R, Z)$; $\{M^*_1, M^*_2\}$ are given in (26), and $q_{M^*_1, M^*_2}(.)$ in (23), with optimal M^*_1, M^*_2 substituted for M_1, M_2.

Theorem 1: *The dropout-NN-based frequentist risk premium predictions* (13) *converge in probability to the posteriors mean of VI-based risk premium densities as the dropout samples $D \to \infty$ and the number of neurons $K \to \infty$, i.e.,*

$$E^*_{it,Dropout} \xrightarrow{p} E_{VI}(\mu^*_{i,t}), \tag{28}$$

where $E_{VI}(\mu^*_{i,t})$ denotes the expectation of $P_{VI}(\mu^*_{i,t}|z^*_{it}, R, Z)$.

Theorem 2: *The dropout-based estimated variances of stock-level risk premiums* (14) *converge in probability to the variances of risk premiums' VI-based approximated posterior densities as the dropped-out samples $D \to \infty$ and the number of neurons $K \to \infty$, i.e.,*

$$\widehat{Var}_t(E^*_{it,Dropout}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2 \xrightarrow{p} Var_{VI}(\mu^*_{i,t}), \tag{29}$$

where $\hat{E}_{i,d,t}$ is given in (15)*; $Var_{VI}(\mu^*_{i,t})$ denotes the variance of $P_{VI}(\mu^*_{i,t}|z^*_{it}, R, Z)$. .*

Now, consider the VI-approximated joint posteriors of a given set of S risk premiums

$$P_{VI}(\mu^*_{1,t}, \mu^*_{2,t}, \ldots, \mu^*_{S,t}|z^*_{it}, R, Z) = \int P(\mu^*_{1,t}, \mu^*_{2,t}, \ldots, \mu^*_{S,t}|z^*_{it}, R, Z, W_1, W_2, b, \sigma^2_\eta) q_{M^*_1, M^*_2}(W_1, W_2) dW_1 dW_2, \tag{30}$$

where the VI-based density $P_{VI}(\mu^*_{1,t}, \mu^*_{2,t}, \ldots, \mu^*_{S,t}|z^*_{it}, R, Z)$ approximates the true posterior $P(\mu^*_{1,t}, \mu^*_{2,t}, \ldots, \mu^*_{S,t}|z^*_{it}, R, Z)$; $\{M^*_1, M^*_2\}$ are given in (26), and $q(.)$ in (23).

Theorem 3: *The dropout-based estimated covariances of stock-level risk premiums* (16) *converge in probability to the covariances of risk premiums' VI-based approximated posterior densities as the* *dropped-out samples $D \to \infty$ and the number of neurons $K \to \infty$, i.e.,*

$$\widehat{Covar}_t(E^*_{it,Dropout}, E^*_{jt,Dropout}) \xrightarrow{p} Covar_{VI}(\mu^*_{i,t}, \mu^*_{j,t}), , \tag{31}$$

where $Covar_{VI}(\mu^*_{i,t}, \mu^*_{j,t})$ denotes the covariance between $\mu^*_{i,t}, \mu^*_{j,t}$ based on the joint density (30).

Theorem 4: *The dropout-based estimated variances of portfolio-level risk premiums (17) converge in probability to variances of the portfolio-level risk premiums' VI-based approximated posterior densities as the dropped-out samples $D \to \infty$ and the number of neurons $K \to \infty$, i.e.,*

$$\widehat{Var}_t(E^*_{Pt,Dropout}) \xrightarrow{p} Var_{VI}(\mu^*_{Pt,Dropout}), \tag{32}$$

where $\mu^*_{Pt,Dropout} = \sum_{i=1}^{S} \omega_{P,i,t}\mu^*_{i,t}$; $\omega_{P,i,t}$ presents the weights that determine the portfolio P; $Var_{VI}(\mu^*_{Pt,Dropout})$ denotes the posterior variance of $\mu^*_{Pt,Dropout}$ based on the joint density (30).

E. Frequentist consistency of dropout-based estimators

Note that theorems 1-4 show that the dropout-based risk premium predictions and their covariances correspond to the VI-based *Bayesian* posterior means and covariances, respectively. The following result shows that the dropout-based estimators exhibit frequentist consistency.

Theorem 5: *Under the assumptions 1-3 in Internet appendix B, as the number of neurons $K \to \infty$ and in the limit of infinite data, for a given finite set of S stock risk premiums*

$$\left|\left| P_{VI}(\mu^*_{1,t}, \mu^*_{2,t}, \ldots, \mu^*_{S,t} | z^*_{it}, R, Z) - MVN\left([\hat{\mu}_{1,t}, \ldots, \hat{\mu}_{S,t}], n^{-1} I(\mu_{1,t}, \ldots, \mu_{S,t})\right) \right|\right|_{TV} \xrightarrow{p} 0, \tag{33}$$

where MVN denotes the multivariate normal density; $[\hat{\mu}_{1,t}, \ldots, \hat{\mu}_{S,t}]$ represents the maximum likelihood estimate (MLE) of true risk premiums; $I(\mu_{1,t}, \ldots, \mu_{S,t})$ denotes the Fisher information matrix evaluated at the true risk premiums; n^{-1} is the total number of observations in the training data.

Theorem 5 shows that Bayesian credible sets formed using the dropout-based or the VI-based risk premium predictions and their (co)variances will asymptotically be confidence intervals obtained using frequentist MLE estimators and their (co)variances. Thus, this paper's dropout-based covariance estimators are justified from the frequentist standpoint.

F. Validating standard errors using simulations

Simulation study in Internet Appendix B (table H) affirms that the estimated variances are well-calibrated in the frequentist sense. Using a high dimensional predictor set, I simulate risk premiums 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.

III. Improved Trading Strategies using Risk Premium Variances

Building on the bias-variance decomposition, this section shows how the previously estimated (co)variances of ML-based risk premiums could be exploited to form improved investment strategies.

A. Bias-variance decomposition

Consider a general additive prediction error model for realized excess stock returns,

$$r_{i,t+1} = E_t(r_{i.t+1}) + \epsilon_{i,t+1},\ E_t(\epsilon_{i,t+1}) = 0,\ V_t(\epsilon_{i,t+1}) = \sigma^2, \tag{34}$$

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. $\epsilon_{i,t+1}$ are iid over time and across stocks.

Like section II, let a flexible model $f(z_{it};\beta)$, involving stock-level predictors $\{z_{it}\}_{(it)}$ and parameters β, estimates unobserved risk premiums. Because the true parameters, β, are unknown, risk premiums are estimated using $\widehat{E_t(r_{i,t+1})} = f(z_{it};\hat{\beta})$, $\forall$ stocks i, where $\hat{\beta}$ are estimated parameters from the past data. The expected squared forecast errors of the model-based risk premium predictions are then given by

$$\left[\left(E_t(r_{i,t+1}) - f(z_{i,t};\hat{\beta})\right)^2\right] = E_t\left[\left(r_{i,t+1} - f(z_{i,t};\hat{\beta})\right)^2\right] - V_t(\epsilon_{i,t+1}),\ \forall i. \tag{35}$$

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 premiums:

$$r_{i,t+1} = f(z_{it};\beta) + \eta_{i,t+1},\ E_t(\eta_{i,t+1}) = 0, \tag{36}$$

where risk premium predictions are given by $\widehat{E_t(r_{i,t+1})} = f(z_{it};\hat{\beta})$. Thus, the expected squared forecast errors of return predictions based on (36) could be decomposed as the sum of three terms:

$$E_t\left[(r_{i,t+1} - f(z_{i,t};\hat{\beta}))^2\right] = \underbrace{\left(E_t(r_{i,t+1}) - E_t(f(z_{i,t};\hat{\beta}))\right)^2}_{Bias^2} + \underbrace{E_t\left(f(z_{i,t};\hat{\beta}) - E_t(f(z_{i,t};\hat{\beta}))\right)^2}_{Variance} + V_t(\epsilon_{i,t+1}). \tag{37}$$

The first term in the right hand side of (37), popularly known as "squared-bias", measures the model misspecification of $f(.)$ in estimating true risk premiums. The second, known as "variance", quantifies parameter uncertainty. The ex-ante risk premium variances derived in the previous section are *consistent* estimators 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 $V_t(\epsilon_{i,t+1})$s are stock-invariant (time-invariant), the squared-bias and variance components wholly determine the cross-sectional (time-series) variation in squared forecast errors.

As a result, the following remarks follow when true risk premiums are non-linear functions of a large set of predictors.

Remark 1: *Ex-ante variances of risk premium predictions based on simple (i.e., models with a sparse set of predictors) linear models are less likely to predict their ex-post squared forecast errors.*

Simple linear models comprise few parameters and thus their parameter uncertainty is more likely dominated by the models' misspecification when true risk premiums are non-linear functions of many predictors. For example, consider predicting risk premiums with zero across all states. By construction, these set of predictions have zero variances. However, these predictions are highly biased, and thus will have high squared forecast errors. Thus, squared-biases rather than variances predominantly predict ex-post squared forecast errors of simple linear models.

Although ex-ante squared-biases will predict simple models' squared forecast errors, measuring them is not possible because true risk premiums are unknown. Thus, it is not possible to construct potentially rewarding Low-bias-HL (analogous to Confident-HL) strategies either.

Remark 2: *In contrast to remark 1, ex-ante variances of ML-based risk premium predictions predict their ex-post squared forecast errors.*

ML-based predictions are less likely to be misspecified because they capture non-linear functions involving a large predictor set.[10] However, these predictions likely have large variances because of parameter uncertainty. Thus, ex-ante variances of ML-based risk premium predictions predominantly determine their squared forecast errors.

Consistent with these remarks, the empirical section documents that the ex-ante variances of the NN-based risk premiums significantly predict their ex-post squared forecast errors, whereas those of the Lewellen-based predictions do not. The following subsection shows how these informative ML-based ex-ante variances could be used in real-time to form improved investment portfolios.

B. Risk Premium Variances and Confident-HL Strategies

Building on example-1 and remark 2, this subsection illustrates why the Confident-HL portfolios yield superior expected returns. I use simulations because computing expected OOS returns of sorting-based HL strategies requires obtaining various moments of "order statistics" of multivariate normal densities, which are not available in the closed-form expressions.

Consider a simple model based on two sets of stocks, viz. S_A, S_B, each containg $2N$ stocks.

[10] GKX note that NNs capture non-linear relations among predictors and thus deliver superior OOS performance.

Let the stocks in S_A and S_B have the true expected risk premiums of μ_A and μ_B, respectively, with $\mu_A > \mu_B$. Because these risk premiums are unknown, consider an econometric model that delivers unbiased, normal, and independent predictions of stock risk premiums. Note that the unbiased assumption suits ML models, which aligns with remark 2. Further suppose that the risk premium predictions of N stocks each in S_A and S_B are relatively precisely (imprecisely) measured with the variance of σ_l^2 (σ_h^2), and $\sigma_l^2 < \sigma_h^2$.

Denote Q_L (Q_S) as the *median* portfolio of stocks containing the top (bottom) $2N$ stocks that have relatively highest (lowest) risk premium predictions. Now, consider the following sorting-based trading strategies formed using risk premium predictions and their variances.

1. EW-HL. This strategy takes EW long (short) positions on all stocks in Q_L (Q_S).

2. Confident-HL. This strategy further sorts stocks in the median portfolio, Q_L (Q_S), based on their confidence levels (i.e., absolute t-ratios) and takes long (short) positions on the subset of top N stocks with relatively higher confidence-levels.

3. Low-Confident-HL. In contrast, this strategy takes EW long (short) positions on the subset of N stocks in Q_L (Q_S) with relatively lower confidence levels.

Thus, Confident-HL and Low-confident-HL are conditional strategies that first sort stocks based on their risk premium predictions and later on their confidence levels. Note that the EW-HL strategy takes EW long (short) positions on $2N$, whereas the Confident-HL and Low-Confident-HL strategies go long (short) only on N stocks. Thus, to make a fair comparison, I also consider the following double-sorted strategy.

4. Double-sorted-HL. This strategy further sorts stocks in the median portfolio, Q_L (Q_S), based on their risk premium predictions and takes long (short) positions on the top N stocks with relatively higher (lower) return predictions. In other words, this strategy takes EW long (short) positions on the top (bottom) N stocks with the highest (lowest) return predictions.

Table I presents the expected OOS monthly returns of all trading strategies formed using 200 stocks for a wide range of parameters (i.e., μ_A, μ_B, σ_l, and σ_h), over 30 years of simulated data. Across all specifications, the Confident-HL strategy outperforms all other trading strategies in terms of expected OOS returns. Because of the estimation uncertainty, strategies that sort solely on return predictions make mistakes by incorrectly going long (short) on the stocks having true expected risk premiums of μ_B (μ_A). The Confident-HL strategy minimizes this misclassification bias by selectively taking positions in the extreme predicted-return stocks that are more precisely measured. Thus, the Confident-HL strategies deliver superior expected OOS returns. In contrast, the Low-Confident-HL portfolio underperforms all other strategies, including EW-HL, as this strategy exclusively comprises stocks that have imprecise risk premium predictions.

Recall that this exercise assumes uncorrelated predictions and forms trading strategies by sorting stocks into median portfolios. Internet Appendix B (table A) presents more comprehensive simulations validating the Confident-HL's superior performance for general cases with correlated return predictions and trading strategies formed using other quantile portfolios (e.g., deciles).

Before economically interpreting the Confident-HLs, it is worth emphasizing a couple of points. First, dropping stocks with imprecise risk premiums improves the expected returns of HL strategies, not necessarily their variances, as it may reduce the diversification benefit. So, this paper also develops mean variance trading strategies that optimally balances between expected HL returns and their (co)variances. However, estimating the entire covariance matrix could lead to additional estimation uncertainty, and thus the mean-variance portfolios trade-off diversification benefits for the accuracy. Determining this trade-off is ultimately an empirical question. The empirical section shows that the Confident-HL portfolios perform on par (and superior in a few cases) with the mean-variance portfolios, suggesting that the diversification benefits arising from incorporating the covariances nearly equal the costs of estimating them.

C. Interpreting Confident-HL Portfolios using Ambiguity-averse Strategies

Now, I show that the Confident-HL portfolios could be further interpreted as *robust* strategies of ambiguity-averse investors, which were extensively discussed by Garlappi et al. (2007).

Let $\{\hat{\mu}_{l1}, \hat{\mu}_{l2}, \dots \hat{\mu}_{lN}\}$ ($\{\hat{\mu}_{s1}, \hat{\mu}_{s2}, \dots \hat{\mu}_{sN}\}$) be the set of N risk premium predictions of stocks in the long (short) leg, with different predictive variances. Suppose that the predicted risk premiums in the long (short) leg are all equal and positive (negative), i.e., $\{\hat{\mu}_{l1} = \hat{\mu}_{l2} = \cdots = \hat{\mu}_{lN}\}$, $\hat{\mu}_{li} > 0$ $\forall i$; and $\{\hat{\mu}_{s1} = \hat{\mu}_{s2} = \cdots = \hat{\mu}_{sN}\}$, $\mu_{si} < 0, \forall i$, which the EW-HL strategies implicitly assume.

Consider the problem of an ambiguity-averse investor who forms optimal long and short portfolios according to the following max-min expected return utilities, respectively

$$\max_{w_{li}} \min_{\{\mu_{li}\}} \sum w_{li}\mu_{li}, \text{ subject to } (\hat{\mu}_{li} - k\sigma_{li}) \leq \mu_{li} \leq (\hat{\mu}_{li} + k\sigma_{li}), \forall i, \text{ and } \sum w_{li} = 1 \quad (38)$$

$$\max_{w_{si}} \min_{\{\mu_{si}\}} \left(-\sum w_{si}\mu_{si}\right), \text{ subject to } (\hat{\mu}_{si} + k\sigma_{si}) \leq \mu_{si} \leq (\hat{\mu}_{si} - k\sigma_{si}), \forall i, \text{ and } \sum w_{si} = 1, \quad (39)$$

where σ_{li} (σ_{si}) denotes the standard error of the risk premium prediction of the the i^{th} stock in the long (short) leg. The utility optimization in (38) and (39) serves two purposes. First, the constraint restricting expected returns to lie within specified confidence intervals shows that the investor acknowledges the estimation uncertainty. Second, the minimization over the choice of expected returns reflects the investor's aversion to ambiguity.

It is straighforward to note that the solutions to (38) and (39) reduce to the Confident-HL strategy that takes long (short) position exclusively on the stock in long (short) leg that has the lowest standard error (or the highest confidence-level). Thus, the Confident-HL strategies could be interpreted as *robust* trading strategies of ambiguity-averse investors with max-min utilities.

D. Mean-variance Strategies

I further generalize the Confident-HL strategies by forming mean-variance strategies that take into account the entire covariance structure (not just variances) of risk premium predictions.

Consider a set of n stock return predictions $\hat{\mu}$, with the covariance matrix $\widehat{\Sigma_r}$. Note that $\widehat{\Sigma_r}$ denotes the covariance of return predictions (not risk premium predictions). Thus, due to (19), $\widehat{\Sigma_r} = \widehat{\Sigma_{er}} + \sigma_\eta^2 I$, where $\widehat{\Sigma_{er}}$ denotes the covariance of risk premium predictions (estimated in (16)), and σ_η^2 denotes the NN model's residual variance. Then the mean-variance efficient weights of Kozak et al. (2019) that explicitly take into account the estimation uncertainty of risk premiums is

$$w = \arg\min_{w} (\hat{\mu} - \widehat{\Sigma_r} w)' \widehat{\Sigma_r}^{-1} (\hat{\mu} - \widehat{\Sigma_r} w) + \gamma_1 w' w + \gamma_2 \sum |w_i|, \quad (40)$$

where the weights $w = [w_1, w_2, \dots, w_n]^{'}$. The realized excess returns of the mean-variance strategy is given by $w^{'}r$, where r denotes the realized excess returns of N stocks. (40) is also equivalent to

$$w = \arg\min_{w} (\hat{\mu} - \widehat{\Sigma_{er}} w)^{'} \widehat{\Sigma_{er}}^{-1} (\hat{\mu} - \widehat{\Sigma_{er}} w) + \gamma_3 w^{'} w + \gamma_2 \sum |w_i|, \tag{41}$$

for a different parameter γ_3 (rather than γ_1).

The regularization parameters (i.e., γ_3, γ_2) solve two purposes. First, because of the estimation uncertainty in risk premiums, the traditional mean-variance portfolio weights often take extreme values and perform poorly OOS. The regularization mitigates this problem by constraining the weights. Second, recall that the paper estimates the covariance of risk premium predictions (Σ_{er}) using 100 dropout samples, rendering it non-invertable when there are more than 100 stocks. The regularization ensures that the regularized covariance matrix is always invertible.

I choose optimal γ_1 and γ_3 so that mean-variance portfolio's Sharpe ratio is maximized in the validation sample. Because any scaled portfolio weights (i.e., λw, where λ is a scalar) deliver the same Sharpe ratio as w, I scale weights so that $\frac{1}{2} \sum |w_i| = 1$. This specification is consistent with the EW HL and Confident-HL strategies, whose portfolio absolute weights always sum to 2.[11]

IV. Empirical results

Recall that section III implies two central predictions. (1) Ex-ante variances of NN-based risk premium predictions predict their ex-post forecast-squared errors, and thus (2) the NN-based Confident-HL and mean-variance strategies will deliver economic gains OOS. This section empirically documents both of these predictions.

[11] In addition, Lintner (1965) notes that paying interest on margin deposits and short-sale proceeds would lead to optimal mean-variance weights that are scaled by $\sum |w_i|$. Also, see Pástor and Stambaugh (2000).

A. Data, Definitions, and Replication Study

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 that are included in the CRSP database. 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).[12] The Treasury-bill rate proxies for the risk-free rate.

2. Models

Neural Network. The paper primarily focuses on a feed-forward NN with three hidden layers (NN-3), with 32, 16, and 18 neurons per layer. This model was previously 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 (7) and the validation sample's MSE (8), 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.

I implement this estimation framework to obtain risk premium predictions, as well as their (co)variances, over the OOS test sample. Whereas GKX and Avramov et al. (2020) mainly apply

[12] 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.

L_1 regularization to estimate the parameters, I use dropout and L_2. As discussed in section II, these regularizations enhance the model's predictive performance and deliver prediction covariances. I retain the other hyperparameters (e.g., SGD learning rate, Adam optimization) used by GKX.

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 the linear characteristics model of with 15 firm-level predictors examined by Lewellen. 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.

3. Definitions of Performance Metrics

I define the ex-ante and ex-post precision measures that I use repeatedly in this section

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

$$EC_{it} = \frac{|\widehat{E_t(r_{i,t+1})}|}{se_t(\widehat{E_t(r_{i,t+1})})}, \tag{42}$$

where EC is ex-ante confidence, $\widehat{E_t(r_{i,t+1})}$ is the risk premium prediction of stock i at period t (for $t+1$) and $se_t(\widehat{E_t(r_{i,t+1})})$ 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. However, my central conclusions are the same when I use inverse standard errors as proxies for precision. Table B in Internet Appendix (B) presents these results.

While the ex-ante confidence levels of NN-3-based risk premium predictions are computed using (15), (16), and (17), 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 (34).[13]

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

$$\text{OOS-}R^2 = 1 - \frac{\sum_{(i,t)\in\mathcal{S}}(r_{i,+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{S}} r_{i,t+1}^2}, \tag{43}$$

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

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 A (B) in the Internet Appendix B 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 a significant OOS return of 2.51% (1.47%) and an annualized Sharpe ratio of 1.56 (0.96). These results qualitatively and quantitatively match with GKX and Avramov et al. (2020), respectively, and reflect the impressive performance of NN-3 OOS.

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 III: The ex-ante precision of NN-based risk premium predictions would predict their ex-post precision, whereas those of Lewellen's predictions need not.

[13] Alternatively, I also consider Fama-Macbeth standard errors for Lewellen's predictions to account for cross-sectional correlations of residuals. The conclusions are the same. I thank Jay Shanken for suggesting this approach.

Figure 3 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 3 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 sizably improved OOS-R^2 of 2.21%, an increment of 172%. This result reinforces that the ex-post precision of NN-based predictions is ex-ante predictable.

Figure 4 repeats the analysis for Lewellen-based predictions, and it supports remark 1. Ex-post OOS-R^2s of Lewellen's predictions, 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 reiterates that "bias" rather than "variance" predominantly determines the ex-post precision of a "simple" model's predictions. Interestingly, though, the lowest ex=ante precision's (decile 1) predictions also register the lowest ex-post OOS-R^2. In fact, OOS-R^2s seem to monotonically increase from decile 1 to decile 7. This result perhaps reflects these deciles' unavoidably large ex-ante "variances", which dominate average "biases" across other predictions, thus yielding a monotonic relation between the ex-ante and the ex-post precision on these subsets. However, it is evident from considering all 10 deciles that the ex-post OOS-R^2s of Lewellen-based predictions are not as uniformly predictable as NN-3-based OOS-R^2s.

Now I describe the procedure for forming various trading strategies.

C. Portfolio Construction

1. EW(VW)-HL. For every month, I sort stocks into deciles according to their next month's return predictions. If L and H represent the lowest and highest return prediction deciles, respectively, the EW(VW)-HL strategy takes EW (VW) long and short positions on H and L, respectively.

2. EW(VW)-Confident-HL. 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. If L_{10} (L_1) and H_{10} (H_1) denote the subsets with the highest (lowest) ex-ante confidence from L and H, respectively, the EW(VW)-Confident-HL strategy takes EW (VW) long and short positions only on H_{10} and L_{10}, respectively.

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

4. PW-HL. Rather than totally excluding low ex-ante confident subsets, the "precision-weighted" strategy disproportionately downweights them while forming portfolios. In particular, this strategy takes 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" strategy takes long (short) positions on each subset H_j (L_j) with the weights proportional to $1/j$.

6. Mean-variance. The mean-variance portfolios are formed using section III.D (41).

7. 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.[14] 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.

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

8. Double-Sorted portfolios. Similar to section III.B, I consider double-sorted strategies matched to contain the same number of stocks as the Confident-HL portfolios. These strategies take long (short) positions on the top (bottom) 1% of the stocks that have relatively higher (lower) risk premium predictions. Despite containing the same number of stocks as the Confident-HL portfolios, these strategies (unlike matching portfolios) may not serve as apt benchmarks for assessing the Confident-HL portfolios' performance because they have higher predicted-returns by construction. Nevertheless, I construct these strategies as an additional robustness check.

D. Economic Gains from Confident-HL and Mean-variance Strategies

This subsection documents that the NN-based Confident-HL and mean-variance strategies significantly outperform the existing NN-based strategies .

1. Main Results.

Table II presents the main results containing various OOS average monthly returns, annualized Sharpe ratios, alphas and information ratios with respect to Fama and French (2015) model added to the momentum factor, and Stambaugh and Yuan (2017) models, respectively, of competing trading strategies. Table III shows whether the pairwise differences in the OOS performance of different strategies are statistically significant using the moving block bootstrap procedure of Allena (2021) that are more conservative tests than DM, as these tests take into account ex-ante parameter uncertainty of risk premium forecasts. Thus, if the bootstrap tests imply that results are significant, the DM tests imply significance too. Internet Appendix A summarizes these bootstrap tests.

Confident-HL and PW-HL strategies outperform HL. Across all performance metrics, the EW(VW)-Confident-HL and the precision-weighted PW-HL portfolios significantly outperform the conventional EW(VW)-HL portfolios. For example, while the traditional EW(VW)-HL portfolio earns an OOS average monthly return of 2.52% (1.48%) and an annualized Sharpe ratio of 1.5 (0.9), 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 economically large 43% (49%) and 17% (21%) increases, which are all statistically significant at the 1% level. Likewise, the PW-HL also dominates the EW-HL with an average return and Sharpe ratio of 2.87% and 1.67, respectively.

Confident-HLs outperform their matching strategies. Note that the matching EW(VW)-HL_{CM} and the EW(VW)-Confident-HL strategies have the same average return predictions. However, the former yields a considerably lower OOS average return and Sharpe ratio than the latter. The 0.54% (0.48%) monthly return difference between the two quantifies the economic value of incorporating the ex-ante precision information into forming NN-3-based HL portfolios, which translates to economically large 6.48% (5.76%) annualized return.

Confident-HLs outperform double-sorted strategies. Table C in Internet Appendix B documents that the EW(VW)-Confident-HL's Sharpe ratio of 1.75 (1.09) is at least 11% (13%) higher than the Sharpe ratio of the EW(VW)-double-sorted strategy. Economically, when the EW(VW)-Confident-HL and EW(VW)-double-sorted strategies are standardized to have the same return variances, this Sharpe ratio improvement translates to a large 4.8% (3.6%) average annual return difference between both strategies.

Low-Confident-HL and LPW-HL underperform. In contrast, the Low-Confident and the low-precision-weighted strategies containing stocks in the extreme predicted-return deciles with imprecise risk premiums deliver significantly lower OOS returns and Sharpe ratios than the traditional HL and Confident-HL portfolios. For example, although the VW-Low-Confident-HL portfolio has higher average predicted-returns than that of the EW(VW)-HL, its 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 reiterates that trading strategies formed using imprecise predictions would lead to large losses.

Note that a seemingly large 0.17% monthly average return difference between the EW(VW)-HL and EW(VW)-Low-Confident-HL is statistically insignificant. This is because the Low-Confident-HL portfolio returns are excessively imprecise (volatile), and thus zero-mean comparison tests with them would have less "power" to reject the null. However, Sharpe ratio tests, which take into account the stock return volatility and thus are more powerful, vividly indicate the underwhelming performance of the Low-Confident-HL portfolios.

Cumulative returns. Figure 5 presents the cumulative log returns of all trading strategies, and it visually demonstrates that the EW-Confident-HL strategy easily beats the EW-HL strategy. In fact, the difference in returns between both strategies is steadily increasing in the recent years. Figure 6 repeats the same after standardizing all strategies to have the same return variances as the EW-HL strategy. It indicates that the improved Sharpe ratio of the EW-Confident-HL portfolio translates to 6.8% (4.9%) annualized *holding period* return difference between than EW-Confident-HL and the EW-HL (EW-double-sorted) strategy.

Mean-variance strategies also deliver large Sharpe ratio improvements. The annualized Sharpe ratio and information ratios of the mean-variance strategy are 1.65, 1.56, and 1.50, respectively. These correspond to 10%-15% enhancements relative to EW-HL portfolio, which are all statistically significant at the 1% level. Note that the mean-variance strategy underperforms the EW-Confident-HL portfolio in terms of OOS Sharpe ratios. This result is due to uncertainty in estimating the covariance of risk premium predictions, as discussed in section III.B: While mean-variance strategies potentially could outperform Confident-HL strategies by incorporating all covariances (not just individual variances) of risk premium predictions, estimating all of them could lead to additional estimation uncertainty. Given that the sample contains 6000 stocks on average per month, uncertainty in estimating the 6000×6000 covariance matrix dominates the diversification benefit. Thus, the EW-Confident-HL portfolio outperforms the mean-variance strategy. However, as will be shown in the next section, the mean-variance strategy performs on par with the EW-Confident-HL portfolios on a smaller subsample containing non-microcap stocks, as estimation uncertainty of prediction covariances will be lower with fewer stocks.

2. Robustness of Confident-HL and mean-variance strategies

Confident-HL and Mean-variance strategies outperform on non-microcaps. To investigate the extent to which microcaps drive the outperformance of this paper's strategies, I retrain NN-3 on non-microcaps by excluding microcaps. Table IV presents the portfolios' OOS performance. Table V shows their statistical significance.

Even on the non-microcap subsample, the Confident-HL and mean-variance portfolios significantly outperform comparable alternative HL strategies. For example, the VW-Confident-HL and its matching VW-HL$_{\mathrm{CM}}$ have the same return prediction averages. However, the difference between the former and the 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 significant at the 1% level. Similarly, the mean-variance strategy delivers an annualized Sharpe ratio of 1.22, which is 22% higher than the EW-strategy. Interestingly, as noted previously, the EW-Confident-HL and the mean-variance strategy deliver similar Sharpe ratios, although the EW-Confident-HL dominates in terms of the information ratios.

Confident-HL and mean-variance strategies are robust to downside risks. Because NN-3-based HL portfolios are known to display positive skewness and excess kurtosis (Avramov et al. (2020)), table VI examines several higher-moment-adjusted performance measures that reflect the portfolios' downside risk. The Omega, Sortio, and upside-potential ratios, typically examined by practitioner-researchers as alternatives for Sharpe ratios, asymmetrically penalize portfolio losses more than rewarding gains.[15] Across all higher-order measures, the Confident-HL, PW-HL, and mean-variance strategies handily outperform the conventional HL and equivalent matching portfolios. Thus, dropping or downweighing stocks with lower ex-ante precision from an investment portfolio also mitigates its downside risk.[16]

Confident-HL's OOS returns are robust to transaction costs. To evaluate whether the economic gains from the Confident-HL portfolios come at the expense of high transaction-costs, the "Turnover" column of table VI calculates their portfolio turnovers. The Confident HL-portfolios deliver economically large transaction-adjusted returns as well. For example, 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. Note that the Confident-HL portfolio turnovers 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

[15] See the following Wikipedia pages for the definitions of these measures: Omega, Sortino, and up-side potential.
[16] The Confident-HL strategies also reduce the drawdowns by more than 11% relative to the HL strategies.

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 relatively much lower, 1.26% (0.79%), respectively.

E. Outperformance of Mean-variance Strategies at the Industry-level

Table G in Internet Appendix B shows that the mean variance strategy formed using the 48 industry portfolios of Fama and French (1997) yields an annualized Sharpe ratio of 1.20, which is almost double the Sharpe ratio (0.66) of the equal-weighted strategy. Thus, the relative outperformance of mean-variance strategies is even more compelling at the industry-level. Moreover, recall that spread portfolios, such as HL and Confident-HL strategies, work only when there is a significant variation in the levels of risk premiums across assets, with a few assets earning more risk premiums than the other (for e.g., the assumption $\mu_A > \mu_B$ in example-1). However, industry portfolios do not seem to exhibit such a cross-sectional variation, with all industries earning nearly equal risk premiums. In particular, tables E and F in Internet Appendix B show that all the 48 industries yield similar monthly risk premiums close to 1.3%. Thus, I do not examine the HL strategies nor the Confident-HL strategies at the 48-industries-level.

F. More robustness checks.

IVOL vs Confident-HL strategies. As emphasized in the introduction, this paper's Confident-HL strategies are fundamentally different from the IVOL strategies. This is because the IVOL strategies take short positions on stocks with high idiosyncratic return volatilities (variances), whereas the Confident-HL strategies totally ignore (i.e., do not take long-short positions) stocks with high risk premium variances. Similarly, one could ask whether the IVOL-based-Confident-HL strategies that first sort stocks based on their predicted returns and later on their IVOL-based-confidence-levels computed using past idiosyncratic volatilities (rather than this paper's ex-ante standard errors), deliver similar economic gains as Confident-HL portfolios. Table D in Internet Appendix B indicates that such IVOL-based-Confident-HL strategies do not outperform the traditional HL strategies. More surprisingly, the IVOL-based-Low-Confident-HL strategy that puts more weights on stocks with relatively high past return idiosyncratic volatilities outperforms the IVOL-based-Confident-HL strategy. Because many stock characteristics, such as 1-month momentum, are not that persistent, a stock exhibiting high past return IVOL does not necessarily imply that its expected return conditional on the current set of characteristics is imprecisely measured. As a result, IVOL-based-Confident-HL strategies need not perform well.

Confident-HL strategies vs *t*-sorted strategies. Recall that the Confident-HLs are conditional strategies that first sort stocks based on their predicted returns and later on their confidence-levels. Rather than these strategies, one could alternatively form *t*-sorted strategies that take long (short) positions on the stocks with the relatively highest (lowest) t-stats (i.e., ratios of risk premium predictions and their standard errors). Such strategies need not deliver large OOS returns because stocks with precise risk premium predictions need not have high expected returns. For example, consider a simple scenario where all stocks with low risk premiums are relatively precisely measured. Then the *t*-strategies take positions only on the subset of stocks with low expected returns, thus delivering low OOS returns. The Confident-HL strategies, consistent with the simulation results in tables (I) and (A), tackle this concern by first sorting on the predicted returns.[17]

G. Comparing NN-3-based and Lewellen-based Trading Strategies

This subsection shows that NN-3-based Confident-HL and mean-variance strategies also outperform competing Lewellen-based strategies. In addition, the subsection also shows that analogous Lewellen-based Confident-HL and mean-variance strategies do not deliver gains OOS. Tables VII and VIII present the results. The portfolio definitions and notations remain the same as in section IV.C. All Lewellen-based HL portfolios are denoted by attaching the subscript "$_{\mathrm{L}}$" to HL. For e.g., the conventional EW-HL portfolio based on the Lewellen model is represented by EW-HL$_{\mathrm{L}}$.

NN-3-Confident-HLs outperform all Lewellen-based strategies. For e.g., the NN-3-based VW-Confident-HL earns 21% higher Sharpe ratio than the Lewellen-based VW-HL portfolios, which is statistically significant at 1%. Also, NN-3-Confident-HL portfolios yield significantly

[17] The *t*-sorted strategies perform on par with the HL strategies in terms of OOS Sharpe ratios. The results, which have not been reported to conserve space, are available upon request.

higher Sharpe ratio and information ratios than Lewellen-Confident-HL portfolios. In contrast, the conventional VW NN-3-HL and Lewellen-HL portfolios deliver similar Sharpe ratios of 0.9 and their difference is statistically insignificant at 5%.[18] In addition, the squared Sharpe ratio 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. Thus, these results emphasize the importance of constructing NN-3-based Confident-HL portfolios.

NN-3-based mean-variance strategy beats all Lewellen-based strategies. The NN-3-based mean-variance strategy delivers 18% higher and 193% larger Sharpe ratios than the Lewellen-based HL and mean-variance strategies, respectively. This result reiterates the relative outperformance of the NN-3-based mean-variance strategies.

Lewellen-based Confident-HL and mean-variance portfolios do not yield gains. The Lewellen-based EW(VW)-Confident-HL strategy delivers similar Sharpe ratio as the Lewellen-based EW(VW)-HL strategy. Moreover, the Lewellen-based mean-variance strategy significantly underperforms OOS by earning nearly one-third the Sharpe ratio (0.57) of the Lewellen-based EW-HL strategy (1.41). Thus, these results confirm that the Confident-HL and mean-variance strategies need not deliver economic gains for simple models, thus validating remarks 1 and 2.

Relative performance of NN-3 monotonically increases with its precision. Finally, figure 7 plots the OOS return and Sharpe ratio differences between both models' VW HL portfolios on various subsamples. The economic gains from the NN-3 monotonically increase with the NN-3-based confidence levels. 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. However, the difference rises to a highly significant 0.82% (0.52) on the subsample comprising the top 10% stocks with the highest NN-3-based ex-ante confidence levels. In contrast, for the bottom 10% of stocks with the lowest NN-3-based confidence, Lewellen statistically outperforms NN-3. The average return (squared-Sharpe ratio) difference between NN-3 and Lewellen HL portfolios is significantly negative -1.2% (-0.58).

[18] My results do not directly compare with GKX, as they examined a different Lewellen model involving 3 characteristics (rather than 15 characteristics).

Overall, this subsection shows that the existing NN-based HL strategies outperform Lewellen-based HL strategies only on subsamples of stocks that have NN-based confident risk premium forecasts. Similarly, the NN-3-based Confident-HL portfolios and mean-variance strategies statistically dominate all competing Lewellen model's strategies. In contrast, Lewellen-based Confident-HL and mean-variance strategies do not deliver gains, as the ex-ante variances of Lewellen-based risk premium forecasts, unlike ex-ante variances of NN-based return predictions, do not predict the ex-post squared forecast errors.

V. Dynamics of Ex-ante Precision

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

To understand the time-series dynamics of the ex-ante precision of risk premium predictions, I compute the cross-sectional average of their ex-ante standard errors and call these "aggregate standard errors". Figure 8 plots the time-series of the aggregate standard errors. The series seem to reflect time-varying financial market uncertainty. For example, Bloom (2009) and Baker, Bloom, and Davis (2016) document that market uncertainty appears to jump up after major shocks, such as Black Monday, the Dotcom Bubble, and the failure of Lehman Brothers. Consistent with these studies, the aggregate standard errors spike after such shocks.

Table IX 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 will also be relatively imprecise. Thus, the aggregate standard errors capture 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.

B. Cross-sectional Variation in Ex-ante Confidence

Table X 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, Yu, and Yuan (2012)). Avramov et al. (2020) note similar observations, albeit examining *ex-post* OOS returns of several ML-based investment strategies' long-legs and short-legs.

Moreover, NN-3 confidently predicting risk premiums 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 X shows why. Because such stock risk premia are more confidently predicted, HL portfolios containing microcaps yield relatively larger economic gains.

Importantly, I also find that a significant proportion of non-microcaps have confidently risk premium predictions. Table XI 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 subsamples containing large-sized stocks.

VI. Conclusions

This paper derives ex-ante (co)variances of risk premium predictions from NNs at the stock-level and portfolio-level. It then provides novel insights showing why and how incorporating the ex-ante precision into trading strategies leads to large OOS returns and Sharpe ratios. The Confident-HL strategies that deliberately exclude stocks in the extreme predicted-return deciles with large imprecise risk premiums deliver superior OOS performance than the traditional HL strategies, as NN-based risk premium predictions with large ex-ante standard errors will have large squared forecast errors. The paper also constructs mean-variance trading strategies that explicitly incorporate the covariances of risk premium predictions, thus providing enhancements to the HL strategies.

The NN-based Confident-HL trading strategies deliver at least 40% higher returns and 15% higher Sharpe ratios than the NN-based conventional HL portfolios. Standardizing both strategies to have the same return variances, this Sharpe ratio improvement translates to a large 6.8% annualized holding period return difference between the Confident-HL and the HL strategies. Likewise, the mean-variance portfolios improves the OOS Sharpe ratios by 10%-22% (from 1.5 (1.00) to 1.65 (1.22)) at the stock-level, and by 86% (from 0.66 to 1.20) at the industry-level, relative to the existing strategies.

The paper also finds that the existing NN-based HL strategies outperform Lewellen-based HL strategies predominantly on subsamples of stock risk premiums that NNs confidently predict. For example, the average monthly return difference between NN and Lewellen VW HL strategies formed using the subsample stocks whose returns are most confidently predicted by NNs is a highly significant 0.82%. In contrast, the corresponding difference is a signficantly negative -1.2% on the subset of stocks whose returns are most imprecisely predicted by NNs. However, the NN-based Confident-HL and mean-variance portfolios uniformly outperform all other competing strategies. Thus, this paper highlights that incorporating ex-ante (co)variances of NN-based risk premium predictions into trading strategies is important.

Table I: Comparing the OOS Performance of Various Trading Strategies: Simulation Evidence This table compares the expected OOS monthly returns of various trading strategies based on several simulated datasets containing 200 stock risk premiums over 30 years. Risk premium predictions are simulated to be unbiased, normal, and independent. 100 stocks yield true expected returns of μ_A, whereas the other 100 yield μ_B. Of the 100 stocks that deliver μ_A expected returns, 50 stocks are relatively precisely (imprecisely) measured with the predicted risk premium variance of σ_l (σ_h). Similarly, of the 100 stocks that deliver μ_B expected returns, 50 stocks are relatively precisely (imprecisely) measured with the predicted risk premium variance of σ_l (σ_h). 'Risk Premium Variances' column presents the variances of risk premium predictions used for simulations. "True Spread Expected Returns" shows simulated μ_A, μ_B, and $\mu_A - \mu_B$. The "Expected OOS Returns" columns present the expected OOS returns of various trading strategies.

		Expected OOS Returns			
Risk Premium Variances	True Spread Expected Returns	EW-HL	Double-sorted-HL	Confident-HL	Low-confident-HL
σ_l=0.001,σ_h=0.02	$\mu_A = 3\%, \mu_B = -3\%, \mu_A - \mu_B = 6\%$	2.46%	2.49%	4.18%	0.74%
σ_l=0.001,σ_h=0.5	$\mu_A = 3\%, \mu_B = -3\%, \mu_A - \mu_B = 6\%$	2.06%	0.77%	3.91%	0.21%
σ_l=0.01, σ_h=0.1	$\mu_A = 3\%, \mu_B = -3\%, \mu_A - \mu_B = 6\%$	0.94%	1.17%	1.57%	0.32%
σ_l=0.01, σ_h=0.5	$\mu_A = 3\%, \mu_B = -3\%, \mu_A - \mu_B = 6\%$	0.80%	0.67%	1.47%	0.12%
σ_l=1, σ_h=5	$\mu_A = 3\%, \mu_B = -3\%, \mu_A - \mu_B = 6\%$	0.08%	0.09%	0.14%	0.01%
σ_l=0.001, σ_h=0.005	$\mu_A = 3\%, \mu_B = -3\%, \mu_A - \mu_B = 6\%$	2.94%	3.90%	4.53%	1.35%
σ_l=0.001,σ_h=0.02	$\mu_A = 2\%$, $\mu_b = -2\%$, $\mu_A - \mu_B = 4\%$	1.20%	1.25%	2.05%	0.36%
σ_l=0.001,σ_h=0.5	$\mu_A = 2\%$, $\mu_b = -2\%$, $\mu_A - \mu_B = 4\%$	0.98%	0.38%	1.88%	0.07%
σ_l=0.01, σ_h=0.1	$\mu_A = 2\%$, $\mu_b = -2\%$, $\mu_A - \mu_B = 4\%$	0.415%	0.530%	0.722%	0.108%
σ_l=0.01, σ_h=0.5	$\mu_A = 2\%$, $\mu_b = -2\%$, $\mu_A - \mu_B = 4\%$	0.38%	0.33%	0.65%	0.11%
σ_l=1, σ_h=5	$\mu_A = 2\%$, $\mu_b = -2\%$, $\mu_A - \mu_B = 4\%$	0.037%	0.061%	0.076%	-0.002%
σ_l=0.001, σ_h=0.005	$\mu_A = 2\%$, $\mu_b = -2\%$, $\mu_A - \mu_B = 4\%$	1.41%	1.93%	2.25%	0.57%
σ_l=0.001,σ_h=0.02	$\mu_A = 1\%, \mu_B = -1\%, \mu_A - \mu_B = 2\%$	0.31%	0.34%	0.54%	0.07%
σ_l=0.001,σ_h=0.5	$\mu_A = 1\%, \mu_B = -1\%, \mu_A - \mu_B = 2\%$	0.27%	0.13%	0.51%	0.03%
σ_l=0.01, σ_h=0.1	$\mu_A = 1\%, \mu_B = -1\%, \mu_A - \mu_B = 2\%$	0.11%	0.15%	0.19%	0.03%
σ_l=0.01, σ_h=0.5	$\mu_A = 1\%, \mu_B = -1\%, \mu_A - \mu_B = 2\%$	0.10%	0.10%	0.18%	0.03%
σ_l=1, σ_h=5	$\mu_A = 1\%, \mu_B = -1\%, \mu_A - \mu_B = 2\%$	0%	0%	0.02%	-0.01%
σ_l=0.001, σ_h=0.005	$\mu_A = 1\%, \mu_B = -1\%, \mu_A - \mu_B = 2\%$	0.35%	0.49%	0.58%	0.12%

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

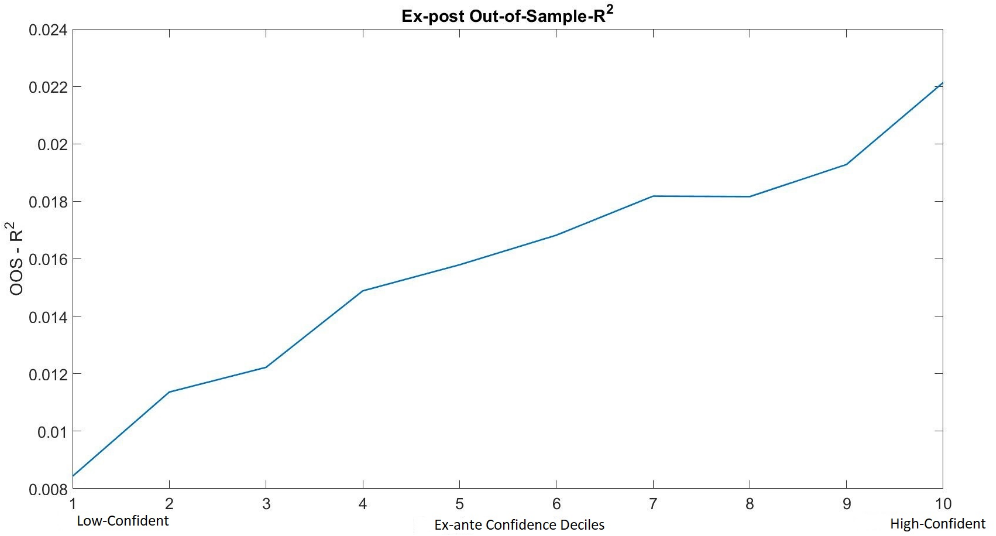
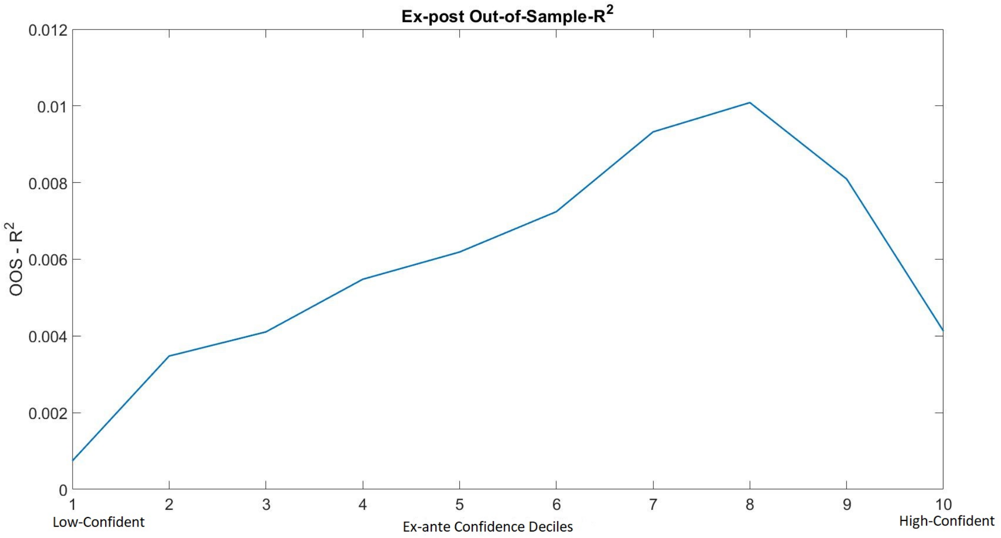


Figure 4. Ex-ante Confidence and Ex-post OOS-R^2: Lewellen-based Predictions and Standard Errors



Note: Figure 3 (4) presents the 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 (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 II: Performance of Various Trading Strategies: 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 in the extreme predicted-return deciles 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$_{\mathrm{LCM}}$ 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$_{\mathrm{CM}}$ is the 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$_{\mathrm{M}}$ is the traditional EW-HL portfolio matched with LPW(PW)-HL. See section IV.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$_{\mathrm{LCM}}$, HL$_{\mathrm{CM}}$ and HL$_{\mathrm{M}}$ 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	1.49	2.18%	7.15	1.44
EW-HL$_{\mathrm{LCM}}$	1.77%	2.64%	8.20	1.50	2.34%	7.7	1.50	2.33%	7.25	1.45
EW-Low-Confident-HL	1.79%	2.35%	6.46	1.18	1.97%	5.65	1.11	1.96%	5.28	1.06
EW-HL$_{\mathrm{CM}}$	1.97%	3.07%	8.65	1.58	2.77%	8.26	1.61	2.75%	7.8	1.56
EW-Confident-HL	1.97%	3.61%	9.58	1.75	3.29%	9.02	1.77	3.27%	8.6	1.73

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	0.67	0.77%	2.68	0.54
VW-HL$_{\mathrm{LCM}}$	1.77%	1.50%	4.61	0.84	0.87%	2.87	0.56	0.76%	2.38	0.48
VW-Low-Confident-HL	1.78%	1.31%	3.02	0.55	0.48%	1.15	0.22	0.39%	0.88	0.18
VW-HL$_{\mathrm{CM}}$	1.90%	1.73%	4.92	0.90	1.12%	3.39	0.66	1.02%	2.95	0.59
VW-Confident-HL	1.90%	2.21%	5.95	1.09	1.79%	4.77	0.93	1.43%	3.82	0.77

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	1.49	2.18%	7.15	1.44
LPW-HL$_{\mathrm{M}}$	1.69%	2.52%	8.21	1.50	2.20%	7.63	1.49	2.18%	7.15	1.44
LPW-HL	1.70%	2.36%	7.63	1.39	2.02%	6.95	1.36	2.00%	6.48	1.30
PW-HL$_{\mathrm{M}}$	1.77%	2.64%	8.20	1.50	2.34%	7.7	1.51	2.33%	7.25	1.46
PW-HL	1.77%	2.87%	9.14	1.67	2.57%	8.68	1.70	2.55%	8.16	1.64

Panel D: Mean-Variance Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
Mean-variance	1.78%	2.01%	9.06	1.65	1.67%	8.54	1.67	1.67%	8.24	1.66

Table III: Statistical Comparison of Various Trading Strategies: 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 A, 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 II and section IV.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$_{\text{LCM}}$, HL$_{\text{CM}}$ and HL$_{\text{M}}$ 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$_{\text{LCM}}$ − 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$_{\text{CM}}$	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$_{\text{LCM}}$ − 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$_{\text{CM}}$	0.48%* (0.086)	0.374*** (0.004)	0.67%*** (0.003)	0.433*** (0.001)	0.41% (0.128)	0.238*** (0.008)

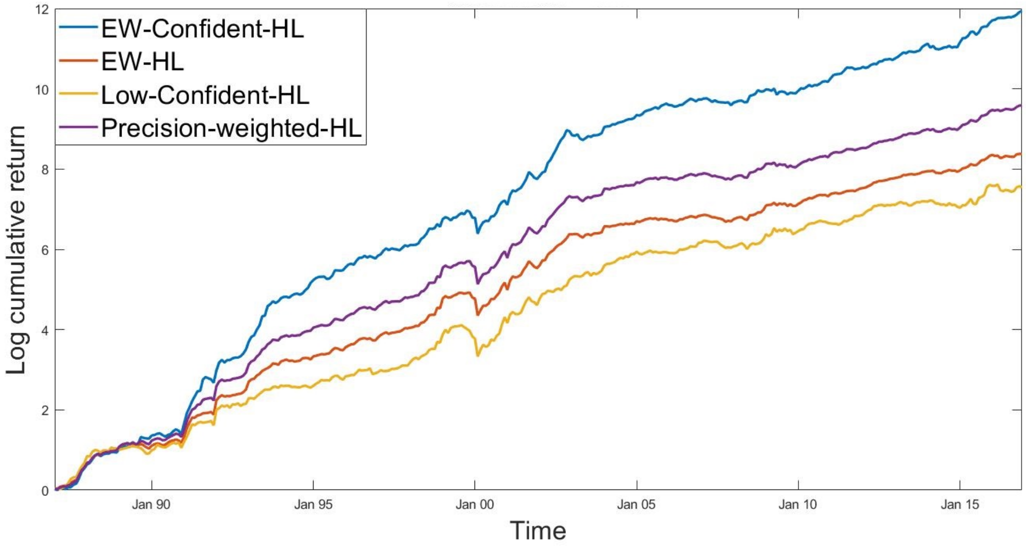
Panel C : OOS Performance Differences of Precision-Weighted Portfolios

	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)	0.18%** (0.013)	0.370*** (0.000)
LPW-HL$_{\text{M}}$ − LPW-HL	0.15%** (0.031)	0.307*** (0)	0.18%*** (0.007)	0.38*** (0)	0.18%** (0.013)	0.370*** (0.000)
PW-HL − EW-HL	0.36%*** (0)	0.535*** (0)	0.37%*** (0)	0.658*** (0)	0.37%*** (0.000)	0.625*** (0.000)
PW-HL − LPW-HL	0.51%*** (0.001)	0.842*** (0)	0.55%*** (0)	1.038*** (0)	0.55%*** (0.000)	0.995*** (0.000)
PW-HL − PW-HL$_{\text{M}}$	0.23%** (0.014)	0.541*** (0)	0.23%*** (0.007)	0.617*** (0)	0.22%** (0.012)	0.568*** (0.000)

Panel D : OOS Performance Differences of Mean-variance Strategies

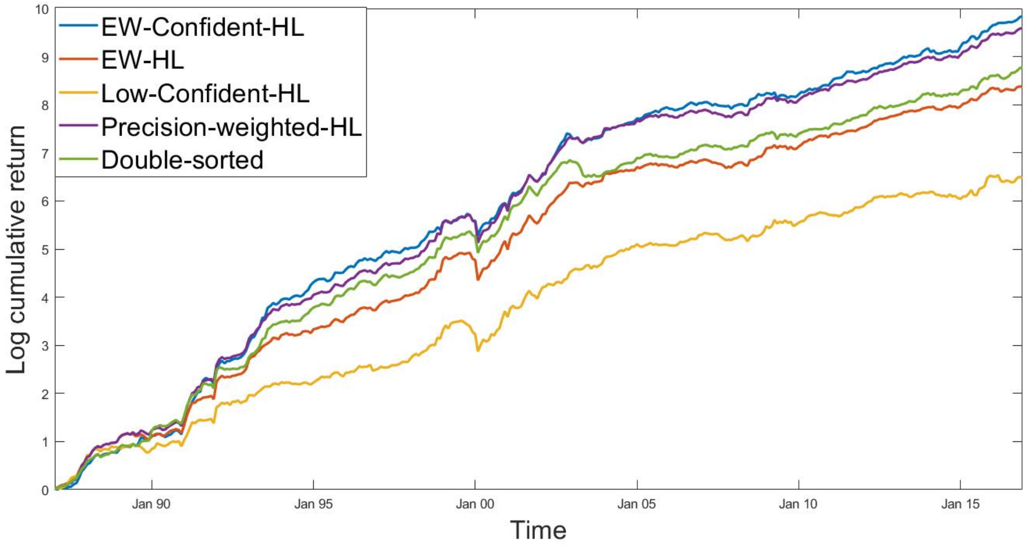
	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
Mean-variance − EW-HL	−0.51% (0.15)	0.488*** (0.002)	−0.54%* (0.068)	0.562*** (0.000)	−0.51%* (0.088)	0.676*** (0.000)
EW-Confident-HL − Mean-variance	1.61%*** (0.001)	0.32** (0.020)	1.63%*** (0.001)	0.321** (0.018)	1.6%*** (0.001)	0.24** (0.038)

Figure 5. Cumulative log OOS returns of various trading strategies



Note: This figure presents the cumulative log OOS returns of various trading strategies, including EW-Confident-HL, EW-HL, EW-Low-Confident-HL, and the precision-weighted PW-HL.

Figure 6. Cumulative log OOS returns of various trading strategies



Note: This figure presents the cumulative log OOS returns of various trading strategies, including EW-Confident-HL, EW-HL, EW-Low-Confident-HL, the precision-weighted PW-HL, and the double-sorted strategies. All returns are standardized to have the same variance as the EW-HL strategy.

Table IV: Performance of Various Trading Strategies: Non-Microcaps

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 20th NYSE size percentile. See table II and section IV.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	0.89	1.24%	3.99	0.80
EW-HL_{LCM}	0.74%	1.83%	5.57	1.02	1.51%	4.76	0.93	1.37%	4.13	0.83
EW-Low-Confident-HL	0.74%	1.50%	3.98	0.73	1.10%	2.96	0.57	0.89%	2.32	0.47
EW-HL_{CM}	0.74%	1.83%	5.57	1.02	1.51%	4.76	0.93	1.37%	4.13	0.83
EW-Confident-HL	0.74%	2.25%	6.68	1.22	2.04%	6.03	1.18	1.93%	5.49	1.10

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	0.70	0.98%	3.1	0.62
VW-HL_{LCM}	0.73%	1.58%	4.76	0.87	1.25%	3.76	0.73	1.10%	3.2	0.64
VW-Low-Confident-HL	0.74%	1.25%	3.13	0.57	0.88%	2.26	0.44	0.74%	1.83	0.37
VW-HL_{CM}	0.73%	1.58%	4.76	0.87	1.25%	3.76	0.74	1.10%	3.2	0.64
VW-Confident-HL	0.72%	2.07%	5.48	1.00	1.84%	4.78	0.93	1.64%	4.14	0.83

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	0.90	1.24%	3.99	0.80
LPW-HL_M	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.90	1.24%	3.99	0.80
LPW-HL	0.69%	1.60%	4.99	0.91	1.26%	4.06	0.79	1.13%	3.47	0.70
PW-HL_{M}	0.68%	1.66%	5.43	0.99	1.35%	4.58	0.90	1.24%	3.99	0.80
PW-HL	0.69%	1.80%	5.93	1.08	1.52%	5.17	1.01	1.41%	4.57	0.92

Panel D: Mean-Variance Portfolios

		Undjusted			FF-5+Mom			SY		
Investment Strategy	pred ret	avg ret	t	SR	α	t	IR	α	t	IR
Mean-variance	1.4%	1.18%	6.70	1.22	0.73%	4.88	0.95	0.68%	4.45	0.89

Table V: Statistical Comparison of Various Trading Strategies: Non-Microcaps

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 A, 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 II and section IV.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$_{\text{LCM}}$, HL$_{\text{CM}}$ and HL$_{\text{M}}$ 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$_{\text{LCM}}$ − 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$_{\text{CM}}$	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$_{\text{LCM}}$ − 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$_{\text{CM}}$	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$_{\text{M}}$ − 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$_{\text{M}}$	0.14%** (0.014)	0.192*** (0.000)	0.17%*** (0.002)	0.222*** (0.000)	0.17%*** (0.001)	0.198*** (0.000)

Panel D : OOS Performance Differences of Mean-variance Strategies

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
Mean-variance − EW-HL	−0.48% (0.108)	0.51*** (0.002)	−0.62%** (0.015)	0.109* (0.091)	−0.56%* (0.053)	0.156** (0.049)
EW-Confident-HL − Mean-variance	1.07%*** (0.002)	−0.008 (0.766)	1.31%*** (0.000)	0.479*** (0.001)	1.25%*** (0.001)	0.416*** (0.005)

Table VI: Transaction Costs and Higher-Moment Adjusted Performance of Various Strategies
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 II and section IV.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$_{\text{LCM}}$, HL$_{\text{CM}}$ and HL$_{\text{M}}$ 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	1.24	4.22	0.98	1.28	1.12	2.46	0.51	0.86
EW-HL$_{\text{LCM}}$	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$_{\text{CM}}$	1.53	4.44	1.05	1.36	1.45	2.49	0.54	0.89
EW-Confident-HL	1.81	4.70	1.28	1.62	1.81	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$_{\text{CM}}$	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$_{\text{M}}$	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$_{\text{M}}$	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

Mean-variance Portfolios: Higher-Moment Adjusted Performance

	All Stocks				Non-Microcaps			
Investment Strategy	Turnover	Omega	Sortino	Upside	Turnover	Omega	Sortino	Upside
Mean-variance	1.30	4.45	1.11	1.43	1.22	3.05	0.69	1.02

Table VII: Comparing various Long-Short Portfolios: NN-3 versus Lewellen (2015)

This table presents various OOS performance metrics of different long-short portfolios based on the NN-3 and Lewellen models. 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 "$\mathrm{HL_L}$" portfolios are based on the NN-3 and Lewellen models, respectively. See table II and section IV.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted ; LPW = low-precision-weighted; PW = precision-weighted; HL=high-low strategy based on NN-3; $\mathrm{HL_L}$=high-low strategy based on Lewellen.

	Raw returns		FF5+UMD		SY	
Investment Strategy	avg ret	Sharpe	α	IR	α	IR
EW-Confident-$\mathrm{HL_L}$	2.62%	1.42	2.04	1.64	1.86%	1.32
EW-$\mathrm{HL_L}$	1.80%	1.41	1.54%	1.41	1.48%	1.27
EW-Confident-HL	3.61%	1.75	3.29%	1.77	3.27%	1.73
EW-HL	2.52%	1.5	2.2%	1.49	2.18%	1.44
VW-Confident-$\mathrm{HL_L}$	1.76%	0.89	0.93%	0.69	0.62%	0.41
VW-$\mathrm{HL_L}$	1.08%	0.9	0.57%	0.64	0.51%	0.50
VW-Confident-HL	2.21%	1.09	1.79%	0.93	1.43%	0.77
VW-HL	1.48%	0.9	0.9%	0.67	0.77%	0.54
Mean-$\mathrm{variance_L}$	1.43%	0.57	0.55%	0.24	0.62%	0.25
Mean-variance	2.01%	1.65	1.67%	1.67	1.67%	1.66

Table VIII: 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 A, 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$_{\text{L}}$" 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 II and section IV.C for a description of the portfolios.

Notes: EW = equal-weighted; VW = value-weighted; HL=high-low strategy based on NN-3; HL$_{\text{L}}$=high-low strategy 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$_{\text{L}}$	0.72%** (0.016)	0.247** (0.036)	0.66%** (0.036)	0.255** (0.025)	0.71%** (0.033)	0.446*** (0.002)
EW-Low-Confident-HL − EW-HL$_{\text{L}}$	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$_{\text{L}}$	1.82%*** (0)	1.055*** (0)	1.75%*** (0)	1.145*** (0)	1.80%*** (0)	1.366*** (0)
EW-Low-Confident-HL − EW-Low-Confident-HL$_{\text{L}}$	1.94%*** (0)	1.33*** (0)	1.64%*** (0)	1.225*** (0)	1.61%*** (0)	1.08*** (0)
EW-Confident-HL − EW-Confident-HL$_{\text{L}}$	0.99%* (0.059)	1.034*** (0.001)	1.25%** (0.02)	2.511*** (0)	1.42%*** (0.009)	1.253*** (0)

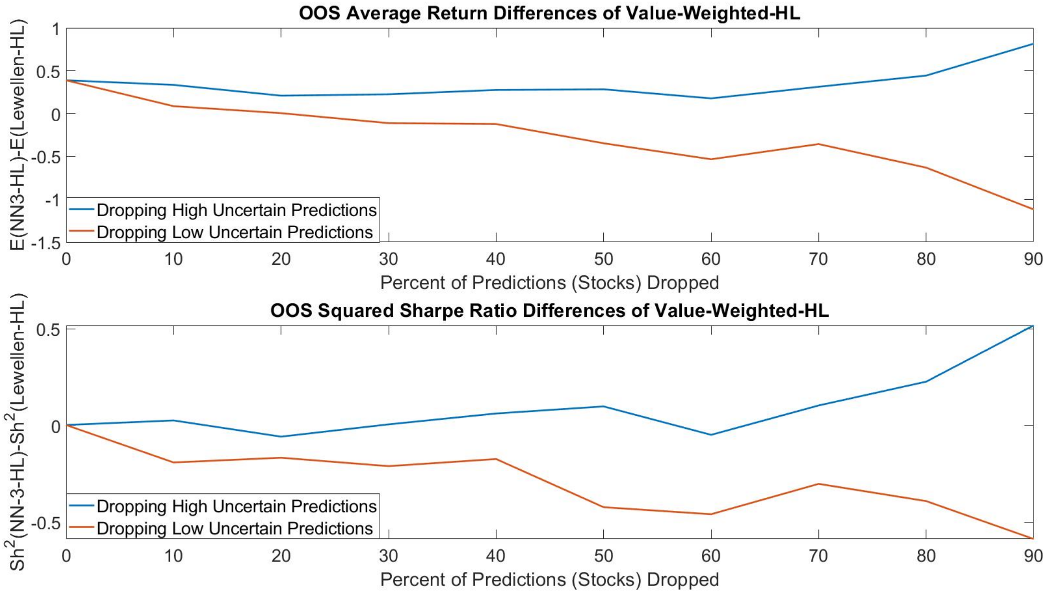
Panel B : Performance Differences of Value-Weighted Portfolios

	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
VW-HL − VW-HL$_{\text{L}}$	0.39% (0.249)	0.002 (0.925)	0.33% (0.256)	0.038 (0.869)	0.27% (0.36)	0.036 (0.423)
VW-Low-Confident-HL − VW-HL$_{\text{L}}$	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$_{\text{L}}$	1.12%*** (0.003)	0.366** (0.013)	1.22%*** (0.001)	0.453*** (0)	0.93%** (0.015)	0.337*** (0.004)
VW-Low-Confident-HL − VW-Low-Confident-HL$_{\text{L}}$	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$_{\text{L}}$	0.44% (0.344)	0.377** (0.014)	0.86%** (0.03)	0.392*** (0.001)	0.81%* (0.072)	0.419*** (0.001)

Panel C : Performance Differences of Mean-variance Portfolios

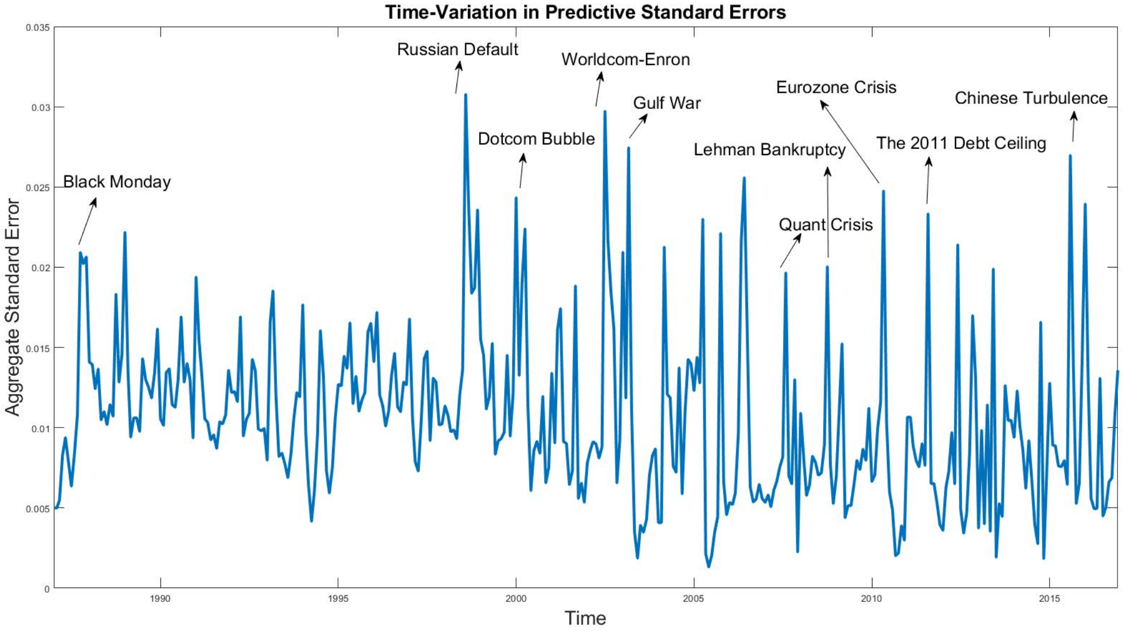
	Raw Returns		FF-5+UMD		SY	
Investment Strategy	avg ret	$Sharpe^2$	α	IR^2	α	IR^2
Mean-variance − Mean-variance$_{\text{L}}$	0.57% (0.423)	2.414*** (0.000)	1.11%* (0.082)	2.739*** (0.000)	1.05% (0.145)	2.679*** (0.000)
EH-HL$_{\text{L}}$ − Mean-variance$_{\text{L}}$	0.36% (0.576)	1.680*** (0.000)	0.98%* (0.088)	1.923*** (0.000)	0.86% (0.160)	1.55*** (0.000)

Figure 7. 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.

Figure 8. 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 IX: 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 X: 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 XI: Characteristic Distributions of the Most Confident Stocks

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%

References

Ahn, Dong-Hyun, Jennifer Conrad, and Robert F. Dittmar, 2009, Basis Assets, *The Review of Financial Studies* 22, 5133–5174.

Allena, Rohit, 2021, Confident Risk Premia: Economics and Econometrics of Machine Learning Uncertainties, SSRN Scholarly Paper ID 3808771, Social Science Research Network, Rochester, NY.

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

Allena, Rohit, and Cesare Robotti, 2021, Out-of-Sample Comparisons of Dynamic Trading Strategies: A Bootstrap Approach, *Working paper, University of Warwick* .

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The Cross-Section of Volatility and Expected Returns, *The Journal of Finance* 61, 259–299.

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.

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

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.

Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu, 2020, Forest Through the Trees: Building Cross-Sections of Stock Returns, SSRN Scholarly Paper ID 3493458, Social Science Research Network, Rochester, NY.

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.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra, 2021, Deep Neural Networks for Estimation and Inference, *Econometrica* 89, 181–213, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16901.

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.

Garlappi, Lorenzo, Raman Uppal, and Tan Wang, 2007, Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach, *The Review of Financial Studies* 20, 41–81.

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.

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

Kleijn, B. J. K., and A. W. van der Vaart, 2012, The Bernstein-Von-Mises theorem under misspecification, *Electronic Journal of Statistics* 6, 354–381, Publisher: Institute of Mathematical Statistics and Bernoulli Society.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2019, Shrinking the cross-section, *Journal of Financial Economics* .

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.

Lintner, John, 1965, The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, *The Review of Economics and Statistics* 47, 13–37, Publisher: The MIT Press.

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

Pástor, Ľuboš, and Robert F. Stambaugh, 2000, Comparing asset pricing models: an investment perspective, *Journal of Financial Economics* 56, 335–381.

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.

Vaart, A. W. van der, 2000, *Asymptotic Statistics* (Cambridge University Press), Google-Books-ID: UEuQEM5RjWgC.

Wager, Stefan, and Susan Athey, 2018, Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, *Journal of the American Statistical Association* 113, 1228–1242.

Wager, Stefan, Trevor Hastie, and Bradley Efron, 2014, Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife, *J. Mach. Learn. Res.* 15, 1625–1651.

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.

A. Internet Appendix: Ex-ante Estimation Uncertainty and Ex-post OOS Inferences

To *statistically* compare the ex-post OOS performance of competing trading strategies, this section discusses the DM tests and the bootstrap tests of Allena (2021) (see also Allena and Robotti (2021)).

A1. Comparing OOS returns of HL strategies.

Consider any two competing model-based HL strategy returns; HL_{1t} and HL_{2t}, where HL_{it} denotes the OOS return of the i^{th} at time time t. Denoting the return differentials $(HL_{1t} - HL_{2t}) = \Delta_t$, the DM test-statistic under the null of equal return means is given by

$$\sqrt{T}\frac{\sum_{t=1}^{T}\Delta_t/T}{se(\Delta_t)} \sim N(0,1), \tag{44}$$

where $se(\Delta_t)$ denotes a heteroskedasticity and autocorrelation consistent estimator of the return differentials Δ_t.

DM emphasized that their tests deliver asymptotically valid inferences only when the differentials are covariance stationary. However, Allena (2021) documents that the ex-ante uncertainties of risk premium predictions would cause the ex-post OOS return differentials to violate the covariance stationarity assumption, thus rendering the DM inferences inadequate. He generalizes the DM tests using a moving block bootstrap procedure that delivers asymptotically valid inferences even when the OOS return differentials violate covariance stationarity.

Moving Block Bootstrap tests of equal return means. Consider a series of return differentials $\{\Delta_t\}_{t=1}^{T}$. 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 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 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 quite similar across other block lengths of 6, 12, 18, and 36.

A2. Comparing Sharpe ratios.

Allena (2021) further shows that the above procedure could be generalized 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

$$Sh_i^2 = \frac{(\frac{1}{T}\sum_{t=1}^{T} HL_{it})^2}{\frac{1}{T}\sum_{t=1}^{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^{T} HL_{it})^2}, \text{ for } i = 1, 2. \tag{45}$$

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,

$$HL_{it}^* = \sqrt{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^{T} HL_{it}) / \sqrt{\sum_{t=1}^{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^{T} HL_{it})^2}, \tag{46}$$

   (c) draw a block bootstrap sample $\{H_{ki}\}$ from the normalized returns;
   (d) compute the bootstrap-based squared Sharpe ratio difference, $Sh_{1i}^2 - Sh_{2i}^2$, where

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

2. Repeat step (1) many times. The p-value equals the proportion of times the absolute value of $(Sh_{1i}^2 - Sh_{2i}^2)$ is greater than the absolute value of $Sh_1^2 - Sh_2^2$.

B. Internet Appendix: Simulation Results and Robustness Checks

Table A: Performance of HL and Confident-HL 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 (B.B1) and equation (52) 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%

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

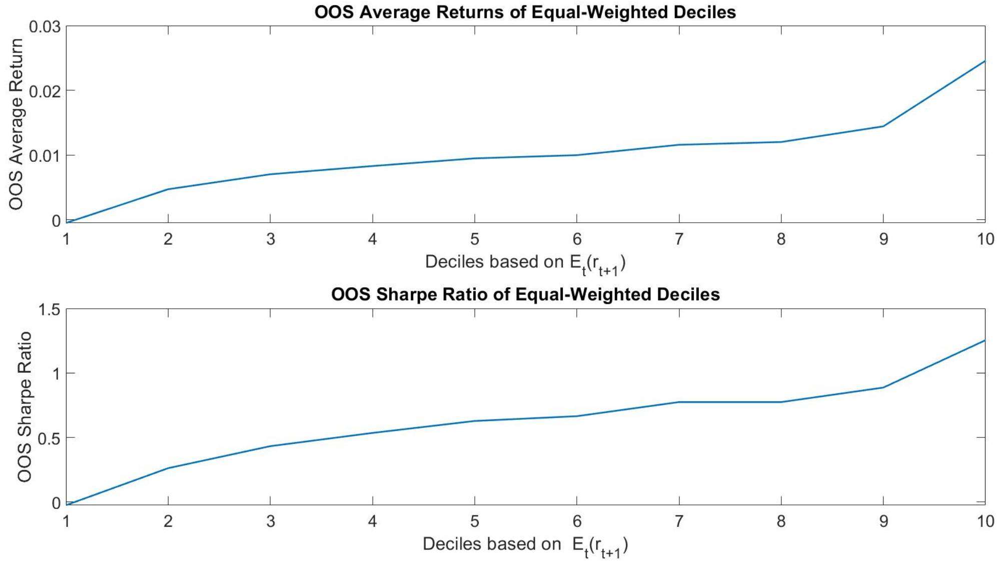
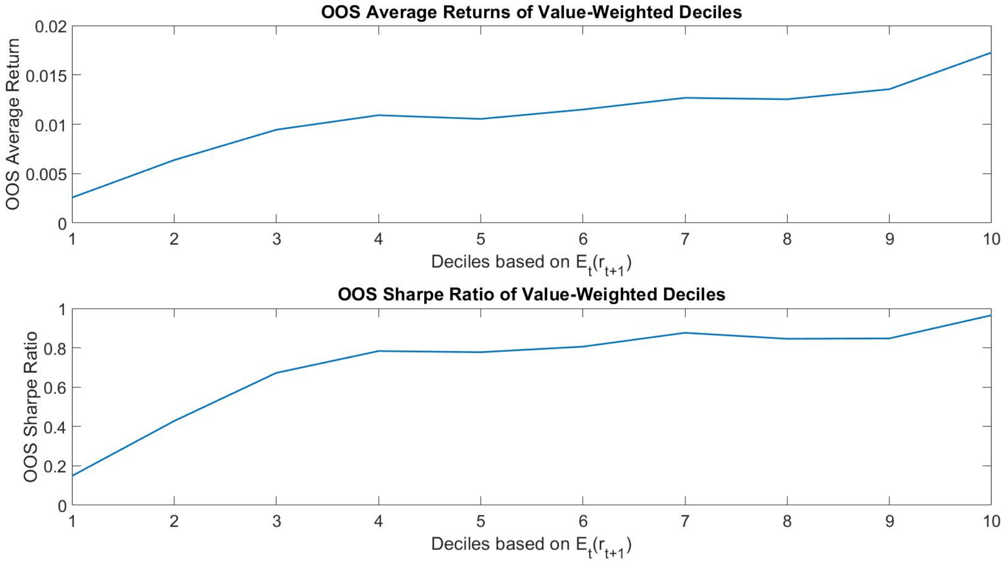


Figure B. Out-of-Sample (OOS) Performance of Value-weighted Deciles Based on NN-3 Predictions.



Note: Figure A (B) 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.

Table B: Performance of Various Long-Short Portfolios: Inverse Standard Errors as Precision
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 II and section IV.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 IV.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)

Table D: Comparing Confident-HL Portfolios with IVOL-based-Confident-HL Portfolios

This table compares the out-of-sample performance of the Confident-HL portfolios with the IVOL-based-Confident-HL portfolios. The IVOL-based-Confident-HL portfolios are similar to the Confident-HL portfolios, with an exception that the IVOL-based-Confident-HLs use past stock return idiosyncratic volatilities (rather than this paper's ex-ante risk premium variances) to compute the confidence levels of risk premium predictions. See section IV.C for a detailed description of the portfolios. 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.

Equal-Weighted

	Undjusted				FF-5+Mom			SY		
Strategy	pred	avg	t	SR	α	t	IR	α	t	IR
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-Confident-HL	1.97%	3.61%	9.58	1.75	3.29%	9.02	1.65	3.27%	8.6	1.57
IVOL-Low-Confident-HL	1.80%	5.75%	8.91	1.63	5.40%	8.14	1.49	5.31%	7.72	1.41
IVOL-Confident-HL	1.67%	1.29%	5.44	0.99	1.27%	5.60	1.022	1.27%	5.4	0.99

Table E: Expected return predictions and their standard errors: 48 industry portfolios of Fama and French (1997)

This table shows the average monthly-level ex-ante standard errors of NN-3-based risk premium predictions of 48 Fama and French industry portfolios. The "pred ret" column presents the average monthly predicted risk premiums. The "std" column shows the average monthly standard errors of the risk premium predictions. The elements in "t-ratio" column are the ratios of the entities in "ret" and "std" columns.

Industry code	Industry name	pred ret	std	t-ratio
47	Fin	1.39%	0.032%	43.32
34	BusSv	1.30%	0.044%	29.25
44	Banks	1.51%	0.047%	31.86
36	Chips	1.31%	0.066%	19.97
42	Rtail	1.37%	0.068%	20.19
30	Oil	1.17%	0.070%	16.69
13	Drugs	1.24%	0.071%	17.40
41	Whlsl	1.38%	0.074%	18.59
45	Insur	1.43%	0.079%	18.01
31	Util	1.30%	0.081%	16.03
32	Telcm	1.28%	0.081%	15.85
35	Comps	1.32%	0.085%	15.54
21	Mach	1.37%	0.086%	15.94
12	MedEq	1.31%	0.087%	15.07
40	Trans	1.30%	0.091%	14.30
22	ElcEq	1.35%	0.097%	13.84
43	Meals	1.37%	0.101%	13.57
11	Hlth	1.35%	0.107%	12.66
14	Chems	1.33%	0.109%	12.21
37	LabEq	1.35%	0.109%	12.39
17	BldMt	1.37%	0.114%	12.02
9	Hshld	1.40%	0.116%	12.01
2	Food	1.38%	0.121%	11.37

Table F: Expected return predictions and their standard errors: 48 industry portfolios of Fama and French (1997)

This table shows the average monthly-level ex-ante standard errors of NN-3-based risk premium predictions of 48 Fama and French industry portfolios. The "pred ret" column presents the average monthly predicted risk premiums. The "std" column shows the average monthly standard errors of the risk premium predictions. The elements in "t-ratio" column are the ratios of the entities in "ret" and "std" columns.

Industry code	Industry name	pred ret	std	t-ratio
48	Other	1.34%	0.122%	10.98
23	Autos	1.37%	0.127%	10.73
7	Fun	1.32%	0.128%	10.32
19	Steel	1.29%	0.130%	9.90
18	Cnstr	1.29%	0.136%	9.53
27	Gold	1.11%	0.138%	8.06
33	PerSv	1.36%	0.138%	9.81
46	RlEst	1.33%	0.143%	9.32
8	Books	1.36%	0.144%	9.44
38	Paper	1.36%	0.146%	9.31
10	Clths	1.37%	0.147%	9.35
6	Toys	1.38%	0.163%	8.48
28	Mines	1.13%	0.179%	6.30
15	Rubbr	1.40%	0.181%	7.74
24	Aero	1.37%	0.221%	6.20
16	Txtls	1.41%	0.224%	6.30
4	Beer	1.38%	0.226%	6.12
39	Boxes	1.37%	0.248%	5.51
1	Agric	1.31%	0.266%	4.92
3	Soda	1.30%	0.266%	4.89
20	FabPr	1.44%	0.271%	5.32
29	Coal	1.22%	0.321%	3.80
25	Ships	1.32%	0.364%	3.64
26	Guns	1.32%	0.365%	3.62
5	Smoke	1.24%	0.379%	3.27

Table G: Performance of Various Trading Strategies: 48 industry portfolios of Fama and French (1997)

This table compares the performance of the EW and mean-variance trading strategies formed using 48 industries of Fama and French (1997) over the 30-year out-of-sample (OOS) period. The "avg ret" column shows the average monthly returns. The "t" column presents the t-stats of average returns of both strategies. The "SR" column shows the Sharpe ratios of both strategies.

Strategy	avg ret	t	SR
EW	1.06%	6.59	0.66
Mean-variance	1.82%	3.66	1.2

B1. 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.[19] I simulate a 3-factor model for excess returns, for $t = 1, 2, \ldots, T$:

$$r_{i,t+1} = g(z_{i,t}) + e_{i,t+1},\ e_{i,t+1} = \beta_{i,t} v_{t+1} + \epsilon_{i,t+1},\ z_{i,t} = (1, x_t)' \otimes c_{i,t},\ \beta_{i,t} = (c_{i1,t}, c_{i2,t}, c_{i3,t}), \quad (47)$$

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

$$c_{ij,t} = \frac{2}{N+1} CSrank(\bar{c}_{ij,t}) - 1,\ \bar{c}_{ij,t} = \rho_j \bar{c}_{ij,t-1} + \epsilon_{ij,t},\ \text{for } 1 \leq i \leq 200,\ 1 \leq j \leq 180, \quad (48)$$

where $CSrank$ denotes the cross-sectional rank.

And the time-series x_t is given by

$$x_t = \rho x_{t-1} + u_t, \quad (49)$$

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

$$g(z_{i,t}) = (c_{i1,t}, c_{i2,t}, c_{i3,t})\theta_0,\ \text{where } \theta_0 = (0.02, 0.02, 0.02)'. \quad (50)$$

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

$$g(z_{i,t}) = (c_{i1,t}^2, c_{i1,t} \times c_{i2,t}, sgn(c_{i3,t} \times x_t))\theta_0,\ \text{where } \theta_0 = (0.04, 0.03, 0.012)'. \quad (51)$$

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

[19] 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 (H) 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

$$r_{i,t+1} = g(z_{i,t}) + e_{i,t+1},\ e_{i,t+1} = \epsilon_{i,t+1},\ z_{i,t} = (1, x_t)^{'} \otimes c_{i,t}, \tag{52}$$

where $\epsilon_{i,t+1} \sim \epsilon_{i,t+1} \sim \mathcal{N}(0, 0.05^2)$, $g(z_{i,t})$ is given by (51) and $c_{i,t}$ is given by (48).

Table H: 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%

C. Internet Appendix: Proofs of theorems 1-4

1. Proof of theorem 1

Using Gal and Ghahramani (2016), I obtain the following expressions for the VI-based approximated predictive distribution of returns, respectively.

$$
\begin{aligned}
&P_{VI}(r^*_{i,t+1}|z^*_{it}, R, Z) = P(r^*_{i,t+1}|z^*_{it}, R, Z, \Omega)q(\Omega)\\
&q(\Omega) = \prod_{k=1}^{K} p_{i,k}, \text{ where each } p_{i,k} \sim Bern(p),\\
&P(r^*_{i,t+1}|z^*_{it}, R, Z, \Omega) = \mathcal{N}(r^*_{i,t+1}; \hat{E}_{i,\Omega,t}, \sigma^2_\eta I),
\end{aligned}
\tag{53}
$$

where $Bern()$ represents Bernoulli distribution. $\hat{E}_{i,\Omega,t}$ is given by (15), with d replaced by Ω.

Note that $r^*_{i,t+1} = \mu^*_{i,t} + \eta_{i,t+1}$, where $\eta_{i,t+1}$ is independent of all information (random variables) at t. Denoting $E_{VI}(.)$ as the expectation under the VI-based approximated posterior, note that

$$
\begin{aligned}
E_{VI}(\mu^*_{i,t}) = E_{VI}(r^*_{i,t+1}|z^*_{it}, R, Z) &= \int r^*_{i,t+1} P_{VI}(r^*_{i,t+1}|z^*_{it}, R, Z) dr^*_{i,t+1}\\
&= \int \left(r^*_{i,t+1} \mathcal{N}(r^*_{i,t+1}; \hat{E}_{i,\Omega,t}, \sigma^2_\eta I) \prod_{k=1}^{K} p_{i,k} \right) dp_{i,1} dp_{i,2} \dots dp_{i,K} dr^*_{i,t+1}\\
&= \int \left(r^*_{i,t+1} \mathcal{N}(r^*_{i,t+1}; \hat{E}_{i,\Omega,t}, \sigma^2_\eta I) \prod_{k=1}^{K} p_{i,k} \right) dp_{i,1} dp_{i,2} \dots dp_{i,K} dr^*_{i,t+1}\\
&= \int \left(r^*_{i,t+1} \mathcal{N}(r^*_{i,t+1}; \hat{E}_{i,\Omega,t}, \sigma^2_\eta I) dr^*_{i,t+1} \right) \prod_{k=1}^{K} p_{i,k} dp_{i,1} dp_{i,2} \dots dp_{i,K}\\
&= \int \left(\hat{E}_{i,\Omega,t} dr^*_{i,t+1} \right) \prod_{k=1}^{K} p_{i,k} dp_{i,1} dp_{i,2} \dots dp_{i,K}
\end{aligned}
\tag{54}
$$

Note that by the weak law of large numbers, as $D \to \infty$, the monte-carlo sum

$$
\frac{1}{D} \sum_{d=1}^{D} (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1id} W_{1,\{\lambda,p\}}))(p_{2id} W_{2,\{\lambda,p\}})) \xrightarrow{p} E_{VI}(\mu^*_{i,t})
\tag{55}
$$

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.

Lastly, (11) $\implies$ $E^*_{it,Dropout} \xrightarrow{p} E_{VI}(\mu^*_{i,t})$

2. Proof of theorem 2

Denoting $Var_{VI}(.)$ as the variance under the VI-based approximated posterior, note that $Var_{VI}\left[(r^*_{i,t+1}|z^*_{it}, R, Z)\right] = E_{VI}\left[Var_{W_1,W_2}(r^*_{i,t+1}|z^*_{it}, R, Z, W_1, W_2)\right] + Var_{VI}\left[E_{W_1,W_2}(r^*_{i,t+1}|z^*_{it}, R, Z, W_1, W_2)\right]$, where E_{W_1,W_2} and Var_{W_1,W_2} represent conditional variance and expectation operations given W_1, W_2, respectively. Further note that $Var_{W_1,W_2}(r^*_{i,t+1}|z^*_{it}, R, Z, W_1, W_2) = \sigma^2_\eta$. Thus,

$$Var_{VI}\left[(r^*_{i,t+1}|z^*_{it}, R, Z)\right] = \sigma^2_\eta + Var_{VI}\left[E_{W_1,W_2}(r^*_{i,t+1}|z^*_{it}, R, Z, W_1, W_2)\right],$$

Similar to (55) in the proof of theorem 1 , and by the weak law of large numbers, as $D \to \infty$

$$\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2 \xrightarrow{p} Var_{VI}\left[E_{W_1,W_2}(r^*_{i,t+1}|z^*_{it}, R, Z, W_1, W_2)\right] \tag{56}$$

Thus,

$$\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2 + \sigma^2_\eta \xrightarrow{p} Var_{VI}\left[(r^*_{i,t+1}|z^*_{it}, R, Z)\right] \tag{57}$$

Denote $Var_{VI}\left[(r^*_{i,t+1}|z^*_{it}, R, Z)\right]$ by $Var_{VI}(r^*_{i,t+1})$, where V_{VI} represents the variance operation under the VI-based probability distribution $P_{VI}(.|z^*_{it}, R, Z)$. Note that by (19), and by the law of total variance,

$$Var_{VI}(r^*_{i,t+1}) = Var_{VI}(E(r^*_{i,t+1}|W_1, W_2)) + E_{VI}(V(r^*_{i,t+1}|W_1, W_2)), \tag{58}$$

where W_1, W_2 are the unknown weight matrices of the NN-1; E_{VI} represents the expectation operation under the probability distribution $P_{VI}(r^*_{i,t+1}|z^*_{it}, R, Z)$; $E()$, $V()$ represents the expectation and variance operations under the likelihood function (19), respectively.

(58) further implies that

$$Var_{VI}(r^*_{i,t+1}) = Var_{VI}(\mu^*_{i,t}) + \sigma^2_\eta, \tag{59}$$

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

Thus, (56) and (58) implies

$$\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2 \xrightarrow{p} Var_{VI}(\mu^*_{i,t}). \tag{60}$$

3. Proof of Theorem-3

To compute covariances, VI-based approximated joint density of return predictions is required. Straightforward algebra implies that it is given by

$$\begin{aligned}
&P_{VI}(r^*_{1,t+1}, r^*_{2,t+1}, \dots r^*_{S,t+1} | \{z^*_{it}\}_{i=1}^S, R, Z) = P(r^*_{1,t+1}, r^*_{2,t+1}, \dots r^*_{S,t+1} | \{z^*_{it}\}_{i=1}^S, R, Z, \Omega) q(\Omega) \\
&q(\Omega) = \prod_{k=1}^{K} p_{i,k}, \text{ where each } p_{i,k} \sim Bern(p), \\
&P(r^*_{1,t+1}, r^*_{2,t+1}, \dots r^*_{S,t+1} | \{z^*_{it}\}_{i=1}^S, R, Z, \Omega) = \mathcal{N}(\hat{E}_{S,\Omega,t}, \sigma^2_\eta I), \text{ where } \hat{E}_{S,\Omega,t} = \begin{bmatrix} \hat{E}_{1,\Omega,t} \\ \hat{E}_{2,\Omega,t} \\ \vdots \\ \hat{E}_{S,\Omega,t} \end{bmatrix}, \qquad (61)
\end{aligned}$$

with each $\hat{E}_{i,\Omega,t}$ given by (15). The key is to use the same Ω across the stocks, as discussed in the main section of the paper.

Then, similar to the proof of (2), the covariance of any two return VI-based posterior predictive densities satisfy

$$\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)\left(\hat{E}_{j,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{j,d,t}\right) \xrightarrow{p} Covar_{VI}\left[(r^*_{i,t+1}, r^*_{j,t+1} | z^*_{it}, R, Z)\right], \qquad (62)$$

for any i, j $(i \neq j)$, where $Covar_{VI}\left[(r^*_{i,t+1}, r^*_{j,t+1} | z^*_{it}, R, Z)\right]$ denotes the covariance between the return predictions $r^*_{i,t+1}$, $r^*_{j,t+1}$ under the VI-based approximated joint posterior density.

Because $r^*_{i,t+1} = \mu^*_{i,t} + \eta_{i,t+1}$ and $r^*_{j,t+1} = \mu^*_{j,t} + \eta_{j,t+1}$, with $\eta_{i,t+1}$ and $\eta_{j,t+1}$ independent of all other random variables, it is immediate that

$$\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)\left(\hat{E}_{j,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{j,d,t}\right) \xrightarrow{p} Covar_{VI}\left[(\mu^*_{i,t}, \mu^*_{j,t} | z^*_{it}, R, Z)\right]. \qquad (63)$$

4. Proof of theorem 4

The proof is straightforward from the proof of theorem 3.

D. Internet Appendix: Frequentist Consistency

This section lays out the conditions under which the dropout-based or the VI-based approximated risk premium predictions from Bayesian NNs satisfy the frequentist consistency, by proving theorem 5.

Suppose that, given a set of characteristics z_{it}, excess returns are given by

$$r_{i,t+1} = b_2 + \phi(b_1 + z_{it}W_{o1})W_{o2} + \eta_{i,t+1}, \ \eta_{i,t+1} = N(0, \sigma_\eta^2), \ \forall i., \tag{64}$$

with the weight matrices W_{o1}, W_{o2} unknown, but b_1, b_2, σ_η^2 known.

Denote the set of true parameters

$$\theta_0 = \{W_{o1}, W_{o2}\} \tag{65}$$

Now consider the Bayesian NN specification similar to (19) in section II.D, with parameters $\theta = \{W_1, W_2\}$. In the spirit of Bernstein-von Mises theorem (Kleijn and Vaart (2012); Vaart (2000); Wang and Blei (2019)), make the following assumptions on the prior and likelihood functions.

Assumption 1: *(Prior mass). The prior density $P(\theta)$ is continuous and positive in a neighborhood of θ_0. There exists a constant $M_p > 0$ such that $|logP(\theta)''| \leq M_p e^{|\theta|^2}$.*

Comment: Assumption 1 states that the prior has some mass around the true parameter θ_0. Assumption 1 also puts a bound on the growth rate of the log prior likelihood. These assumptions are very mild, which many commonly used priors, including this paper's priors (21), satisfy.

Assumption 2: *(Consistent testability). For every $\epsilon > 0$, $\exists$ a sequence of tests ϕ_n such that*

$$\int \phi_n(R) \left[\Pi p_0(r_{it})\right] dR \to 0, \tag{66}$$

$$\sup_{\theta:||\theta-\theta_0||\geq\epsilon} \int (1 - \phi_n(R)) \left[\Pi p_0(r_{it})\right] dR \to 0, \tag{67}$$

where R denotes the panel of excess returns for a given set of stocks over a given period of time; $p(r_{it}|\theta)$ represents the likelihood of r_{it} given θ; $p_0(r_{it})$ denotes the likelihood of r_{it} given θ_0.

Comment: Assumption 2 requires that θ_0 is identifiable from the likelihood function $p_0(r_{it})$, which this paper's likelihood satisfies. In particular, to meet assumption 2, it suffices to show that $\frac{p(R|\theta_1)}{p(R|\theta_2)}$ is a continuous function of R, for all θ_1 θ_2.

Assumption 3: *(Local asymptotic normality). For every compact set $K \subset R^d$, $\exists$ random vectors*

Δ_{n,θ_0} bounded in probability and nonsingular matrices V_{θ_0} such that

$$\sup_{h\in K}\left|\log\frac{p(R|\theta_0+\delta_n h)}{P(R|\theta_0)}-h^T V_{\theta_0}\Delta_{n,\theta_0}+\frac{1}{2}h^T V_{\theta_0}h\right|\xrightarrow{P_0}0, \tag{68}$$

where δ_n is a $d\times d$ diagonal matrix that describes how fast each dimension of the θ posterior converges to a point mass, with $\delta_n\to 0$ as $n\to\infty$.

Comment. This assumption determines the limiting normal distribution of the VI-based approximated posterior. The quantities Δ_{n,θ_0} and V_{θ_0} determine the normal distribution that the VI-approximated posterior will converge to. The constant δ_n determines the convergence rate of the VI-approximated posterior to a point mass.

Result 1: *Given the parameteric specification of excess returns in* (64)*, and the iid assumption of excess returns given the parameters, Theorem 7.2 of Vaart (2000) implies that*

$$\sup_{h\in K}\left|\log\frac{p(R|\theta_0+h/\sqrt{num})}{P(R|\theta_0)}-\frac{1}{num}\sum_{i,t}h^T p'_{\theta_0}(r_{it})+\frac{1}{2}h^T I_{\theta_0}h\right|\xrightarrow{P_0}0, \tag{69}$$

where num is the total number of stocks and time periods; $p'_{\theta_0}(r_{it})$ is the derivative of the likelihood function evaluated at θ_0; I_{θ_0} is the Fisher information matrix evaluated at θ_0.

Thus, result 1 shows that this paper's framework satisfies assumption 3. Moreover, note that VI-based posteriors would eventually converge to the multivariate normal centered around the maximum likelihood estimator, with a convergence rate of $\sqrt{num}$.

Result 2: *Under assumptions 1-3, the optimal variational density converges in total variation to the KL minimizer of the multivariate normal with mean equal to the MLE of θ and variance equal to the information matrix (evaluated at θ_0).*

$$\left|\left|q_{M_1^*,M_2^*}(.)-\arg\min_q KL\left(q()||MVN\left(\hat{\theta}_{MLE},(1/\sqrt{num})I_{\theta_0}\right)\right)\right|\right|_{TV}\xrightarrow{p}0, \tag{70}$$

where $q_{M_1^*,M_2^*}(.)$ are given in (23)*, with optimal M_1^*,M_2^* substituted for M_1,M_2; $\{M_1^*,M_2^*\}$ are given in* (26)*; $\hat{\theta}_{MLE}$ denotes the MLE of θ.*

Proof. The proof follows directly from result 1 and theorem 5 (5.2) of Wang and Blei (2019). □

Comment. Note that the KL minimizer $\arg\min_q KL\left(q()||MVN(\hat{\theta}_{MLE},I_{\theta_0})\right)$ is the member in the variational family containing the mixture of Gaussian distributions (23) that is closest to the multivariate normal centered around MLE. Thus, even the KL minimizer is a mixture of Gaussians.

However, given that the weight matrices W_1 and W_2 are independent across columns or neurons K, as $K \to \infty$, the KL minimizer converges to a multivariate normal. The reason is that the entropy of a mixture of Gaussians with a large enough dimensionality and randomly distributed means tends towards to the sum of Gaussians' volumes.[20]. In addition, note that the VI family q_{M_1,M_2} specifies the rows of W_1 and W_2 to be correlated, thus capturing all significant correlations between NN weights. Although the VI family ignores the correlations across columns, such correlations would be negligible as $K \to \infty$. For example, Gal and Ghahramani (2016) note that the variational family induces strong joint correlations over the rows of matrices W_i, which correspond to the frequencies in sparse spectrum Gaussian Process (equivalent to Bayesian NN) approximation. Thus, the KL minimizer could be approximated by

$$\arg\min_{q} KL\left(q()||MVN(\hat{\theta}_{MLE}, I_{\theta_0})\right) \approx MVN\left(\hat{\theta}_{MLE}, (1/\sqrt{num})I_{\theta_0}\right) \tag{71}$$

Thus, due to (71), the following result follows.

Result 3: *Under assumptions 1-3, the optimal variational density converges in total variation to the multivariate normal with mean equal to the MLE of θ and variance equal to the information matrix (evaluated at θ_0).*

$$\left|\left|q_{M_1^*,M_2^*}(.) - MVN\left(\hat{\theta}_{MLE}, (1/\sqrt{num})I_{\theta_0}\right)\right|\right|_{TV} \xrightarrow{p} 0. \tag{72}$$

Note that the paper focuses on the joint density of risk premium predictions (rather than NN weights). Because risk premiums could be expressed as smooth functions of θ given a set of characteristics, i.e.,

$$\mu_{i,t}^* = b_2 + \phi(b_1 + z_{it}^* W_1)W_2, \tag{73}$$

applying the delta method to (72) proves theorem 5.

[20] For a detailed proof, see the appendix (page 7) of Gal and Ghahramani (2016).