Confident Risk Premiums and Investments using Machine Learning Uncertainties

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

Current draft: March 1, 2023

Abstract

This paper derives ex-ante confidence intervals of stock risk premium forecasts that are based on a wide range of linear and Machine Learning models. Exploiting the cross-sectional variation in 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 outperform traditional high-low strategies in delivering superior out-of-sample returns and Sharpe ratios across all models. The outperformance increases (decreases) with the model complexity (bias). The confident-high-low strategies are economically interpretable as trading strategies of ambiguity-averse investors who account for confidence intervals around risk premium forecasts.

Keywords: Investment Strategies, Stock Return Forecasts, Confidence Intervals, Risk Premiums, Lasso, Elastic Net, Ridge, Neural Networks, Standard Errors, Machine Learning

*I thank Stefano Giglio (RFS Editor) and three anonymous reviewers for many insightful comments. The current 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". I previously presented this research in seminars at the business schools of Yale University, Boston College, Rice University, Boston University, Copenhagen, HEC Paris, HKUST, Florida, Georgia, Houston, Tulane, and Virgina Tech, Indian School of Business and National University of Singapore. It was also presented at the SoFiE (2022), North American Econometric Society (2022), Auckland Finance (2022), 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, Helen Lu (discussant), 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

1. Introduction

Modern empirical asset pricing literature applies machine learning (ML) 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 outperform linear characteristic-based models examined by Lewellen (2015) (henceforth Lewellen) in forecasting stock risk premiums out-of-sample (OOS). However, nothing is known about the ex-ante precision (i.e., standard errors and confidence intervals) of risk premium forecasts that are based on these models. Fama and French (1997) and Pástor and Stambaugh (1999) show that expected return estimates from traditional, linear factor models are unavoidably imprecise due to uncertainty about unknown parameters. Given that forecasting with ML models entails estimating a massive number of parameters, determining the precision of ML-based risk premium forecasts is important.

This paper estimates ex-ante (co)variances and confidence intervals of expected return forecasts that are based on a wide range of linear and ML models, including the Lewellen model, penalized linear models (Lasso, Ridge and Elastic Net) and Neural Networks (NN). The ex-ante confidence intervals capture estimation uncertainty related to risk premium forecasts. Whereas variances of risk premium estimates from linear factor models are available in the literature, those of highly complex ML-based risk premium forecasts are not. I tackle this challenge by proving that the risk premium forecasts that are based on various ML models have different Bayesian interpretations, whose posterior densities are easily estimable. Thus, I obtain the confidence intervals of ML-based risk premium forecasts using the comparable Bayesian posterior densities. The obtained (co)variances are then statistically justified and empirically validated using Monte-Carlo simulations.

Establishing Bayesian interpretations for ML-based risk premium forecasts has three main advantages. First, ex-ante confidence intervals are easily and simultaneously obtained along with the risk premium forecasts without additional computational costs.[1] Other procedures like Bootstrap that estimate forecast variances require retraining ML models a large number of times, rendering

[1] Confidence intervals are easily obtained using only a few lines of code that I make publicly available.

them computationally infeasible.[2] Second, risk premium forecasts (and their confidence intervals) are allowed to explicitly take into account the cross-sectional correlations of stock returns, which the majority of ML studies in finance ignore. Third, it is possible to simultaneously obtain the the confidence intervals of the portfolio-level risk premium forecasts (e.g., 48 industry portfolios of Fama and French (1997)), not just at the stock level.

The estimated ex-ante precision measurements are useful for numerous 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 significant 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 forecasts, 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 first sort stocks based on their risk premium forecasts and then take long-short positions exclusively on the subset of stocks in the extreme return-forecast-deciles that have relatively more confident risk premium forecasts.

The Confident-HL strategies formed using unbiased return forecasts deliver superior OOS returns and Sharpe ratios. Whereas a risk premium forecast proxies for the next period's return, its standard error proxies for its squared forecast error. Alternatively, when the standard error of a stock's risk premium forecast is large, so will its squared forecast error. The reason is that the expected squared forecast errors equal the sum of ex-ante "variances" that capture estimation uncertainty and squared "biases" that quantify model misspecification. Thus, the ex-ante variances of unbiased risk premium forecasts completely predict their ex-post squared forecast errors. The Confident-HL strategies exploit this predictability. By deliberately dropping ex-ante imprecise forecasts, they minimize ex-post OOS misclassification of stocks into appropriate return deciles, and thus they deliver superior returns OOS. A simple example provides the central intuition.

[2] In addition, bootstrap methods do not deliver reliable standard errors for zero coefficients in penalized linear regressions Kyung, Gill, Ghosh, and Casella (2010).

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 forecasts that are normal, unbiased, and uncorrelated with the measurement error variance σ^2. Then the expected OOS return of the HL strategy that takes a long (short) position on the stock with the highest (lowest) risk premium forecast 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 forecasts. In other words, between any two sets of stocks with the same levels of risk premiums, the HL strategy formed from more precise forecasts yields higher OOS expected returns. For example, when the forecasts 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 forecasts 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 forecasts, the precision helps to better determine the cross-sectional ranking among stocks. Thus, the Confident-HL strategies exclusively containing stocks with precise risk premium forecasts generate higher returns OOS.

The magnitude of improvements provided by the Confident-HL portfolios depends on the model bias or how well the ex-ante standard errors of risk premium forecasts predict their ex-post squared forecast errors. Since forecasts from NNs capture non-linear interactions amongst a high-dimensional set of predictors, they are relatively less biased than the penalized linear model forecasts that ignore non-linearity, which further are relatively less biased than Lewellen forecasts that ignore non-linearity and many useful return predictors. Thus, the relative portfolio gains delivered by ex-ate confidence intervals will be in the decreasing order of the model biases, with the most significant gains for NNs, followed by the penalized linear, and then followed by Lewellen. In addition, the predictive magnitudes that quantify how well ex-ante variances predict ex-post

squared forecast errors will also be in decreasing order of model biases.

Consistent with this intuition, the empirical section documents large economic gains from the Confident-HL portfolios across all the models, where the gains are in the decreasing order of the model biases. It also documents that the ex-ante standard errors significantly predict ex-post squared forecast errors across the models, where the predictability decreases with the model bias. On a large sample of US stock returns between 1957 and 2020, I examine three forecasting models: Lewellen, Lasso, and a 3-layer Neural Network (NN-3). And it uses a high-dimensional set of cross-sectional and macroeconomic predictors, previously examined by GKX and Avramov et al. (2020), to obtain OOS risk premium forecasts and their confidence intervals.

For NN-3 based risk premium forecasts, the conventional equal-weighted (value-weighted), EW (VW), HL portfolio earns an average monthly OOS return of 2.21% (1.29%), with an annualized Sharpe ratio of 1.36 (0.78). However, the EW (VW) Confident-HL portfolio delivers corresponding measures of 3.84% (2.70%) and 1.78 (1.26), respectively. Thus, dropping imprecise forecasts leads to enormous improvements in the OOS average return and Sharpe ratio. For perspective, the Sharpe ratio improvement translates to 9.14% (7.07%) *annualized holding period return* difference between the EW (VW) Confident-HL and EW (VW) HL strategies, standardizing both to have the same return variances. In contrast to the Confident-HL strategies, the EW (VW) "Low-Confident" portfolio that instead takes long-short positions on the subset of stocks in the extreme return-forecast-deciles with the most imprecise risk premium forecasts yields relatively much lower OOS average monthly return and Sharpe ratio, 1.43% (0.76%) and 0.68 (0.33), respectively.

The Confident-HL strategies also significantly outperform two other benchmark strategies: 1%-HL strategies and Low-Ivol-HL strategies. 1%-HL strategies take long (short) positions on the top (bottom) 1% of the stocks with the highest (lowest) risk premium forecasts, thus containing exactly the same number of stocks as the Confident-HL strategies. The average annual OOS return difference between the (EW) VW Confident-HL and EW (VW) 1%-HL is 6.24% (6.28%), standardizing both strategies to have the same variances. 1%-HL strategies underperform because they ignore the precision of risk premium forecasts, leading them to misclassify stocks into inappropriate return-forecast deciles. The other benchamark, Low-Ivol-HL, is a double-sorted strategy that first

sorts on return forecasts and then takes long (short) positions on the subset of stocks in extreme return-forecast-deciles with low idiosyncratic return volatilities . Even the EW (VW) Low-Ivol-HL strategy delivers substantially lower monthly return of 1.24% (0.16%) and the annualized Sharpe ratio of 0.83 (0.12), relative to those of the Confident-HL strategies. Whereas this paper's ex-ante confidence intervals are conditional measures that explicitly depend on the current predictor set, the IVOL measures do not incorporate the up-to-date information from the predictor set, thereby not providing good predictions of the ex-post squared forecast errors. Thus, the Low-Ivol strategies underperform.

The two main results, the outperformance of the Confident-HL portfolios and the underpeformance of the Low-Confident-HL strategies, also hold for risk premium forecasts that are based on Lewellen and Lasso models. For instance, the average annualized holding period return difference between the EW Confident-HL and the EW HL strategies that are based on Lasso (Lewellen) is 8.22% (6.55%). The differences in Squared Sharpe ratios, Information ratios with respect to the 6-factor model that adds the momentum factor to the 5 factors of Fama and French (2015), and the average monthly returns between the Confident-HL and the other benchmark strategies are statistically significant across all the three models. The Confident-HL portfolios significantly outperform even on the subsample of non-microcaps. And the outperformance is robust to transaction costs, to drawdowns, and to higher-moment risks that penalize losses more than rewarding gains.

To accurately measure the economic value of incorporating confidence intervals into trading strategies, I construct counterfactual matching strategies that resemble the Confident-HL strategies but ignore confidence intervals. The economic value, measured by the average return difference between both strategies, is significant across all the three models and decreases with the model bias. The values are additional annual OOS returns of 11.78% for NN-3, 10.70% for Lasso, and 8.68% for Lewellen. Recall that the Confident-HL portfolios outperform because the ex-ante standard errors of risk premium forecasts predict their ex-post squared forecast errors. Consistent with this result, I document that the ex-ante precision and the ex-post mean squared errors (MSEs) are monotonically related across all the models. For NN-3 (Lasso, Lewellen) model, the bottom decile of stocks with the most imprecise return forecasts attains an OOS MSE of 8.52% (7.59%, 6.38%).

In contrast, the top decile of stocks with most precise forecasts delivers significantly lower MSE 2.26% (1.57%, 2.52%). And the steepness of the monotonicity decreases with the model bias.

Economically, the Confident-HL strategies are interpretable as the trading strategies of ambiguity-averse investors with max-min utility who explicitly take into account the confidence intervals around risk premium forecasts. In contrast, the traditional HL portfolios are the strategies of ambiguity-neutral investors who ignore confidence intervals. While Garlappi, Uppal, and Wang (2007) discuss the trading strategies of ambiguity-averse investors, the novelty of this paper is to show why such strategies sorted on the ex-ante confidence intervals can deliver superior returns OOS (see Example-1). The Confident-HL strategies are also interpretable as a stylized class of mean-variance portfolios that regularize the mean-variance weights depending on the ex-ante variances of risk premium forecasts to mitigate estimation uncertainty.

In the time-series, ex-ante standard errors reflect stock market uncertainty, with standard errors increasing by a factor of more than two after major shocks such as Black Monday, Lehman Bankruptcy, and Covid. Since many individual predictors (e.g., price trends) deviate from their usual distributions when markets are uncertain, risk premium forecasts based on these unusual predictors will also be imprecise. Thus, ex-ante standard errors can capture market uncertainty. Cross-sectionally, the ex-ante standard errors are associated with a multitude of characteristics, such as size, book-to-market, and momentum, and a few characteristics cannot explain them. This explains why sorting on confidence intervals provides more portfolio gains than sorting on a single characteristic like size or IVOL (e.g., 1%-HL and Low-Ivol-HL). At the portfolio-level, ML models more precisely forecast risk premiums of industry portfolios related to Financial Services, Business Services, Banks, Chips, Retail and Oil, suggesting that the existing predictors and models are more suited for estimating risk premiums of these industries. Possible mechanisms that explain the cross-sectional variation in the precision of risk premium forecasts warrants a future study.

Overall, this paper estimates the ex-ante precision of risk premium forecasts that are based on a wide range of linear and Machine Learning models. It then shows that the Confident-HL strategies that incorporate the ex-ante precision deliver superior OOS performance across all models, where the improvements increase with model complexity.

1.1. Contribution

This is the first paper to estimate the (co)variances of risk premium forecasts that are based on penalized linear models and NNs, both at the stock-level and at the portfolio-level. I also contribute to the literature on ML applications in asset pricing. I discuss each in turn.

Penalized linear models. Kyung et al. (2010) provide a Bayesian framework to estimate otherwise intractable standard errors of forecasts that are based on penalized linear models, including LASSO, Ridge, and Elastic-net, assuming an iid data generating process. Simiarly, most ML studies in finance (e.g., GKX and Avramov et al. (2020)) also implicitly make the iid assumption to generate risk premium forecasts. Since this assumption ignores cross-sectional dependence of stock returns, it violates existing empirical evidence and theoretical models like the CAPM. So, this paper explicitly takes into account the cross-sectional correlation structure of stock returns to estimate risk premium forecasts and their (co)variances more robustly.

Neural Networks. I make methodological advancements relative to Gal and Ghahramani (2016) (GG), who show how to estimate standard errors of NN-based forecasts, in *three dimensions*. First, this paper incorporates the cross-sectional correlation of stock returns while estimating risk premium forecasts and their (co) variances, whereas GG assume that data are iid. Second, I show how to compute the variances of "prediction means" (i.e., risk premium forecasts), which are more relevant in the finance literature, whereas GG compute those of individual "raw" predictions (i.e., excess return forecasts). Last, I show how to compute the marginal and joint densities of NN-based risk premium forecasts that GG do not provide but are necessary for computing portfolio-level variances.

Asset Pricing studies in ML. Several prominent researchers apply ML to address various asset pricing questions. For instance, ML has been used by Feng, Giglio, and Xiu (2020) to understand the contribution to asset pricing of a new factor, above and beyond what existing factors explain; by Chen, Pelger, and Zhu (2020) to estimate the stochastic discount factor implied by the no-arbitrage restriction; by Bryzgalova, Pelger, and Zhu (2020) to construct a set of statistically-motivated test assets, and by Jensen et al (2022) to obtain transaction-cost-adjusted

efficient portfolios. Instead of using ML to address an asset pricing question, I quantify the uncertainties of forecasts made by ML models to understand the scenarios when these ML models make precise return forecasts and for what stocks. I also show how this precision information helps to form better investments. Finally, the Confident-HL strategies are fundamentally different 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 forecast variances.

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 semiparametric 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 forecasts that are based on random forests. However, their approach does not have a Bayesian interpretation and thus cannot yield portfolio-level forecast variances. A similar exercise with random forests is left for the future.

2. Risk Premium Forecasts and their Ex-ante Precision

This section presents the statistical framework to estimate ex-ante confidence intervals and (co)variances of risk premium forecasts that are based on linear regression, LASSO, Ridge, Elastic Net, and Neural Network models. It proves that the forecasts from these models have different Bayesian interpretations whose posterior densities are easily estimable. Thus, the (co)variances of risk premium forecasts are obtained using the comparable Bayesian forecasts' posterior densities. Although Bayesian posterior variances and frequentist variances philosophically represent different entities, the Internet Appendix discusses the *frequentist consistency* of the estimated (co)variances.

Notations: Throughout this section, $r_{i,t+1}$ denotes stock i's excess return at period $t+1$; $\{z_{it}\}$ denotes a large set of p stock i's raw predictors, such as size, book-to-market, 1-month momentum

returns, at time t; and $\{\eta_{i,t+1}\}$ denotes the set of model residuals with zero mean.

2.1. Linear, Lasso, Ridge, and Elastic Net:

Under the penalized linear model specification, the excess returns are modeled as

$$r_{it+1} = z_{it}^T \beta + \eta_{i,t+1}, \tag{2}$$

where the parameters $\beta = \{\beta_1, \beta_2, \ldots, \beta_p\}$ are unknown and must be estimated.

2.1.1. Parameter and Confidence Interval Estimation under the IID Assumption.

Most studies in finance estimate the unknown parameters by minimizing the following penalized MSE over the training sample:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{N_{Tr} N_S} \sum_{t \in Tr} \sum_{i \in S} \Big(r_{i,t+1} - z_{it}^T \beta) \Big)^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2, \tag{3}$$

where Tr is the training sample over N_{Tr} periods; S is the total set of N_S stocks; $||.||_1$ represents the L_1 norm operator; $||.||_2$ the L_2 norm operator; λ_1 and λ are the L_1 and L_2 "hyperparameters" that prevent overfitting. The model is linear if $\lambda_1 = 0$ and $\lambda_2 = 0$; LASSO if $\lambda_1 \neq 0$ and $\lambda_2 = 0$; Ridge if $\lambda_1 = 0$ and $\lambda_2 \neq 0$; and Elastic Net if both $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$. Risk premiums are then forecasted using $z_{it}^T \hat{\beta}$. The estimated parameters and forecasts enjoy the Oracle property under the iid assumption of residuals (Zou (2006)).

Bayesian Interpretation of Penalized Linear Models. When the residuals are further assumed to be $N(0, \sigma_\eta^2)$, Kyung et al. (2010) prove that the estimated $\hat{\beta}$ is identical to the (Bayesian) posterior mode of β under the following exponential prior on β:

$$\Pi(\beta | \sigma_\eta^2) \propto \exp\Big\{ -\frac{\lambda_1}{\sigma_\eta} \sum_{j=1}^{p} |\beta_j| - \frac{\lambda_2}{2\sigma_\eta^2} \sum_{j-1}^{p} |\beta_j|^2 \Big\} \tag{4}$$

Note that the prior in (4) collapses to the standard diffuse prior under a linear model. The

(co)variances of risk premium forecasts could then be obtained using the posterior variances of the Bayesian predictive density, which is easily estimated using a simple Gibbs Sampling procedure in the spirit of Kyung et al. (2010).

However, the iid specification of the residuals ignores the cross-sectional dependence of monthly stock returns, clearly violating existing empirical evidence and theoretical models like the CAPM.[3] Thus, the following subsection relaxes the iid specification in the estimation of risk premium forecasts and their (co)variances.

2.1.2. Parameter and Confidence Interval Estimation under the non-IID Assumption.

To take into account the cross-sectional dependence of stock returns, I model the residuals using the following factor model that captures the cross-sectional dependence of stock returns.

$$\eta_{it} = \Lambda_i f_t + \epsilon_{it}, \;\; \epsilon_{it} \sim N(0, \sigma_\epsilon^2), \tag{5}$$

where the idiosyncratic errors, ϵ_{it}, are assumed to be iid, and the factor f_t has zero mean.

While multiple factors could be allowed to model the residuals, Smith and Timmermann (2021, 2022) document that the single common factor, innovations in the market portfolio, sufficiently captures the cross-sectional correlation structure in the linear regression of monthly stock returns on a set of lagged characteristics. So, I consider a one-factor model with the monthly market return innovations as the factor.

I further specify the factor loadings as a function of few characteristics,

$$\Lambda_i = a_0 + a_1 Size_{i,t-1} + a_2 BM_{i,t-1} + a_3 Mom_{i,t-1}, \tag{6}$$

where $Size_{i,t-1}$ denotes the market cap of firm i at time $t-1$; $BM_{i,t-1}$ denotes the book-to-market of firm i at time $t-1$; $Mom_{i,t-1}$ denotes the 1-year momentum return of firm i at time $t-1$.

The specification in (6) is motivated for two reasons. First, it is consistent with the conditional

[3] Ignoring time-series correlations is expected to not pose serious problems because monthly stock returns are documented to exhibit insignificant or weak auto-correlations (Fama (1971)).

CAPM (Ferson and Harvey (1999) and Kelly, Pruitt, and Su (2019)) and allows for time-varying factor loadings. Second, it is not possible to estimate Λ_i in unbalanced panels for stocks that are newly listed or that have limited return history. The specification in (6) tackles this problem by exploiting the entire pool to estimate the factor loadings of all stocks, including those with limited trading history. Empirically, the null hypothesis that the residuals satisfy (5) and (6) is not rejected.[4]

While (5) and (6) specify a known factor model with loadings that take a specific paramteric form, the paper's estimation framework generally applies to the specification that has multiple latent factors and loadings in the spirit of Pesaran (2006), Bai (2009), and Lu and Su (2016).

Importantly, when the residuals are non-iid, the usual risk premium forecasts from penalized linear models that are obtained by minimizing the regularized mean squared error (MSE), as in GKX, would not only be inefficient, but they also would not enjoy the Oracle property (i.e., *consistent* variable selection). The desirable forecasts that are efficient and satisfy the Oracle property will be those that instead maximize the penalized log-likelihood (Fan and Li (2001), Fan and Peng (2004), and Lee and Liu (2012)).[5] These estimators are given by

$$\begin{aligned}\hat{\beta}_l &= \arg\min_{\beta} \frac{1}{N_{Tr}N_S} \sum_{t\in Tr}\sum_{i\in S}\left(r_{i,t+1} - z_{it}^T\beta - \Lambda_i f_{t+1}\right)^2 (f_{t+1}^2) + \lambda_1||\beta||_1 + \lambda_2||\beta||_2 \\ &= \arg\min_{\beta} \frac{1}{N_{Tr}N_S} \sum_{t\in Tr}\sum_{i\in S}\left(\bar{r}_{i,t+1} - z_{it}^T\beta\right)^2 (f_{t+1}^2) + \lambda_1||\beta||_1 + \lambda_2||\beta||_2 \end{aligned} \tag{7}$$

where Λ_i is given in (6); and $\bar{r}_{i,t+1} = r_{i,t+1} - \Lambda_i f_{t+1}$, which represents othorgonalized returns with respect to the market factor.

Thus, the main difference between the iid-case estimator $\hat{\beta}$ in (3) and the non-iid-case estimator $\hat{\beta}_l$ in (7) is that the latter focuses on forecasting the orthogonalized returns, whereas the former targets the raw returns. In fact, $\hat{\beta}_l$ can be interpreted as the penalized WLS coefficients in the regression of orthogonalized returns on $\{z_{it}\}$, where the weights are proportional to the market

[4] I validate this by estimating the average pairwise correlations between the idiosyncratic errors $\{\epsilon_{i,t}\}$ and testing their significance using the cross-sectional dependence test proposed by Pesaran (2021). The null of non dependence is not rejected by the data. See also page 560 in Smith and Timmermann (2022).

[5] See sections 3 and 5 in Lee and Liu (2012) for a theoretical proof and a simulation evidence, respectively.

factor. While the orthogonalization captures the cross-sectional dependence of the model residuals, the factor weighting accounts for the residual heterogeneity. The risk premium forecasts are then given by $z_{it}\hat{\beta}_l$, whose (co)variances are estimated by establishing the following Bayesian result.

Theorem 1: *When excess returns are modeled using a linear model as in* (1) *with the model residuals satisfying the factor structure in* (5)*, the penalized log-likelihood estimator in* (7) *is identical to the Bayesian posterior mode of* β *under the double exponential prior of* β *in* (4)*.*

Appendix provides the proof. Thus, the confidence intervals of risk premium forecasts could be estimated using the posterior density of β under the double exponential prior. In particular, the (co)variance of risk premium forecasts equal the posterior (co)variances of the risk premium predictive density, which is easily obtained using the Gibbs Sampling algorithm described in Appendix A. The algorithm involves repeated sampling of the parameters, i) β; ii) the scale parameters that govern the penalty; iii) and the residual variance parameters $\{\sigma_\epsilon, a_0, a_1, a_2\}$, from their respective conditional posterior densities given the other parameters.

2.2. Neural Networks

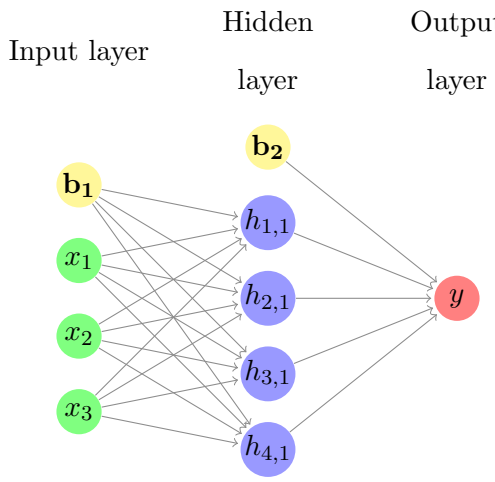
GG prove that forecasts from NNs that impose "dropout" regularization have equivalent Bayesian interpretation, whose posterior densities are easily estimable.[6] Building on this insight, the section extends GG in three dimensions. First, it computes the variances of "prediction means" (i.e., risk premium forecasts), which are more relevant in the finance literature, whereas GG compute those of individual "raw" predictions (i.e., excess return forecasts). Second, it computes the marginal and joint densities of NN-based risk premium forecasts that GG do not provide but are necessary for computing portfolio-level variances. Last, it incorporates the cross-sectional correlation of stock returns while estimating risk premium forecasts and their (co) variances, whereas GG assume that data are iid. I present all proofs and technical details in Appendix and Internet Appendix, respectively. Below I directly discuss how NN-based risk premium forecasts and their

[6] More specifically, NNs that impose dropout are mathematically equivalent to approximating a Bayesian Gaussian Process using a sophisticated methodology known as Variational Inference (VI). See Allena and Chordia (2022) for a detailed discussion on VI and its application in finance.

(co)variances are estimated.

Like GKX, I consider "feed-forward" NNs that 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. Example of a 1-layer Neural Network



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

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

$$h_{k,1} = \phi\left(b_{1k} + \sum_{j=1}^{3} w_{1jk} x_j\right), \text{ for } k = 1, 2, 3, 4, \tag{8}$$

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

$$\phi(x) = ReLU(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ otherwise.} \end{cases} \tag{9}$$

Likewise, the final output is given by

$$y_{output} = b_2 + \sum_{j=1}^{4} w_{2j}h_{j,1}, \tag{10}$$

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

$$y_{output} = b_2 + \phi(b_1 + xW_1)W_2, \tag{11}$$

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.

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. Excess returns are modeled using NN-1 as:

$$r_{it+1} = \mathcal{F}(z_{it}; \beta) + \eta_{i,t+1}, \ E(\eta_{i,t+1}) = 0 \forall i, t \tag{12}$$

where $\mathcal{F}$ is a flexible non-linear model that takes the parametric form in (11) with $\beta = \{W_1, W_2, b_1, b_2\}$ and $x = \{z_{it}\}$. Then the risk premiums are measured using $\mathcal{F}(z_{it}; \hat{\beta})$, where $\hat{\beta}$ are estimated parameters of β.

2.2.1. Parameter and Confidence Interval Estimation under the IID Assumption.

Under the iid assumption, the literature typically estimates the parameters by minimizing the following regularized MSE:

$$\hat{\beta}_\lambda = \arg\min_{\beta} \frac{1}{N_{Tr} N_S} \sum_{t \in Tr} \sum_{i \in S} \left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it} W_1) W_2)\right)^2 + \lambda \left[||W_1||_2^2 + ||W_2||_2^2 + ||b_1||_2^2 + ||b_2||_2^2\right], \tag{13}$$

where λ is the L_2 regularization hyperparamter.

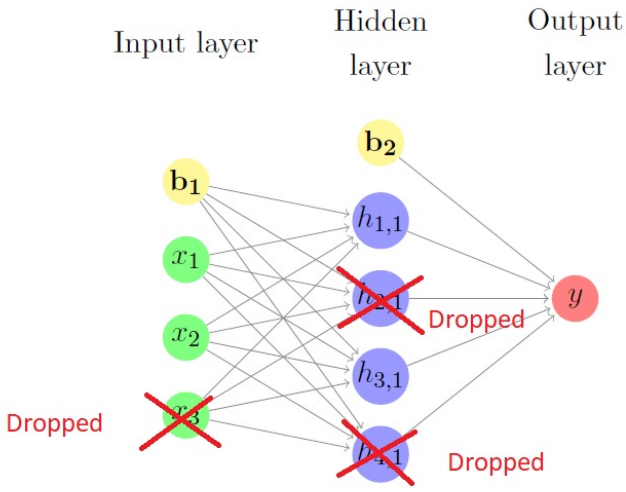
Because minimizing (13) is not possible in closed-forms, numerical algorithms start with an initial estimate (guess), and then iteratively update the parameters by feeding each observation into the training data one-by-one. Since this procedure could be computationally intensive, literature uses stochastic 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.[7]

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.[8] Dropout is useful because it not only boosts the performance of NN models but also delivers forecast variances simultaneously.

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, 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

[7] See GKX for a detailed review of parameter estimation using SGD and other regularizations such as L_1.
[8] 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.

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

$$\hat{\beta}_{\lambda,p} = \arg\min_{\beta} \frac{1}{N_{Tr}N_S} \sum_{t\in Tr} \sum_{i\in S} \left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it}(p_{1it}W_1))(p_{2it}W_2))\right)^2 + \lambda \left[||W_1||^2 + ||W_2||^2 + ||b_1||^2 + ||b_2||^2\right], \tag{14}$$

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

Stock-level risk premium forecasts. Given newly observed "test data" (Te) of raw pre-

dictors that do not overlap with the training and validation data sets, a dropout NN-1-based risk premium forecast is given by

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

where the parameters, $\{b_{2,\{\lambda,p\}}, b_{1,\{\lambda,p\}}, W_{1,\{\lambda,p\}}, W_{2,\{\lambda,p\}}\}$, are given in (14). $E^*_{it,Dropout}$ represents the dropout NN-1-based risk premium forecast of stock i at period t. Note that no neurons are dropped out while making predictions on the test data.

Portfolio-level risk premium forecasts. The risk premium forecast, $E^*_{Pt,Dropout}$, 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

$$\hat{E}_t(r^*_{P,t+1}) = E^*_{Pt,Dropout} = \sum_{i=1}^{S} \omega_{P,i,t} E^*_{it,Dropout},\ r^*_{i,t+1} \in Te, \tag{16}$$

where $r^*_{P,t+1} = \sum_{i=1}^S \omega_{P,i,t} r^*_{i,t+1}$, and $E^*_{it,Dropout}$ is given in (15).

Theorems in Appendix formally prove that the above estimators have a Bayesian interpretation. Based on these results, I now discuss how to instantly obtain (co)variances of dropout NN-based risk premium forecasts.

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

$$E^*_{it,Dropout} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}W_{1,\{\lambda,p\}})W_{2,\{\lambda,p\}}),\ r_{i,t+1}, z^*_{it} \in Te. \tag{17}$$

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

$$\widehat{Var}_t(E^*_{it,Dropout}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t+1} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t+1}\right)^2, \tag{18}$$

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

$$\hat{E}_{i,d,t} = (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1d}W_{1,\{\lambda,p\}}))(p_{2d}W_{2,\{\lambda,p\}})),\ z^*_{it} \in Te. \tag{19}$$

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.[9]

Intuition. To summarize, after estimating an NN-1 model's weights using the training and validation data sets, variances of risk premium forecasts on the test data are quickly available by measuring the sample variance of different forecasts that are obtained by deliberately assigning 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 that has the same number of neurons and hidden layers as the considered NN-1.

Stock-level risk premium forecast covariances. The predictive covariance between any two estimated stock risk premium forecasts $E^*_{it,Dropout}$ and $E^*_{jt,Dropout}$ is estimated by

$$\widehat{Covar}_t(E^*_{it,Dropout}, E^*_{jt,Dropout}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t+1} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t+1}\right)\left(\hat{E}_{j,d,t+1} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{j,d,t+1}\right), \tag{20}$$

where $\hat{E}_{i,d,t}$ and $\hat{E}_{i,d,t}$ are given in (19).

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

$$\widehat{Var}_t(E^*_{Pt,Dropout}) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{P,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{P,d,t}\right)^2, \tag{21}$$

where

$$\hat{E}_{P,d,t} = \sum_{i=1}^{S}\omega_{P,i,t}\left(b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1d}W_{1,\{\lambda,p\}}))(p_{2d}W_{2,\{\lambda,p\}})\right),\ z^*_{it} \in Te, \tag{22}$$

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

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 (18) and (21) generally applies to all forecasts from 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. Although the parametric forms of NNs could be mathematically formulated, forecasting with NNs involves imposing a number of other computational regularizations like early-stopping and batch normalization that are difficult to formulate. In that sense, forecasts from NNs remain black-box and their econometrics are difficult to comprehend. While addressing these concerns is beyond the scope of the paper, I take the first step in quantifying the estimation uncertainties related to NN-based forecasts and show how these uncertainties could be exploited to form improved trading strategies.

It is also worth emphasizing that (18) and (21) yield variances of risk premium forecasts and not excess return forecasts. 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 excess return forecasts could be easily estimated as well.

The main advantage of the dropout procedure is that the confidence intervals of risk premium forecasts could be obtained by training NNs only once without additional computational costs. While a bootstrap procedure could be employed to estimate the confidence intervals, it involves training numerous NNs, rendering it computationally infeasible.[10] Simulations in table J in Internet

[10] A shortcoming with the dropout is that it does not take into account the model uncertainty and assumes that

Appendix E.E1 assert that the dropout procedure delivers well-calibrated confidence intervals in small samples even when the residuals are allowed to be correlated in the cross-section and in the time-series.

2.2.2. Parameter and Confidence Interval Estimation under the non-IID Assumption.

When the excess returns are not iid and described as in (5) and (6), it follows that

$$r_{it+1} - \Lambda_i f_{t+1} = \mathcal{F}(z_{it}; \beta) + \epsilon_{i,t+1}, \ \epsilon_{i,t+1} \sim N(0, \sigma_\epsilon^2). \tag{23}$$

Thus, valid risk premium forecasts and their (co)variances could be estimated by minimizing the penalized squared residual sum of orthogonolized returns ($\bar{r}_{i,t+1} = r_{it+1} - \Lambda_i f_t$) rather than the excess returns. In particular, these measures are estimated by using equations (15)-(22), but by trading $\bar{r}_{i,t+1}$ for $r_{i,t+1}$. For computing $\bar{r}_{i,t+1}$, I estimate the unknown factor loadings Λ_i by first estimating excess returns with a NN using the iid assumption, as described in the previous subsection, and then regressing the model residuals $\{\eta_{i,t+1}\}$ on the market factor, as in (5) and (6).

This procedure delivers valid risk premium forecasts and confidence interval measures when Λ_i is assumed to be given. If it is unknown and must be estimated, valid measures could be obtained using Gibbs Sampling that draws conditional posteriors of i) Λ_i given the other NN parameters (β); and ii) β given Λ_i. Since this procedure requires fitting NNs over a large number of iterations, which is computationally challenging, I estimate risk premium forecasts and confidence intervals under the assumption that Λ_i is known. Empirical results suggest that the measures obtained using the iid assumption and the non-iid assumption with known Λ_i, respectively, deliver significantly enhanced trading strategies. It is expected that the measures estimated using the non-iid assumption with unknown Λ_i would provide even better results.

the model is *fixed* (Osband 2016). While incorporating model uncertainty is worth pursuing in the future, the biggest advantage of dropout is that it delivers confidence intervals without additional computational costs.

3. Improved Trading Strategies using Forecast Variances

This section shows how the previously estimated confidence intervals of risk premium forecasts could be exploited to form improved investment strategies.

3.1. Bias-variance decomposition

The squared forecast error of a risk premium forecast $\mathcal{F}(z_{it}, \beta)$ equals

$$E\left[(r_{i,t+1} - \mathcal{F}(z_{i,t}; \hat{\beta}))^2\right] = \underbrace{\Big(E(r_{i,t+1}) - E(\mathcal{F}(z_{i,t}; \hat{\beta}))\Big)^2}_{Bias^2} + \underbrace{E\left(\mathcal{F}(z_{i,t}; \hat{\beta}) - E(\mathcal{F}(z_{i,t}; \hat{\beta}))\right)^2}_{Variance} + V(\epsilon_{i,t+1}), \tag{24}$$

where $\mathcal{F}(z_{i,t}) = z_{it}^T \hat{\beta}$ for linear models; $\mathcal{F}(z_{i,t})$ takes the structural form in (11) for NN models. The first term in the right hand side of (24), popularly known as "squared-bias", measures the model misspecification of $\mathcal{F}(.)$ in estimating risk premiums. The second, known as "variance", quantifies parameter uncertainty. The ex-ante variances of risk premium forecasts 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.

When the residuals $\{\epsilon_{i,t+1}\}$ are iid and if the model is not missspecified, the ex-ante variances of risk premium forecast errors completely predict the cross-sectional (and time-series) variation in the squared forecast errors. The Confident-HL strategies exploit this predictability. By down weighting the stocks with high risk premium forecast variances, the Confident-HL strategies minimize the squared forecast errors. However, the ability of ex-ante variances predicting ex-post squared forecast errors diminishes for misspecified models because squared biases (rather than ex-ante variances) predominantly predict squared forecast errors. Thus, the Confident-HL strategies would deliver relatively more gains for models that are less biased. The following remark formalizes this intuition.

Remark 1: *The predictive ability of ex-ante variances in predicting squared forecast errors decreases with the model bias. Thus, it is expected that the improvements provided by trading strategies that utilize ex-ante confidence intervals decrease with the model bias.*

When the model residuals $\{\epsilon_{i,t}\}$ are not iid, the bias variance decomposition similar to (24) holds for the orthogonolized return forecasts (rather than excess return forecasts). Here, the ex-ante variances of the orthogonalized return forecasts predict their ex-post squared forecast errors. And this predictability diminishes with the model bias. Given that the non-iid risk premium forecasts discussed in the previous section are mathematically equivalent to ortghogonalized return forecasts, remark 1 holds under the non-iid assumption for the orthogonalized forecasts.

Recall from the introduction that the forecasts from NNs are relatively less biased than the penalized linear model forecasts, which further are relatively less biased than Lewellen forecasts . Thus, remark 1 implies that the predictive ability of ex-ante variances in predicting squared forecast errors is in the decreasing order of model biases, with the most significant predictability for NNs, followed by the penalized linear and then followed by the Lewellen model. Similarly, it is expected that the relative gains provided by the Confident-HL strategies is in the decreasing order of the model biases. The empirical section validates both these results.

3.2. Why Confident-HL Strategies improve expected returns OOS?

Recall that ex-ante variances of risk premium forecasts proxy for the ex-post squared forecast errors. Thus, the Confident-HL strategies that deliberately exclude (ex-ante) imprecise risk premium forecasts minimize the ex-post misclassification of stocks into inappropriate return-forecast deciles. As a consequence, they deliver superior OOS returns and Sharpe ratios.

Simulations in table K in Internet Appendix E.E2 validate this result. Since return forecasts have estimation error, the HL and 1%-HL strategies that rely solely on the return forecasts can incorrectly assign stocks into inappropriate return-forecast deciles. As a result, these strategies yield substantially lower returns than the maximum possible expected return of the high-low spread portfolio that is attainable when the stock returns are measured with infinite precision. However, the Confident-HL strategies selectively take long and short positions only on the subset of stocks in the extreme return-forecast deciles that have relatively more precise risk premium forecasts, and thus they earn superior returns OOS by minimizing the ex-post misclassification errors.

While the simulation results in table K assume that the return forecasts are uncorrelated, table L presents more comprehensive simulations validating the Confident-HL's superior performance for general cases with correlated return forecasts and trading strategies formed using various quantile-sorted portfolios (e.g., portfolios sorted on 30th percentile, portfolios sorted on 70th percentile, and decile-sorted portfolios, etc.). Table M further extends these simulations to a even more general framework that models stock returns using a NN-3. Across all the simulations, the Confident-HL strategies outperform the existing strategies that ignore ex-ante confidence intervals.

3.3. Economic interpretation of the Confident-HL strategies

3.3.1. Strategies of ambiguity-averse investors

The Confident-HL portfolios could be interpreted as the strategies of ambiguity-averse investors with a max-min utility.[11] For instance, let $\{\hat{\mu}_{l1}, \hat{\mu}_{l2}, \ldots \hat{\mu}_{lN}\}$ ($\{\hat{\mu}_{s1}, \hat{\mu}_{s2}, \ldots \hat{\mu}_{sN}\}$) be the set of N risk premium forecasts of stocks in the long (short) leg, with different forecast variances. Suppose that the risk premium forecasts 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. Now consider an ambiguity-averse investor who forms optimal long and short portfolios according to the following max-min expected return utility functions, respectively

$$\max_{w_{li}} \min_{\{\mu_{li}\}} \sum w_{li}\mu_{li}, \text{ subject to } (\hat{\mu}_{li} - k\sigma_{li}) \leq \mu_{li} \leq (\hat{\mu}_{li} + k\sigma_{li}), \forall i, \text{ and } \sum w_{li} = 1 \quad (25)$$

$$\max_{w_{si}} \min_{\{\mu_{si}\}} \left(-\sum w_{si}\mu_{si}\right), \text{ subject to } (\hat{\mu}_{si} + k\sigma_{si}) \leq \mu_{si} \leq (\hat{\mu}_{si} - k\sigma_{si}), \forall i, \text{ and } \sum w_{si} = 1, \quad (26)$$

where σ_{li} (σ_{si}) denotes the standard error of the risk premium forecast of i^{th} stock in the long (short) leg. The utility optimizations in (25) and (26) serve 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.

[11] See Garlappi et al. (2007) for an extensive discussion on the trading strategies of ambiguity-averse investors.

It is straighforward to note that the solutions to (25) and (26) 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 the trading strategies of ambiguity-averse investors with a max-min utility.

3.3.2. Regularized mean-variance strategies

Note that the Confident-HL strategies that drop stocks with imprecise risk premium forecasts improves the expected returns of HL strategies, not necessarily their variances, as dropping stocks may reduce the diversification benefit. So, Internet Appendix H also examines regularized mean variance trading strategies, in the spirit of Kozak, Nagel, and Santosh (2019), that optimally balances between expected HL returns and their (co)variances. However, estimating the entire covariance matrix is a high-dimensional problem which could lead to significant estimation uncertainty. Thus, the regularized mean-variance portfolios, considered in (26), may not deliver significant OOS improvements relative to the Confident-HL strategies.

In fact, Internet Appendix H shows that the Confident-HL strategies are interpretable as the regularized mean-variance strategies that impose adaptive Lasso penalties on the mean-variance weights, where the penalties are proportional to the risk premium forecast variances of stocks. In contrast, the regularized mean-variance strategies considered in (26) impose the standard Lasso penalty. Since the adaptive Lasso estimators typically outperform the standard Lasso estimators, it turns out that the Confident-HL portfolios outperform the regularized mean-variance portfolios (see Internet Appendix).

4. Empirical results

The empirical section validates three central predictions from section 3. First, the Confident-HL strategies outperform existing strategies that do not incorporate ex-ante confidence intervals OOS across all models. Second, the outperformance is due to the result that the ex-ante variances of risk premium forecasts predict ex-post squared forecast errors. Last, the magnitude of ex-ante

standard errors in predicting ex-post squared errors decreases with the model bias, and thus the relative gains provided by the Confident-HL strategies also decrease with the model bias.

4.1. Data, Definitions, and Replication Study

4.1.1. 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 2020 that are included in the CRSP database. The data include 31530 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.

4.1.2. Models and Estimation

Lasso and Neural Network. I examine Lasso and a feed-forward NN with three hidden layers (NN-3), with 32, 16, and 18 neurons per layer. These models were previously examined by GKX and Avramov et al. (2020), respectively. I precisely mimic their "recursive scheme" to estimate the model parameters and hyperparameters. The scheme first divides the data into 18 years of training (1957-1974), 12 years of validation (1975-1986), and 34 years (1987-2020) of OOS test samples. It then estimates the parameters and hyperparameters using objective functions to minimize the training sample's regularized penalized likelihoods and the validation sample's MSE, 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

[12] Besides these 176 predictors, GKX and Avramov et al. (2020) also consider (94×8) interactions between the stock characteristics and macroeconomic variables. Since NNs automatically capture such interactions, this paper excludes those additional variables.

year's data, maintaining the same size. I implement this recursive estimation framework to obtain risk premium forecasts, as well as their confidence intervals, over the OOS test sample.

My estimation procedure differs from GKX and Avramov et al. (2020) in two important aspects. Whereas GKX and Avramov et al. (2020) implicitly assume iid model residuals and estimate parameters by minimizing penalized MSEs, I estimate them by minimizing penalized likelihoods to allow for non-iid model residuals. However, the empirical results are quite robust to estimation with the iid specification as well, which are available in the Internet Appendix and a previous version of the paper. Second, GKX and Avramov et al. (2020) mainly apply L_1 regularization to estimate NN parameters, but I use dropout and L_2 because they enhance predictive performance and deliver confidence intervals simultaneously. I retain the other NN hyperparameters (e.g., SGD learning rate, Adam optimization) that GKX use.

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

4.1.3. Definitions of Performance Metrics

The following ex-ante and ex-post precision measures are used repeatedly in the paper.

Ex-ante Confidence (EC) of a risk premium forecast is computed as

$$EC_{it} = \frac{|\widehat{E_t(r_{i,t+1})}|}{se_t(\widehat{E_t(r_{i,t+1})})}, \tag{27}$$

where $\widehat{E_t(r_{i,t+1})}$ is the risk premium forecast of stock i at period t (for $t+1$) and $se_t(\widehat{E_t(r_{i,t+1})})$ is its ex-ante standard error, both of which are estimated in 2. I use EC as a proxy for the ex-ante precision because an estimate's standard error must always be understood relative to its mean. However, the main results still hold when inverse standard errors are instead as used as proxies for the precision. Table D in Internet Appendix (D) presents these results.

Ivol-based-Confidence (EC^{IVOL}). To assess the importance of EC, I also construct a benchmark precision measure based on stocks' past idiosyncratic volatility (rather than this paper's ex-ante standard errors) as

$$EC_{it}^{IVOL} = \frac{|\widehat{E_t(r_{i,t+1})}|}{IVOL_{it}}, \tag{28}$$

where $IVOL_{it}$ denotes the past IVOL measure of stock i at period t, which is the estimated residual standard error in the regression of stock i's daily returns on the value-weighted market index (Ali, Hwang, Trombley (2003)). Unlike EC, EC^{IVOL} is not a conditional measure because it does not depend on the predictor set. Thus, EC^{IVOL} measures will not predict ex-post squared forecast errors, and the strategies formed using them will not deliver gains OOS.

Ex-post OOS R^2 and MSE. The ex-post OOS R^2 and OOS MSE are given by

$$\begin{aligned} \text{OOS } R^2 &= 1 - \frac{\sum_{(i,t)\in\mathcal{S}}(r_{i,+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{S}} r_{i,t+1}^2}, \\ \text{OOS } MSE &= \frac{\sum_{(i,t)\in\mathcal{S}}(r_{i,+1} - \hat{r}_{i,t+1})^2}{\sum_{(i,t)\in\mathcal{S}} 1}. \end{aligned} \tag{29}$$

Now I describe the portfolio formation procedure.

4.1.4. Portfolio Construction

EW(VW) HL. Stocks are sorted into deciles every month based on their next month's return forecasts. If L and H represent the lowest and highest return forecast deciles, respectively, the EW(VW) HL strategy takes EW (VW) long and short positions on H and L, respectively.

EW(VW) Confident-HL. The extreme return-forecast-deciles, L and H, are further (double) sorted into deciles, $\{L_1, L_2, \ldots, L_{10}\}$ and $\{H_1, H_2, \ldots, H_{10}\}$ based on their EC levels. If L_{10} (L_1) and H_{10} (H_1) denote the subsets of stocks with relatively highest (lowest) EC values 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. In other words, the Confident-HL strategies take long-short positions exclusively on the subset of stocks in the extreme return-forecast-decile that have relatively more confident risk premiums.

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.

To fairly assess the Confident-HL portfolios' performance, I also construct two benchmark strategies: Low-Ivol-strategies and 1%-HL strategies.

EW (VW) Low-Ivol-strategies. This strategy mimics EW (VW) Confident-HL except for the fact that the double sort is based on EC^{IVOL} rather than EC. Thus, the EW (VW) Low-Ivol-HL strategy takes EW (VW) long-short positions exclusively on the subset of stocks in the extreme return-forecast-deciles that have relatively more EC^{IVOL}.

EW (VW) 1% HL strategies. These strategies take long (short) positions on the top (bottom) 1% of the stocks that have relatively higher (lower) risk premium forecasts. Thus, they contain the same number of stocks as the Confident-HL strategies but ignore confidence intervals.

4.1.5. Replication of Gu, Kelly, and Xiu (2020)

To ensure that this paper's risk premium measurements are comparable with GKX and Avramov et al. (2020), I replicate their studies. Figure A (B) in the Internet Appendix D presents the EW (VW) average OOS returns and Sharpe ratios of the decile portfolios that are sorted on return forecasts. All of these monotonically increase from decile-1 through decile-10, thereby confirming that the realized OOS returns align with their forecasts. And these results qualitatively and quantitatively match with GKX and Avramov et al. (2020), respectively.

Having outlined the data and showing that this paper's risk premium forecasts match those of the previous studies, I move on to test the theoretical predictions.

4.2. Main Results: Economic Gains from the Confident-HL strategies

Table I presents a wide range of performance metrics, including OOS average monthly returns, annualized Sharpe ratios, alphas and information ratios (with respect to Fama and French (2015) model added to the momentum factor), of competing trading strategies. Table II shows whether the pairwise differences in the OOS performance metrics between different trading strategies are

statistically significant using moving block bootstrap tests that are more conservative than Diebold and Mariano (2002), as they take into account ex-ante parameter uncertainty of risk premium forecasts. Thus, if the bootstrap tests imply that results are significant, the DM tests also imply significance. Internet Appendix C summarizes these bootstrap tests. And Figure 1 plots the log cumulative returns of the competing strategies across the models.

In terms of all the performance metrics and across all the models, the Confident-HL strategies beat the other benchmark strategies. For example, the NN-3-based VW Confident-HL strategy yields an average OOS monthly return of 2.70%, whereas VW HL, VW Low-Confident-HL, VW 1%-HL, and VW Low-Ivol-HL deliver 1.29%, 0.86%, 1.76%, and 0.16%, respectively. The respective monthly return differences between the the VW Confident-HL and the benchmarks, 1.41%, 1.84%, 0.94%, and 2.54%, are all statistically and economically significant. In addition, the NN-3-based VW Confident-HL strategy improves the Sharpe ratio by at least 40% relative to the benchmarks. For perspective, this Sharpe ratio improvement translates to 12% annualized holding period return difference between the NN-3-based VW Confident-HL and the VW 1%-HL strategies. Similarly, the Confident-HL strategies that are based on Lewellen and Lasso models also outperform respective benchmarks. The annualized holding period return difference between the VW Confident-HL and the VW HL strategies that are based on the Lasso (Lewellen) model is 10% (8%).

In contrast to the Confident-HL strategies, the Low-Confident-HL strategies underperform the other strategies across all models. For example, the NN-3-based (Lewellen-based, Lasso-based) VW Low-Confident-HL attains an annualized Sharpe ratio of 0.33 (0.19, -0.04), whereas the conventional VW HL yields substantially higher Sharpe ratio of 0.78 (0.34, 0.43). In terms of cumulative returns, the annualized holding period return difference between the above two strategies is -8% (-5%, -6%). These results are consistent with the insight that strategies based on imprecise risk premium forecasts misclassify stocks into inappropriate deciles and thus earn relatively lower returns and Sharpe ratios OOS.

Robustness on non-microcaps. To investigate the extent to which microcaps drive the outperformance of the Confident-HL strategies, I retrain linear and ML models on non-microcaps, excluding the stocks that fall below the 20th percentile of the NYSE size distribution. Tables III,

IV, and Figure 2 repeat the previous analysis on the non-microcap sample.

The Confident-HL strategies significantly outperform the benchmark strategies even on the sample that excludes microcaps. For instance, the average monthly return difference between the NN-3-based (Lewellen-based, Lasso-based) EW Confident-HL and the corresponding benchmark strategies is at least 0.86% (0.5%, 0.92%). Except for the return differences between the Lewellen-based Confident-HL and 1%-HL, all the other return differences are highly statistically significant. The Confident-HL strategy gains are relatively less pronounced under Lewellen because the value of incorporating ex-ante confidence intervals into trading strategies decreases with the model bias. The next section validates this intuition in more detail. The differences in the squared Sharpe ratios and the squared Information ratios between the Confident-HL and all the other strategies are highly significant across the models. In terms of the holding period returns, the Confident-HL strategies dominate the other strategies by a large margin across the three models.

Robustness to downside risks. Because ML-based strategies are known to display positive skewness and excess kurtosis (Avramov et al. (2020)), table V 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.[13] Across all the higher-order measures, the Confident-HL, strategies handily outperform the benchmark strategies. Thus, dropping imprecise risk premium forecasts from trading portfolios also mitigates the portfolios' downside risk.[14]

Robustness 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 V 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

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

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

Confident-HL strategies vs t-sorted strategies. Rather than the Confident-HL strategies that are based on conditional double sorting, one could alternatively form single sorted strategies that take long (short) positions on the stocks with the relatively highest (lowest) t-stats (i.e., ratios of risk premium forecasts and their standard errors). Such strategies need not deliver large OOS returns because stocks with precise risk premium forecasts need not necessarily have high expected returns. For example, consider a simple scenario where all stocks that have 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 (K) and (L), tackle this concern by first sorting on the return forecasts.[15]

4.3. Validating why Confident-HL strategies outperform.

Recall that the Confident-HL strategies outperform because the ex-ante variances of risk premium forecasts predict the ex-post squared errors. Figure 2 confirms this result. Under each model, I sort stocks into deciles every month based on their ex-ante variances. I then calculate the ex-post OOS MSEs attained by these decile subsamples over the OOS period. Figure 2 reveals that the ex-post OOS MSEs monotonically decrease with the level of ex-ante precision. The bottom decile (i.e., decile-1), containing stock return forecasts that are most imprecisely measured by NN-3 (Lewellen, Lasso), attains an OOS MSE of 8.42% (6.38%, 7.59%). In contrast, the top decile (i.e., decile-10) with the most precise risk premium forecasts delivers significantly lower OOS MSE of of

[15] 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.

2.26% (2.52%, 1.57%).

While the monotonic relationship between the ex-ante variances and ex-post square errors holds across all the three models, the steepness of the monotonicity increases with the model complexity, from Lewellen to NN-3. This result is consistent with remark-1 that the magnitude with which ex-ante variances predict ex-post squared forecast errors decreases with the model bias. Since NN-3 return forecasts capture non linear non-linear interactions amongst a high-dimensional set of predictors, they are relatively less biased than Lasso forecasts, which further are relatively less biased than Lewellen forecasts. Thus, the monotonicity is relatively steeper for NN-3 forecasts.

Consequently, the relative gains provided by the Confident-HL strategies are also expected to be in the decreasing order of the model biases. Figure 3 visually validates this result. The gap between the cumulative holding period return curves of the Confident-HL strategies and the Low-Confident-HL strategies is relatively wider for ML-based forecasts than for Lewellen-based forecasts. For perspective, the difference between the annualized cumulative holding period returns of the above two strategies is 16% for NN-3, 12% for Lasso, and 8% for Lewellen.

5. Dynamics of Ex-ante Precision

5.1. 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 3 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 VI 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.

5.2. Cross-sectional Variation in Ex-ante Confidence

Table VII 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 VII 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 VIII presents the results. It shows that 34% of the stocks with the most precise risk premium predictions have market caps greater than the median size across all individual stocks. Thus, NN-3-based Confident-HL portfolios yield impressive gains even on sub-samples containing large-sized stocks.

6. Conclusions

This paper estimates ex-ante confidence intervals of risk premium forecasts that are based on a wide range of linear and ML models, including Lewellen, Lasso, and NNs. Incorporating ex-ante confidence intervals, besides risk premium forecasts, into trading strategies is important, especially for ML models, as it significantly improves OOS returns and Sharpe ratios. The reason is that ex-ante variances of risk premium forecasts predict ex-post squared forecast errors, and thus they help to minimize ex-post misclassification of stocks into inappropriate deciles, resulting in significant gains in OOS returns and Sharpe ratios. Since the expected squared forecast errors equal the sum of ex-ante variances and squared biases, the magnitude with which ex-ante variances predict ex-post squared forecast errors increases (decreases) with the model complexity (bias). Thus, the relative portfolio gains obtained by incorporating ex-ante confidence intervals are also in the decreasing order of the model biases, with the most significant gains for NNs, followed by the penalized linear, and then followed by Lewellen.

Consistent with this intuition, across all return forecasting models, the Confident-HL strategies that discard stocks with imprecise risk premium forecasts significantly outperform existing benchmark strategies. The average annual return difference between the Confident-HL strategy and the counterfactual matching strategy that rely solely on return forecasts and ignore the forecast confidence intervals is 12.78% for NN-3, 10.70% for Lasso, and 8.68% for Lewellen, standardizing both strategies to have the same variances. The outperformance is robust to transaction costs, downside risks, excluding microcaps, and different model specification for the return residuals.

Economically, the Confident-HL strategies are interpretable as the trading strategies of ambiguity averse investors with a max-min utility. The ex-ante variances of risk premium forecasts exhibit significant variation across different return forecasting models, suggesting that a unified asset pricing model may not be sufficient to precisely forecast risk premiums of all stocks.

A. Appendix: Penalized Linear Models

This section presents the algorithms for obtaining the risk premium forecasts and their associated confidence intervals under the Lasso and the linear models.

Algorithm 1 : LASSO-based risk premium forecasts and their Confidence Intervals

Given $0 \leq \lambda_1 \leq 1, 0 \leq \lambda_2 \leq 1$, simulate iteratively

1. $\beta|\{\sigma_\epsilon, a_0, a_1, a_2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, Z, \bar{R}\} \sim N_p\left(\left(Z^TZ + D_\tau^{*-1}\right)^{-1} Z^T\bar{R}, \sigma_\epsilon^2\left(Z^TZ + D_\tau^{*-1}\right)^{-1}\right)$,
2. $1/\tau_j^2 = \gamma_j|\{\beta, \sigma_\epsilon^2, a_0, a_1, a_2, Z, \bar{R}\} \sim$ inverse Gaussian $\left(\sqrt{\frac{\lambda_1\sigma_\epsilon^2}{\beta_j^2}}, \lambda_1^2\right) I(\gamma_j) > 0$, for $j = 1, 2, \ldots, p$,
3. $\sigma_\epsilon^2|\{\beta, a_0, a_1, a_2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, Z, \bar{R}\} \sim$ inverted gamma $\left(\frac{n-1+p}{2}, \frac{1}{2}(\bar{R} - Z\beta)'(\bar{R} - Z\beta) + \frac{1}{2}\beta' D_\tau^{*-1}\beta\right)$
4. $a_0, a_1, a_2|\{\sigma_\epsilon, \beta, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, Z, \bar{R}\} \sim N_p\left(h_1\left(X_e^TX_e\right)^{-1} X_e^T\bar{R}, h_2\left(X_e^TX_e\right)^{-1}\right)$,

where n is the total number of observations in the training sample; D_τ^* is a diagonal matrix with diagonal elements $(\tau_i^{-2} + \lambda_2)^{-1}$; Z represents the $(n \times p)$ matrix of all characteristics z_{it}; $\bar{R}$ is the $(n \times 1)$ matrix of orthogonalized returns $\{r_{it} - \Lambda_i f_t\}$; X_e is the error matrix with columns $\{Size_{i,t-1} \times f_t, Mom_{i,t-1} \times f_t, Size_{i,t-1} \times f_t\}$; h_1 and h_2 are hyperparameters that are estimated using the validation sample. The algorithm is derived under the following prior for $\{a_0, a_1, a_2, \sigma_\epsilon^2\}$:

$$a_0, a_1, a_2|X_e, \sigma_\epsilon^2 \sim N\left(0, \nu\left(X_e^TX_e\right)^{-1}\right),\ P(\sigma_\epsilon^2) \propto \frac{1}{\sigma_\epsilon^2}. \tag{30}$$

When the model is a standard linear model with no penalty, the algorithm reduces to:

Algorithm 2 : Linear-based risk premium forecasts and their Confidence Intervals

Simulate iteratively

1. $\beta|\{\sigma_\epsilon, a_0, a_1, a_2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, Z, \bar{R}\} \sim N_p\left(\left(Z^TZ\right)^{-1} Z^T\bar{R}, \sigma_\epsilon^2\left(Z^TZ\right)^{-1}\right)$,
2. $\sigma_\epsilon^2|\{\beta, a_0, a_1, a_2, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, Z, \bar{R}\} \sim$ inverted gamma $\left(\frac{n-1+p}{2}, \frac{1}{2}(\bar{R} - Z\beta)'(\bar{R} - Z\beta)\right)$
3. $a_0, a_1, a_2|\{\sigma_\epsilon, \beta, \tau_1^2, \tau_2^2, \ldots, \tau_p^2, Z, \bar{R}\} \sim N_p\left(h_1\left(X_e^TX_e\right)^{-1} X_e^T\bar{R}, h_2\left(X_e^TX_e\right)^{-1}\right)$.

B. Appendix: Neural Networks

This section 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

$$r_{i,t+1} = b_2 + \phi(b_1 + z_{it}W_1)W_2 + \eta_{i,t+1}, \;\; E_t(\eta_{i,t+1}^2) = \sigma_\eta^2 \tag{31}$$

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.[16] Denote the risk premiums by μ_{it}, where

$$\mu_{i,t} = E_t(r_{it+1}) = b_2 + \phi(b_1 + z_{it}W_1)W_2. \tag{32}$$

Specify the unknown weight matrices with the standard Gaussian priors,

$$[W_1, W_2] = \mathcal{N}(0, l^{-2}I), \tag{33}$$

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

$$P(\mu_{i,t}^*|z_{it}^*, R, Z) = \int P(\mu_{i,t}^*|z_{it}^*, R, Z, W_1, W_2, b, \sigma_\eta^2)P(W_1, W_2|R, Z, b, \sigma_\eta^2)dW_1dW_2, \tag{34}$$

where $P(W_1, W_2|R, Z, b, \sigma_\eta^2)$ 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 (13), 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

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

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.[17] For reference in the finance literature, see Allena and Chordia (2022), 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:

$$q_{\{M_1,M_2\}}(W_1,W_2) = q_{M_1}(W_1)q_{M_2}(W_2), \text{ with } q_{M_1}(W_1) = \prod_{k=1}^{Q} q_1(w_{1q}),\ q_{M_2}(W_2) = \prod_{k=1}^{K} q_2(w_{2q}),$$
$$\text{where } q_i(w_{iq}) = p\mathcal{N}(m_{iq},\sigma^2 I_i) + (1-p)\mathcal{N}(0,\sigma^2 I_i) \text{ for } i = 1,2, \tag{35}$$

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

$$\{M_1^*, M_2^*\} = \arg\min_{\{M_1,M_2\}} KL\left(q_{M_1}(W_1)q_{M_2}(W_2)||P(W_1,W_2|R,Z_b,\sigma_\eta^2)\right), \tag{36}$$

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 (36) minimize the loss function that resembles a dropout NN's frequentist-based loss function (14).

$$\{M_1^*, M_2^*\} = \arg\min_{\{M_1,M_2\}} \frac{1}{N_{Tr}N_S}\sum_{t\in Tr}\sum_{i\in S}\left(r_{i,t+1} - (b_2 + \phi(b_1 + z_{it}(p_{1it}M_1))(p_{2it}M_2))\right)^2$$
$$+ \mu_1||M_1||^2 + \mu_2||M_2||^2 + \mu_3||b_1||^2 + \mu_4||b_2||^2, \tag{37}$$

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

Thus, for an appropriate choice of the prior's hyper-parameter l, the variational parameters,

[17] 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?

$\{M_1^*, M_2^*\}$, that best approximate the (Bayesian) NN weight matrices' posterior density are identical to the frequentist estimation of the dropout NN's weights. This implies

$$M_1^* = W_{1,\{\lambda,p\}}, \text{ and } M_2^* = W_{2,\{\lambda,p\}}. \tag{38}$$

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

$$P_{VI}(\mu_{i,t}^*|z_{it}^*, R, Z) = \int P(\mu_{i,t}^*|z_{it}^*, R, Z, W_1, W_2, b, \sigma_\eta^2) q_{M_1^*, M_2^*}(W_1, W_2) dW_1 dW_2, \tag{39}$$

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 (38), and $q_{M_1^*, M_2^*}(.)$ in (35), with optimal M_1^*, M_2^* substituted for M_1, M_2.

Theorem 2: *The dropout-NN-based frequentist risk premium predictions* (17) *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.,*

$$E_{it,Dropout}^* \xrightarrow{p} E_{VI}(\mu_{i,t}^*), \tag{40}$$

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

Theorem 3: *The dropout-based estimated variances of stock-level risk premiums* (18) *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.,*

$$\widehat{Var}_t(E_{it,Dropout}^*) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t} - \frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2 \xrightarrow{p} Var_{VI}(\mu_{i,t}^*), \tag{41}$$

where $\hat{E}_{i,d,t}$ *is given in* (19)*;* $Var_{VI}(\mu_{i,t}^*)$ *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

$$P_{VI}(\mu_{1,t}^*, \mu_{2,t}^*, \ldots, \mu_{S,t}^*|z_{it}^*, R, Z) = \int P(\mu_{1,t}^*, \mu_{2,t}^*, \ldots, \mu_{S,t}^*|z_{it}^*, R, Z, W_1, W_2, b, \sigma_\eta^2) q_{M_1^*, M_2^*}(W_1, W_2) dW_1 dW_2, \tag{42}$$

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 (38), and $q(.)$ in (35).

Theorem 4: *The dropout-based estimated covariances of stock-level risk premiums* (20) *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.,*

$$\widehat{Covar}_t(E^*_{it,Dropout}, E^*_{jt,Dropout}) \xrightarrow{p} Covar_{VI}(\mu^*_{i,t}, \mu^*_{j,t}),, \tag{43}$$

where $Covar_{VI}(\mu^*_{i,t}, \mu^*_{j,t})$ *denotes the covariance between* $\mu^*_{i,t}, \mu^*_{j,t}$ *based on the joint density* (42)*.*

Theorem 5: *The dropout-based estimated variances of portfolio-level risk premiums* (21) *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.,*

$$\widehat{Var}_t(E^*_{Pt,Dropout}) \xrightarrow{p} Var_{VI}(\mu^*_{Pt,Dropout}), \tag{44}$$

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* (42)*.*

B1. Frequentist consistency of dropout-based estimators

Note that theorems 2-5 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 6: *Under the assumptions 1-3 in Internet appendix D, 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*

$$\left|\left| P_{VI}(\mu^*_{1,t}, \mu^*_{2,t}, \ldots, \mu^*_{S,t} | z^*_{it}, R, Z) - MVN\left([\hat{\mu}_{1,t}, \ldots, \hat{\mu}_{S,t}], n^{-1} I(\mu_{1,t}, \ldots, \mu_{S,t})\right) \right|\right|_{TV} \xrightarrow{p} 0, \tag{45}$$

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 6 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.

Table I: Performance of Various Trading Strategies: All Stocks

This table reports the performance of different long-short portfolios that are constructed using monthly stock return forecasts and their estimated confidence intervals over the 34-year out-of-sample (OOS) period January 1987- February 2020. Panel A presents the equal-weighted strategies and Panel B shows the value-weighted strategies. Expected return forecasts and their respective confidence intervals are simultaneously estimated under each of the Lewellen, Lasso, and NN-3 models. HL strategies are the conventional spread portfolios that exploit information only from the return forecasts but not their confidence intervals. These strategies take long (short) positions on the top (bottom) decile of stocks with relatively highest (lowest) return forecasts. In contrast, Confident-HL strategies use information from both return forecasts and their confidence intervals by taking long (short) positions only on the subset of stocks in the decile of highest (lowest) risk premium forecasts that have relatively more confident return predictions. Similarly, Low-Confident-HL strategies take equal-weighted or value-weighted long (short) positions on the subset of stocks in the decile of highest (lowest) risk premium forecasts that are relatively imprecisely measured. Two other benchmark strategies are also considered. 1%-HL strategies take long (short) positions on the top (bottom) 1% of stocks that have relatively highest (lowest) return forecasts. These strategies contain the same number of stocks as Confident-HL strategies but use the information only from return forecasts, ignoring their confidence intervals. Low-Ivol-HL strategies take long (short) positions only on the subset of stocks in the decile of highest (lowest) risk premium forecasts that have relatively low idiosyncratic volatility (rather than this paper's variance measures of return forecasts). See section 4.4.1.4.1.4 for a detailed description of the portfolios. All portfolio returns are also adjusted for Fama-French 5-factors plus momentum (FF-5+UMD). 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.

		Panel A: Equal-weighted Strategies				Panel B: Value-weighted Strategies			
				FF-5+Mom				FF-5+Mom	
Model	Strategy	avg ret	Sh	α	IR	avg ret	Sh	α	IR
Lewellen	Confident-HL	2.40%	1.05	1.86%	1.26	1.75%	0.77	1.05%	0.65
	HL	1.21%	0.81	0.79%	0.76	0.39%	0.34	0.23%	0.26
	Low-Confident-HL	0.36%	0.19	0.49%	0.27	0.34%	0.19	0.36%	0.24
	1%-HL	1.91%	0.83	1.37%	0.69	0.25%	0.11	-0.23%	-0.12
	Low-ivol-HL	1.24%	0.83	1.07%	0.88	0.16%	0.12	0.14%	0.12
Lasso	Confident-HL	2.71%	1.28	2.54%	1.29	1.76%	0.77	1.44%	0.65
	HL	1.67%	1.12	1.24%	1.03	0.60%	0.43	0.16%	0.16
	Low-Confident-HL	0.49%	0.25	0.31%	0.17	-0.09%	-0.04	-0.43%	-0.21
	1%-HL	1.98%	0.94	1.54%	0.81	0.84%	0.37	0.53%	0.26
	Low-ivol-HL	1.03%	0.56	0.45%	0.31	0.42%	0.24	-0.18%	-0.13
NN-3	Confident-HL	3.84%	1.78	3.72%	1.81	2.70%	1.26	2.49%	1.26
	HL	2.21%	1.36	2.13%	1.36	1.29%	0.78	0.97%	0.64
	Low-Confident-HL	1.43%	0.68	1.40%	0.69	0.86%	0.33	0.63%	0.26
	1%-HL	3.07%	1.42	2.96%	1.41	1.76%	0.83	1.59%	0.79
	Low-ivol-HL	1.24%	0.83	1.07%	0.88	0.16%	0.12	0.14%	0.12

Table II: Statistical Comparison of Various Trading Strategies

This table conducts pairwise statistical comparisons of the out-of-sample (OOS) performance of various long-short portfolios that are constructed using monthly stock return forecasts and their estimated confidence intervals over the 34-year out-of-sample (OOS) period January 1987- February 2020. The tests are based on the moving block bootstrap procedure discussed in the Internet Appendix (C), with a block-length of 12. The Strategy column shows the comparing pair of portfolios. The $Sharpe^2$ columns show the annualized Sharpe ratio squared differences between the investment portfolios. The IR^2 columns show the annualized information ratio squared 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 I and section 4.4.1.4.1.4 for a description of the portfolios.

			Equal Weighted				Value-Weighted			
			Raw returns		FF-5+Mom		Raw returns		FF-5+Mom	
Model	Strategy		avg ret	Sh^2	α	IR^2	avg ret	Sh^2	α	IR^2
Lewellen	Confident-HL –	HL	1.2%*** (0.000)	0.45*** (0.001)	1.07%*** (0.001)	1*** (0.000)	1.37%*** (0.004)	0.47*** (0.004)	0.0.83%*** (0.0002)	0.35*** (0.003)
		Low-Confident-HL	2.05%*** (0.000)	1.06*** (0.002)	1.14%*** (0.007)	1.5*** (0.000)	1.38%*** (0.002)	0.55*** (0.003)	0.69%*** (0.04)	0.36*** (0.002)
		1%-HL	0.5% (0.27)	0.41 *** (0.001)	0.49% (0.175)	1.11*** (0.001)	1.5%*** (0.008)	0.58*** (0.000)	1.28*** (0.003)	0.41*** (0.002)
		Low-Ivol-HL	1.16%*** (0.002)	0.41*** (0.002)	0.8%** (0.003)	0.8*** (0.001)	1.59%*** (0.001)	0.57*** (0.001)	0.92%*** (0.000)	0.41*** (0.000)
Lasso	Confident-HL –	HL	1.04%*** (0.000)	0.40*** (0.000)	1.29%*** (0.000)	0.61*** (0.000)	1.17%*** (0.001)	0.41*** (0.001)	1.28%*** (0.000)	0.39*** (0.000)
		Low-Confident-HL	2.22%*** (0.000)	1.59*** (0.000)	2.22%*** (0.001)	1.63*** (0.000)	1.85%*** (0.000)	0.6*** (0.000)	1.87*** (0.000)	0.37*** (0.002)
		1%-HL	0.73%** (0,03)	0.77*** (0.000)	1.00%** (0.011)	1*** (0.000)	0.92%** (0.048)	0.46*** (0.001)	0.91%** (0.041)	0.35*** (0.005)
		Low-Ivol-HL	1.68%*** (0.000)	1.33*** (0.000)	2.09%*** (0.000)	1.57*** (0.000)	1.34%*** (0.004)	0.54*** (0.000)	1.62%*** (0.000)	0.4*** (0.000)
NN-3	Confident-HL –	HL	1.63%*** (0.000)	1.32*** (0.000)	1.59%*** (0.000)	1.4*** (0.000)	1.42%*** (0.001)	0.99*** (0.000)	1.52%*** (0.000)	1.17*** (0.000)
		Low-Confident-HL	2.41%*** (0.000)	2.7*** (0.000)	2.32%*** (0.000)	2.79*** (0.000)	1.84%*** (0.000)	1.49*** (0.000)	1.86%*** (0.002)	1.52*** (0.000)
		1%-HL	0.77%** (0.023)	1.14*** (0.000)	0.76%** (0.019)	1.26*** (0.000)	0.94%** (0.015)	0.92*** (0.000)	0.9%** (0.025)	0.96*** (0.000)
		Low-Ivol-HL	2.6%*** (0.000)	2.47*** (0.000)	2.65%*** (0.000)	2.48*** (0.000)	2.54%*** (0.000)	1.58*** (0.000)	2.36%*** (0.000)	1.57*** (0.000)

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

This table reports the performance of different long-short portfolios that are constructed using monthly stock return forecasts and their estimated confidence intervals over the 34-year out-of-sample (OOS) period January 1987- February 2020. Every period, the sample excludes microcap stocks with market capital smaller than the 20^{th} NYSE size percentile. Panel A presents the equal-weighted strategies and Panel B shows the value-weighted strategies. Expected return forecasts and their respective confidence intervals are simultaneously estimated under each of the Lewellen, Lasso, and NN-3 models. HL strategies are the conventional spread portfolios that exploit information only from the return forecasts but not their confidence intervals. These strategies take long (short) positions on the top (bottom) decile of stocks with relatively highest (lowest) return forecasts. In contrast, Confident-HL strategies use information from both return forecasts and their confidence intervals by taking long (short) positions only on the subset of stocks in the decile of highest (lowest) risk premium forecasts that have relatively more confident return predictions. Similarly, Low-Confident-HL strategies take equal-weighted or value-weighted long (short) positions on the subset of stocks in the decile of highest (lowest) risk premium forecasts that are relatively imprecisely measured. Two other benchmark strategies are also considered. 1%-HL strategies take long (short) positions on the top (bottom) 1% of stocks that have relatively highest (lowest) return forecasts. These strategies contain the same number of stocks as Confident-HL strategies but use the information only from return forecasts, ignoring their confidence intervals. Low-Ivol-HL strategies take long (short) positions only on the subset of stocks in the decile of highest (lowest) risk premium forecasts that have relatively low idiosyncratic volatility (rather than this paper's variance measures of return forecasts). See section 4.4.1.4.1.4 for a detailed description of the portfolios. All portfolio returns are also adjusted for Fama-French 5-factors plus momentum (FF-5+UMD). 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.

		Panel A: Equal-weighted Strategies				Panel B: Value-weighted Strategies			
				FF-5+Mom				FF-5+Mom	
Model	Strategy	avg ret	Sh	α	IR	avg ret	Sh	α	IR
Lewellen	Confident-HL	1.76%	0.82	1.21%	0.89	1.36%	0.61	0.83%	0.53
	HL	0.92%	0.62	0.40%	0.42	0.61%	0.42	0.33%	0.33
	Low-Confident-HL	-0.08%	-0.05	-0.15%	-0.10	0.70%	0.36	0.54%	0.32
	1%-HL	1.26%	0.58	0.54%	0.32	0.92%	0.41	0.45%	0.25
	Low-ivol-HL	0.77%	0.49	0.63%	0.51	0.52%	0.30	0.47%	0.33
Lasso	Confident-HL	1.72%	0.86	1.23%	0.80	1.48%	0.73	1.02%	0.63
	HL	0.79%	0.55	0.37%	0.45	0.77%	0.50	0.33%	0.39
	Low-Confident-HL	0.05%	0.03	-0.38%	-0.25	0.47%	0.26	-0.10%	-0.06
	1%-HL	0.41%	0.21	0.08%	0.06	0.81%	0.36	0.39%	0.22
	Low-ivol-HL	0.67%	0.33	0.16%	0.09	0.71%	0.36	0.27%	0.17
NN-3	Confident-HL	2.19%	1.30	1.88%	1.16	2.00%	1.05	1.56%	0.86
	HL	1.33%	0.82	0.82%	0.55	1.11%	0.70	0.62%	0.42
	Low-Confident-HL	0.99%	0.52	0.62%	0.34	0.87%	0.40	0.55%	0.27
	1%-HL	1.15%	0.68	0.78%	0.50	0.91%	0.48	0.54%	0.30
	Low-ivol-HL	1.29%	0.80	0.86%	0.57	1.22%	0.71	0.80%	0.50

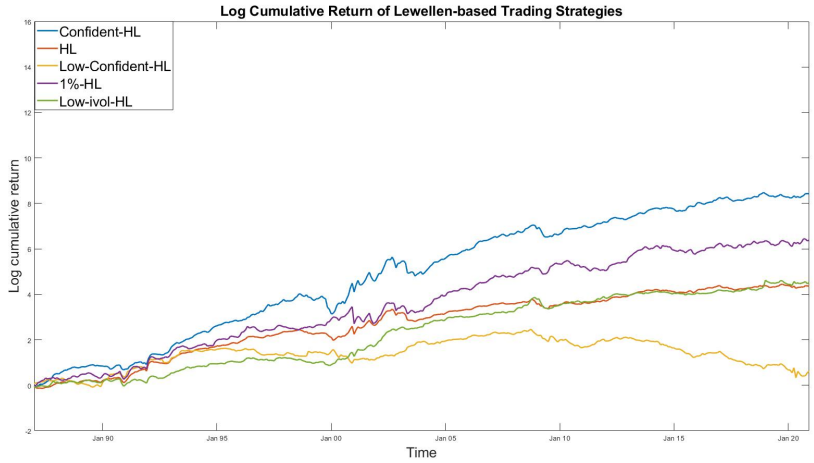
Table IV: Statistical Comparison of Various Trading Strategies: Non-Microcaps

This table conducts pairwise statistical comparisons of the out-of-sample (OOS) performance of various long-short portfolios that are constructed using monthly stock return forecasts and their estimated confidence intervals over the 34-year out-of-sample (OOS) period January 1987- February 2020. The tests are based on the moving block bootstrap procedure discussed in the Internet Appendix (C), with a block-length of 12. The Strategy column shows the comparing pair of portfolios. The $Sharpe^2$ columns show the annualized Sharpe ratio squared differences between the investment portfolios. The IR^2 columns show the annualized information ratio squared 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 I and section 4.4.1.4.1.4 for a description of the portfolios.

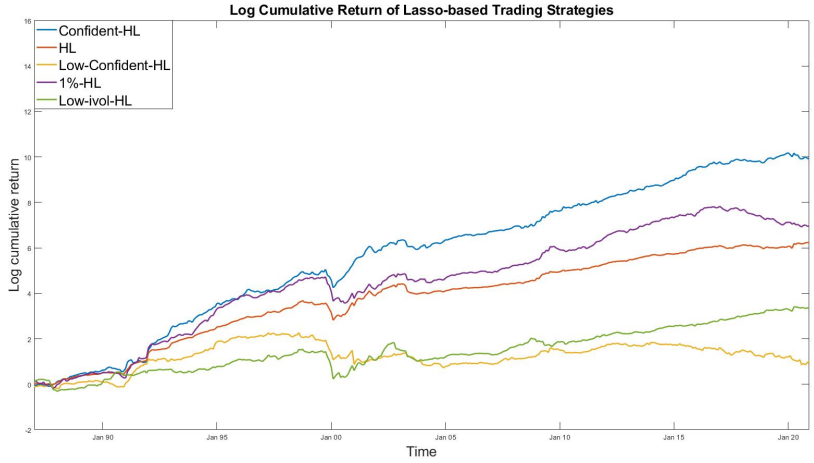
			Equal Weighted				Value-Weighted			
			Raw returns		FF-5+Mom		Raw returns		FF-5+Mom	
Model	Strategy		avg ret	Sh^2	α	IR^2	avg ret	Sh^2	α	IR^2
Lewellen	Confident-HL −	HL	0.84%*** (0.000)	0.28*** (0.006)	0.81%*** (0.000)	0.61*** (0.000)	0.75%** (0.031)	0.19** (0.018)	0.5%** (0.046)	0.17** (0.038)
		Low-Confident-HL	1.84%*** (0.000)	0.67*** (0.000)	1.37%*** (0.000)	0.78*** (0.000)	0.66% (0.215)	0.24** (0.022)	0.29% (0.34)	0.18** (0.031)
		1%-HL	0.5% (0.127)	0.31*** (0.003)	0.67%** (0.048)	0.69*** (0.000)	0.44% (0.14)	0.2** (0.015)	0.39% (0.21)	0.22** (0.028)
		Low-Ivol-HL	1.00%*** (0.000)	0.43*** (0.000)	0.58%** (0.013)	0.53*** (0.000)	0.84%** (0.032)	0.27*** (0.01)	0.36% (0.21)	0.17** (0.033)
Lasso	Confident-HL −	HL	0.92%*** (0.000)	0.36*** (0.000)	0.85%*** (0.000)	0.43*** (0.000)	0.71%*** (0.002)	0.29*** (0.001)	0.69%*** (0.003)	0.25*** (0.001)
		Low-Confident-HL	1.66%*** (0.000)	1.59*** (0.000)	1.61%*** (0.000)	0.57*** (0.000)	1.01%*** (0.001)	0.47*** (0.000)	1.12%** (0.000)	0.4*** (0.000)
		1%-HL	1.3%*** (0.000)	0.77*** (0.000)	1.14%*** (0.000)	0.63*** (0.000)	0.76%** (0.015)	0.41*** (0.001)	0.68%** (0.02)	0.35*** (0.000)
		Low-Ivol-HL	1.05%*** (0.000)	1.33*** (0.000)	1.07%*** (0.000)	0.63*** (0.000)	0.77%** (0.016)	0.41*** (0.000)	0.75%** (0.018)	0.37*** (0.001)
NN-3	Confident-HL −	HL	0.86%*** (0.000)	1.01*** (0.000)	1.06%*** (0.000)	1.04*** (0.000)	0.88%*** (0.000)	0.61*** (0.000)	0.94%*** (0.000)	0.56*** (0.000)
		Low-Confident-HL	1.2%*** (0.000)	1.41*** (0.000)	1.26%*** (0.000)	1.23*** (0.000)	1.12%*** (0.002)	0.94*** (0.000)	1.00%*** (0.003)	0.66*** (0.000)
		1%-HL	1.04%*** (0.004)	0.75*** (0.000)	1.1%*** (0.001)	1.1*** (0.000)	1.09%*** (0.008)	0.87*** (0.000)	1.02%** (0.004)	0.65*** (0.000)
		Low-Ivol-HL	1.42%*** (0)	1.04*** (0.000)	1.01%*** (0.000)	1.03*** (0.000)	1.48%*** (0.000)	0.6*** (0.000)	0.75%*** (0.005)	0.49*** (0.000)

Figure 1. Log Cumulative OOS returns of NN-3-based trading strategies

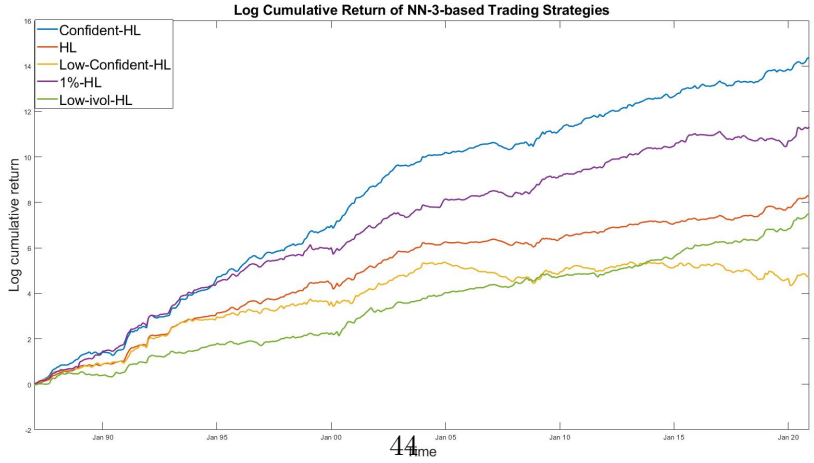
(a) Lewellen-based strategies



(b) Lasso-based strategies

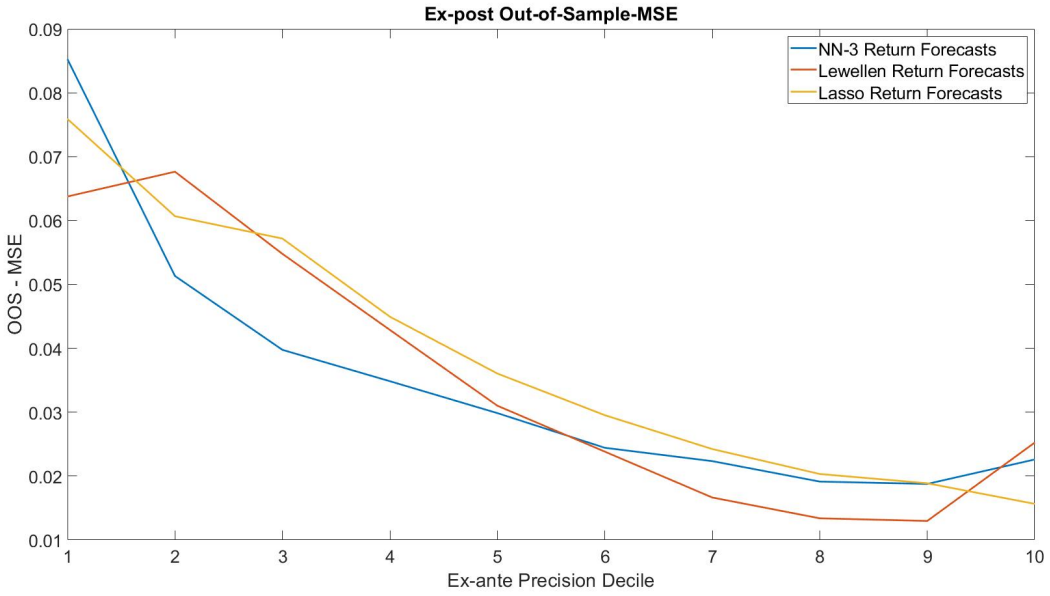


(c) NN-3-based strategies



Note: This figure presents the log cumulative OOS returns of various trading strategies.

Figure 2. Ex-ante Confidence and Ex-post OOS-MSE

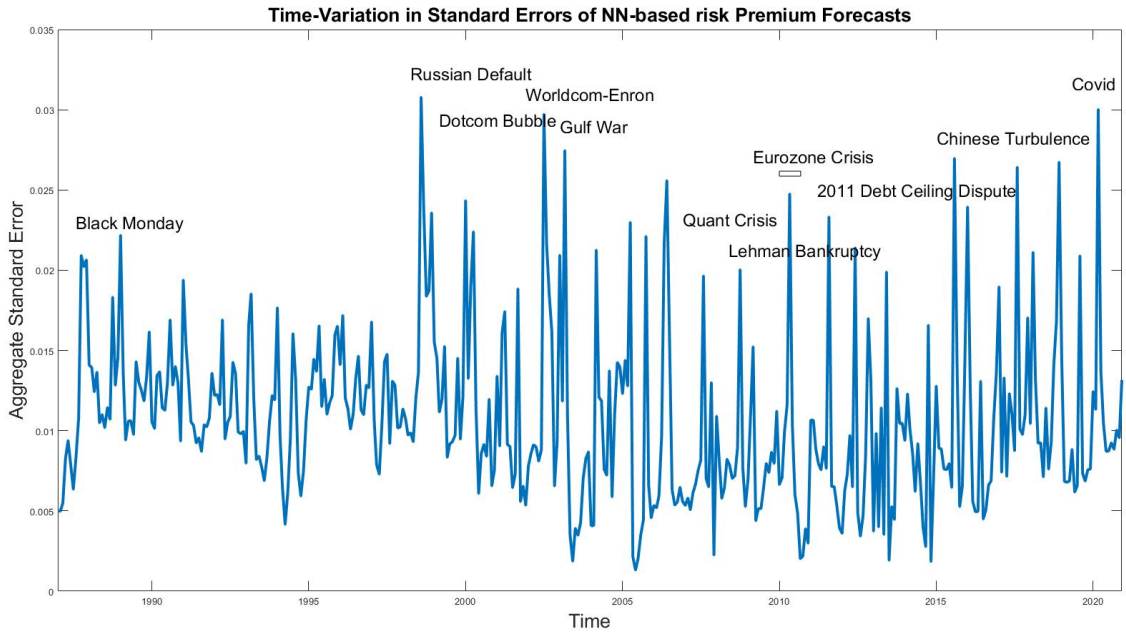


Note: This figure confirms that the ex-ante precision of risk premium forecasts and their realized out-of-sample precision are monotonically related across all the three return forecasting models. Every period stocks are first sorted into deciles based on their risk premium forecasts. Each of these ten predicted-return deciles are further sorted into deciles based on the confidence levels of their risk premium forecasts. Ex-ante Precision Decile 10 (1) comprises the top (bottom) 10% stocks with the lowest (highest) confidence levels in forecasting risk premiums, combined across all the ten predicted-return deciles. The y-axis represents the ex-post OOS MSEs attained by the ex-ante precision deciles.

Table V: Transaction Costs and Higher-Moment Adjusted Performance of Various Strategies
This table reports the transaction costs and higher-moment-risk-adjusted performance of different long-short portfolios that are constructed using monthly stock return forecasts and their estimated confidence intervals over the 34-year out-of-sample (OOS) period January 1987- February 2020. 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 Drawdown, Omega, and Sortino columns respectively represent the maximum drawdown, Omega, and Sortino ratios. These ratios measure the higher-moment-risk-adjusted performance of portfolios, explicitly penalizing losses more than realizing gains. See table I and section 4.4.1.4.1.4 for a description of the portfolios.

		Equal Weighted				Value Weighted			
Model	Strategy	Turnover	Drawdown	Omega	Sortino	Turnover	Drawdown	Omega	Sortino
	Confident-HL	0.85	0.88	2.43	0.52	0.90	0.88	1.88	0.41
	HL	0.51	0.55	2.03	0.39	0.46	0.79	1.30	0.15
Lewellen	Low-Confident-HL	0.85	2.13	1.18	0.09	0.91	0.95	1.17	0.09
	1%-HL	0.71	0.76	1.98	0.40	0.75	3.26	1.09	0.05
	Low-Ivol-HL	0.6	0.51	2.03	0.45	0.76	1.19	1.10	0.05
	Confident-HL	1.3	0.80	3.06	0.78	1.19	0.69	1.91	0.45
	HL	0.79	0.87	2.97	0.61	0.68	0.87	1.46	0.20
Lasso	Low-Confident-HL	1.16	1.55	1.24	0.11	1.40	3.27	0.96	-0.01
	1%-HL	0.84	1.16	2.21	0.48	0.92	1.79	1.34	0.16
	Low-Ivol-HL	0.76	1.31	1.65	0.23	0.78	0.98	1.23	0.10
	Confident-HL	1.78	0.32	4.53	1.47	1.86	0.50	2.79	0.83
	HL	1.19	0.37	3.34	0.98	1.35	0.55	1.92	0.42
NN-3	Low-Confident-HL	1.81	1.05	1.73	0.36	1.83	1.42	1.32	0.17
	1%-HL	1.78	0.67	3.42	1.04	1.86	0.67	1.94	0.45
	Low-Ivol-HL	1.86	0.51	2.03	0.45	1.73	1.19	1.10	0.05

Figure 3. Time-Series Variation in Standard Errors of NN-based Risk Premium Forecasts



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 VI: 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
Covid shock in the US	3.01%	March 2029 to March 2020
Crisis Period Average	2.37%	
Non-Crisis Period Average	1.07%	

Table VII: 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 VIII: 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

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, 2022, True Liquidity and Equilibrium Prices: US Tick Pilot, *Working Paper, SSRN* .

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.

Bai, Jushan, 2009, Panel Data Models With Interactive Fixed Effects, *Econometrica* 77, 1229–1279.

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.

Chen, Luyang, Markus Pelger, and Jason Zhu, 2020, Deep Learning in Asset Pricing, SSRN Scholarly Paper ID 3350138, 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.

Fan, Jianqing, and Runze Li, 2001, Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association* 96, 1348–1360.

Fan, Jianqing, and Heng Peng, 2004, Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics* 32, 928–961.

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.

Ferson, Wayne E., and Campbell R. Harvey, 1999, Conditioning Variables and the Cross Section of Stock Returns, *The Journal of Finance* 54, 1325–1360, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/0022-1082.00148.

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.

Kelly, Bryan T., Seth Pruitt, and Yinan Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* .

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.

Lee, Wonyul, and Yufeng Liu, 2012, Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood, *Journal of Multivariate Analysis* 111, 241–255.

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.

Lu, Xun, and Liangjun Su, 2016, Shrinkage estimation of dynamic panel data models with interactive fixed effects, *Journal of Econometrics* 190, 148–175.

Pesaran, M. Hashem, 2006, Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure, *Econometrica* 74, 967–1012.

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.

Smith, Simon C, and Allan Timmermann, 2021, Break Risk, *The Review of Financial Studies* 34, 2045–2100.

Smith, Simon C., and Allan Timmermann, 2022, Have risk premia vanished?, *Journal of Financial Economics* 145, 553–576.

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.

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.

Zou, Hui, 2006, The Adaptive Lasso and Its Oracle Properties, *Journal of the American Statistical Association* 101, 1418–1429, Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214506000000735.

C. 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)).

C1. 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

$$\sqrt{T}\frac{\sum_{t=1}^{T}\Delta_t/T}{se(\Delta_t)} \sim N(0,1), \tag{46}$$

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.

C2. 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

$$Sh_i^2 = \frac{(\frac{1}{T}\sum_{t=1}^{T} HL_{it})^2}{\frac{1}{T}\sum_{t=1}^{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^{T} HL_{it})^2}, \text{ for } i = 1, 2. \tag{47}$$

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,

$$HL_{it}^* = \sqrt{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^{T} HL_{it}) / \sqrt{\sum_{t=1}^{T}(HL_{it} - \frac{1}{T}\sum_{t=1}^{T} HL_{it})^2}, \tag{48}$$

   (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 $Sh_1^2 - Sh_2^2$.

D. Internet Appendix: Robustness Checks

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

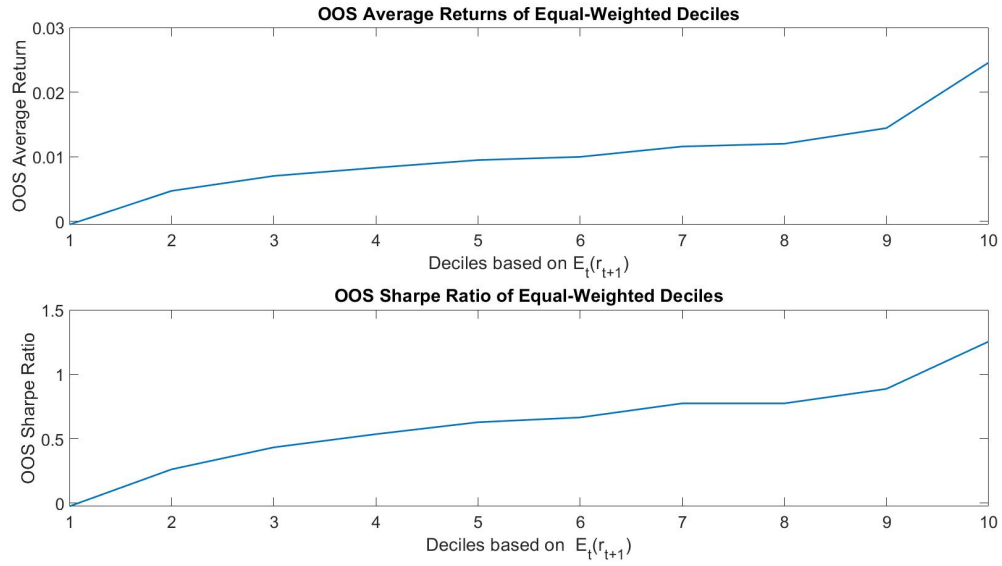
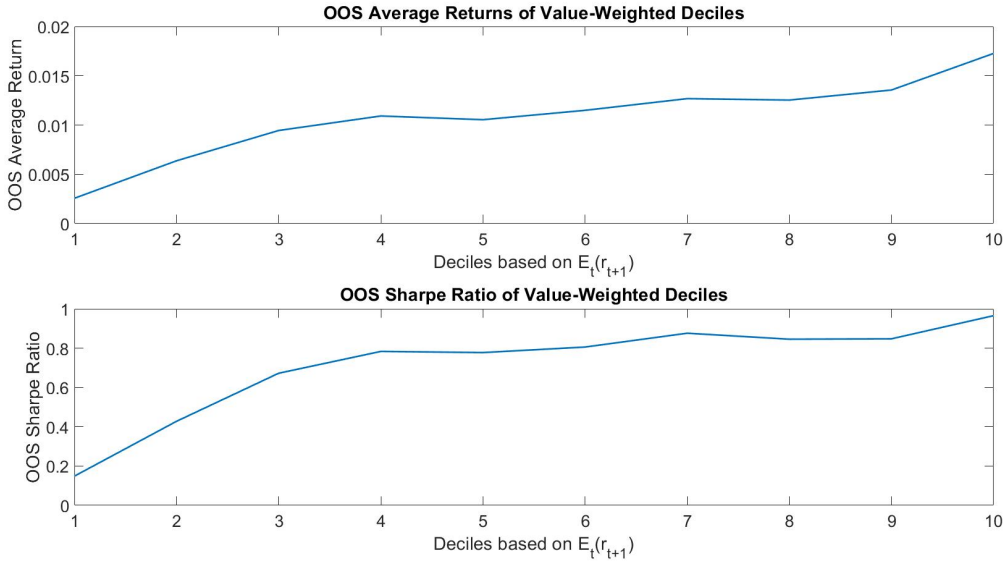


Figure B. Out-of-Sample (OOS) Performance of Value-weighted Deciles Based on NN-3 Predictions.



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 A: Performance of Various Trading Strategies under the IID Assumption: All Stocks
This table reports the performance of different long-short portfolios that are constructed using monthly stock return forecasts and their estimated confidence intervals over the 34-year out-of-sample (OOS) period January 1987- February 2020. Panel A presents the equal-weighted strategies and Panel B shows the value-weighted strategies. Expected return forecasts and their respective confidence intervals are simultaneously estimated under each of the Lewellen, Lasso, and NN-3 models. The confidence intervals are estimated under the assumption that the model residuals are iid. HL strategies are the conventional spread portfolios that exploit information only from the return forecasts but not their confidence intervals. These strategies take long (short) positions on the top (bottom) decile of stocks with relatively highest (lowest) return forecasts. In contrast, Confident-HL strategies use information from both return forecasts and their confidence intervals by taking long (short) positions only on the subset of stocks in the decile of highest (lowest) risk premium forecasts that have relatively more confident return predictions. Similarly, Low-Confident-HL strategies take equal-weighted or value-weighted long (short) positions on the subset of stocks in the decile of highest (lowest) risk premium forecasts that are relatively imprecisely measured. Two other benchmark strategies are also considered. 1%-HL strategies take long (short) positions on the top (bottom) 1% of stocks that have relatively highest (lowest) return forecasts. These strategies contain the same number of stocks as Confident-HL strategies but use the information only from return forecasts, ignoring their confidence intervals. Low-Ivol-HL strategies take long (short) positions only on the subset of stocks in the decile of highest (lowest) risk premium forecasts that have relatively low idiosyncratic volatility (rather than this paper's variance measures of return forecasts). See section 4.4.1.4.1.4 for a detailed description of the portfolios. All portfolio returns are also adjusted for Fama-French 5-factors plus momentum (FF-5+UMD). 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.

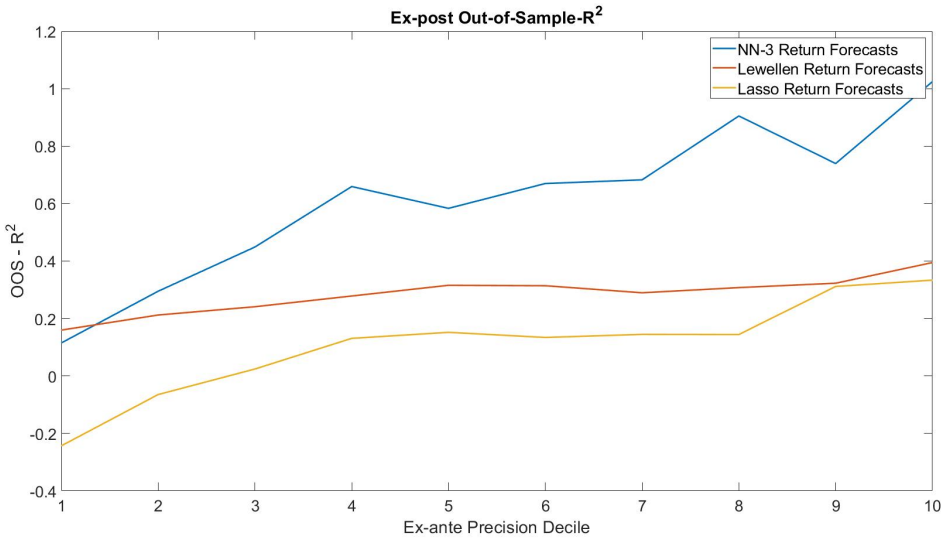
		Panel A: Equal-weighted Strategies				Panel B: Value-weighted Strategies			
				FF-5+Mom				FF-5+Mom	
Model	Strategy	avg ret	Sh	α	IR	avg ret	Sh	α	IR
Lewellen	Confident-HL	2.27%	0.91	1.45%	0.90	1.87%	0.81	1.16%	0.70
	HL	1.12%	0.69	0.54%	0.55	0.44%	0.47	0.30%	0.38
	Low-Confident-HL	0.37%	0.22	0.41%	0.25	0.14%	0.09	0.15%	0.10
	1%-HL	1.46%	0.58	0.78%	0.42	0.24%	0.11	-0.04%	-0.02
	Low-Ivol-HL	0.79%	0.55	0.56%	0.53	0.14%	0.12	0.13%	0.13
Lasso	Confident-HL	2.16%	1.11	1.82%	1.01	1.28%	0.59	0.58%	0.29
	HL	1.01%	0.70	0.51%	0.41	0.58%	0.40	-0.09%	-0.08
	Low-Confident-HL	0.67%	0.33	0.18%	0.10	0.02%	0.01	-0.54%	-0.31
	1%-HL	1.79%	0.92	1.37%	0.80	0.54%	0.25	-0.21%	-0.11
	Low-Ivol-HL	0.75%	0.64	0.65%	0.59	0.62%	0.44	0.40%	0.30
NN-3	Confident-HL	3.42%	1.68	3.22%	1.74	2.22%	1.11	1.95%	1.03
	HL	2.34%	1.42	2.13%	1.46	1.42%	0.88	0.99%	0.70
	Low-Confident-HL	2.07%	1.04	1.80%	1.00	1.08%	0.45	0.37%	0.17
	1%-HL	2.90%	1.43	2.62%	1.45	1.68%	0.84	1.20%	0.67
	Low-Ivol-HL	0.79%	0.55	0.56%	0.53	0.14%	0.12	0.13%	0.13

Table B: Performance of Various Trading Strategies under the IID Assumption: Non-microcaps

This table reports the performance of different long-short portfolios that are constructed using monthly stock return forecasts and their estimated confidence intervals over the 34-year out-of-sample (OOS) period January 1987-February 2020. Every period, the sample excludes microcap stocks with market capitall smaller than the 20^{th} NYSE size percentile. Panel A presents the equal-weighted strategies and Panel B shows the value-weighted strategies. Expected return forecasts and their respective confidence intervals are simultaneously estimated under each of the Lewellen, Lasso, and NN-3 models. The confidence intervals are estimated under the assumption that the model residuals are iid. HL strategies are the conventional spread portfolios that exploit information only from the return forecasts but not their confidence intervals. These strategies take long (short) positions on the top (bottom) decile of stocks with relatively highest (lowest) return forecasts. In contrast, Confident-HL strategies use information from both return forecasts and their confidence intervals by taking long (short) positions only on the subset of stocks in the decile of highest (lowest) risk premium forecasts that have relatively more confident return predictions. Similarly, Low-Confident-HL strategies take equal-weighted or value-weighted long (short) positions on the subset of stocks in the decile of highest (lowest) risk premium forecasts that are relatively imprecisely measured. Two other benchmark strategies are also considered. 1%-HL strategies take long (short) positions on the top (bottom) 1% of stocks that have relatively highest (lowest) return forecasts. These strategies contain the same number of stocks as Confident-HL strategies but use the information only from return forecasts, ignoring their confidence intervals. Low-Ivol-HL strategies take long (short) positions only on the subset of stocks in the decile of highest (lowest) risk premium forecasts that have relatively low idiosyncratic volatility (rather than this paper's variance measures of return forecasts). See section 4.4.1.4.1.4 for a detailed description of the portfolios. All portfolio returns are also adjusted for Fama-French 5-factors plus momentum (FF-5+UMD). 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.

		Panel A: Equal-weighted Strategies				Panel B: Value-weighted Strategies			
				FF-5+Mom				FF-5+Mom	
Model	Strategy	avg ret	Sh	α	IR	avg ret	Sh	α	IR
	Confident-HL	1.40%	0.58	0.60%	0.43	1.28%	0.56	0.50%	0.32
	HL	0.95%	0.71	0.34%	0.47	0.53%	0.63	0.25%	0.40
Lewellen	Low-Confident-HL	0.61%	0.40	0.44%	0.33	0.07%	0.04	0.12%	0.11
	1%-HL	1.04%	0.43	0.36%	0.21	0.49%	0.21	-0.17%	-0.08
	Low-Ivol-HL	0.48%	0.35	0.44%	0.37	0.13%	0.12	0.12%	0.12
	Confident-HL	1.25%	0.75	0.92%	0.63	1.43%	0.70	1.17%	0.63
	HL	0.66%	0.55	0.49%	0.45	0.68%	0.54	0.51%	0.46
Lasso	Low-Confident-HL	0.30%	0.17	0.38%	0.23	0.17%	0.09	0.24%	0.13
	1%-HL	0.96%	0.57	0.63%	0.42	0.66%	0.32	0.41%	0.21
	Low-Ivol-HL	0.64%	0.59	0.66%	0.44	0.75%	0.59	0.68%	0.47
	Confident-HL	2.12%	1.16	1.90%	1.12	1.98%	0.97	1.78%	0.92
	HL	1.55%	0.95	1.27%	0.87	1.33%	0.81	1.02%	0.68
NN-3	Low-Confident-HL	1.43%	0.71	1.06%	0.57	1.24%	0.58	0.89%	0.45
	1%-HL	1.73%	0.95	1.50%	0.89	1.71%	0.83	1.49%	0.77
	Low-Ivol-HL	1.12%	0.92	1.05%	0.88	0.89%	0.62	0.89%	0.67

Figure C. Ex-ante Confidence and Ex-post OOS-R^2



Note: Figure 2 (C) 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 C: Transaction Costs and Higher-Moment Adjusted Performance of Various Strategies: Non-microcaps

This table reports the transaction costs and higher-moment-risk-adjusted performance of different long-short portfolios that are constructed using monthly stock return forecasts and their estimated confidence intervals over the 34-year out-of-sample (OOS) period January 1987- February 2020. Every period, the sample excludes microcap stocks with market capital smaller than the 20^{th} NYSE size percentile. 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 Drawdown, Omega, and Sortino columns respectively represent the maximum drawdown, Omega, and Sortino ratios. These ratios measure the higher-moment-risk-adjusted performance of portfolios, explicitly penalizing losses more than realizing gains. See table I and section 4.4.1.4.1.4 for a description of the portfolios.

		Equal Weighted				Value Weighted			
		Equal Weighted				Value Weighted			
Model	Strategy	Turnover	Drawdown	Omega	Sortino	Turnover	Drawdown	Omega	Sortino
	Confident-HL	0.92	0.74	2.00	0.39	1.01	0.74	2.00	0.39
	HL	0.52	1.12	1.79	0.26	0.50	1.12	1.79	0.26
Lewellen	Low-Confident-HL	0.93	1.63	0.96	-0.02	0.99	1.63	0.96	-0.02
	1%-HL	0.72	2.28	1.66	0.26	0.78	2.28	1.66	0.26
	Low-Ivol-HL	0.64	0.92	1.53	0.22	0.73	0.92	1.53	0.22
	Confident-HL	1.21	1.17	2.02	0.39	1.08	0.76	1.79	0.37
	HL	0.67	0.76	1.66	0.24	0.50	0.84	1.58	0.23
Lasso	Low-Confident-HL	1.09	1.53	1.03	0.01	1.24	0.82	1.24	0.11
	1%-HL	0.69	1.12	1.19	0.08	0.72	0.99	1.35	0.15
	Low-Ivol-HL	0.64	1.14	1.32	0.14	0.64	1.19	1.36	0.16
	Confident-HL	1.76	0.31	2.87	0.83	1.84	0.43	2.36	0.59
	HL	1.09	0.89	2.05	0.39	1.22	0.62	1.82	0.34
NN-3	Low-Confident-HL	1.78	1.14	1.54	0.24	1.82	1.04	1.40	0.17
	1%-HL	1.40	1.09	1.75	0.30	1.51	1.34	1.49	0.21
	Low-Ivol-HL	1.55	0.92	1.53	0.22	1.62	1.12	1.31	0.13

Table D: 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 I and section 4.4.1.4.1.4 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 E: 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 4.4.1.4.1.4 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 F: 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 4.4.1.4.1.4 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 G: 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 H: 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 I: 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%	3.66	0.66
Mean-variance	1.82%	6.59	1.2

E. Simulation results

E1. Validating standard errors using simulations

Using simulations, this section (table J) 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.

5.1.1. 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.[18] I simulate a 3-factor model for excess returns, for $t = 1, 2, \ldots, T$:

$$r_{i,t+1} = g(z_{i,t}) + e_{i,t+1},\ e_{i,t+1} = \beta_{i,t} v_{t+1} + \epsilon_{i,t+1},\ z_{i,t} = (1, x_t)' \otimes c_{i,t},\ \beta_{i,t} = (c_{i1,t}, c_{i2,t}, c_{i3,t}), \quad (49)$$

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

$$c_{ij,t} = \frac{2}{N+1} CSrank(\bar{c}_{ij,t}) - 1,\ \bar{c}_{ij,t} = \rho_j \bar{c}_{ij,t-1} + \epsilon_{ij,t},\ \text{for } 1 \le i \le 200,\ 1 \le j \le 180, \quad (50)$$

where $CSrank$ denotes the cross-sectional rank.

And the time-series x_t is given by

$$x_t = \rho x_{t-1} + u_t, \quad (51)$$

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

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

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

$$g(z_{i,t}) = (c_{i1,t}, c_{i2,t}, c_{i3,t})\theta_0, \text{ where } \theta_0 = (0.02, 0.02, 0.02)^{'}. \tag{52}$$

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

$$g(z_{i,t}) = (c_{i1,t}^2, c_{i1,t} \times c_{i2,t}, sgn(c_{i3,t} \times x_t))\theta_0, \text{ where } \theta_0 = (0.04, 0.03, 0.012)^{'}. \tag{53}$$

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

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

$$r_{i,t+1} = g(z_{i,t}) + e_{i,t+1}, \; e_{i,t+1} = \epsilon_{i,t+1}, \; z_{i,t} = (1, x_t)^{'} \otimes c_{i,t}, \tag{54}$$

where $\epsilon_{i,t+1} \sim \epsilon_{i,t+1} \sim \mathcal{N}(0, 0.05^2)$, $g(z_{i,t})$ is given by (53) and $c_{i,t}$ is given by (50).

Table J: 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%

E2. Simulations demonstrating why Confident-HL strategies outperform

Building on the result that the ex-ante variances predict their squared forecast errors, this subsection illustrates why the Confident-HL portfolios that utilize confidence intervals deliver superior *expected returns* OOS. I use simulations because computing the expected OOS returns of sorting-based HL strategies requires obtaining various moments of "order statistics", which are not available in the closed-form expressions. The exercise that follows resembles example-1.

Consider a simple model based on two sets of stocks, viz. S_A, S_B, each containg $2N$ stocks. 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 forecasts of stock risk premiums. Further suppose that the risk premium forecasts 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 forecasts and their variances.

1. 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 forecasts 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. 1%-HL. This strategy further sorts stocks in the median portfolio, Q_L (Q_S), based on their risk premium forecasts 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 forecasts.

Table K 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 forecasts 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 subset of stocks in Q_L and Q_S that have relatively more precise risk premium forecasts. Thus, the Confident-HL strategies deliver superior expected OOS returns. In contrast, the Low-Confident-HL portfolio delivers relatively lower expected OOS returns than all the other strategies. The reason is that it exclusively comprises stocks with imprecise risk premium forecasts, thus it induces significant misclassification bias.

Table K: 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%

Table L: 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 (E.E1.5.1.1) and equation (54) 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%

F. Internet Appendix: Proofs of theorems 1-4

1. Proof of theorem 1

Under the double exponential prior specification, the posterior log-density of β is

$$\Pi(\beta|\{r_{it}\},\{z_{it}\}) \propto -\Big\{\frac{\lambda_1}{\sigma_\eta}\sum_{j=1}^{p}|\beta_j|+\frac{\lambda_2}{2\sigma_\eta^2}\sum_{j-1}^{p}|\beta_j|^2+\frac{1}{N_{Tr}N_S}\sum_{t\in Tr}\sum_{i\in S}\Big(r_{i,t+1}-z_{it}^T\beta-\Lambda_i f_{t+1}\Big)^2(f_{t+1}^2)\Big\}. \tag{55}$$

Posterior mode of β maximizes (55), which is equivalent to minimizing

$$\hat{\beta}_l = \arg\min_{\beta}\frac{1}{N_{Tr}N_S}\sum_{t\in Tr}\sum_{i\in S}\Big(r_{i,t+1}-z_{it}^T\beta-\Lambda_i f_{t+1}\Big)^2(f_{t+1}^2)+\mu_1||\beta||_1+\mu_2||\beta||_2, \tag{56}$$

where $\mu_1=\frac{\lambda_1}{\sigma_\eta}$, and $\mu_2=\frac{\lambda_2}{2\sigma_\eta^2}$.

Comparing (55) and (56) proves theorem (1).

2. Proof of theorem 2

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

$$\begin{aligned}
&P_{VI}(r_{i,t+1}^*|z_{it}^*,R,Z) = P(r_{i,t+1}^*|z_{it}^*,R,Z,\Omega)q(\Omega)\\
&q(\Omega)=\prod_{k=1}^{K}p_{i,k},\ \text{where each } p_{i,k}\sim Bern(p),\\
&P(r_{i,t+1}^*|z_{it}^*,R,Z,\Omega)=\mathcal{N}(r_{i,t+1}^*;\hat{E}_{i,\Omega,t},\sigma_\eta^2 I),
\end{aligned} \tag{57}$$

where $Bern()$ represents Bernoulli distribution. $\hat{E}_{i,\Omega,t}$ is given by (19), 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

$$\begin{aligned}
E_{VI}(\mu^*_{i,t}) = E_{VI}(r^*_{i,t+1}|z^*_{it}, R, Z) &= \int r^*_{i,t+1} P_{VI}(r^*_{i,t+1}|z^*_{it}, R, Z) dr^*_{i,t+1} \\
&= \int \left(r^*_{i,t+1} \mathcal{N}(r^*_{i,t+1}; \hat{E}_{i,\Omega,t}, \sigma^2_\eta I) \prod_{k=1}^{K} p_{i,k} \right) dp_{i,1} dp_{i,2} \dots dp_{i,K} dr^*_{i,t+1} \\
&= \int \left(r^*_{i,t+1} \mathcal{N}(r^*_{i,t+1}; \hat{E}_{i,\Omega,t}, \sigma^2_\eta I) \prod_{k=1}^{K} p_{i,k} \right) dp_{i,1} dp_{i,2} \dots dp_{i,K} dr^*_{i,t+1} \\
&= \int \left(r^*_{i,t+1} \mathcal{N}(r^*_{i,t+1}; \hat{E}_{i,\Omega,t}, \sigma^2_\eta I) dr^*_{i,t+1} \right) \prod_{k=1}^{K} p_{i,k} dp_{i,1} dp_{i,2} \dots dp_{i,K} \\
&= \int \left(\hat{E}_{i,\Omega,t} dr^*_{i,t+1} \right) \prod_{k=1}^{K} p_{i,k} dp_{i,1} dp_{i,2} \dots dp_{i,K} \qquad (58)
\end{aligned}$$

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

$$\frac{1}{D} \sum_{d=1}^{D} (b_{2,\{\lambda,p\}} + \phi(b_{1,\{\lambda,p\}} + z^*_{it}(p_{1id} W_{1,\{\lambda,p\}}))(p_{2id} W_{2,\{\lambda,p\}})) \xrightarrow{p} E_{VI}(\mu^*_{i,t}) \qquad (59)$$

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

3. Proof of theorem 3

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 (59) in the proof of theorem 2 , and by the weak law of large numbers, as $D \to \infty$

$$\frac{1}{D} \sum_{d=1}^{D} \left(\hat{E}_{i,d,t} - \frac{1}{D} \sum_{d=1}^{D} \hat{E}_{i,d,t} \right)^2 \xrightarrow{p} Var_{VI}\left[E_{W_1,W_2}(r^*_{i,t+1}|z^*_{it}, R, Z, W_1, W_2)\right] \qquad (60)$$

Thus,

$$\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t}-\frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2+\sigma_\eta^2\xrightarrow{p}Var_{VI}\left[(r^*_{i,t+1}|z^*_{it},R,Z)\right] \tag{61}$$

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 (31), and by the law of total variance,

$$Var_{VI}(r^*_{i,t+1})=Var_{VI}(E(r^*_{i,t+1}|W_1,W_2))+E_{VI}(V(r^*_{i,t+1}|W_1,W_2)), \tag{62}$$

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 (31), respectively.

(62) further implies that

$$Var_{VI}(r^*_{i,t+1})=Var_{VI}(\mu^*_{i,t})+\sigma_\eta^2, \tag{63}$$

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

Thus, (60) and (62) implies

$$\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t}-\frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)^2\xrightarrow{p}Var_{VI}(\mu^*_{i,t}). \tag{64}$$

4. 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

$$P_{VI}(r^*_{1,t+1},r^*_{2,t+1},\dots r^*_{S,t+1}|\{z^*_{it}\}_{i=1}^S,R,Z)=P(r^*_{1,t+1},r^*_{2,t+1},\dots r^*_{S,t+1}|\{z^*_{it}\}_{i=1}^S,R,Z,\Omega)q(\Omega)$$

$$q(\Omega)=\prod_{k=1}^{K}p_{i,k},\ \text{where each } p_{i,k}\sim Bern(p),$$

$$P(r^*_{1,t+1},r^*_{2,t+1},\dots r^*_{S,t+1}|\{z^*_{it}\}_{i=1}^S,R,Z,\Omega)=\mathcal{N}(\hat{E}_{S,\Omega,t},\sigma_\eta^2 I),\ \text{where } \hat{E}_{S,\Omega,t}=\begin{bmatrix}\hat{E}_{1,\Omega,t}\\ \hat{E}_{2,\Omega,t}\\ \vdots\\ \hat{E}_{S,\Omega,t}\end{bmatrix}, \tag{65}$$

with each $\hat{E}_{i,\Omega,t}$ given by (19). 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 (3), the covariance of any two return VI-based posterior predictive densities satisfy

$$\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t}-\frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)\left(\hat{E}_{j,d,t}-\frac{1}{D}\sum_{d=1}^{D}\hat{E}_{j,d,t}\right)\xrightarrow{p} Covar_{VI}\left[(r^*_{i,t+1},r^*_{j,t+1}|z^*_{it},R,Z)\right], \quad (66)$$

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

$$\frac{1}{D}\sum_{d=1}^{D}\left(\hat{E}_{i,d,t}-\frac{1}{D}\sum_{d=1}^{D}\hat{E}_{i,d,t}\right)\left(\hat{E}_{j,d,t}-\frac{1}{D}\sum_{d=1}^{D}\hat{E}_{j,d,t}\right)\xrightarrow{p} Covar_{VI}\left[(\mu^*_{i,t},\mu^*_{j,t}|z^*_{it},R,Z)\right]. \quad (67)$$

5. Proof of theorem 5

The proof is straightforward from the proof of theorem 4.

G. 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 6.

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

$$r_{i,t+1} = b_2 + \phi(b_1 + z_{it}W_{o1})W_{o2} + \eta_{i,t+1},\ \eta_{i,t+1} = N(0, \sigma_\eta^2),\ \forall i., \tag{68}$$

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

Denote the set of true parameters

$$\theta_0 = \{W_{o1}, W_{o2}\} \tag{69}$$

Now consider the Bayesian NN specification similar to (31) in section B, 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 (33), satisfy.

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

$$\int \phi_n(R) \left[\Pi p_0(r_{it})\right] dR \to 0, \tag{70}$$

$$\sup_{\theta: ||\theta - \theta_0|| \geq \epsilon} \int (1 - \phi_n(R)) \left[\Pi p_0(r_{it})\right] dR \to 0, \tag{71}$$

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*

$$\sup_{h\in K}\left|\log\frac{p(R|\theta_0+\delta_n h)}{P(R|\theta_0)}-h^T V_{\theta_0}\Delta_{n,\theta_0}+\frac{1}{2}h^T V_{\theta_0}h\right|\xrightarrow{P_0}0, \tag{72}$$

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* (68)*, and the iid assumption of excess returns given the parameters, Theorem 7.2 of Vaart (2000) implies that*

$$\sup_{h\in K}\left|\log\frac{p(R|\theta_0+h/\sqrt{num})}{P(R|\theta_0)}-\frac{1}{num}\sum_{i,t}h^T p'_{\theta_0}(r_{it})+\frac{1}{2}h^T I_{\theta_0}h\right|\xrightarrow{P_0}0, \tag{73}$$

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*).*

$$\left|\left|q_{M_1^*,M_2^*}(.)-\arg\min_q KL\left(q()||MVN\left(\hat{\theta}_{MLE},(1/\sqrt{num})I_{\theta_0}\right)\right)\right|\right|_{TV}\xrightarrow{p}0, \tag{74}$$

where $q_{M_1^*,M_2^*}(.)$ *are given in* (35)*, with optimal* M_1^*,M_2^* *substituted for* M_1,M_2*;* $\{M_1^*,M_2^*\}$ *are given in* (38)*;* $\hat{\theta}_{MLE}$ *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 (35) 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.[19]. 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

$$\arg\min_{q} KL\left(q()||MVN(\hat{\theta}_{MLE}, I_{\theta_0})\right) \approx MVN\left(\hat{\theta}_{MLE}, (1/\sqrt{num})I_{\theta_0}\right) \tag{75}$$

Thus, due to (75), 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).*

$$\left|\left|q_{M_1^*,M_2^*}(.) - MVN\left(\hat{\theta}_{MLE}, (1/\sqrt{num})I_{\theta_0}\right)\right|\right|_{TV} \xrightarrow{p} 0. \tag{76}$$

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.,

$$\mu_{i,t}^* = b_2 + \phi(b_1 + z_{it}^* W_1)W_2, \tag{77}$$

applying the delta method to (76) proves theorem 6.

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

H. Mean-variance Strategies

This section discusses the regularized mean-variance strategies that take into account the entire covariance structure (not just variances) of risk premium forecasts.

Consider a set of n stock return forecasts $\hat{\mu}$, with the covariance matrix $\widehat{\Sigma_r}$. Note that $\widehat{\Sigma_r}$ denotes the covariance of return predictions (not risk premium forecasts). Thus, due to (31), $\widehat{\Sigma_r} = \widehat{\Sigma_{er}} + \sigma_\eta^2 I$, where $\widehat{\Sigma_{er}}$ denotes the covariance of risk premium predictions (estimated in (20)), 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

$$w = \arg\min_w (\hat{\mu} - \widehat{\Sigma_r} w)' \widehat{\Sigma_r}^{-1} (\hat{\mu} - \widehat{\Sigma_r} w) + \gamma_1 w' w + \gamma_2 \sum |w_i|, \tag{78}$$

where the weights $w = [w_1, w_2, \ldots, w_n]'$. The realized excess returns of the mean-variance strategy is given by $w' r$, where r denotes the realized excess returns of N stocks. (78) is also equivalent to

$$w = \arg\min_w (\hat{\mu} - \widehat{\Sigma_{er}} w)' \widehat{\Sigma_{er}}^{-1} (\hat{\mu} - \widehat{\Sigma_{er}} w) + \gamma_3 w' w + \gamma_2 \sum |w_i|, \tag{79}$$

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.[20]

Interpreting Confident-HL strategies

Note that the previously considered regularized mean-variance strategy uses a unified L_1 regularization coefficient, γ_2, across all stocks. Alternatively, consider the following strategy that imposes the adaptive L_1 regularization

$$w = \arg\min_w (\hat{\mu} - \widehat{\Sigma_{er}} w)' \widehat{\Sigma_{er}}^{-1} (\hat{\mu} - \widehat{\Sigma_{er}} w) + \gamma_3 w' w + \sum \gamma_{2i} |w_i|, , \tag{80}$$

[20] 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).

where the L_1 regularization coefficients γ_{2i} vary across assets. If γ_{2i} is allowed to be proportional to stock i's risk premium forecast variances, the resultant strategy discards all stocks with high risk premium forecast variances and thus reduces to the Confident-HL strategy.