The Uncertainty of Machine Learning Predictions in Asset Pricing

Yuan Liao
Department of Economics
Rutgers University

Xinjie Ma
NUS Business School
National University of Singapore

Andreas Neuhierl
Olin School of Business
Washington University in St. Louis

Linda Schilling
Olin School of Business
Washington University in St. Louis

November 11, 2024

Abstract

We develop novel methodology to construct forecast confidence intervals (FCI) for machine learning predictions in asset pricing. We show FCIs for machine learning predictions obtained from sophisticated ML methods, such as neural networks, can be accurately approximated by simpler nonparametric methods such as B-splines. We prove that these FCIs provide correct coverage probabilities. In addition, we also establish the validity of a version of the wild bootstrap. We illustrate the practical use of the obtained confidence intervals in the context of a portfolio selection application for an uncertainty averse investor.

1 Introduction

The asset pricing literature has seen a rapid growth of papers that applies machine learning models in return prediction, portfolio selection and stochastic discount factor estimation. The vast majority of studies in this field makes use of out-of-sample predictions obtained from machine learning models. While he gains in forecasting performance are impressive (e.g. Gu et al. (2020); Bianchi et al. (2021)), the theoretical properties of these forecasts are studied to a much lesser extent. Instead, the literature typically focuses on point predictions, which are silent about the associated uncertainty.

In this paper, we develop novel methods to quantify the uncertainty of return predictions obtained from machine learning models. In particular, we show how to construct forecast confidence intervals for return predictions obtained from neural networks. We provide two methods for obtaining such forecast confidence intervals. The first relies on closed-form approximations and the second on the bootstrap. For the first method, we prove the following fundamental result: *ML-specific forecast methods do not possess an ML-specific asymptotic distribution.* This claim will be defined clearly and proved rigorously in our formal analysis. This result will significantly simplify the analysis and pave the way for relatively easy approximations for the FCI. Armed with these results, we can derive the formula for the machine learning forecast standard error, where the impact of strong cross-sectional dependence appears explicitly. For the bootstrap approach, we show that a simple time-series bootstrap provides asymptotically valid inference for ML-forecast of expected returns while also effectively addressing strong cross-sectional dependencies. Through extensive simulations, we demonstrate that alternative bootstrap methods, such as bootstrapping across firms or jointly across firms and time series, fail to produce valid FCIs for ML forecasts. We show that both methods provide the correct coverage. While these results are of great theoretical interest, the bootstrap may appear to be less attractive from a computational point of view. However, we illustrate an implementation of the so-called k-step bootstrap to also obtain forecast confidence intervals.

In general, statistical inference for ML models such as neural networks is technically demanding and rarely available in "plug-in" form. Researchers often resort to rather crude heuristics without formal justification. In a seminal paper Chen and White (1999) set the stage for a rigorous analysis of the distributional property of neural network predictions.

In our context, however, we cannot merely invoke their results as they are developed for independent and identically distributed (iid) data. This assumption is grossly violated in asset pricing. Asset returns are driven by common factors and thus exhibit strong cross-sectional dependence. This dependence is at the heart of the theoretical challenge. Our approach explicitly accounts for this dependence, moving beyond simple extensions of FCI for weakly dependent time series like autoregressive models.

Standard mean-variance portfolio theory relies on expected returns and (co)variances to compute optimal portfolios. Theoretical studies often assume that agents know the population parameters to streamline the analysis. In many empirical studies, researchers obtain point predictions and act as-if these were the true parameters. By now, it is however well recognized in portfolio selection that estimated parameters should be be treated as the true population parameters (e.g. Garlappi et al. (2007); DeMiguel et al. (2009)). These findings confront researchers with choice between unsatisfactory alternatives. Either use inferior return predictions from simple models with well-studied forecast confidence intervals or use superior return predictions from machine learning estimators and ignore the associated uncertainty. The central objective of this paper is therefore to provide the a sound theoretical justification for the uncertainty associated with machine learning predictions and incorporating them in portfolio selection. To formally incorporate estimation uncertainty in portfolio selection, we adopt the view of an uncertainty averse (UA) investor. We characterize two distinct approaches of UA-portfolios. The first leads to a "no-holding" position, i.e. the uncertainty averse investor may decide not to invest in a risky asset at all if the uncertainty exceeds a certain level.

We characterize two distinct forms of UA-portfolio allocations: one leading to a "no-holding" position in the risky asset. Building Garlappi et al. (2007), we formally show that the resulting optimal portfolio is a solution to an ℓ_1-regularized optimization problem, known as the "Lasso" in the ML literature. The second UA-portfolio follows the proposal in Hansen and Sargent (2008). It yields a "shrinkage" solution which is not sparse, i.e. it does not set many portfolio weights to zero.

Empirically, we find that incorporating the uncertainty into portfolio selection decisions helps to improve the risk-reward trade-off.

Related Literature

The literature on machine learning in asset pricing has seen a rapid growth over the past years. Kozak et al. (2020), Freyberger et al. (2020), Gu et al. (2020), Chen et al. (2020) are concerned with improving predictions in various panel applications. While their objectives differ subtly, they all find improvements from applying ML methods relative to simple parametric alternatives.

Recently, several papers in asset pricing make progress in the theoretical analysis of machine learning predictions in asset pricing. Fan et al. (2022) derive rates of convergence for neural network estimators in return predictions and also show that any regression of excess returns on characteristics can be interpreted through the lens of characteristics based factor models. Jagannathan et al. (2023) develop FCI for the so-called "period-by-period" ML, which relies on estimating the latent risk factors. We instead focus exclusively on the popular "pooled machine learning" approach which is the most popular approach in forecasting applications that do not involve in estimating factors. Kelly et al. (2021); Didisheim et al. (2023) derive exciting results about the out-of-sample properties of portfolios in the context of overparametrized models, i.e. models in which the number of parameters is (much) larger than the sample size. In a very different setting, Liao et al. (2024) also study overparametrized models. All these recent contributions find stronger support for dense rather than sparse models. The paper by Allena (2021) also deserves special mention. He develops confidence intervals for risk premia from a Bayesian perspective in a regression setting.

The statistical theory of machine learning predictors typically focuses on convergence properties such as rates of convergences and approximation error bounds, e.g., Bartlett et al. (2019); Bauer and Kohler (2019); Schmidt-Hieber (2020). Fewer papers deal with quantifying forecast uncertainty, and research on objects such as confidence intervals, forecast standard errors, and forecast distributions is still in its infancy. We briefly review some of the primary studies in this field in below.

In the field of econometric program evaluation, there is a popular method known as "doubly robust machine learning inference", which aims to develop asymptotic confidence intervals of some "structural parameters" in econometric models, e.g., Chernozhukov et al. (2018). These methods develop sophisticated procedures that require so-called *orthogonal moment conditions* and *cross-fitting*, which are not the usual way of implementing the machine learning forecast in asset pricing. In a seminal contribution, Chen and White (1999) derive a theoretical distribution theory for the neural network regression. However, both the "doubly robust machine learning" and Chen-White approach rely on the assumption of i.i.d. (or weakly dependent) data, which is not a valid assumption in the context of asset pricing due to the strong cross-sectional dependence driven by common risk factors.

Our proposed method for computing the FCI is also valuable for areas of economics in which quantifying forecasting precision and uncertainty is needed. For instance, in financial accounting issuing earnings forecasts is an important channel that managers use to convey information to investors. Ciconte et al. (2014) documented that over eighty percent of quarterly earnings forecasts issued between 2002 and 2010 are range forecasts, and there has been a dramatic shift towards issuing range forecasts since then. Prior research has shown that managers have incentives to report forecast precision and range forecasts to investors, and forecast precision significantly affects the sensitivity of market prices to forecast news (Cheng et al., 2013).

Finally, there is a strand of literature in asset pricing that studies the model uncertainty, Avramov (2002); Avramov et al. (2023); Anderson and Cheng (2016). In this line of research, no stance is taken of the "correct model" and the goal is often to achieve robust portfolio allocations and predictions via Bayesian model averages. Researchers therefore often specify a probability distribution over the different models, but do not derive the forecast uncertainty within a given model.

2 Background and Intuition

2.1 Pooled Machine Learning

Let $y_{i,t}$ denote the observed excess return for asset i at period t. In addition, researchers also observe asset-specific "characteristics" (features), $x_{t-1} = (x_{1,t-1}, ..., x_{N,t-1})$, where $x_{i,t-1}$ is a vector of characteristics for asset i, such as momentum, volatility, financial liabilities, etc; for instance, Jensen et al. (2022) provide a large dataset of more than one hundred firm-level characteristics. The goal is to forecast a portfolio return z_{t+1}. This contains the special case of an individual asset, i.e., $z_{t+1} = y_{1,t+1}$, or a broad market index. In either case, z_{t+1} can be represented as a portfolio:

$$z_{t+1} = \sum_{i=1}^{N} w_i y_{i,t+1},$$

where the portfolio weights w_i are assumed to be known and may also vary over time.

In this setting, "pooled machine learning" Gu et al. (2020); Bianchi et al. (2021) has become a very successful and popular methodology to obtain point predictions. Pooled ML builds on the nonparametric model,

$$y_{i,t} = g(x_{i,t-1}) + e_{i,t}$$

with an unknown function $g(\cdot)$, where $e_{i,t}$ is the error term. The unknown function g is learned by *pooling* all observed data (cross-sectionally and over time) and solving a least squares problem:

$$\widehat{g}(\cdot) = \arg\min_{g \in \mathcal{G}_{\mathrm{ML}}} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - g(x_{i,t-1}))^2, \tag{2.1}$$

where the optimal solution is searched for in a space $\mathcal{G}_{\mathrm{ML}}$ that often corresponds to a fixed machine learning method. For instance, for deep neural networks (DNN), $\mathcal{G}_{\mathrm{ML}}$ includes all possible neural network functions with predetermined width and length of the layers as well as the activation function for each neuron; then $\widehat{g}(\cdot)$ is found to be the one with the optimal "bias" and "weights" for the neurons. Once $\widehat{g}(\cdot)$ has been computed, the future return z_{T+1} is predicted by plugging in the most recent characteristic $x_{i,T}$ and constructing a portfolio:

$$\widehat{z}_{T+1|T} := \sum_{i=1}^{N} w_i \widehat{g}(x_{i,T}). \tag{2.2}$$

This method is very popular in academic studies and industry application. It will thus be the primary forecasting method analyzed in this paper. Since all observations are pooled in the cross-section and over all time periods, we call it "pooled ML".

Despite the great popularity and empirical success of pooled ML, little work has been devoted to understanding the structure and sources of predictability. However, understanding the structure of the prediction, $\widehat{z}_{T+1|T}$, is crucially important to quantifying the prediction uncertainty. Recently, Fan et al. (2022) provide an insightful analysis by creating a natural bridge between the machine learning model and factor models, commonly used in empirical asset pricing. Suppose the excess return $y_{i,t}$ can be represented as:

$$y_{i,t} = \alpha_{i,t-1} + \beta_{i,t-1}' f_t + u_{i,t}$$

where $\alpha_{i,t-1}$ and $\beta_{i,t-1}$ are respectively the "alpha" and "beta" of the asset; f_t is the set of (possibly latent) risk factors, and $u_{i,t}$ is the idiosyncratic return. In addition, suppose characteristics are informative about factor loadings (betas) and mispricing (alpha), i.e., there are functions g_α and g_β so that we can rewrite alpha and beta as:[1]

$$\alpha_{i,t-1} = g_\alpha(x_{i,t-1}), \quad \beta_{i,t-1} = g_\beta(x_{i,t-1}).$$

We can thus rewrite the asset pricing model as

$$y_{i,t} = g_\alpha(x_{i,t-1}) + g_\beta(x_{i,t-1})'\mathbb{E}f_t + g_\beta(x_{i,t-1})'v_t + u_{i,t}, \tag{2.3}$$

where $v_t := f_t - \mathbb{E}f_t$. Both v_t and $u_{i,t}$ are mean-zero processes contributing to the error term:

$$e_{i,t} := g_\beta(x_{i,t-1})'v_t + u_{i,t}, \tag{2.4}$$

then the first term is the exposure to factor shocks, while the second term is the idiosyncratic return. Next, suppose that the price of risk, i.e. $\mathbb{E}f_t$ does not change over time. We can define

$$g(x) := g_\alpha(x) + g_\beta(x)'\mathbb{E}f_t. \tag{2.5}$$

Then indeed, (2.3) can be formulated as the machine learning model $y_{i,t} = g(x_{i,t-1}) + e_{i,t}$, with $g(\cdot)$ and $e_{i,t}$ defined in (2.4) and (2.5) respectively. Therefore, by applying pooled ML within the context of this model, Fan et al. (2022) show that the learned ML function $\widehat{g}(x)$ is estimating $g_\alpha(x) + g_\beta(x)'\mathbb{E}f_t$. More formally, let $\mathbb{E}_t$ denote the conditional expectation given information up to t. Write

$$z_{T+1|T} := \mathbb{E}_T z_{T+1}.$$

[1] This formulation is formalizing the notion that firm characteristics are informative about risk exposures which is also documented in Jagannathan and Wang (1996); Ferson and Harvey (1999); Gagliardini et al. (2016); Kelly et al. (2019).

The pooled ML function has the following (probability) limit:

$$\widehat{z}_{T+1|T} \rightarrow^P z_{T+1|T} := \sum_{i=1}^N w_i[g_\alpha(x_{i,T}) + g_\beta(x_{i,T})'\mathbb{E}f_t].$$

This illustrates the source of the pooled ML predictability in forecasting the future expected return: it forecasts portfolio alpha, $\sum_{i=1}^N w_i\alpha_{i,T}$ and the risk premium of the portfolio, $\sum_{i=1}^N w_i\beta_{i,T}'\mathbb{E}f_t$; both are key components of expected returns.

Understanding the forecast object of $\widehat{z}_{T+1|T}$ is an essential first step towards understanding machine learning predictability. The next and equally crucial problem is quantifying the uncertainty of $\widehat{z}_{T+1|T}$. For instance, what is the standard error of $\widehat{z}_{T+1|T}$? Can the forecast standard error be used to interpret economists' behaviors?

2.2 The Challenge of Non-iid Data

First, we discuss a main technical difficulty in deriving the forecast confidence interval for pooled ML – strong cross-sectional dependence. Recall the standard ML model:

$$y_{i,t} = g(x_{i,t-1}) + e_{i,t}. \tag{2.6}$$

If the error term, $e_{i,t}$, were independent over both i and t, then obtaining the FCI of the ML forecast is considerably less complicated. We could merge i, t and treat the stacked data as a long i.i.d. series. The model then would be equivalent to a standard forecast model:

$$y_i = g(x_i) + e_i, \quad i = 1, ..., NT.$$

In this formulation, the asymptotic distribution of Chen and White (1999) can potentially be adapted to derive a forecast standard error.

However, the problem in the asset pricing context is far more sophisticated: the "noise" term $e_{i,t}$ is far from being independent. The errors are not even weakly dependent (in the near-epoch dependence sense). Take two firms, i, j then we can characterize their covariance by using equation (2.4). We have the following expression,

$$\operatorname{Cov}(e_{i,t}, e_{j,t}) = \mathbb{E}g_\beta(x_{i,t-1})' \operatorname{Cov}(f_t)g_\beta(x_{j,t-1}) \neq 0.$$

Clearly, we expect to see strong correlation because the firms are exposed to the same sources of systematic risk. Hence, the pooled model (2.6) is not a usual ML model with i.i.d. errors. This is the main reason why new methods are needed to incorporate the strong cross-sectional dependence structure explicitly.[2]

3 The ML Forecast Confidence Interval

We shall propose two methods to derive a valid FCI for $z_{T+1|T}$, when expected returns are predicted by the ML model of equation (2.6). Our approach builds on two insights which we will formally establish in Section 3.3.

The first is a relatively surprising theoretical result. In Theorem 1 we will show that the forecast distribution is "not ML-specific", in the sense that the asymptotic distribution of $\widehat{g}(x_{i,T})$ is *the same* regardless of the specific ML method – chosen from a relatively broad class. This insight allows us to approximate the forecast standard error using an "easier ML" method.

The second insight establishes that the dominant sources of uncertainty comes from the time series. This is also far from obvious. The errors have strong cross sectional correlation, but the asymptotic distribution of the ML predictor is largely driven by the time series variation. Since asset returns are almost serially independent and the factor shocks can be modelled as a martingale difference sequence we can apply the time-series bootstrap. In the following, we illustrate both methods.

3.1 Method I: "Easier ML" Approximation

We will show that there exists a function, $\zeta^*(\cdot)$, so that we can write, asymptotically,

$$\widehat{z}_{T+1|T} - z_{T+1|T} \quad = \quad \frac{1}{T}\sum_{t=1}^{T}\mathcal{A}_t + o_P(T^{-1/2})$$

where

[2] One possibility of avoiding the strongly dependent noise is to treat factors as "interactive fixed effects" as in Bai (2009) and explicitly estimate them. However, the method in Bai (2009) or Freyberger (2018) does not cover the case of sophisticated machine learning methods, nor is it the "standard" implementation in the applied forecasting literature. We will therefore not pursue this approach.

$$\mathcal{A}_t \quad := \quad \frac{1}{N}\sum_{i=1}^{N}\zeta^*(x_{i,t-1})\beta'_{i,t-1}(f_t-\mathbb{E}f_t). \tag{3.1}$$

The expression of $\mathcal{A}_t$ has three important implications:

First, the prediction error is mainly driven by the common factor shocks rather than the idiosyncratic errors. Thus, the rate of convergence is $O_P(1/\sqrt{T})$, which is much slower than $O_P(1/\sqrt{NT})$ for the usual panel data models. This is consistent with the previous discussion that the errors of the pooled ML are cross-sectionally strongly dependent. The theoretical analysis that leads to (3.1) is novel and cannot be deduced from existing results in the ML literature.

Second, the precision of the ML forecast, measured by the standard error of $\mathcal{A}_t$, is negatively associated with the aggregated betas, the systematic risk exposures. Forecasts conducted at times with higher systematic risk exposures are less precise than those conducted at times with lower systematic risk exposures. This observation has broad implications as it explains the mechanisms of some empirical findings. For instance, in financial accounting, Choi et al. (2011) find that the width of disclosed firms' earnings forecast range is positively associated with the magnitude of the forecast surprise, which is also associated with periods of high systematic risk.

The third implication is the most fundamental and perhaps quite surprising: while the closed-form expression for the function $\zeta^*(\cdot)$ is very difficult to derive, it is completely determined by the quantities from the asset pricing model (2.3), and does not depend on the specific machine learning model $\mathcal{G}_{\mathrm{ML}}$ in (2.1). That is, *the ML-specific* $\widehat{z}_{T+1|T}$ *does not have an ML-specific asymptotic distribution.* This is a powerful result. We can thus apply various kinds of machine learning methods, neural networks, random forests, XG-boosting, B-splines, etc., to estimate $g(\cdot)$, and they all have *the same asymptotic distribution* as determined by (3.1). The main intuition is that the predictor is obtained by optimizing a *regular loss function*; it is the loss function rather than $\mathcal{G}_{\mathrm{ML}}$ that completely determines the asymptotic distribution.This seemingly surprising result is not entirely unfamiliar in econometrics. For instance, the asymptotic distribution of neural networks in the i.i.d. case, derived by Chen and White (1999), also has the same insight that the asymptotic variance is "neural network free".

As the explicit expression of $\zeta^*(\cdot)$ is very difficult to derive, it is challenging to estimate the forecast standard error (FSE) directly. Fortunately, the insight that $\mathcal{A}_t$ is "not ML-specific" motivates us to seek an "easier machine learning" method only for the purpose of deriving its FSE, which is much easier to derive but asymptotically will be the same as that of $\mathcal{A}_t$. One of the simplest nonlinear ML methods are B-splines, which require specifying a set of B-spline basis functions: $\Phi(x) = (\phi_1(x), ..., \phi_J(x))$. Then B-spline regression searches for the optimal function within a much smaller space:

$$\mathcal{G}_B := \{\Phi(x)'\theta : \theta \in \mathbb{R}^J\}.$$

Hence, the B-spline predictor is:

$$\begin{aligned} \widehat{g}_B(\cdot) &= \Phi(x)'\widehat{\theta}, \quad \text{where } \widehat{\theta} := \arg\min_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{i,t} - \Phi(x_{i,t-1})'\theta)^2, \\ \widehat{z}_{T+1|T,B} &:= \sum_{i=1}^{N} w_i \widehat{g}_B(x_{i,T}). \end{aligned} \tag{3.2}$$

The fact that B-spline is an easier machine learning method arises from its OLS-type analytic solution:

$$\widehat{g}_B(x) = \Phi(x)'(\Psi'\Psi)^{-1} \sum_{i,t} \Phi(x_{i,t-1}) y_{i,t}$$

where Ψ is the $NT \times J$ matrix stacking all $\Phi(x_{i,t-1})$ from the data. Therefore, we can easily derive:

$$\widehat{z}_{T+1|T,B} - z_{T+1|T} = \sum_{t=1}^{T} H'\Phi_{t-1}'\beta_{t-1} v_t + o_P(T^{-1/2}) \tag{3.3}$$

where $W = (w_1, ..., w_N)$, $\Phi_{t-1} = (\Phi(x_{1,t-1}), ..., \Phi(x_{N,t-1}))$, and $H' = W'\Phi_T(\Psi'\Psi)^{-1}$. This immediately gives us an asymptotic standard error:

$$\mathrm{SE}(\widehat{z}_{T+1}) := \sqrt{\sum_{t=1}^{T} H'\Phi_{t-1}'\beta_{t-1} \operatorname{Cov}(f_t) \beta_{t-1}' \Phi_{t-1} H}. \tag{3.4}$$

This is straightforward to estimate by: (for $\widehat{e}_t$ be the vector of $\widehat{e}_{i,t} := y_{i,t} - \widehat{g}(x_{i,t-1})$)

$$\widehat{\mathrm{SE}}(\widehat{z}_{T+1}) := \sqrt{\sum_{t=1}^{T} H'\Phi_{t-1}'\widehat{e}_t\widehat{e}_t'\Phi_{t-1}H}. \tag{3.5}$$

It is important to keep in mind that we employ the B-spline only to compute the forecast standard error. The forecast itself is still from a sophisticated neural network. The fact that B-spline is also an ordinary machine learning method ensures that $\widehat{z}_{T+1|T,B}$ should have the same "not ML-specific" expression (3.1). This means the standard error $\mathrm{SE}(\widehat{z}_{T+1})$ also applies to $\mathcal{A}_t$. Formally, we prove the following result:

$$\widehat{\mathrm{SE}}(\widehat{z}_{T+1})^{-1}(\widehat{z}_{T+1|T} - z_{T+1|T}) \to^d \mathcal{N}(0,1), \tag{3.6}$$

where $\widehat{z}_{T+1}$ is obtained via neural networks and $\mathrm{SE}(\widehat{z}_{T+1})$ is obtained via B-splines.

Why sophisticated machine learning?

The above discussion may suggest that sophisticated ML methods, once they satisfy some conditions, lead to the same standard error as the "simpler ML" does. It is therefore tempting to ask why we should employ more sophisticated methods such as DNN in the first place? The answer is that the benefits from using DNNs or related methods *do not* arise from a smaller standard error, but from fewer constraints in handling highly nonlinear functions and the capability of approximating a much larger class of functions. For instance, the so-called "curse of dimensionality" has been a long-time challenge for usual nonparametric methods (e.g., B-splines) even when the number of input features is only mildly large. The simpler ML methods also have larger biases or slower rates of convergence due to more constraints on their required assumptions, limiting their application. In contrast, DNN can automatically adapt to the "intrinsic dimension" of the input features so that it converges to the true unknown function with the fastest possible approximation rate, e.g., Schmidt-Hieber (2020); Kohler and Langer (2021) for detailed theoretical analysis and Gu et al. (2020); Fan et al. (2022) for more in-depth discussion of the advantages of using ML based forecasts in asset pricing. Meanwhile, all we need from the simpler ML method is its standard error, which requires much weaker conditions than using it as a predictor, so the aforementioned constraints are no longer concerns.

3.2 Method II: The k-step Bootstrap

The expression in equation (3.1) also admits the bootstrap as an alternative way of constructing a forecast confidence interval. The bootstrap requires fewer conditions than the easy-ML-approximation method. For instance, it does not require that the simpler method such as the B-spline approximates the true underlying function well. The bootstrap computes the critical value by repeatedly resampling from the original dataset and using the quantile of the recomputed ML-predictor from the resampled data. By doing so, it allows avoiding explicitly estimating the standard error.

However, a limitation of the bootstrap procedure is its high computational demand. For instance, estimation using $B = 100$ bootstrap datasets within each estimation window would require training 100 separate neural networks, one for each bootstrap sample. Fully training these neural networks is computationally very costly, thus limiting the applicability of bootstrap-based DNN inference. To address this challenge, we propose a k-step bootstrap method for DNN inference, which significantly reduces the computational burden.

The k-step bootstrap was originally proposed and studied by Davidson and MacKinnon (1999); Andrews (2002). The idea is that, instead of fully training the neural network for each bootstrap sample, we only train it iteratively for k epochs, with a relatively small k such as 10 or 20. This approach leverages the observation that the fully trained DNN function using the original data, $\widehat{g}(\cdot)$, should provide an excellent starting point for training on the bootstrap data. Thus, we initialize with $\widehat{g}_0^*(\cdot) = \widehat{g}(\cdot)$ and then proceed with k epochs of training, for example, using the `Adam` optimizer.

An important next question is how to generate bootstrap data. It is well known that the validity of the bootstrap procedure crucially depends on how the bootstrap data is generated, which in the asset pricing context, should properly capture the primary sources of uncertainty in predicting expected returns. By studying the expression,

$$\widehat{z}_{T+1|T} - z_{T+1|T} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{A}_t + o_P(T^{-1/2})$$

we find that the sampling uncertainty of pooled ML is mainly driven by the time series variation. Thus, we can cluster at the time-level, by applying the wild bootstrap to mimic the sampling distribution of $\widehat{z}_{T+1|T}$.[3] Specifically, let η_t^* denote an i.i.d. sequence of standard normal random variables and let $\widehat{g}(\cdot)$ denote the learned prediction function using pooled-ML. Define the bootstrap residuals as:

$$e_{i,t}^* = (y_{i,t} - \widehat{g}(x_{i,t-1}))\eta_t^*. \tag{3.7}$$

We then apply pooled ML (e.g., deep neural network) to the resampled excess return $y_{i,t}^* = \widehat{g}(x_{i,t-1}) + e_{i,t}^*$ and $x_{i,t-1}$ to estimate the function $\widehat{g}$ in the bootstrap world, and repeat this step many times to obtain $\widehat{g}^{*,1}(x_{i,T}), ..., \widehat{g}^{*,M}(x_{i,T})$ for a large number M. The forecast critical value, q_α^* would then be the $1-\alpha$ quantile of

$$\left|\sum_i w_i \widehat{g}^{*,1}(x_{i,T}) - \widehat{z}_{T+1|T}\right|, \quad ..., \quad \left|\sum_i w_i \widehat{g}^{*,M}(x_{i,T}) - \widehat{z}_{T+1|T}\right|.$$

The full algorithm is given as follows:

k-step Bootstrap Algorithm.

Step 1. Generate $\eta_t^* \sim \mathcal{N}(0,1)$ independently; generate

$$\begin{aligned} e_{i,t}^* &= (y_{i,t} - \widehat{g}(x_{i,t-1}))\eta_t^* \\ y_{i,t}^* &= \widehat{g}(x_{i,t-1}) + e_{i,t}^*. \end{aligned}$$

Step 2. Run pooled-ML on the bootstrap resampled data: initialize at $\widehat{g}(\cdot)$, and iterate over k epochs. Obtain $\widehat{g}^*(\cdot)$.

Step 3. Repeat Steps 1-2, M times to get $\widehat{g}^{*^1}(\cdot), ..., \widehat{g}^{*^M}(\cdot)$. Let q_α^* be the $1-\alpha$ quantile of

$$\left|\sum_i w_i \widehat{g}^{*,b}(x_{i,T}) - \widehat{z}_{T+1|T}\right|, \quad b = 1, ..., M.$$

[3] If serial correlations are to be allowed, then block-bootstrap (Kunsch (1989)) or the stationary bootstrap (Politis and Romano (1994)) is also applicable.

The bootstrap $1-\alpha$ level FCI for $z_{T+1|T}$ is

$$[\widehat{z}_{T+1|T} - q^*_\alpha, \quad \widehat{z}_{T+1|T} + q^*_\alpha].$$

Although the bootstrap is conceptually easy to apply, justifying its validity to approximate the asymptotic distribution correctly is typically hard. Theorem 3 in Section 3.3 does just that. It is important to note that the bootstrap may fail, and it is not a panacea. For instance, it is now well known that the usual bootstrap fails to provide valid inference when the Lasso is used. Also, in the panel data setting, whether one should bootstrap cross-sectional units or time series units is also crucial to determine the success of bootstrap, which does not have a universal solution and has to be decided model-by-model. In our context, expression (3.1) clearly shows that it is the time series variation that determines the sampling distribution of the pooled machine learning.[4] In contrast, clustering at the firm-level (i.e., cross-sectional bootstrap) would fail to capture the strong cross-sectional dependence and would lead to inconsistent results. We shall illustrate this in the simulation.

Therefore, our procedure's novelty lies in the mathematically solid analysis of the proposed bootstrap procedure. It is critical to let our theory guide the bootstrap, and to properly reflect that the forecast uncertainty should be driven by the time-variation of factor-shocks instead of the idiosyncratic noise. Due to the strong cross-sectional dependence, the bootstrap residual, $e^*_{i,t}$, should depend on η^*_t, the time series residuals. Resampling, η^*_i, the cross-sectional residuals or even, η^*_{it}, the time-series and cross-sectional residuals will lead to an incorrect implementation, and inference will no longer be valid. Intuitively, this happens because these implementations will vastly understate the uncertainty. We will illustrate these failures in simulation in Section 6.

3.3 Theory

In this section we formally characterize the asymptotic distribution of predictor. It will be shown that the predictor, learned from a sophisticated ML space $\mathcal{G}_{\text{ML}}$, has the same asymptotic distribution of the predictor learned from an "easier" nonparametric predictor

[4] In an asset pricing setting, the bootstrap is also applied in Kosowski et al. (2006); Fama and French (2010); Harvey and Liu (2020); Chordia et al. (2020); Giglio et al. (2021).

from $\mathcal{G}_B$. Formally, by an "easier" nonparametric predictor, we mean a space $\mathcal{G}_B$ that collects functions of the form $\phi(x)'\theta$ with a growing number of basis functions in $\phi(x)$.

The key technical assumption of our analysis is that both $\mathcal{G}_{\text{ML}}$ and $\mathcal{G}_B$ should have well controlled complexity should approximate the true underlying function well. The formal assumption is stated below. Let $\mathcal{G}$ be a generic space of functions which can be either $\mathcal{G}_{\text{ML}}$ or $\mathcal{G}_B$. Define the *pseudo dimension* $p(\mathcal{G})$ as the Vapnik-Chervonenkis dimension of the subgraph class $\{f(x,y) := \text{sgn}(h(x) - y) : h \in \mathcal{G}\}$. We shall require

$$p(\mathcal{G}) \log(NT) = o(T^{1/2})$$

Hence the machine learning method being used should not be over-parametrized. Assumption 2 in the appendix provides a formal condition on the complexity of the ML methods.

Kelly et al. (2021); Didisheim et al. (2023) work with over-parametrized models in the context of portfolio construction and show that complex models often perform better than relatively simple models. Extending the theory to over-parametrized models as in their setting for neural networks is a challenging an interesting research question, which we leave for future research.

The following theorem characterizes the asymptotic expansion of the DNN index predictor $\widehat{z}_{T+1|T}$, and shows that it is the same as if the index were predicted using a simpler method, e.g. such as B-splines.

Theorem 1. *Suppose $\sum_i |w_i| < \infty$ and Assumption 1 and Assumption 2 in the appendix hold. There is a function $\zeta^*(\cdot)$ so that*

$$\widehat{z}_{T+1|T} - z_{T+1|T} = \frac{1}{T}\sum_t \frac{1}{N}\sum_i \zeta^*(x_{i,t-1})\beta'_{i,t-1}v_t + o_P(T^{-1/2}). \tag{3.8}$$

The function ζ^* only depends on the joint distribution of $(x_{i,t-1})$ and the realization $x_{i,T}$, but does not depend on whether $\mathcal{G}_{\text{ML}}$ or $\mathcal{G}_B$ were used for constructing $\widehat{z}_{T+1|T}$. In other words, the same asymptotic expansion holds with the same function $\zeta^*(\cdot)$ if the "easier" ML space $\mathcal{G}_B$ were used in place of $\mathcal{G}_{\text{ML}}$ in the definition of $\widehat{z}_{T+1|T}$.

The important implication of this theorem is that, while it might be challenging to directly compute the asymptotic standard error of the DNN predictor from the expansion, we can easily adopt the standard error of the B-spline predictor $\widehat{\text{SE}}(\widehat{z}_{T+1})$ as given in (3.5).

The following theorem is one of our main results. It shows that this simple formula of FSE can be used as the standard error for the DNN predictor to construct forecast confidence intervals.

Theorem 2 (Easier ML approximation)**.** *Suppose* $\sum_i |w_i| < \infty$*. Let* $\widehat{z}_{T+1|T}$ *be the index predictor using* DNN*. Then*

$$\widehat{\mathrm{SE}}(\widehat{z}_{T+1})^{-1}(\widehat{z}_{T+1|T} - z_{T+1|T}) \to^d \mathcal{N}(0,1).$$

In contrast to the "easier-ML", the bootstrap does not require computing the analytical standard error. For ease of technical proofs however, Theorem 3 below establishes the asymptotic validity of the fully trained bootstrap neural networks. That is, let $(y^*_{i,t})$ denote the bootstrap data. We prove for the case when the bootstrap DNN $\widehat{g}^*(\cdot)$ is defined as:

$$\widehat{g}^*(\cdot) = \arg\min_{g \in \mathcal{G}_{\mathrm{ML}}} \sum_{i=1}^{N} \sum_{t=1}^{T} (y^*_{i,t} - g(x_{i,t-1}))^2, \tag{3.9}$$

where $\mathcal{G}_{\mathrm{ML}}$ is the pooled-ML space, such as DNN.

Theorem 3 (Bootstrap)**.** *Suppose Assumptions 2 in the appendix holds, but it is only required to hold for the sophisticated ML* $\mathcal{G}_{\mathrm{ML}}$*, and the B-spline machine learning does not need to satisfy it. Also Assumption 1 holds. Then for any* $\alpha \in (0,1)$*,*

$$P(|\widehat{z}_{T+1|T} - z_{T+1|T}| < q^*_\alpha) \to 1-\alpha,$$

where q^*_α *is the* $1-\alpha$ *bootstrap sample as in Step 3 in the bootstrap algorithm.*

4 Uncertainty-Aversion Under Deep-Learning Forecasts

In classic portfolio theory, it assumed that the investor knows the population moments that determine her portfolio decision. One of the major challenges in operationalizing the theory has been that these parameters have to be estimated and that estimation error can often dominate the portfolio decision, see e.g. Jagannathan and Ma (2003); DeMiguel et al. (2009) for classic examples.

If the economist is aware of the forecast uncertainty, she will deviate from the population solution of the mean-variance problem and transition to *uncertainty-aversion* (UA) positions. Our goal of this section is to embrace the merits of good point predictions obtained from machine learning models, while at the same time incorporating the estimation uncertainty into the portfolio section problem.

While risk aversion and uncertainty-aversion share similarities in applications, economic theory carefully distinguishes between the two, see Ellsberg (1961); Gilboa and Schmeidler (1989, 1993); Schmeidler (1989); Epstein and Wang (1994). In the context of portfolio selection, a risk-averse expected utility investor would behave exactly as if she knew the expected return and (co)variances. An uncertainty averse investor however, would take the estimation uncertainty into consideration and incorporate it into the portfolio selection problem.

In this section, we show how uncertainty-aversion investors behave when using DNN forecasts of expected returns with the quantified uncertainty we derived in the previous section. We separately discuss two types of uncertainty-aversion implementations. The first is a "no-holding" approach, e.g. (Dow and da Costa Werlang, 1992). This approach implies that when the forecast uncertainty is too high, the dominating strategy would be not to invest at all in risky assets, i.e. not hold a given risky asset at all. The other strategy stems from the framework of robust optimization (Hansen and Sargent, 2008), which is less conservative, but still leads to more cautious holdings of risk assets compared to the standard mean-variance solution.

4.1 No-holding positions

We consider the allocation among multiple factor portfolios. Therefore, in this section we shall use $z_{T+1|T}$ to denote a multivariate expected return of a set of factor portfolios, predicted by $\widehat{z}_{T+1|T}$ using pooled ML:

$$z_{T+1|T} = (z_{1,T+1|T}, ..., z_{R,T+1|T})', \quad \widehat{z}_{T+1|T} = (\widehat{z}_{1,T+1|T}, ..., \widehat{z}_{R,T+1|T})'.$$

For instance,

$$\widehat{z}_{1,T+1|T} = \sum_{i=1}^{N} w_{i,1}\widehat{g}(x_{i,T}),$$

where $w_{i,1}$ are the weights in the individual stock returns for the first asset.

Let Σ_T denote the covariance matrix of a set of portfolios $z_{T+1} = (z_{1,T+1}, ..., z_{R,T+1})$. We focus only on the uncertainty of $z_{T+1|T}$, setting aside the estimation uncertainty in the covariance matrix, Σ_T.

Consider the following portfolio problem:

$$\text{MV problem} \qquad \max_{\boldsymbol{\omega}} \boldsymbol{\omega}'(\widehat{z}_{T+1|T} - \mathbf{1}r_f) - \frac{\gamma}{2}\boldsymbol{\omega}'\Sigma_T\boldsymbol{\omega} \tag{4.1}$$

where r_f is the return on the risk-free asset, and $\gamma > 0$ is the coefficient of risk aversion. The MV problem is often used for determining the mean-variance efficient portfolio. However, it takes $\widehat{z}_{T+1|T}$ as given without accounting for its associated uncertainty. When the forecast uncertainty is taken into account, $\widehat{z}_{T+|T}$ yields a forecast confidence interval for the true expected return

$$\text{FCI} = [\widehat{z}_{1,T+1|T} - q_{1,\alpha}, \widehat{z}_{1,T+1|T} + q_{1,\alpha}] \times \cdots \times [\widehat{z}_{R,T+1|T} - q_{R,\alpha}, \widehat{z}_{R,T+1|T} + q_{R,\alpha}]$$

where $q_{i,\alpha}$ is the critical value for $\widehat{z}_{i,T+1|T} - z_{i,T+1}$ under significance level α obtained using either the analytic forecast standard error or bootstrap. For instance, using the "easier ML" FSE, we can take

$$q_{i,\alpha} = \text{SE}(\widehat{z}_{i,T+1}) \times \epsilon_\alpha$$

where ϵ_α is the quantile of the standard normal distribution. If the investor has high uncertainty aversion, she could use $\epsilon_\alpha = 2.5758$, corresponding to a 99% confidence interval. For the bootstrap, we can use

$$q_{i,\alpha} = q^*_{i,\alpha}$$

which is the $1 - \alpha$ quantile of the bootstrap critical value for $\widehat{z}_{i,T+1|T}$. In practice, it is also desirable to control for the Type I error rate for multiple testing, in which case, one can set $\alpha = 0.05/R$ as in the Bonferroni correction.

To incorporate the FCI into the portfolio selection problem, (4.1), consider the max-min problem as follows.

$$\text{UA-MV problem} \qquad \max_{\boldsymbol{\omega}} \min_{\mu \in \text{FCI}} \boldsymbol{\omega}'(\mu - r_f) - \frac{\gamma}{2}\boldsymbol{\omega}'\Sigma_T\boldsymbol{\omega}. \tag{4.2}$$

It aims to find the optimal portfolio weights under the "worst scenario" of the predicted expected return within the FCI. Therefore, unlike the classical mean-variance portfolio selection in which the predicted $\widehat{z}_{T+1|T}$ in place of $\mathbb{E}_T z_{T+1}$, the UA-constraint problems admit more possible values the expected return might take due to forecast uncertainty.

In this section we shall characterize the solutions to (4.2). A key result is that the solution yields a "no-holding" position.

4.1.1 One risky and one risk-free asset

We will start with the classic example of allocating between one risky asset and the risk-free asset. Denote the portfolio weight in the risky assets as ω. The estimated expected return of the risky asset is $\widehat{z}_{T+1|T}$, the true but unknown expected return is μ and its variance is denoted σ^2. With the additional uncertainty from estimation, the mean-variance (MV) problem is studied by Garlappi et al. (2007) and is given by:

$$\max_{\omega} \min_{\mu} \omega\mu + (1-\omega)r_f - \frac{\gamma}{2}\omega^2\sigma^2, \tag{4.3}$$

subject to:

$$|\mu - \widehat{z}_{T+1|T}| \leq q_\alpha. \tag{4.4}$$

Here σ^2 is the variance of z_{T+1}. There are also two scalar parameters: γ is the coefficient of risk aversion and q_α is the constraint that reflects the investor's aversion to uncertainty, which also corresponds to the $(1-\alpha)$-confidence level for the predicted expected return.

The uncertainty-averse MV problem has an insightful representation, as shown as in the following theorem. We denote $(x)_+ = \max\{x, 0\}$ and $\mathrm{sgn}(x)$ denote the sign of x. Extending the study of Garlappi et al. (2007), the following theorem shows that the solution can be characterized using ℓ_1-penalization, known as "Lasso" regularization:

Theorem 4. *The uncertainty-constraint MV problem (4.3)- (4.4) is equivalent to the following Lasso-problem:*

$$\omega^{\mathrm{NH}} = \arg\min_{\omega} \frac{1}{2}\left(\omega - \omega^{\mathrm{MV}}\right)^2 + \lambda_\alpha|\omega|$$

and the optimal solution is

$$\omega^{\mathrm{NH}} = \mathrm{sgn}(\omega^{\mathrm{MV}})(|\omega^{\mathrm{MV}}| - \lambda_\alpha)_+$$

where

$$\omega^{\mathrm{MV}} = \frac{\widehat{z}_{T+1|T} - r_f}{\gamma\sigma^2}, \quad \lambda_\alpha = \frac{q_\alpha}{\gamma\sigma^2}.$$

In the above theorem, ω^{MV} is the classic mean-variance portfolio without uncertainty constraints, i.e. the solution to

$$\max_{\omega} \omega\widehat{z}_{T+1|T} + (1-\omega)r_f - \frac{\gamma}{2}\omega^2\sigma^2.$$

The theorem explains the reason why the UA-MV problem yields no-holding positions in the solution, as solving the uncertainty-constraint problem is equivalent to solving a Lasso-type ℓ_1-penalized regression problem, which has an analytic solution ω^{NH} that yields zero as one of the solutions. This facilitates the practical computations.

More importantly, this result gives a direct economic interpretation of the UA-MV problem. First, the optimal ω^{NH} is zero ("no-holding") whenever $|\widehat{z}_{T+1|T} - r_f| \leq q_\alpha$. That is, when the investor's prediction $\widehat{z}_{T+1|T}$ is close enough to the risk-free rate, whose difference is smaller than her uncertainty tolerance, she would hold no position in the risky asset. The intuition lies in the nature of uncertainty aversion: we note that $|\widehat{z}_{T+1|T} - r_f| \leq q_\alpha$ means the investor finds insignificance for testing the following null hypothesis:

$$H_0: \quad z_{T+1|T} = r_f.$$

Her preference is then investing solely in the risk-free rate without bearing any risk.[5]

Second, the optimal ω^{NH} continuously evolves when $|\widehat{z}_{T+1|T} - r_f| > q_\alpha$, that is, the investor predicts that the expected return of the risky asset is significantly different from the risk-free rate. She will then start investing in the risky asset, and the decision of whether short- or long- the risky asset (the sign of ω^{NH}) is determined by the sign of $\widehat{z}_{T+1|T} - r_f$. But even in this case, the investor will still invest more cautiously in the risky asset, by *shrinking* her investment towards the risk-free rate. For instance, suppose the investor finds that $\widehat{z}_{T+1|T} > r_f + q_\alpha$, then her allocation in the risky asset is

$$\omega^{\mathrm{NH}} = \omega^{\mathrm{MV}} - \frac{q_\alpha}{\gamma\sigma^2} > 0.$$

[5] Bessembinder (2018) performs an extensive empirical investigation of this hypothesis and finds that many stocks indeed do not outperform treasuries ex-post.

Instead of adopting the classical mean-variance portfolio, she reduces her allocation to the risky asset, and the amount of reduction, $q_\alpha/(\gamma\sigma^2)$, reflects her tolerance of uncertainty, which is also due to the nature of risk aversion.

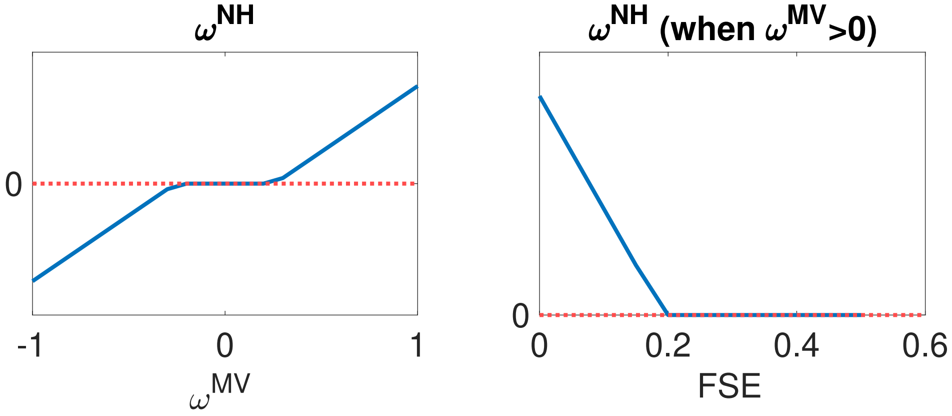

Figure 1: The ω^{NH} position. The left panel fixes the level of FSE and compares ω^{NH} with the regular ω^{MV}, whereas the right panel fixes the value of ω^{MV} and plots ω^{NH} as a function of FSE.

To illustrate ω^{NH}, the left panel of Figure 1 plots the lasso function with respect to various magnitudes of the classical MVE portfolio. Here we fix the level of FSE and thus q_α. As we can see, ω^{NH} is zero for small magnitudes of $|\omega^{\mathrm{MV}}|$, i.e. the investor does not allocate towards this asset. It starts to increase but with cosntant shrinkage relative to the MVE as the later deviates from zero. The right panel plots ω^{NH} with respect to the change of forecast standard error and a fixed ω^{MV}. As FSE increases meaning that the economists starts to takes into account the forecast uncertainty, her position in the risky asset decays linearly, and eventually becomes zero.

In their seminal work, Dow and da Costa Werlang (1992) extend the expected-utility framework to incorporate the uncertainty arising from unknown probabilities for portfolio choices. Their study reveals that, under uncertain distributions of present values, there exists a range of prices within which investors have no position in the asset, and that this range of prices depends only on the attitudes and beliefs about uncertainty. Crucially, their findings elucidate an inherent economic intuition – uncertainty aversion induces a "no-holding" effect in investors' positions, reflecting a range of estimated payoffs in which no-holding is strictly better than investing in the asset.

Similarly, we have established the effect of "no-holding" position in the context of forecast uncertainty in portfolio allocation: there exists a range of forecast expected returns within which uncertainty-averse investors would not invest in the risky asset. Notably, this range of expected returns is determined by the sign of $|\widehat{z}_{T+1|T} - r_f| - q_\alpha$, i.e., the significance of comparing the predicted mean return and the risk-free rate. It is then clear that uncertainty-averse investors would reduce their exposure to the risky asset by shrinking their investment in the risky asset towards the risk-free asset.

4.1.2 Multiple risky assets

We now consider the problem where at the end of period T, the investor would like to allocate her portfolio into R risky asset whose expected returns are $z_{T+1|T}$. To incorporate her uncertainty constraints for allocating her portfolio onto these R risky assets, the investor then formulates a multivariate, constrained, mean-variance problem:

$$\max_{\boldsymbol{\omega}=(\omega_1,...,\omega_R)} \min_{\mu_1,...,\mu_R} \sum_{i=1}^{R} \omega_i \mu_i - \frac{\gamma}{2} \boldsymbol{\omega}' \Sigma_T \boldsymbol{\omega}$$

subject to the constraint on $(\boldsymbol{\omega}, \mu_1, ..., \mu_R)$:

$$\sum_{j=1}^{R} \omega_j = 1, \quad |\mu_i - \widehat{z}_{T+1|T,i}| \leq q_{\alpha,i}, i = 1, ..., R.$$

The following theorem shows that the uncertainty constraint MV-problem can also be formulated as a Lasso-type problem, which shrinks the portfolio weights $w_1, ..., w_R$ towards zero. In addition, the Lasso penalty equals $q_{\alpha,i}$, which is proportional to the length of the ML-based forecast confidence interval.

Theorem 5. *The multivariate constraint MV portfolio allocation can be cast as a Lasso-problem as follows:*

$$\min_{\boldsymbol{\omega}=(\omega_1,...,\omega_R)} \frac{\gamma}{2} \boldsymbol{\omega}' \Sigma_T \boldsymbol{\omega} - \sum_{i=1}^{R} \omega_i \widehat{z}_{T+1|T,i} + \sum_{i=1}^{R} q_{\alpha,i} |\omega_i|, \text{ subject to } \sum_{i} \omega_i = 1.$$

To see how the portfolio shrinkage is determined by the investor's risk aversion, let us focus on the case when there are two risky assets ($R = 2$). Then the optimal portfolio has a closed form solution: (the proof is given in the appendix) for $k, j \in \{1, 2\}$ and $k \neq j$, and $c_0 = \gamma^{-1} \operatorname{Var}(z_{T+1,1} - z_{T+1,2})^{-1}$

$$\omega_k^* = \begin{cases} 0 & \text{when } -c_0(q_{\alpha,1} + q_{\alpha,2}) < \omega_k^{\mathrm{MV}} < c_0(q_{\alpha,k} - q_{\alpha,j}) \\ \omega_k^{\mathrm{MV}} + c_0(q_{\alpha,1} + q_{\alpha,2}) & \text{when } \omega_k^{\mathrm{MV}} < -c_0(q_{\alpha,1} + q_{\alpha,2}) \\ \omega_k^{\mathrm{MV}} - c_0(q_{\alpha,k} - q_{\alpha,j}) & \text{when } \omega_1^{\mathrm{MV}} > c_0(q_{\alpha,1} - q_{\alpha,2}) \text{ and } \omega_2^{\mathrm{MV}} > c_0(q_{\alpha,2} - q_{\alpha,1}) \end{cases} . \tag{4.5}$$

Here ω_k^{MV} is the usual mean-variance efficient portfolio: $\omega_2^{\mathrm{MV}} = 1 - \omega_1^{\mathrm{MV}}$ and

$$\omega_1^{\mathrm{MV}} \quad := \quad \arg\max_{\omega \in \mathbb{R}} \omega \widehat{z}_{T+1|T,1} + (1 - \omega)\widehat{z}_{T+1|T,2} - \frac{\gamma}{2}\boldsymbol{\omega}' \Sigma_T \boldsymbol{\omega}, \quad \boldsymbol{\omega} = (\omega, 1 - \omega)'.$$

As the expression of the optimal portfolios ω_1^* and ω_2^* is symmetric, we focus on ω_1^*. To explain the economic insights, let us consider and compare the behavior of two investors: one MV-investor, who would adopt the classical unconstrained mean-variance portfolio ω_1^{MV}; and one UA-investor, who is uncertainty averse and would adopt ω_1^*.

Consider the first case of (4.5). The “no-holding” position $\omega_1^* = 0$ appears whenever

$$|\omega_1^{\mathrm{MV}} + q_{\alpha,2} c_0| \leq q_{\alpha,1} c_0. \tag{4.6}$$

To understand the intuition, recall that $q_{\alpha,k}$ measures the UA-investor’s degree of uncertainty aversion. The higher $q_{\alpha,k}$, the more cautious she is when allocating to asset k, and vice versa. Suppose $q_{\alpha,2} \to 0$ and is much smaller than $q_{\alpha,1}$. When (4.6) holds, then

$$|\omega_1^{\mathrm{MV}}| \leq q_{\alpha,1} c_0 [1 + o(1)],$$

meaning that the MV-investor will not invest much in $z_{T+1,1}$, whose allocated portfolio weight is no more than approximately $q_{\alpha,1} c_0$. Meanwhile, as $q_{\alpha,2}$ is very small, the UA-investor has little uncertainty about the predictor $\widehat{z}_{T+1|T,2}$, so she would be better off allocating all her assets in $z_{T+1,2}$, and thus $\omega_1^* = 0$.

Next, consider the second case of (4.5). Suppose $\omega_1^{\mathrm{MV}} < -c_0(q_{\alpha,1} + q_{\alpha,2})$, then $\omega_1^{\mathrm{MV}} < 0$, meaning that the MV-investor would short $z_{T+1,1}$. Meanwhile, the UA-investor would also short $z_{T+1,1}$, because $\omega_1^* = \omega_1^{\mathrm{MV}} + c_0(q_{\alpha,1} + q_{\alpha,2}) < 0$, but she would short less on $z_{T+1,1}$

and shrink her allocation towards $z_{T+1,2}$ as $\omega_1^{\mathrm{MV}} < \omega_1^* < 0$. The amount of shrinkage is $c_0(q_{\alpha,1} + q_{\alpha,2})$. Therefore, the UA-investor is more cautious than the MV-investor when shorting $z_{T+1,1}$.

Finally, consider the third case of (4.5). Then

$$\omega_1^* = \omega_1^{\mathrm{MV}} - c_0(q_{\alpha,1} - q_{\alpha,2}), \text{ which is } \begin{cases} < \omega_1^{\mathrm{MV}} & \text{if } q_{\alpha,1} > q_{\alpha,2} \\ > \omega_1^{\mathrm{MV}} & \text{if } q_{\alpha,1} < q_{\alpha,2}. \end{cases}$$

When $q_{\alpha,1} > q_{\alpha,2}$, this case implies $\omega_1^{\mathrm{MV}} > c_0(q_{\alpha,1} - q_{\alpha,2}) > 0$, so the MV-investor would long $z_{T+1,1}$. Meanwhile, because the UA-investor is more uncertain about $z_{T+1,1}$ than $z_{T+1,2}$, so her allocation in the first asset would satisfy $0 < \omega_1^* < \omega_1^{\mathrm{MV}}$, meaning that she would shrink her allocation toward $z_{T+1,2}$, and the amount of shrinkage is $c_0(q_{\alpha,1} - q_{\alpha,2})$, proportional to the difference of the degrees of uncertainty. The case when $q_{\alpha,1} < q_{\alpha,2}$ follows from a similar insight due to the symmetry between ω_1^* and ω_2^*.

In summary, the UA-constrained portfolio will contain many zero positions, reflecting the caution of the uncertainty averse investor. In general, the UA-investor allocates her portfolios more cautiously than the MV-investor by shrinking her allocation towards assets that she is less uncertain about their predicted mean returns. Such a *shrinkage effect* reduces the allocation compared to the classical mean-variance portfolios, and the amount of reductions depends on the investors' degree of uncertainty aversion.

4.2 Risk-Sensitive Optimization

The UA-constraint portfolio allocation, as we studied in the previous subsection, takes a typical no-holding position, which may be considered too conservative by some economists. In this subsection we consider a less conservative framework that also accounts for the forecast uncertainty, following closely to the ambiguity literature in Hansen and Sargent (2008), among many others.

4.2.1 One risky and one riskless asset

Anderson and Cheng (2016) formulate a mean-variance optimization for choosing portfolio allocations that are robust to misspecifications in the predictive model. To take an example where there is one risky asset with return z_{T+1} and a risk-free asset r_f, their problem is formulated as:

$$\max_{\omega} \min_{p_T} \omega \mathbb{E}_{p_T}(z_{T+1}) + (1-\omega) r_f - \frac{\gamma}{2}\omega^2 \operatorname{Var}_{p_T}(z_{T+1}) + \frac{1}{\tau} D(p_T || f_T) \tag{4.7}$$

where both p_T and f_T are probability density functions, and $\mathbb{E}_{p_T}(z_{T+1})$ and $\operatorname{Var}_{p_T}(z_{T+1})$ respectively denote the expectation and variance of the future risky return with respect to the distribution p_T; $D(p_T||f_T)$ denotes the Kulback-Leibler divergence from f_T to p_T. Therefore, different from the UA-constrained problem, (4.7) introduces an inner optimization with respect to the density function p_T with an additional measure $D(p_T||f_T)$. This problem is also called "risk-sensitive optimization" by Hansen and Sargent (2008).

The idea is that the investor takes f_T as the benchmark predictive density of z_{T+1}, but she is concerned that f_T might be misspecified. So she considers an alternative predictive density p_T for the risky return, and constructs portfolio choices to maximize utility on the worst specification of p_T. Meanwhile, the investor also has the belief that f_T is "reasonably specified", so by introducing the penalization term $D(p_T||f_T)$, she focuses on alternatives that are close to f_T. The scalar parameter τ is a measure of the investor's uncertainty aversion. One suggestion of Anderson and Cheng (2016) is to use the benchmark predictive density

$$f_T \sim \mathcal{N}(\widehat{z}_{T+1|T}, \sigma_T^2), \tag{4.8}$$

which directly uses the predicted return $\widehat{z}_{T+1|T}$ as the mean of the predictive density, and is interpreted as "the agent's best approximation for the distribution (of returns)".

Importantly, in the Anderson-Cheng model, the incorporated uncertainty is from the density f_t of the true future return, rather than the uncertainty from the prediction $\widehat{z}_{T+1|T}$. It can be easily shown that the solution is equivalent to the MV-portfolio with an increased risk-aversion parameter $\tau + \gamma$. Hence benchmarking against (4.8) does not incorporate the prediction uncertainty, which could yield portfolios that are too aggressive to be robust to sudden changes of market and idiosyncratic risk.

To account for the uncertainty of the prediction $\widehat{z}_{T+1|T}$, which in our applications is the pooled ML (e.g., DNN) forecast, we suppose the investor as a prior distribution of the expected return:

$$p(z_{T+1|T}) \sim \mathcal{N}(\pi, v)$$

where (π, v) respectively denote the prior mean and prior variance. Meanwhile, the asymptotic distribution of the ML forecast is approximately (as proved in Theorem 2):

$$p(\widehat{z}_{T+1|T}|z_{T+1|T}) \sim \mathcal{N}(z_{T+1|T}, \text{FSE}^2)$$

where $\text{FSE} = \text{SE}(\widehat{z}_{T+1})$ is the forecast standard error obtained using easy-ML. This distribution serves as the likelihood function, which then yields a posterior distribution of the future raw return z_{T+1}:

$$f_T(z_{T+1}) = \int p(z_{T+1}|z_{T+1|T})p(z_{T+1|T})p(\widehat{z}_{T+1|T}|z_{T+1|T})dz_{T+1|T} \sim \mathcal{N}(\widetilde{z}_{T+1}, \widetilde{\sigma}_T^2)$$

where $\widetilde{z}_{T+1}$ and $\widetilde{\sigma}_T^2$ are the posterior mean and variance, respectively given by

$$\widetilde{z}_{T+1} = (1-W_1)\widehat{z}_{T+1|T} + W_1\pi, \quad \widetilde{\sigma}_T^2 = vW_1 + \sigma_T^2, \quad W_1 = \frac{\text{FSE}^2}{\text{FSE}^2 + v}.$$

In addition, we focus on p_T taking the form:

$$p_T \sim \mathcal{N}(\mu, \widetilde{\sigma}_T^2),$$

where $\mu \in \mathbb{R}$ is unspecified and we search for the optimal μ in the inner optimization of (4.7). This allows us to concentrate on the mean forecast as the main source of misspecification. In this case, the Kullback-Leibler divergence for two normal distributions is simply $D(p_T||f_T) = \frac{1}{2\widetilde{\sigma}_T^2}(\widetilde{z}_{T+1|T} - \mu)^2$. Therefore, the optimal portfolio can be obtained as:

$$\omega^{\text{RS}} := \arg\max_{\omega}\min_{\mu} \omega\mu + (1-\omega)r_f - \frac{\gamma}{2}\omega^2\widetilde{\sigma}_T^2 + \frac{1}{2\tau\widetilde{\sigma}_T^2}(\widetilde{z}_{T+1|T} - \mu)^2$$

To characterize the solution, let

$$\omega^{\text{MV}} = \frac{\widehat{z}_{T+1|T} - r_f}{\sigma_T^2\gamma}, \quad \omega_\pi^{\text{MV}} = \frac{\pi - r_f}{\sigma_T^2\gamma}$$

be the MV portfolios based on the predicted mean $\widehat{z}_{T+1|T}$ and the prior mean π. Then it can be shown that the solution is

$$\omega^{\mathrm{RS}} = \left[\omega^{\mathrm{MV}}(1 - W_1) + \omega_{\pi}^{\mathrm{MV}} W_1\right] \frac{\sigma_T^2}{vW_1 + \sigma_T^2} \frac{\gamma}{\tau + \gamma}. \tag{4.9}$$

Therefore, when taking into account the forecast uncertainty, ω^{RS} conducts a *double shrinkage*: First, it shrinks ω^{MV} towards $\omega_{\pi}^{\mathrm{MV}}$ due to the weight W_1. Second, it shrinks the overall portfolio towards zero due to the factor $\frac{\sigma_T^2}{vW_1+\sigma_T^2}$.

To further shed light on the effect of incorporating forecast uncertainty to the shrinkage portfolio, we recall that $W_1 = \mathrm{FSE}^2 / (\mathrm{FSE}^2 + v)$, hence ω^{RS} can be written as an explicit function of FSE that shrinks the mean-variance portfolio: $\omega^{\mathrm{RS}} = g(\mathrm{FSE})$, where

$$\begin{aligned} g(s) &:= \left[\omega^{\mathrm{MV}}(1 - W(s)) + \omega_{\pi}^{\mathrm{MV}} W(s)\right] \frac{\sigma_T^2}{vW(s) + \sigma_T^2} \frac{\gamma}{\tau + \gamma} \\ W(s) &:= \frac{s^2}{s^2 + v}. \end{aligned}$$

The monotonicity of $g(\cdot)$ depends on the relative magnitude between ω^{MV} and $a \times \omega_{\pi}^{\mathrm{MV}}$, where $a = \frac{\sigma_T^2}{v+\sigma_T^2}$. If $\omega^{\mathrm{MV}} > a\omega_{\pi}^{\mathrm{MV}}$, then $g(\cdot)$ is decreasing; otherwise $g(\cdot)$ is increasing. As SE increases, the ML-forecast becomes more uncertain, which means we should rely less on the mean-variance portfolio ω^{MV} because it is built on $\widehat{z}_{T+1|T}$. This prompts uncertainty-averse investors to adopt strategies that move in the opposite direction relative to the traditional MVE-position that does not incorporate the forecast uncertainty.

As an illustration, the left panel of Figure 2 compares ω^{RS} with ω^{MV} as the latter increases with a fixed FSE. Clearly ω^{RS} increases linearly in ω^{MV} but at a slower rate because of the linear shrinkage effect. The right panel plots the $g(\cdot)$ function with respect to the forecast standard error. in the case $\omega^{\mathrm{MV}} > a\omega_{\pi}^{\mathrm{MV}}$. This case intuitively means that the traditional MVE-position suggests holding a "large" position in the risky asset. As FSE increases, the uncertainty-averse investor would decrease her position on the risky asset, i.e., moving in the opposite direction relative to ω^{MV}. There is however, no zero-position. As forecast uncertainty increases, ω^{RS} converges to a "discounted" MV portfolio that relies solely on prior beliefs:

$$g(\infty) = \omega_{\pi}^{\mathrm{MV}} \frac{\sigma_T^2}{v + \sigma_T^2} \frac{\gamma}{\tau + \gamma},$$

Essentially, when forecast uncertainty is extremely high, the portfolio is predominantly

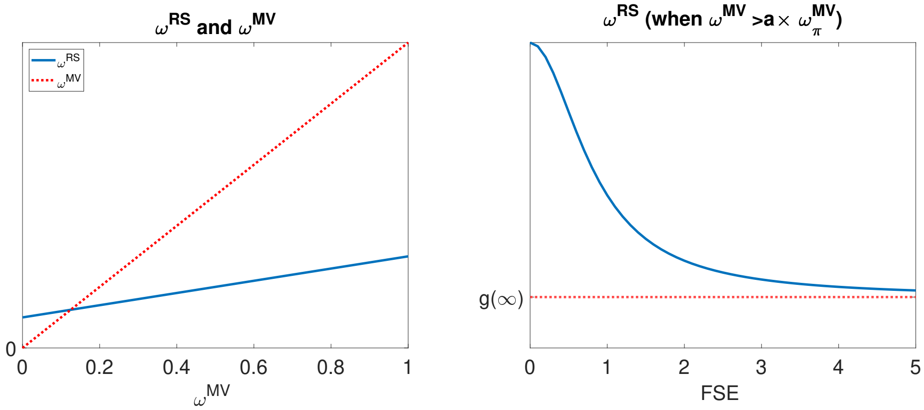

Figure 2: The ω^{RS} position. The left panel fixes the level of FSE and compares ω^{RS} with the regular ω^{MV}, whereas the right panel fixes the value of ω^{MV} and plots ω^{RS} as a function of FSE for the case $\omega^{\text{MV}} > a\omega_\pi^{\text{MV}}$.

guided by prior beliefs, minimizing the impact of the predicted expected returns.

4.2.2 Multiple risky assets

The framework outlined in the previous subsection can be extended to the scenario of multiple risky assets. We obtain the pooled ML forecast $\widehat{z}_{T+1|T}$ for the R-dimensional expected return $z_{T+1|T}$. Let SE^2 be the forecast covariance matrix of $\widehat{z}_{T+1|T}$, and let Σ_T be $R \times R$ covariance of $z_{T+1} - z_{T+1|T}$. We impose a multivariate normal prior $z_{T+1|T} \sim \mathcal{N}(\pi, v)$, where v is $R \times R$ prior covariance matrix, set to $v = g\Sigma_T$ for some $g > 0$ as in Zellner's g-prior. Then the posterior of the predictive density for the N-dimensional z_{T+1} is $f_T(z_{T+1}) \sim \mathcal{N}(\widetilde{z}_{T+1}, \widetilde{\Sigma})$, where for $W_1 = \text{SE}^2(\text{SE}^2 + v)^{-1}$,

$$\widetilde{z}_{T+1} = (I - W_1)\widehat{z}_{T+1|T} + W_1\pi, \quad \widetilde{\Sigma} = \Sigma_T + vW_1'.$$

Now consider the problem: for $p_T \sim \mathcal{N}(\mu, \widetilde{\Sigma})$:

$$\begin{aligned} \boldsymbol{\omega}^{\text{RS}} &:= \arg\max_{\boldsymbol{\omega} \in \mathbb{R}^N} \min_{\mu \in \mathbb{R}^N} \boldsymbol{\omega}'(\mu - 1_N r_f) - \frac{\gamma}{2}\boldsymbol{\omega}'\widetilde{\Sigma}\boldsymbol{\omega} + \frac{1}{\tau}D(p_T||f_T) \\ &= \frac{\gamma}{\tau + \gamma}\widetilde{\Sigma}^{-1}\Sigma_T\left[(I - W_1)'\boldsymbol{\omega}^{\text{MV}} + W_1'\boldsymbol{\omega}_\pi^{\text{MV}}\right] \qquad (4.10) \end{aligned}$$

where

$$\boldsymbol{\omega}^{\mathrm{MV}} = \Sigma_T^{-1}(\widehat{z}_{T+1|T} - 1_N r_f)\frac{1}{\gamma}, \quad \boldsymbol{\omega}_{\pi}^{\mathrm{MV}} = \Sigma_T^{-1}(\pi - 1_N r_f)\frac{1}{\gamma}$$

are respective the mean-variance portfolios based on $\widehat{z}_{T+1|T}$ and π. Hence ω^{RS} still conducts a *double shinkage*: First, it shrinks $\boldsymbol{\omega}^{\mathrm{MV}}$ towards $\boldsymbol{\omega}_{\pi}^{\mathrm{MV}}$ due to the weight W_1. Secondly, it shrinks the overall portfolio towards zero due to the factor $\widetilde{\Sigma}^{-1}\Sigma_T$.

Note that as $g \to \infty$, the prior becomes more diffusive, then $W_1 \to 0$, hence more weights are imposed on $\boldsymbol{\omega}^{\mathrm{MV}}$. The resulting portfolio $\boldsymbol{\omega}^{\mathrm{RS}}$ becomes less robust to the forecast uncertainty in $\widehat{z}_{T+1|T}$. We shall see the impact of this in the empirical study.

4.3 Comparing the two uncertainty-aversion strategies

It is interesting to note that unless FSE= 0, ω^{RS} will never hold zero positions in the risky asset. That is, the "no-holding" position never appears in the risk-sensitive optimization framework. As such, while both ω^{NH} and ω^{RS} are uncertainty-averse, the latter is less conservative in the sense of always holding a non-zero position on the risky asset, even though the position monotonically varies with the FSE. This feature can distinguish the application scopes of the two strategies for which each strategy is more suitable. The risk-sensitive strategy ω^{RS}, being less conservative, is suitable to scenarios when economists prefer to always investing on the risky asset to stimulate the economy growth. In contrast, the ω^{NH}, due to the no-holding position, is suitable when the economist is highly concerned about the forecast uncertainty.

Quantifying the uncertainty in machine learning forecasts is also crucial beyond financial economics, especially in applications that involve high-stakes decision-making. In autonomous driving (AD), for instance, complex sensing systems installed on autonomous vehicles predict the real-time locations of surrounding vehicles to assess whether it is safe to change its driving trajectory (as illustrated in the leftmost panel of Figure 3). Uncertainty estimation in such ML predictions helps enhance the safety and reliability of autonomous vehicles, as it allows the system to account for potential errors or anomalies in its predictions (Han et al., 2022; Su et al., 2023).

An uncertainty-averse approach in AD aims to handle prediction uncertainty by establishing an "uncertainty circle," similar to a confidence interval, around the forecast position of nearby vehicles. The AD system then plans a robust trajectory that completely avoids this uncertainty circle. By doing so, it takes a max-min approach, optimizing the trajectory to account for the worst-case scenario: the possibility that the front vehicle might be at any point within this uncertainty circle. This approach takes a conservative risky action that reduces the risk of collision or other hazards (as depicted in the middle panel of Figure 3).

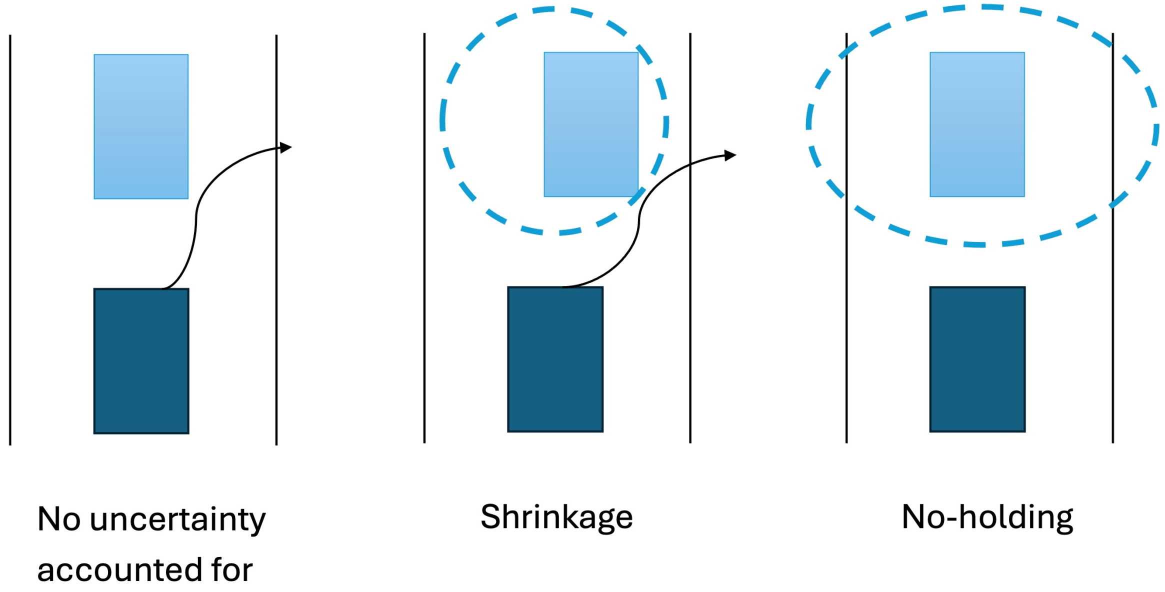

Figure 3: Uncertainty-aversion of the AD system. Left panel: AD designs a trajectory based on the forecast position of the front vehicle without considering uncertainty; middle panel: AD adjusts to a "shrinkage" trajectory for the worst-case position within the uncertainty circle; right panel: AD opts for a "no-holding" position when uncertainty is too high.

Furthermore, if the uncertainty circle becomes large enough to significantly overlap with an adjacent lane, the AD system will conclude that the optimal strategy is to remain in the current lane rather than adjusting its trajectory. In this scenario, the system opts for a no-action position or no-holding position (as shown in the rightmost panel). This conservative approach reflects a cautious, uncertainty-averse approach where the system avoids any risky actions associated with a lane change.

5 Empirical Analysis

5.1 Data

We take the dataset of Jensen et al. (2022) as our starting point. It uses stock returns, volume and price data from the Center for Research in Security prices (CRSP) monthly stock file. Following standard conventions in the literature, we restrict the analysis to common stocks of firms incorporated in the US trading on NYSE, Nasdaq or Amex. Balance sheet data is obtained from Compustat. In order to avoid potential forward looking biases, we lag all characteristics that build on Compustat annual by at least six months and all that build on Compustat quarterly by at least four months. In order to mitigate a potential back-filling bias as noted by Banz and Breen (1986), we discard the first 24 months for each firm.

We use moving windows for estimation and predicting the market excess returns, in which we fix the window size at $T = 240$ for estimation. The first prediction occurs for December 1964, and the last prediction is for December 2021.

5.2 Implementation

In each estimation window, we fit a DNN function $\widehat{g}(\cdot)$, which is a three-layer feedforward neural networks with 32, 16 and 8 neurons the hidden layers. After fitting the neural network, we substitute in the firm level characteristics $x_{i,T}$, and predict the individual stocks at month $T + 1$ as

$$\widehat{y}_{i,T+1|T} = \widehat{g}(x_{i,T}).$$

Note that the DNN function is fit using an unbalanced panel in the 240-month estimation window. We then focus on R risky assets $\boldsymbol{z}_{T+1} = (z_{T+1,1}, ..., z_{T+1,R})'$, where each asset is a portfolio: $z_{T+1,k} = W_k' Y_{T+1}$ with $Y_t = (y_{1,t}, ..., y_{N,t})'$. These risky asset returns are predicted by

$$\widehat{z}_{T+1|T,k} = W_k' \widehat{Y}_{T+1|T}, \quad \widehat{Y}_{T+1} = (\widehat{g}(x_{1,T}), ..., \widehat{g}(x_{N,T}))'. \tag{5.1}$$

To determine W_k, we use the principal components of sorted portfolios. To this end, we first construct $d = 110$ sorted portfolios $S_{j,t}' Y_{t+1}$, one from each characteristic:

$$S_{j,t}' = \left(x_{t,j}' M_t x_{t,j}\right)^{-1} x_{t,j}' M_t, \quad j = 1, ..., d$$

where $x_{t,j}$ is the N_t- dimensional characteristic vector of the j th characteristic, and $M_t = I - 1_{N_t}1'_{N_t}/N_t$ with N_t being the number of stocks at period t. Let $S_t = (S_{1,t}, ..., S_{d,t})_{N_t \times d}$ be the matrix of sorted portfolio weights. We then take $z_{t+1,k}$ as the k th principal component from $\{S'_t Y_{t+1}, t = 1, ..., T\}$ by computing ξ_k as the d-dimensional eigenvector corresponding to the k th largest eigenvalue. Hence

$$W'_k = \xi'_k S'_T / \sqrt{d}, \quad k = 1, ..., R.$$

We implement the "easy ML approximation" method to compute the forecast standard error, which uses a Fourier basis expansion $\phi(x) = (\sin(2j\pi x), \cos(2j\pi x), j = 1...4)$ and W_k as the portfolio weighting vector for the k the index. The standard error $\widehat{\mathrm{SE}}(\widehat{z}_{T+1,k})$ is then used for the uncertainty averse portfolio allocation. We set the risk-averse parameter $\gamma = 1$ throughout the study.

As for the volatility Σ_T, we apply the factor-based covariance estimator of Fan et al. (2013) to $S'_t Y_t$ and denote it by V. Then

$$\Sigma_T = \frac{1}{d} \xi' V \xi, \quad \xi = (\xi_1, ..., \xi_R)_{d \times R}.$$

We do not consider the uncertainty from estimating Σ_T in this paper, leaving the FCI of ML forecasts of the volatility for future research.

5.3 The out-of-sample performance of the no-holding strategy

The out-of-sample performance is assessed by the annualized Sharpe ratio, computed by the actual out-of-sample excess return $\boldsymbol{\omega}^{*'}(\boldsymbol{z}_{T+1} - r_f)$ from January 1983 to December 2021.

We implement the uncertainty-portfolio $\boldsymbol{\omega}^*$, setting the uncertainty level respectively as 25%, 50%, 75% and 95% confidence levels. As the confidence level becomes larger, the investor has higher uncertainty aversion, and is more cautiously investing based on the forecast mean return. The benchmark portfolio is taken as the usual mean-variance (MV) efficient portfolio, which is a normalized vector of $\Sigma_T^{-1} \widehat{z}_{T+1|T}$. Each chosen portfolio is normalized so that the annualized in-sample standard deviation is fixed to twenty percent in each estimation window.

Table 1 shows the annualized mean, standard deviation and Sharpe ratio of the out-of-

Table 1: Annualized Results of No-holding uncertainty-averse portfolios

This table shows annualized mean, standard deviation (STD), and Sharpe ratio (SR) of the out-of-sample return $\boldsymbol{\omega}'\boldsymbol{z}_{T+1}$. Here MV represents the mean-variance efficient portfolio and UA-MV represents the no-holding uncertainty-averse portfolios with confidence levels 25%, 50%, 75% and 95%. The specified R is the number of risky factor to invest. The forecast standard error is computed using 240 months in-sample data for each monthly forecast.

	MV and UA-MV portfolios									
	MV	25%	50%	75%	95%	MV	25%	50%	75%	95%
			$R = 20$					$R = 40$		
mean	0.715	0.703	0.681	0.651	0.583	0.932	0.916	0.887	0.837	0.721
STD	0.390	0.376	0.355	0.331	0.289	0.580	0.556	0.524	0.480	0.399
SR	1.831	1.871	1.917	1.968	2.018	1.607	1.648	1.693	1.742	1.805
			$R = 60$					$R = 110$		
mean	1.213	1.193	1.155	1.099	0.958	1.994	1.957	1.902	1.843	1.608
STD	0.740	0.713	0.677	0.621	0.524	1.147	1.111	1.052	0.970	0.817
SR	1.639	1.671	1.706	1.768	1.827	1.738	1.761	1.806	1.900	1.967

sample excess returns of these portfolios. In general, as we increase the confidence level, the investment is indeed more cautious, yielding less expected returns. Take $R = 60$ as an example, the out-of-sample annualized excess return is 1.213 for the MV-portfolio, and reduces to 0.958 for the 95% uncertainty-averse portfolio. Meanwhile, the uncertainty-averse investment yields reduced risk, with standard deviation reduces from 0.74 of the MV-portfolio to 0.524 of the 95% uncertainty-averse portfolio. The annualized Sharpe ratio increases as the level of uncertainty aversion increases.

The upper panels of Figure 4 plots the annualized mean, standard deviation (STD) and Sharpe Ratios of out-of-sample excess returns of the UA-portfolios versus various confidence levels. The label "MV" corresponds to the usual mean-variance portfolio directly using the DNN predicted return. Both mean and STD of the returns decrease as the level of confidence increases. The Sharpe ratio monotonically increases with the confidence level. The bottom panels of Figure 4 plots the same quantities with respect to the and number of indices R. Both mean and STD of excess returns increase as R increases. Meanwhile, the Sharpe ratio exhibits a "W" shape as R increases across all confidence levels. From the bottom right panel, the amount of change in Sharpe ratio with respect to R is less pronounced at higher confidence levels, indicating that tolerating more uncertainties help to stabilize the economic

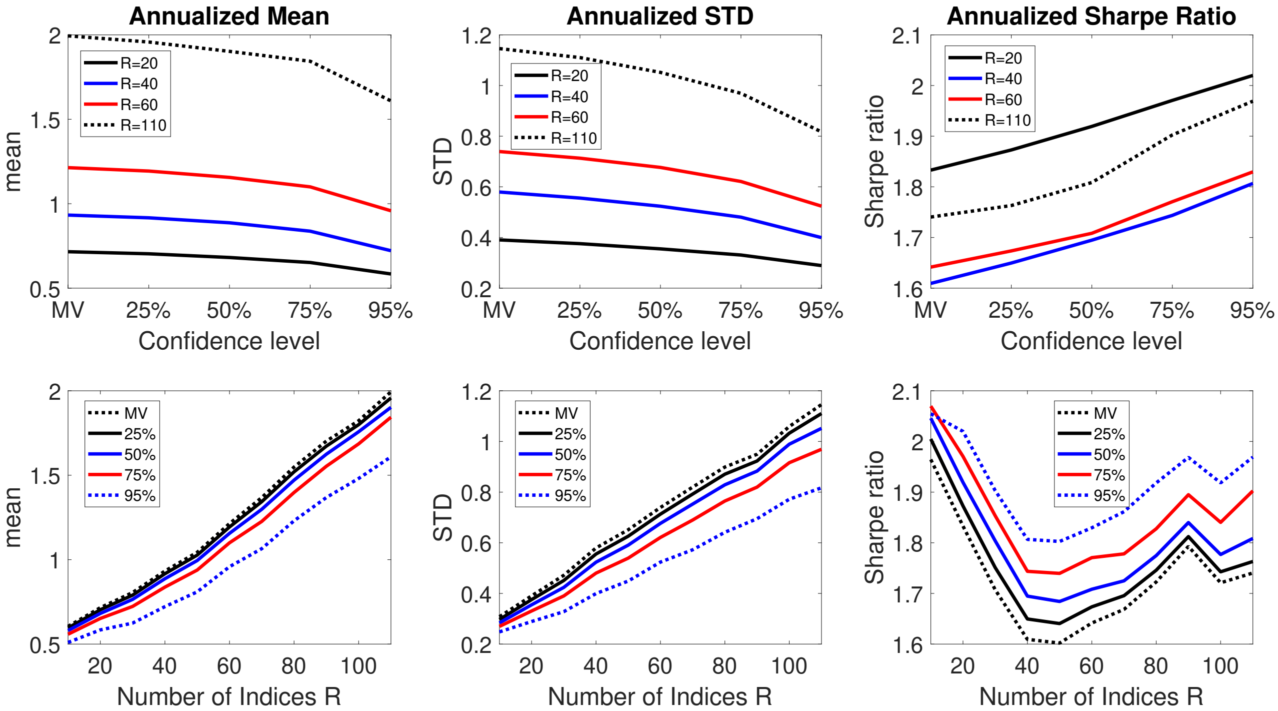

Figure 4: Upper panels: Annualized Mean, STD and Sharpe Ratios versus various confidence level; Bottom panels: Annualized Mean, STD and Sharpe Ratios versus R.

performance in managed portfolios.

Next, we analyze the impact on the no-holding position from forecast standard errors. We take $R = 110$ as an example, and compute the average forecast standard error, as well as the percentage of no-holding positions in the τ th estimation window:

$$\mathrm{FSE}_\tau = \frac{1}{R}\sum_{k=1}^{R} \widehat{\mathrm{SE}}(\widehat{z}_{T_\tau+1,k}), \quad \mathrm{NH}_\tau = \frac{1}{R}\sum_{k=1}^{R} 1\{|\omega^*_{T_\tau,k}| < c\},$$

where $z_{T_\tau+1,k}$ and $\omega^*_{T_\tau,k}$ respectively denote the forecast return and the UA-MV portfolio of the k the asset of the τ th estimation window. We set $c = 10^{-4}$ as the threshold to determine whether a position is "nearly zero". Figure 5 plots the 24-month moving averages of FSE_τ and NH_τ. Except for a brief period in 2003, the two series demonstrate roughly similar trend over most time periods.

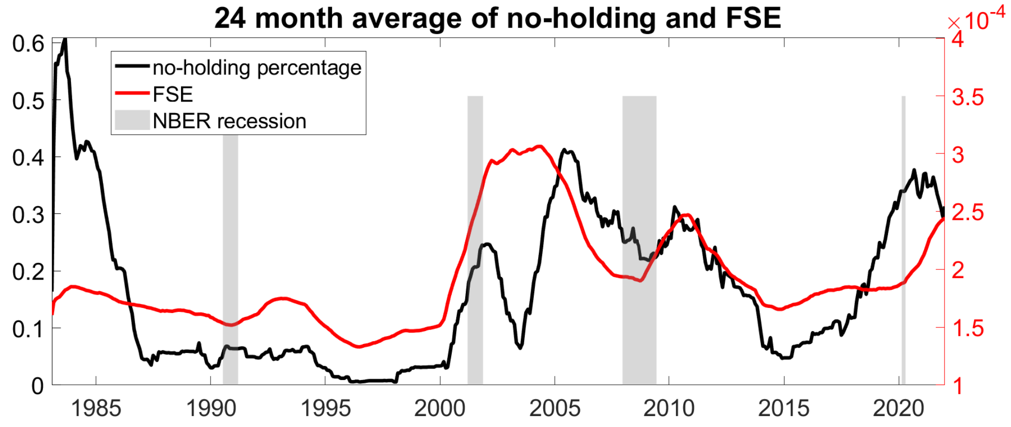

Figure 5: Moving averages of no-holding percentages and averaged FSE. The no-holding percentage is defined as NH_τ with $R = 110$ portfolios.

5.4 The Risk-Sensitive approach

We now implement the alternative uncertainty-averse portfolio using the Risk-Sensitive approach as described in (4.10). In each estimation window of size 240 months, we set the prior mean π as the in-sample average return of the portfolio z_t with Zellner's $g \in \{0.5, 1, 2\}$. Table 2 reports the annualized results.

Recall that as g increases, the UA-RS portfolio is closer to the regular MV- portfolio, hence becomes less robust to the forecast uncertainty in the predicted returns. This is clearly illustrated in Table 2: larger g in the prior yields larger out-of-sample risk in the managed portfolio, resulting smaller annualized Sharpe ratio in all settings.

6 Simulation

We conduct Monte Carlo simulations to assess the constructed confidence intervals, using a data generating process (DGP) calibrated from real excess return data, the monthly returns of 3184 firms from January 2015 through December 2017 (calibrated period). Simulated data are generated from a conditional three–factor model, with d characteristics as follows:

$$x_{i,t,k} = \frac{1}{N}\mathrm{rank}(\bar{x}_{i,t,k}), \quad \bar{x}_{i,t,k} = 0.7\bar{x}_{i,t-1,k} + 0.5\epsilon_{i,t,k}, \quad \epsilon_{i,t,k} \sim \mathcal{N}(0,1).$$

Table 2: Annualized Results of Risk-Sensitive uncertainty-averse portfolios

This table shows annualized mean, standard deviation (STD), and Sharpe ratio (SR) of the out-of-sample return $\boldsymbol{\omega}'\boldsymbol{z}_{T+1}$. Here MV represents the mean-variance efficient portfolio and UA-RS represents the uncertainty-averse Risk-sensitive portfolios, in (4.10) with prior $\mathcal{N}(\pi, g\Sigma_T)$. The specified R is the number of risky factor to invest. The forecast standard error is computed using 240 months in-sample data for each monthly forecast.

	MV and UA-RS portfolios							
	MV	$g = 0.5$	$g = 1$	$g = 2$	MV	$g = 0.5$	$g = 1$	$g = 2$
	$R = 20$				$R = 40$			
mean	0.715	0.732	0.726	0.724	0.932	0.946	0.937	0.932
STD	0.390	0.338	0.346	0.352	0.580	0.452	0.469	0.480
SR	1.831	2.162	2.096	2.056	1.607	2.091	1.997	1.939
	$R = 60$				$R = 110$			
mean	1.213	1.206	1.196	1.189	1.994	1.913	1.913	1.914
STD	0.740	0.548	0.570	0.584	1.147	0.716	0.743	0.765
SR	1.639	2.199	2.099	2.036	1.738	2.670	2.573	2.502

The characteristics are generated via AR(1), then normalized by taking the cross-sectional ranking. Characteristics within firm i have strong temporal dependence over time, but they are independent across firms. The β-functions are generated as follows:

$$g_{\beta,1}(x) = x_1 x_2, \quad g_{\beta,2}(x) = \frac{1}{d}\sum_{j=1}^{d} x_j^2, \quad g_{\beta,3}(x) = \text{median}\{x_1, ..., x_d\}.$$

The three factors are generated from a multivariate normal distribution whose mean-vector and covariance matrix are calibrated from the monthly return of Fama-French-three factors in the calibrated period. Finally, the idiosyncratic noises are generated from a heteroskedastic normal distribution: $u_{i,t} \sim \mathcal{N}(0, s_i^2\sigma^2)$, and $s_i \sim \text{Unif}[0.1, 0.9]$. Here we set σ so that $\text{Median}(s_i^2\sigma^2/\operatorname{Var}(y_{i,t})) = 50\%$. Therefore, the idiosyncratic variances are determined so that the overall signal-noise ratio is fifty percent.

Throughout we fix $N = 500$ firms, $T = 240$ periods and $d = 80$ characteristics. The goal is to forecast $z_{T+1} := \frac{1}{N}\sum_i y_{i,T+1}$ using pooled neural network, and examine the forecast distribution using proposed two methods. We train three-layer feedforward neural networks with 4 neurons on each layer.[6]

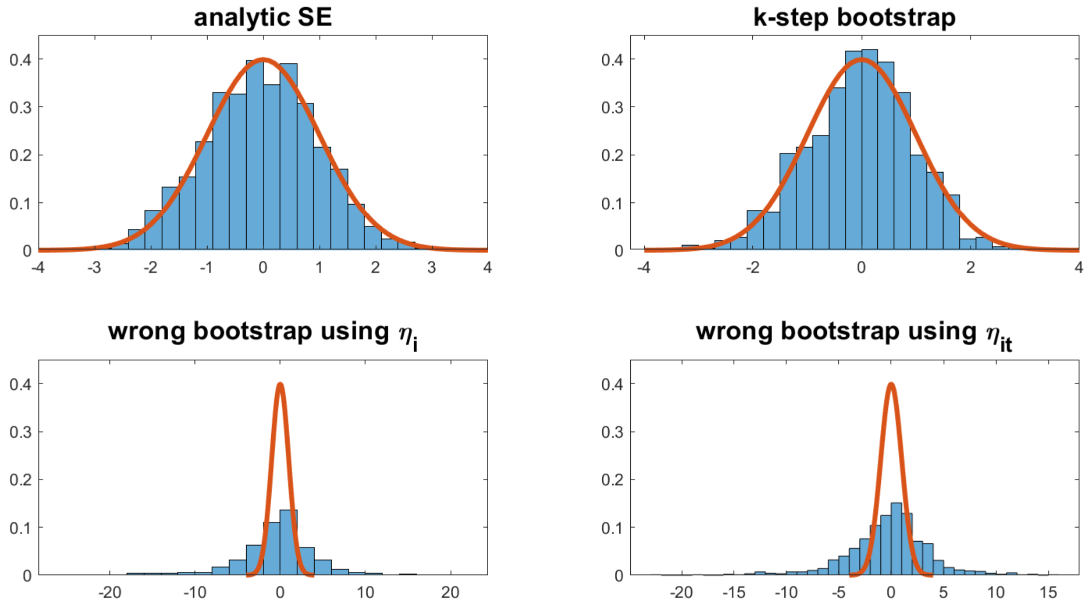

Figure 6: Histogram of t-statistics over 1000 simulation replications, and the standard normal density. The t-statistics are standardized by either the analytical standard error $\widehat{\mathrm{SE}}(\widehat{z}_{T+1})$ (top left panel) or the bootstrap interquartile range σ^* (top right panel). The bottom panels use η_i and η_{it} to generate bootstrap residuals.

As for quantifying the forecast uncertainty, we compute the forecast standard error of the neural networks, using either of the proposed methods. For method I "easier ML", we compute the t-statistic

$$\frac{\widehat{z}_{T+1|T} - z_{T+1|T}}{\widehat{\mathrm{SE}}(\widehat{z}_{T+1})},$$

where the standard error $\widehat{\mathrm{SE}}(\widehat{z}_{T+1})$ has an analytical form (3.5). Here we use five Fourier bases for $\Phi(x)$. For method II "k-step bootstrap", we generate the wild residual η_t from the standard normal, bootstrap 100 times, and implement the k-step DNN bootstrap with

[6] The training algorithm is `Adam` with learning rate 0.01. All codes are written in `Flux` on Julia.

$k = 10$. Then we compute the interquartile range of bootstrap, defined as

$$\sigma^* := \frac{q^*_{0.75} - q^*_{0.25}}{z_{0.75} - z_{0.25}}$$

where q^*_α denotes the α-quantile of bootstrap samples $\sum_i w_i \widehat{g}^{*,b}(x_{i,T}) - \widehat{z}_{T+1|T}$, and z_α denotes the α-quantile of the standard normal distribution. Then we also compute the t-statistic using σ^* in place of $\widehat{\mathrm{SE}}(\widehat{z}_{T+1})$. The interquantile range is a good proxy to the standard error obtained using bootstrap distribution, which is often used instead of the usual bootstrap standard error, because the former is guaranteed to be consistent but the latter is not.

The top two panels of Figure 6 plot the histograms of the t statistics over 1000 simulations, superimposed by the standard normal density function. The t-statistics are standardized by either $\widehat{\mathrm{SE}}(\widehat{z}_{T+1})$ (left panel) or σ^* (right panel). We see that although there are only 100 replications, the histograms of the t-statistics fit reasonably well to the standard normal density. Hence both proposed methods for quantifying the forecast uncertainty seem promising.

It is critical to let the bootstrap be guided by the theory of the proposed research, and to properly reflect that the forecast uncertainty should be driven by time-variations of factor-shocks. So we also compare the outcome if the bootstrap is misued. The bottom two panels of Figure 6 are the histograms of the bootstrap t-statistics (standardized by the bootstrap interquartile range), but the bootstrap residual is generated as $e^*_{i,t} = (y_{i,t} - \widehat{g}(x_{i,t-1}))\eta^*_i$ (the bottom left panel) and $e^*_{i,t} = (y_{i,t} - \widehat{g}(x_{i,t-1}))\eta^*_{i,t}$ (the bottom right panel), where $\eta^*_i, \eta^*_{i,t} \sim \mathcal{N}(0,1)$. These bootstraps mistreat the forecast uncertainty as mainly driven by the cross-sectional variations of idiosyncratic noises. Indeed, we see from Figure 6 that the misuse of bootstrap vastly understates the uncertainty.

7 Conclusion

Investors are often uncertainty-averse in asset pricing, seeking optimal portfolio allocations under ambiguity, particularly when faced with uncertainty in forecasting future returns using ML. A critical factor in this context is the forecast confidence interval (FCI) of ML predictions, which quantifies forecast uncertainty and plays a key role in uncertainty-averse asset pricing. In this paper, we develop a novel methodology for constructing FCIs for ML predictions of expected returns and establish their asymptotic validity. Remarkably, we find that the FCIs of sophisticated ML methods can be closely approximated by simpler nonparametric approaches. We then apply the proposed FCI methodology to two portfolio allocation frameworks under uncertainty aversion, rigorously characterizing the solutions as “no-holding” and “shrinkage” positions. These positions lead to different optimal behaviors compared to traditional approaches that do not account for uncertainty aversion.

References

Allena, R. (2021). Confident risk premiums and investments using machine learning uncertainties. *Available at SSRN 3956311*.

Anderson, E. W. and A.-R. Cheng (2016). Robust bayesian portfolio choices. *The Review of Financial Studies 29*(5), 1330–1375.

Andrews, D. W. (2002). Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators. *Econometrica 70*(1), 119–162.

Anthony, M. and P. L. Bartlett (2009). *Neural network learning: Theoretical foundations.* cambridge university press.

Avramov, D. (2002). Stock return predictability and model uncertainty. *Journal of Financial Economics 64*(3), 423–458.

Avramov, D., S. Cheng, L. Metzker, and S. Voigt (2023). Integrating factor models. *The Journal of Finance 78*(3), 1593–1646.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*, 1229–1279.

Banz, R. W. and W. J. Breen (1986). Sample-dependent results using accounting and market data: some evidence. *the Journal of Finance 41*(4), 779–793.

Bartlett, P. L., N. Harvey, C. Liaw, and A. Mehrabian (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res. 20*, 63–1.

Bauer, B. and M. Kohler (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics 47*(4), 2261–2285.

Bessembinder, H. (2018). Do stocks outperform treasury bills? *Journal of financial economics 129*(3), 440–457.

Bianchi, D., M. Büchner, and A. Tamoni (2021). Bond risk premiums with machine learning. *The Review of Financial Studies 34*(2), 1046–1089.

Chen, L., M. Pelger, and J. Zhu (2020). Deep learning in asset pricing. *Available at SSRN 3350138*.

Chen, X. and D. Pouzo (2015). Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica 83*(3), 1013–1079.

Chen, X. and X. Shen (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 289–314.

Chen, X. and H. White (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory 45*(2), 682–691.

Cheng, Q., T. Luo, and H. Yue (2013). Managerial incentives and management forecast precision. *The Accounting Review 88*(5), 1575–1602.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., W. Newey, and R. Singh (2018). De-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*.

Choi, J.-H., L. A. Myers, Y. Zang, and D. A. Ziebart (2011). Do management eps forecasts allow returns to reflect future earnings? implications for the continuation of management's quarterly earnings guidance. *Review of Accounting Studies 16*, 143–182.

Chordia, T., A. Goyal, and A. Saretto (2020). Anomalies and false rejections. *The Review of Financial Studies 33*(5), 2134–2179.

Ciconte, W., M. Kirk, and J. W. Tucker (2014). Does the midpoint of range earnings forecasts represent managers' expectations? *Review of Accounting Studies 19*, 628–660.

Davidson, R. and J. G. MacKinnon (1999). Bootstrap testing in nonlinear models. *International Economic Review 40*(2), 487–508.

DeMiguel, V., L. Garlappi, and R. Uppal (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The review of Financial studies 22*(5), 1915–1953.

Didisheim, A., S. B. Ke, B. T. Kelly, and S. Malamud (2023). Complexity in factor pricing models. Technical report, National Bureau of Economic Research.

Dow, J. and S. R. da Costa Werlang (1992). Uncertainty aversion, risk aversion, and the optimal choice of portfolio. *Econometrica 60*, 197–204.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics 75*(4), 643–669.

Epstein, L. G. and T. Wang (1994). Intertemporal asset pricing under knightian uncertainty. *Econometrica 62*(2), 283–322.

Fama, E. F. and K. R. French (2010). Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance 65*(5), 1915–1947.

Fan, J., Z. T. Ke, Y. Liao, and A. Neuhierl (2022). Structural deep learning in conditional asset pricing. *Available at SSRN 4117882*.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B 75*, 603–680.

Ferson, W. E. and C. R. Harvey (1999). Conditioning variables and the cross section of stock returns. *The Journal of Finance 54*(4), 1325–1360.

Freyberger, J. (2018). Non-parametric panel data models with interactive fixed effects. *The Review of Economic Studies 85*(3), 1824–1851.

Freyberger, J., A. Neuhierl, and M. Weber (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies 33*(5), 2326–2377.

Gagliardini, P., E. Ossola, and O. Scaillet (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica 84*(3), 985–1046.

Garlappi, L., R. Uppal, and T. Wang (2007). Portfolio selection with parameter and model uncertainty: A multi-prior approach. *The Review of Financial Studies 20*(1), 41–81.

Giglio, S., Y. Liao, and D. Xiu (2021). Thousands of alpha tests. *The Review of Financial Studies 34*(7), 3456–3496.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of mathematical economics 18*(2), 141–153.

Gilboa, I. and D. Schmeidler (1993). Updating ambiguous beliefs. *Journal of economic theory 59*(1), 33–49.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Han, S., S. Su, S. He, S. Han, H. Yang, and F. Miao (2022). What is the solution for state-adversarial multi-agent reinforcement learning? *arXiv preprint arXiv:2212.02705*.

Hansen, L. P. and T. J. Sargent (2008). *Robustness*. Princeton university press.

Harvey, C. R. and Y. Liu (2020). False (and missed) discoveries in financial economics. *The Journal of Finance 75*(5), 2503–2553.

Jagannathan, R., Y. Liao, and A. Neuhierl (2023). Robust market index forecast using deep learning.

Jagannathan, R. and T. Ma (2003). Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The journal of finance 58*(4), 1651–1683.

Jagannathan, R. and Z. Wang (1996). The conditional capm and the cross-section of expected returns. *The Journal of finance 51*(1), 3–53.

Jensen, T. I., B. Kelly, and L. H. Pedersen (2022). Is there a replication crisis in finance? *The Journal of Finance*.

Kelly, B. T., S. Malamud, and K. Zhou (2021). The virtue of complexity in return prediction. *Swiss Finance Institute Research Paper* (21-90).

Kelly, B. T., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Kohler, M. and S. Langer (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics 49*(4), 2231–2249.

Kosowski, R., A. Timmermann, R. Wermers, and H. White (2006). Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis. *The Journal of Finance 61*(6), 2551–2595.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, 1217–1241.

Liao, Y., X. Ma, A. Neuhierl, and Z. Shi (2024). Does noise hurt economic forecasts?

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 1349–1382.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical association 89*(428), 1303–1313.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, 571–587.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics 48*(4), 1875–1897.

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics 25*(6), 2555–2591.

Su, S., Y. Li, S. He, S. Han, C. Feng, C. Ding, and F. Miao (2023). Uncertainty quantification of collaborative detection for self-driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5588–5594. IEEE.

van der Vaart, A. and J. Wellner (1996). *Weak convergence and empirical processes* (The First Edition ed.). Springer.

A Assumptions and Technical Intuitions

A.1 Assumption on the ML space

We shall use two machine learning spaces: the "forecast ML" $\mathcal{G}_{\mathrm{ML}}$, which is the space for deep neural networks, and the "easier ML" $\mathcal{G}_B$, which is the B-spline to compute the standard error. Let $\mathcal{N}(\delta, \mathcal{G}, \|.\|_\infty)$ denote the entropy cover of $\mathcal{G}$, which is the smallest number of $\|.\|_\infty$-balls of radius δ to cover $\mathcal{G}$.

Assumption 1. *Conditioning on X, $(v_t, u_{i,t})$ and are independent over time and are sub-Gaussian. Also suppose $x_{i,t-1}$ are independent across i.*

Assumption 2. *The following conditions hold for both $\mathcal{G} = \mathcal{G}_{\mathrm{ML}}$ and $\mathcal{G} = \mathcal{G}_B$:*

(i) There is $p(\mathcal{G})$ so that the covering number satisfies: for any $\delta > 0$,

$$\mathcal{N}(\delta, \mathcal{G}, \|.\|_\infty) \leq \left(\frac{CT}{\delta}\right)^{p(\mathcal{G})}. \tag{A.1}$$

(ii) Let

$$\varphi_{\mathcal{G}}^2 = \inf_{v \in \mathcal{G}} \sup_{h \in \{\tau_T g + \mathcal{G}\} \cup \{g\}} \|v - h\|_\infty^2.$$

Suppose $\varphi_{\mathcal{G}} = o(T^{-1})$, $p(\mathcal{G}) \log(NT) = o(T^{1/2})$ and $\varphi_{\mathcal{G}} p(\mathcal{G}) \log(NT) = o(1)$.

(iii) Define the best approximation to the true g function under the $\|.\|_{L^2}$-norm:

$$g_{NT,\mathcal{G}} := \arg\min_{h \in \mathcal{G}} \frac{1}{NT} \sum_{it} \mathbb{E}(h(x_{i,t-1}) - g(x_{i,t-1}))^2.$$

Suppose $\sqrt{T} \max_{j \leq N} |g_{NT,\mathcal{G}}(x_{j,T}) - g(x_{j,T}))| = o_P(1)$.

(iv) Let J be the number of basis adopted by the "easier ML" method $\mathcal{G}_B$. Then $J^{5/2} = o(T)$ and $(\frac{p(\mathcal{G}_{\mathrm{ML}}) \log(T)}{T} + \varphi_{\mathcal{G}_{\mathrm{ML}}})J^2 \leq o(1)$.

The above assumption is well known to be satisfied by many well known classical nonparametric methods as well as modern machine learning spaces. For instance, if DNN is used as the sophisticated ML space $\mathcal{G}_{\mathrm{ML}}$, Schmidt-Hieber (2020) showed that a multilayer feedforward network with ReLu activation functions at each layer can well approximate a rich class of functions with compact support. It also follows from Anthony and Bartlett

(2009) that (A.1) holds with $p(\mathcal{G}_{\mathrm{ML}})$ as the pseudo dimension of the neural network, with $p(\mathcal{G}_{\mathrm{ML}}) = O(J^2M^2)$, where J, L are respectively the maximum width and depth of the network. In addition, if B-splines are used as the "easier ML" $\mathcal{G}_B$, it is well known that it can well approximate a Hölder class of smooth functions; and (A.1) holds with $p(\mathcal{G})$ being the number of basis functions.

A.2 Technical intuitions

In this subsection we briefly explain the technical intuition on how we derive the expression

$$\widehat{z}_{T+1|T} - z_{T+1|T} = \frac{1}{T}\sum_{t=1}^{T}\mathcal{A}_t + o_P(T^{-1/2})$$

and why the leading term $\mathcal{A}_t$ does not depend on the specific choice of the ML space.

The Riesz-representation plays a key role in the asymptotic analysis, and has been commonly used in the inferene for semiparametric models, e.g., Newey (1994); Shen (1997); Chen and Shen (1998); Chen and White (1999); Chernozhukov et al. (2018), among many others. The use of Riesz-representation allows to directly span the ML forecast using the least squares loss function, which also requires studying the object of interest in a Hilbert space. To do so, define an inner product:

$$\langle h_1, h_2 \rangle := \frac{1}{NT}\sum_{it}\mathbb{E}h_1(x_{i,t-1})h_2(x_{i,t-1})$$

where the expectation is taken jointly with respect to the serial and cross-sectional distribution of $x_{i,t-1}$, treating h_1, h_2 as fixed functions. Define

$$g_{NT,\mathcal{G}} := \arg\min_{h\in\mathcal{G}_{\mathrm{ML}}\cup\mathcal{G}_B}\|h-g\|_{L^2}^2,$$

which is the best approximation to the true g on the space $\mathcal{G}_{\mathrm{ML}} \cup \mathcal{G}_B$ under the norm $\|_{L^2}$. Then $\mathcal{A}_{NT} := \mathrm{span}(\mathcal{G}_{\mathrm{ML}} \cup \mathcal{G}_B - \{g_{NT,\mathcal{G}}\})$ is a finite dimensional Hilbert spaced endowed with the inner product $\langle \cdot, \cdot \rangle$. Next, evaluated at the out-of-sample characteristics $x_{i,T}$, define a sequence of linear functionals:

$$\mathcal{T}_i(h) := h(x_{i,T}), \quad i = 1, ..., N.$$

Because $\mathcal{T}_j$ is a linear functional it is always bounded on the finite dimensional Hilbert space $\mathcal{A}_{NT}$. The Riesz representation theorem then implies that there is a function $m_j^* \in \mathcal{A}_{NT}$, called Riesz representer, so that

$$\mathcal{T}_j(h) = \langle h, m_j^* \rangle, \quad \forall h \in \mathcal{A}_{NT}.$$

The key fact to our argument is that m_j^* does not depend on the specific machine learning space being used (whether $\mathcal{G}_{NT,\mathcal{G}}$ or $\mathcal{G}_B$) for $\widehat{z}_{T+1|T}$. It only depends on the joint distribution of $\{x_{i,t-1}\}$, the realization $\{x_{j,T}\}$ and the union space $\mathcal{G}_{\text{ML}} \cup \mathcal{G}_B$.

Next, using an argument for M-estimations (e.g. Theorem 3.2.5 of van der Vaart and Wellner (1996)), we show in Lemma 8 that uniformly for $j \leq N$,

$$\langle \widehat{g} - g, m_j^* \rangle = \frac{1}{NT} \sum_{it} e_{i,t} m_j^*(x_{i,t-1}) + o_P(T^{-1/2}).$$

Then heuristically,

$$\begin{aligned} \widehat{z}_{T+1|T} - z_{T+1|T} &= \sum_j w_j [\widehat{g}(x_{j,T}) - g(x_{j,T})] = \sum_j w_j \mathcal{T}_j(\widehat{g} - g) \\ &\approx \sum_j w_j \langle \widehat{g} - g, m_j^* \rangle \\ &\approx \frac{1}{NT} \sum_{it} \sum_j w_j e_{i,t} m_j^*(x_{i,t-1}). \end{aligned}$$

This yields the desired expansion with $\mathcal{A}_t \approx \frac{1}{N} \sum_i \sum_j w_j e_{i,t} m_j^*(x_{i,t-1})$. It is clear from this expression that $\mathcal{A}_t$ does not depend on the specific choice of the ML method.

Many papers in the literature, when adopting the Riesz representation theorem on the linear functional, lacks the rigor that the Riesz representer may not exist on an infinite-dimensional Hilbert space unless the linear functional is bounded with respect to the endowed norm from the inner product. The boundedness of the functional requires additional assumptions and technical arguments to verify. In our case, $\mathcal{A}$ is indeed infinite dimensional because the true g is. Meanwhile, the Riesz representer always exists on an finite dimensional Hilbert space, because linear functionals are always bounded on the finite dimensional space. This requires a careful argument in the proof, where we follow the guidelines in Chen and Pouzo (2015).