Robust Stock Index Return Predictions Using Deep Learning[*]

Ravi Jagannathan[†] Yuan Liao[‡] Andreas Neuhierl[§]

Abstract

We introduce a conditional machine learning approach to forecast the stock index return. Our approach is designed to work well for short-horizon forecasts to address the well-documented instability in predicting aggregate stock returns in long panels. We formally characterize the forecast standard errors to assess the uncertainty associated with our cross-sectional neural network predictions, which also enables us to explain the predictability of our model. The explainability covers both correctly and incorrectly assumed forecasting models, and stems from the forecast standard errors and out-of-sample R square. To explain the economic impacts of the economy's stability on the forecast quality, we introduce a "CDI" index defined as the correlation between firms' market value share and sales share, and show that it can well explain the forecast uncertainties, thus provides economic insights of the success and failure of machine learning based forecasting models.

Key words: machine learning explainability, uncertainty, forecast confidence intervals

[*] We thank Ken Judd, Jun Pan, Alex Horenstein, Maarten Meeuwis, Todd Clark, Francis Diebold, Andrew Patton, Seth Pruitt, Alberto Rossi, Linda Schilling, Allan Timmermann, Dacheng Xiu and conference and seminar participants at the Conference in Honor of Torben G. Andersen, the Society of Financial Econometrics Meetings, the 2023 SITE New Frontiers of Asset Pricing Meetings, SAIF, University of Miami, Washington University in St. Louis, the 2023 NBER time series conference, the 2023 Midwest Econometric Group meeting.

[†] Kellogg School of Management, Northwestern University, rjaganna@kellogg.northwestern.edu

[‡] Department of Economics, Rutgers University, yuan.liao@rutgers.edu

[§] Olin Business School, Washington University in St. Louis, andreas.neuhierl@wustl.edu

1 Introduction

This paper proposes a robust conditional machine learning (CML) approach that delivers asset return forecasts specifically for the case of short time series. To address forecast instability, it allows for variation in the relationship of variables over time. Our approach draws on the rich information in the cross-section rather than relying on long and stable time series relationships.

Cochrane (2011) emphasizes that understanding discount rate variation is the central question in current asset-pricing research. While discount rate variation is widely accepted, there is little consensus to what extent discount rates can be predicted. The forces that drive variation in discount rates are too vast to be all captured in simple model and instability in predictive relationships can be caused by fundamental factors such as shifts in technological innovation, changes in stock market participation, tax codes, macroeconomic uncertainty and expectations or improved risk sharing. Although the literature has extensively documented the variation in the relationships between variables across time, it poses a significant problem for many asset return prediction models that rely on time-stability. For example, principal component analysis or partial least squares (Kelly and Pruitt, 2013, 2015) both require a long time series to estimate parameters consistently.[1] The dependence of these approaches on long samples however makes them reliant on stable predictive relationships. The applicability of these methods, hence, often breaks down because most forecast models work within short-lived periods, then having modest return predictability, as concluded in Timmermann (2008). Prediction instability across time, as documented in Goyal and Welch (2003), Rapach and Wohar (2006), Lettau and Van Nieuwerburgh (2008) and Ang and Bekaert (2007) or unstable relationships between returns and financial ratios likewise constitute a challenge for common models in the academic return prediction literature since time-variation causes poor out of sample forecasting performance, see Welch and Goyal (2008); Goyal et al. (2021) and Dangl and Halling (2012). Also see Rossi (2021) for an excellent recent survey on forecast instability.

Modern machine learning (ML) methods have been successful at predicting asset returns (Gu et al. (2020); Kelly et al. (2021)) because ML estimators typically make fewer strong

[1] In informal terms, long time series are necessary to ensure that idiosyncratic noise will average out over time and not pollute the estimates.

functional form assumptions than traditional parametric models. Moreover, they may almost avoid the curse of dimensionality which is typically a challenge in high dimensional nonparametric models.[2] The standard machine learning approach ("unconditional machine learning"), however, trains the model using pooled data that aggregates returns on many assets over several years. Because these estimates rely on having a long time series of observations, unconditional machine learning models may not be rich enough to allow for sufficient variation in the relation between returns and predictors.

We develop a robust conditional forecasting method that does not require long time series of observations to estimate forecasting factors. By focusing on the aspect of robustness, we emphasize that the market index forecasts should not be too sensitive to possible instabilities/varyingness of firm level betas and idioscynratic volatilities. The key idea is to train machine learning models period-by-period, drawing on the rich information of a large cross-section of asset returns and observable firm characteristics. Building on the recent contributions in the econometrics literature (Fan et al., 2016, 2022), we estimate factors and loadings consistently over short samples, with the key intuition that these quantities should be estimated using expected rather than realized returns, which produce more stable predictions over time than most existing methods that rely on realized returns and long time series.

We apply period-by-period machine learning to estimate firm level expected returns. But unlike the usual time series approaches, our estimated expected returns also preserve the factor structure of the realized returns, which yields high-quality estimates of stock betas. Then following Kelly and Pruitt (2013), we use the statistical factors learned from valuation ratios. One novel implementation step in our procedure is that we *do not* directly estimate the book-to-market betas (BM-betas), but use the estimated stock betas as good *instrumental variables* for the BM-betas, even though these are generally different. The use of instrumental variables methods for estimating the forecasting factors, the key difference from methods in the existing literature, serves for our central purpose of forecasting robust to the use of small-T. Therefore, unlike the usual aggregated indices that utilize firms' market capitalizations as the weights, our BM-factor is constructed as a weighted average of firms' book-to-market, whereas the stock-betas are being used as the weights. The so-constructed

[2] Gu et al. (2020) compare numerous machine learning methods to forecast returns of individual assets and portfolios. See also Bianchi et al. (2021) for an application in forecasting bond risk premiums.

index is much less noisy and produces more stable forecasts over short horizons.

The economic significance of our approach is the *explainability* of the CML-based forecast, which stems from the forecast uncertainties. Unlike the usual linear forecasts, machine learning-based forecasting models are highly nonlinear, often used as black-boxes whose predictability is hard to explain. As the predictability of any given forecasting model varies over time, understanding the evolution of its success or failure is crucial for explaining the financial insights of the predictability. We provide a theoretical framework of the forecast standard errors (FSE) and out-of-sample R square (OOSR2) for CML, both yield the explainability that covers both correctly and incorrectly assumed forecasting models: the FSE measures the forecast uncertainties of CML-based methods for correctly specified models, whereas the OOSR2 shows the impact of model specifications. By examining the decomposition of both FSE and OOSR2, we show that the predictability can be explained by the evolution of various volatilities and forecast coefficients.

We rigorously develop the forecast standard error of neural-network based forecasters, with a solid theoretical foundation. This enables us to quantify the forecast uncertainty of the machine learning based method. The FSE has an intuitive interpretation, it arises from: the uncertainty for using estimated forecast coefficients and the uncertainty for using estimated factors. We derive the asymptotic distribution for the estimated factors that come from neural network estimation. In particular, the uncertainty from the estimate factors, which also depends on idiosyncratic volatility, is different from the parametric case where factors are estimated using standard PCA (Bai and Ng, 2006). In conventional PCA, a long time series is needed to remove the effect of idiosyncratic shocks. In contrast, when factors are estimated by neural networks, we solely require that the number of cross-sectional units shall be large, but the length of the time series can be finite. The distribution theory underlying our confidence intervals is an important next step in fostering our understanding of neural networks or general high dimensional models in finance, as brought forward by the well-known PPCA/IPCA methods, Fan et al. (2016); Kim et al. (2021); Kelly et al. (2019); Fan et al. (2022), and the autoencoder Gu et al. (2019).

Our analysis uncovers the relation between the forecast stability and *volatilitis of uncertainties*. To explain the economic insight of this relationship, we introduce a new measurement of the economy's stability, which we call "Creative Destruction Index" (CDI), defined as the cross-sectional correlation between the firms' market value share and sales share. It captures the changes of firms' exposure to different economic shocks, which may or may not coincide with peaks and drops of business cycles. Empirically we find that the CDI is well connected to the evolution of factor volatilities and FSE, thus provides economic insights of the success and failure of machine learning based forecasting models.

Our methodology is also related to the recent contribution made by Farmer et al. (2022), with the intuition that any single forecasting model is likely going to fail after some time due to economic competition. It is therefore likely that *no single model* will produce successful predictions all the time. They developed a "pocket forecast" framework to document that forecasting algorithms, when working by themselves, are successful only for a brief period of time, so they will perform in "pockets". Bianchi et al. (2023) also studied the effect of model averaging on the forecastability, and showed that model averaging can compete with economically motivated predictive regressions. In our context, for instance, our method may lose predictive power in some periods if the contemporaneous stock factors depend very weakly on lagged book-to-market factors, or if the contemporaneous idiosyncratic volatility is abnormally large, but may gain the predictability in other periods. Therefore, we rely on the forecast uncertainty, i.e., the forecast standard error, to balance two machine learning forecasts: the proposed conditional ML and the widely used unconditional ML.

Notation

Throughout the paper, we will use the notation $X_n \to^P X$ if random variable sequence X_n converges in probability to a limit X. We also denote by $X_n = o_P(1)$ if $X_n \to^P 0$. Finally, we denote by $X_n = O_P(a_n)$ if the stochastic order of X_n is a_n; specifically, for any $\epsilon > 0$, there is $C > 0$, such that $P(|X_n| > Ca_n) < \epsilon$.

2 The Model

The goal is to construct a forecasting model for the market index return y_{t+1}, defined as a weighted average of individual stock returns:

$$y_{t+1} = \sum_{i=1}^{N} w_{i,t} x_{i,t+1}.$$

where $x_{i,t+1}$ are firm level realized returns, each associated with a possibly time-varying weight $w_{i,t}$. We assume that each firm i is also associated with a firm level characteristic $v_{i,t}$, and both $x_{i,t+1}$ and $v_{i,t}$ are driven by conditional factor models as follows:

$$x_{i,t+1} = \beta_{i,t}' f_{t+1} + u_{i,t+1}, \tag{2.1}$$

$$v_{i,t} = \lambda_{i,t-1}' g_t + \eta_{i,t}. \tag{2.2}$$

We observe data of $(y_t, x_{i,t}, v_{i,t})$ for $t = 1, ..., T$, but the factors (f_t, g_t) are latent.

We specify $v_{i,t}$ as the firm-level book-to-market (BM), whose predictability on the market index has been well realized in the literature, e.g., Kelly and Pruitt (2013).[3] Here f_{t+1} and g_t (possibly overlapping) are respectively the factors that are driving stock returns and book-to-market ratios, whose dimensions are K_f and K_g, and we assume that $K_f \geq K_g$. Meanwhile, the idiosyncratic shocks, u_{it} and η_{it} can be correlated. We allow intercepts in both factor models, which are absorbed in the coefficients of $\beta_{i,t}$ and $\lambda_{i,t-1}$. These coefficients respectively denote the loadings of the two factor models. We consider a conditional factor model in which both $\beta_{i,t}$ and $\lambda_{i,t}$ may change over time, which is essential to achieving forecasts that are robust to market instabilities because it has been well known that asset pricing models can hold only conditionally, and a changing investment opportunity set can induce time-varying systematic risk exposures of assets, see e.g., Merton (1973); Hansen and Richard (1987). Therefore, incorporating the time variation is important for evaluation and testing of asset pricing models, see for example Shanken (1990), Jagannathan and Wang (1996) and Ferson and Harvey (1999).

Substituting (2.1) to the definition of y_{t+1}, we have

$$y_{t+1} = \widetilde{\rho}_{f,t}' f_{t+1} + \widetilde{\epsilon}_{t+1}, \tag{2.3}$$

where $\widetilde{\rho}_{f,t} = \sum_{i=1}^N \beta_{i,t} w_{i,t}$ and $\widetilde{\epsilon}_{t+1} = \sum_{i=1}^N u_{i,t+1} w_{i,t}$. Hence the market index is also driven by the stock-factors. Model (2.3) is however, not feasible for practical forecasts, because it depends on the contemporaneous factors.

We now discuss the main assumption of our model. First, there is temporal persistence

[3] We present the main model in terms of book-to-market ratios in model (2.2), but naturally any valuation ratio could also be employed. Our model can also admit multiple valuation ratios.

between the risk factors:

$$f_{t+1} = \Phi_0 + \Phi_g g_t + e_{t+1}, \tag{2.4}$$

with a coefficient matrix Φ_g. This equation shows that the common stock-factors are also driven by lagged BM-factors. [4] Substituting (2.4) for f_{t+1} in (2.3),

$$\begin{aligned} y_{t+1} &= \rho_{0,t} + \rho_{g,t}' g_t + \epsilon_{t+1}, \qquad (y_{t+1}: \text{ market index}) \qquad (2.5) \\ x_{i,t} &= \beta_{i,t-1}' f_t + u_{i,t}, \qquad (f_t: \text{ stock-factors}) \qquad (2.6) \\ v_{i,t} &= \lambda_{i,t-1}' g_t + \eta_{i,t}, \qquad (g_t: \text{ BM-factors}) \qquad (2.7) \end{aligned}$$

where $\rho_{0,t} = \widetilde{\rho}_{f,t}'\Phi_0$, $\rho_{g,t}' = \widetilde{\rho}_{f,t}'\Phi_g$, and $\epsilon_{t+1} = \widetilde{\rho}_{f,t}' e_{t+1} + \widetilde{\epsilon}_{t+1}$. As we do not observe either $x_{i,t+1}$ or f_{t+1} when forecasting y_{t+1}, this assumption allows us to apply a feasible forecasting model based on lagged factors. The key assumption here is that $(\rho_{0,t}, \rho_{g,t})$ is either constant or slowly moving over time.

Secondly, we assume that the factor loadings are functions of observable firm characteristics. More formally, there exists a (possibly time-varying) nonparametric function, $h_{\beta,t}(\cdot)$, such that

$$\beta_{i,t-1} = h_{\beta,t}(z_{i,t-1}). \quad \text{(stock betas)} \tag{2.8}$$

This assumption is also made in models of Connor et al. (2012); Fan et al. (2016); Kelly et al. (2020), who develop estimation procedures and asset pricing tests in characteristic based factor models.[5]

The third main assumption is that $\beta_{i,t-1}$ of stock returns is a good instrument to the $\lambda_{i,t-1}$ of book-to-market. Formally, we assume that the rank of the covariance matrix between

[4] A more general assumption is to allow lagged stock-factors: $f_{t+1} = \Phi_0 + \Phi_f f_t + \Phi_g g_t + e_{t+1}$, which would result in a prediction equation: $y_{t+1} = \rho_0 + \rho_f' f_t + \rho_g' g_t + \epsilon_{t+1}$ with $\rho_f' = \widetilde{\rho}_f' \Phi_f$. In essence, we apply a constraint that $\Phi_f = 0$ to only include the BM-factor as a predictor. Our motivation stems from the fact that stock-factors have very little persistence, whose inclusion in fact worsens the predictive performance.

[5] The intuition to link factor loadings to observable firm characteristics is already formulated in Rosenberg and McKibben (1973) and is also part of the Fama and French (2015) and related models. In addition, Ferson and Harvey (1999) find that the lagged characteristics have explanatory power for factor betas because they pick up time-variation in the factor loadings. Further evidence for the usefulness of firm characteristics to model systematic risk exposure is also part of Jagannathan and Wang (1996), Lettau and Ludvigson (2010).

stock and BM-betas is equal to the number of common BM-factors, K_g:

$$\frac{1}{N}\sum_{i=1}^{N}\beta_{i,t-1}\lambda_{i,t-1}', \quad K_f \times K_g.$$

This condition implies that $\beta_{i,t}$ and $\lambda_{i,t}$ are correlated, which is also the key benefit from including $x_{i,t}$ in the forecast model: the fact that the stock-beta carries information regarding the BM-beta allows us to use the former as the *instrumental variables* to estimate the BM-factors. This plays a central role of our innovative forecast methodology for robust short-horizon forecasts.

3 The Robust Forecast

3.1 Large-T versus small-T

Our forecasting model is based on

$$\begin{aligned} y_{t+1} &= \rho_{0,t} + \rho_{g,t}' g_t + \epsilon_{t+1}, \\ v_{i,t} &= \lambda_{i,t-1}' g_t + \eta_{i,t}, \quad t = 1, ..., T. \end{aligned}$$

Similar to Stock and Watson (2002) and Kelly and Pruitt (2013), we need to respectively estimate the latent factors g_t and the unknown forecasting coefficients $(\rho_{0,t}, \rho_{g,t})$. In the usual wisdom of forecasting y_{t+1} using time series however, the error-in-variables (EIV) problem arising from estimating these quantities would depend on the time series length through two kinds of volatilities:

$$\begin{aligned} &\text{estimating } (\rho_{0,t}, \rho_{g,t}): \quad \frac{1}{T}\operatorname{Var}(\epsilon_{t+1}) \\ &\text{estimating } g_t: \quad \frac{1}{T}\operatorname{Var}(\eta_{i,t}). \end{aligned}$$

So large-T has been critically required to offset the impact of both volatilities. In particular, estimating the factor g_t would require a large T, due to the EIV in estimating its factor loadings using most existing methods.

A key innovation in our methodology is to show that at least for estimating the latent factors, the “large-T” requirement can be relaxed for, meaning that the effect of the idiosyncratic volatility $\mathrm{Var}(\eta_{i,t})$ can be offset even if T is small. Taking advantage of conditional forecasts based on cross-sectional regressions using deep neural networks, we show that this is the case as long as the cross-sectional dimension is sufficiently large, regardless of the size of T.

As for estimating the forecasting coefficients $(\rho_{0,t}, \rho_{g,t})$, it appears that large- T is still needed. While estimating these coefficients requires time series regressions, nevertheless, having a more precise estimate of factors improves the finite sample properties. The empirical relevance of relaxing the large-T requirement when estimating factors is profound: on one hand, the impact of $\mathrm{Var}(\eta_{i,t})$ is much more severe than $\mathrm{Var}(\epsilon_{t+1})$, as the idiosyncratic volatility is less stable over time, and may have structural breaks much more often. On the other hand, allowing short-T to estimate BM-factors makes the forecast be also robust to changes in betas (of either stocks, characteristics, or both) over time. Hence being robust to the instability of the idiosyncratic volatility is critical for achieving the robustness of market index forecasts.

3.2 Formal algorithm

We propose a conditional machine learning approach (CML) to forecasting aggregate returns builds on neural networks embedded in a characteristic based factor models and follows four steps:

I Estimate expected returns by applying deep neural network regression of stock returns onto characteristics each period.

II Apply “local principal component analysis” (local PCA) on the estimated expected returns to estimate the stock betas.

III Estimate the book-to-market factors by using the stock betas as *instrumental variables*, and conduct forecasts.

IV Construct forecast confidence intervals to quantify the uncertainty of the predictions.

While steps I and II build on the theory developed in Fan et al. (2016, 2022), our primary methodological innovations are in steps III and IV, i.e. we show how to estimate g_t using neural networks without estimating its factor loadings directly. In addition, step IV develops novel forecast confidence intervals for cross-sectional deep neural network (DNN) forecasts in asset pricing.

We now detail the four steps.

Algorithm. Construct a forecast for future aggregate returns (conditional on time T information), $\widehat{y}_{T+1|T}$, as follows:

step I (Conditional DNN) Compute the expected stock returns

$$\widehat{x}_{i,t} = \widehat{m}_t(z_{i,t-1})$$

where $\widehat{m}_t(\cdot)$ is constructed using cross sectional neural network regression at each period $t = 1, ..., T$:

$$\widehat{m}_t(\cdot) = \arg\min_{m \in \text{DNN}} \sum_{i=1}^{N} (x_{i,t} - m(z_{i,t-1}))^2.$$

step II (Local-PCA) Let

$$S_t := \frac{1}{T} \sum_{s=1}^{T} \widehat{x}_s \widehat{x}_s' K_{s,t}$$

where $K_{s,t}$ is a time-dependent weight, and $\widehat{x}_s$ denotes the N-dimensional vector of the expected returns with elements $\widehat{x}_{i,s}$.

Estimate Stock-betas using $\widehat{\beta}_{t-1}$, which equals $\sqrt{N}$ times the $N \times K_f$ eigenvector matrix of S_t, corresponding to the top K_f eigenvalues.

step III (Forecast y_{T+1})

Factors: Let $\widehat{\lambda}^{\text{IV}}_{t-1}$ denote the first K_g columns of $\widehat{\beta}_{t-1}$, ($K_f \geq K_g$). Then estimate BM-factors by:

$$\widehat{g}_t = \frac{1}{N} \sum_{i=1}^{N} \widehat{\lambda}^{\text{IV}}_{i,t-1} v_{i,t}. \tag{3.1}$$

Index: Forecast the index by

$$\widehat{y}_{T+1|T} := \widehat{\rho}_0 + \widehat{\rho}_g' \widehat{g}_T$$

where $(\widehat{\rho}, \widehat{\rho}_g)$ are obtained by the following time-series regression:

$$y_{t+1} = \widehat{\rho}_0 + \widehat{\rho}_g' \widehat{g}_t + u_{i,t+1}$$[6]

step IV (Forecast uncertainty) Construct the forecast confidence interval for the expected index return $y_{T+1|T} := \rho_0 + \rho_g' g_T$ as:

$$\left[\widehat{y}_{T+1|T} - z_\tau \mathrm{SE}(\widehat{y}_{T+1|T}), \quad \widehat{y}_{T+1|T} + z_\tau \mathrm{SE}(\widehat{y}_{T+1|T})\right] \tag{3.2}$$

where z_τ is the $1-\tau$ critical value for the standard normal distribution; $\mathrm{SE}(\widehat{y}_{T+1|T})$ is the forecast standard error, whose exact expression is given in Section 4.1.

The following subsections give detailed explanations of each step in the algorithm.

3.3 An illustrative example

To illustrate how the method works in practice, consider a scenario of a single factor in both $y_{i,t}$ and $v_{i,t}$. In the first step, at each period t, we conduct cross-sectional neural network regression to reach $\widehat{x}_{i,t} = \widehat{m}_t(z_{i,t-1})$, where $\widehat{m}_t$ is the DNN function learned in this period, using cross-sectional returns $y_{i,t-1}$ and $z_{i,t-1}$ for all $i \leq N$. Now suppose the kernel function is simply:

$$K_{s,t} = \begin{cases} 1 & \text{if } s = t \\ 0 & \text{if } s \neq t. \end{cases}$$

That is, we only use fitted expected return $\widehat{x}_t = (\widehat{x}_{1,t}, ..., \widehat{x}_{N,t})$ at one observation, and obtain the "conditional covariance"

$$S_t = \frac{1}{T} \sum_s \widehat{x}_s \widehat{x}_s' K_{s,t} = \frac{1}{T} \widehat{x}_t \widehat{x}_t'.$$

[6] The time series regression in this step treats the cofficients (ρ_0, ρ_g) to be time-invariant. One can also apply kernel smoothing regression to estimate time-varying coefficients. Empirically, we find that treating (ρ_0, ρ_g) to be time-invariant leads to better forecasts than treating them to be time-invariant.

In the second step, let $\widehat{\beta}_{t-1}$ be the first eigenvector of S_t, which is given as

$$\widehat{\beta}_{t-1} = \frac{\sqrt{N}}{\|\widehat{x}_t\|}\widehat{x}_t.$$

Then in step III, we estimate the BM-factors as

$$\widehat{g}_t = \frac{1}{\|\widehat{x}_t\|\sqrt{N}}\sum_{i=1}^{N}\widehat{x}_{i,t}v_{i,t},$$

and forecast $y_{T+1|T}$ by conducting time series regression of y_{t+1} onto $\widehat{g}_t$ with intercept. Then straightforward calcuations yield

$$\widehat{y}_{T+1|T} = Y\widehat{G}(\widehat{G}'\widehat{G})^{-1}\widehat{G}_T$$

where $\widehat{G}$ is $(T-1)\times 2$ matrix of $(\widehat{g}_t, 1)$ and $\widehat{G}_T = (\widehat{g}_T, 1)'$.

3.4 Intuitions of the algorithm

3.4.1 Expected returns from deep neural networks

While firm level stock returns carry valuable information for predicting market returns, as emphasized by Polk et al. (2006), firm-level returns can be very noisy. This typically leads to EIV problems for estimating betas and factors and severely distorts predictive information. Our solution is work with the firm-level *conditional expected returns* (CER) given the characteristics, which is free of idiosyncratic noise but preserves the factor structure:

$$\mathbb{E}_t(x_{i,t}|z_{i,t-1}) = h_{\beta,t}(z_{i,t-1})'f_t. \text{ [7]} \tag{3.3}$$

To estimate expected returns we apply deep neural networks by regressing the excess

[7] The conditional expectation $\mathbb{E}_t(\cdot|z_{i,t-1})$ is taken with respect to the cross-sectional distributions at a fixed and give period t, so it is subscripted by t.

stock returns onto characteristics:

$$\widehat{m}_t(\cdot) = \arg\min_{m \in \text{DNN}} \sum_{i=1}^{N} (x_{i,t} - m(z_{i,t-1}))^2, \quad t = 1, ..., T \tag{3.4}$$

where "DNN" denotes a set of feedforward neural networks with predetermined layers and number of neurons in each layer. The optimization (3.4) is carried out to train the parameters in the DNN. We then estimate $\mathbb{E}_t(x_{i,t}|z_{i,t-1})$ by substituting the characteristics to the learned function:

$$\widehat{x}_{i,t} := \widehat{m}_t(z_{i,t-1}). \tag{3.5}$$

It is important to note that the neural networks in (3.4) are trained period-by-period, so each network is fitted using cross-sectional regressions only. This would entail that

$$\widehat{x}_{i,t} \to^P h_{\beta,t}(z_{i,t-1})' f_t, \quad \text{as } N \to \infty,$$

hence preserving the factor structure in the CER. The preserved factor structure is valuable to construct factor-based forecasts, which we shall explain next.

3.4.2 Local PCA

Given that the expected returns have the same factor structure as realized returns, we have:

$$\widehat{x}_{i,t} \approx h_{\beta,t}(z_{i,t-1})' f_t = \beta_{i,t-1}' f_t.$$

Since factor-loadings are varying over time, we apply local principal components analysis (local-PCA) to estimate factors and betas. The local-PCA estimates $\beta_{i,t}$ as eigenvectors of the weighted covariance matrix:

$$S_t := \frac{1}{T} \sum_{s=1}^{T} \widehat{x}_s \widehat{x}_s' K_{s,t}$$

where $K_{s,t}$ is a time-dependent weight, and $\widehat{x}_s$ denotes the N-dimensional vector of the expected returns $\widehat{x}_s = (\widehat{x}_{1,s}, ..., \widehat{x}_{N,s})'$. Then the estimated stock betas at period $t-1$, denoted by $\widehat{\beta}_{t-1}$, are the $N \times K_f$ eigenvector matrix of S_t, corresponding to its largest K_f eigenvalues.

The idea behind the local-PCA is that it gives more weights to observations during periods closer to the time of estimation interest. We create S_t being the weighted average of $\widehat{x}_s\widehat{x}_s'$ where $K(s,t)$ is chosen to be close to zero when s and t are far apart. So expected returns $\widehat{x}_s$ are effectively contributing to the calculation of S_t only if $s \approx t$. For those "close periods" s,

$$\widehat{x}_s = \beta_{t-1}' f_s + o_P(1), \quad s \approx t, \tag{3.6}$$

which approximately is also a factor model, but only when s is close to t. Let $S_{F,t} := \frac{1}{T}\sum_{s=1}^T f_s f_s' K_{s,t}$. Then (3.6) implies

$$S_t = \beta_{t-1} S_{F,t} \beta_{t-1}' + o_P(1).$$

This equation shows that the idiosyncratic term is almost negligible, the remainder is $o_P(1)$, i.e. it vanishes in the limit. More concretely, the noise is negligible so long as the cross-sectional deep neural network yields a good approximation to the CER, which is typically the case as long as $N \to \infty$. Therefore, taking $\widehat{\beta}_{t-1}$ as eigenvectors of S_t leads to a good estimate for β_{t-1}. It is clear that this approach also takes into account the time-varying nature of betas.

As for the kernel, define

$$K_{s,t} = \frac{1}{h} K\left(\frac{s-t}{Th}\right) A_t^{-1}, \quad A_t := \frac{1}{Th}\sum_{l=1}^T K\left(\frac{l-t}{Th}\right). \tag{3.7}$$

Here $K(\cdot)$ is a predetermined baseline kernel function with h being the bandwidth, which is a well established technique in nonparametric econometrics. We apply the two-sided quartic kernel:

$$K(x) = \frac{15}{16}(1-x^2)^2, \quad -1 \leq x \leq 1.$$

To avoid the use of forward-looking information in out-of-sample forecasts, we apply a boundary adjustment for $t = T$. The boundary adjustments is described in greater detail in the appendix (see also Li and Racine (2007)). By virtue of this adjustment our forecasts are strictly out-of-sample and can be made in real time.

Farmer et al. (2022) applied kernel smoothing regression to identify "pockets", i.e. short episodes of predictability, which is essentially a robust time-series approach. Our adoption of kernel smoothing builds on their intuition that parameters stay close within a short period of time, and is also related to Ang and Kristensen (2012).

Meanwhile, we develop this idea further by adapting the kernel smoothing to the period-by-period machine learning, which eliminates most of the idiosyncratic noise, followed by local-PCA. We show that local-PCA leads to a consistent estimator for beta, i.e.:

$$\|\widehat{\beta}_{i,t-1} - \beta_{i,t-1}H\| = o_P(1)$$

for some rotation matrix H. Importantly, the "$o_P(1)$" term is vanishing as long as $N \to \infty$, regardless of the size of T.

3.5 Constructing the BM-factors

We rely on the predictability from $\widehat{g}_t$ to forecast the market index. Step III in our forecast algorithm constructs it via:

$$\widehat{g}_t \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\widehat{\lambda}_{i,t-1}^{\mathrm{IV}} v_{i,t}. \tag{3.8}$$

where $\widehat{\lambda}_{i,t-1}^{\mathrm{IV}}$ are the estimated stock-betas. This subsection explores the predictability of this step.

3.5.1 The instrumental variable approach

Note that the BM-factors g_t is the common factor of book-to-market ratios (or more generally some valuation ratio), from the model:

$$v_{i,t} = \lambda_{i,t-1}' g_t + \eta_{i,t}, \qquad (g_t: \text{ BM-factors}). \tag{3.9}$$

Our methodological innovation is to estimate g_t without the large-T requirement, while producing a factor estimator that is robust to instabilities in the idiosyncratic volatility.

The key idea is to avoid estimating λ_{t-1} directly, but use the estimated stock betas β_{t-1}

as *instrumental variables* for the true λ_{t-1}. Specifically, we recall that in the previous step, K_f stock-betas are estimated in the local PCA step (we also assume $K_f \geq K_g$). We then use a sub-vector of $\widehat{\beta}_{i,t-1}$ consisting of its first K_g elements to construct a vector of IV:

$$\widehat{\lambda}^{\text{IV}}_{i,t-1} := (\widehat{\beta}_{i,t-1,1}, ..., \widehat{\beta}_{i,t-1,K_g}).$$

Next, use $\widehat{\lambda}^{\text{IV}}_{i,t-1}$ to estimate g_t as follows:

$$\widehat{g}_t = \left(\sum_{i=1}^{N} \widehat{\lambda}^{\text{IV}}_{i,t-1} \widehat{\lambda}^{\text{IV}\,\prime}_{i,t-1} \right)^{-1} \sum_{i=1}^{N} \widehat{\lambda}^{\text{IV}}_{i,t-1} v_{i,t} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\lambda}^{\text{IV}}_{i,t-1} v_{i,t}. \tag{3.10}$$

where the second equality follows since $\widehat{\lambda}^{\text{IV}}_{i,t-1}$ are also eigenvectors.

Clearly, we are not estimating the true λ_{t-1} to construct the BM-factors, and it is important to note λ_{t-1} need *not* lie in the span of β_{t-1}. Instead, we are using $\widehat{\lambda}^{\text{IV}}_{t-1}$, or essentially β_{t-1}, as IV for the true λ_{t-1}, which is the key difference of our approach compared to existing methods in the literature. Like the usual IV in linear regressions, all is needed is to satisfy the two conditions commonly imposed on IV:

(1) relevance: β_{t-1} should be correlated with λ_{t-1}.

(2) exogeneity: β_{t-1} should be orthogonal to the error term $\eta_{i,t}$.

The relevance is a plausible condition because both the stock and BM-betas are assumed to depend on the same firm-specific characteristics, whereas the exogeneity condition is also reasonably satisfied if these characteristics are exogenous.

We now illustrate how the use of IV approach helps address the error-in-variables problem. From (3.9), we have

$$\widehat{g}_t = \frac{1}{N} \sum_{i=1}^{N} \widehat{\lambda}^{\text{IV}}_{i,t} \lambda'_{i,t} g_t + \underbrace{\frac{1}{N} \sum_{i=1}^{N} \widehat{\lambda}^{\text{IV}}_{i,t} \eta_{i,t}}_{\text{statistical error}} \tag{3.11}$$

Let

$$H_g := \frac{1}{N} \sum_{i=1}^{N} \widehat{\lambda}^{\text{IV}}_{i,t-1} \lambda'_{i,t-1}, \quad \widetilde{\eta}_t := \frac{1}{N} \sum_{i=1}^{N} \widehat{\lambda}^{\text{IV}}_{i,t} \eta_{i,t}.$$

The relevance condition would ensure that H_g is a full rank matrix, and the exogeneity condition also ensures $\widetilde{\eta}_t$ disappears as $N \to \infty$. Hence equation (3.11) readily implies:

$$\widehat{g}_t - H_g g_t = o_P(1). \tag{3.12}$$

This means $\widehat{g}_t$ consistently estimates the true BM-factor g_t up to a rotation matrix H_g. Yet, the statistical error $\widetilde{\eta}_t$ vanishes without requiring large-T, as it depends on the idiosyncratic volatility only through:

$$\frac{1}{N}\operatorname{Var}(\eta_{i,t})$$

and the EIV inside $\widehat{\lambda}^{\mathrm{IV}}_{t-1}$ also vanishes as $N \to \infty$, thanks to the use of cross-sectional deep learning. Even in periods when $\operatorname{Var}(\eta_{i,t})$ is large, the impact of EIV and idiosyncratic volatility is still well controlled as long as there are sufficiently many individual stocks.

3.5.2 Economic Interpretation of the $\widehat{g}_t$-factor

Financial indices/aggregated state variables are often constructed as weighted averages, where one of the commonly used weights is the firm level market capitalization. For instance, one can construct the "aggregate BM" by taking a weighted average

$$\widetilde{g}_{t,abm} = \sum_{i=1}^{N} \widetilde{w}_{it,mk} v_{i,t}, \quad w_{it,mk} = \text{ market capitalization for firm } i.$$

In contrast, our instrumental variable approach constructs $\widehat{g}_t$ as an average, weighted by the stock-betas. Therefore in essence our predictor differs from the market-cap aggregation in the choice of the weights.

One economic interpretation of using the stock-betas as the weights stems from the intuition of utilizing expected stock returns instead of realized returns for predictions. The market capitalization contains realized stock returns also idiosyncratic noise. These noises often do not carry predictive signals of the market index, making the aggregate BM-factor a very noisy predictor. In contrast, the stock-betas are important components constituting to the conditional expected returns, which by definition is much less noisy. As such, our constructed BM-factors is much cleaner than market-cap weighted aggregations, and yet retains the predictability of market index returns.

3.5.3 Large-T robust concerns of existing methods

We now explain from the perspective of robust forecasts, why we should estimate the BM-factors by incorporating the information from $x_{i,t}$ (through the stock-betas). Suppose otherwise the BM-beta/factors are estimated directly from:

$$v_{i,t} = \lambda_{i,t}' g_t + \eta_{i,t}. \tag{3.13}$$

using either ordinary PCA or partial least squares (PLS, e.g., Kelly and Pruitt (2013)).[8] To illustrate the main issues, suppose for now that $\lambda_{i,t}$ does not vary over time. Denote by $\widetilde{\lambda}_i, \widetilde{g}_t$ as the estimated λ_i, g_t. The statistical error in λ_i would depend on $\breve{\eta}_i = \frac{1}{T}\sum_t g_t \eta_{i,t}$, and as a result, the EIV in the estimated BM-betas would critically depend on the idiosyncratic volatility through:

$$\widetilde{\lambda}_i - \lambda_i = O_P\left(\sqrt{\frac{\operatorname{Var}(\eta_{i,t})}{T}}\right).$$

This EIV carries over to estimating g_t, whose statistical error can be shown as

$$\frac{1}{N}\sum_i \eta_{i,t}\breve{\eta}_i \approx \frac{1}{T}\left(\frac{1}{N}\sum_{i=1}^N g_t \operatorname{Var}(\eta_{i,t})\right).$$

So it would require a large T to offset the EIV problem arising from the impact of the idiosyncratic volatility. Hence estimating the BM-factor without the information from $x_{i,t}$ leads to forecasts non-robust to the volatility instabilities.

4 Forecast Uncertainty and Explainability

Quantifying the uncertainty around point forecasts is crucial because in an unstable environment, forecasts may be far away from their target and could thus be highly unreliable. For example, the Bank of England Inflation Report routinely reports confidence intervals around their predictions via fan charts, where the fan charts are obtained as percentiles

[8] We acknowledge that our model is designed for forecasting index returns, while the approach in Kelly and Pruitt (2013) also applies to forecasting other outcomes such as the cash flow growth predictions. Nevertheless, the key idea of using CML-based forecasts to alleviate the limitation of short-horizon forecasts can still be generally applicable.

from a sequence of forecast densities. It is however, difficult to construct fan charts when sophisticated neural networks are used in forecasting. Statistical inference for DNN is technically very demanding and typically not readily available. While neural network based forecasts are used frequently in asset pricing, to the best of our knowledge, there is currently no paper that quantifies forecast uncertainty formally.[9]

In this section, we construct forecast intervals by rigorously deriving the forecast standard errors (FSE) for the estimated factors that come from neural network estimations, while being robust to the time-variation in betas and idiosyncratic volatility.[10] The emphasis of our analysis is that the uncertainty study also provides a natural economic explanation of the machine learning forecastability.

The importance of quantifying forecast uncertainty is for economically explaining the success and failure of machine learning predictability. We distinguish two kinds of periods for the economic explainability: when the factor model is corrected specified and when it is not. For the former, we shall introduce a "creative destruction index" which can well explain the evolution of FSE. For the latter, we shall conduct a dynamic out-of-sample R^2 analysis, which is useful to explain the impact of model misspecification on the predictability.

4.1 Forecast Confidence Intervals

As discussed earlier, the central goal of this paper is to achieve forecasts that are robust to changes in idiosyncratic volatilities. As we will see below, the forecast confidence interval explicitly shows that the impact of idiosyncratic volatilities is offset by using large-N, instead of large-T.

In the context of our model, let $\rho_t = (\rho_{0,t}, \rho_{g,t}')'$, $F_T = (1, g_T')'$. Proposition 1 below shows:

$$\widehat{y}_{T+1|T} \to^P y_{T+1|T} := \rho_T' F_T. \tag{4.1}$$

The standard error of $\widehat{y}_{T+1|T}$ arises from two sources of uncertainty: a) the error in estimating

[9] In the forecast literature, non-confidence based uncertainty measures have been also introduced, e.g., Rossi and Sekhposyan (2015); Carriero et al. (2018); Clark et al. (2020).

[10] Statistical convergence theory for DNN is still in its infancy. Since the pioneering contribution of Chen and White (1999), important recent contributions are due to Schmidt-Hieber (2020) and Kohler and Langer (2021). But the forecast confidence interval derived in this paper, or results of this kind, are not available prior to our work.

the coefficients of the prediction equation. Quantifying this error follows directly from linear regression theory and is straightforward. b) The uncertainty from estimating factors. This is challenging in our setting as these factors are estimated with neural networks.

Despite the technical sophistication of DNN-based forecasts, we obtain a relatively simple and intuitive expression for the forecast confidence interval, which is established in the following Proposition.

Proposition 1. *Suppose Assumptions 1-4 in the appendix hold. Then* $\widehat{y}_{T+1|T} \to^P y_{T+1|T} := \rho' F_T$. *In addition,*

$$\frac{\widehat{y}_{T+1|T} - y_{T+1|T}}{\sqrt{\frac{1}{Th} F_T' \overline{\mathrm{Var}}_\rho F_T + \frac{1}{N} \rho_{g_T}' \overline{\mathrm{Var}}_F \rho_{g_T}}} \to^d \mathcal{N}(0,1)$$

for some covariance matrices $\overline{\mathrm{Var}}_\rho$ *and* $\overline{\mathrm{Var}}_F$ *that can be consistently estimated by (4.4) below.*

With this result in hand, we can construct a forecast confidence interval for $y_{T+1|T}$ using the estimated forecast standard error:

$$\left[\widehat{y}_{T+1|T} - z_\tau \mathrm{SE}(\widehat{y}_{T+1|T}), \quad \widehat{y}_{T+1|T} + z_\tau \mathrm{SE}(\widehat{y}_{T+1|T})\right], \tag{4.2}$$

where z_τ is the $1-\tau$ critical value for the standard normal distribution. Let $\widehat{\rho}_g$ and $\widehat{F}_T$ denote the estimators for ρ_{gT} and F_T obtained in steps 2-3 of the main algorithm in Section 3.2. The squared standard error is then given by:

$$\mathrm{SE}(\widehat{y}_{T+1|T})^2 := \frac{1}{Th} \widehat{F}_T' \widehat{\mathrm{Var}}_\rho \widehat{F}_T + \frac{1}{N} \widehat{\rho}_g' \widehat{\mathrm{Var}}_F \widehat{\rho}_g \tag{4.3}$$

where

$$\begin{aligned} \widehat{\mathrm{Var}}_\rho &= \left(\frac{1}{T}\sum_t \widehat{F}_t \widehat{F}_t' K_{t,T}\right)^{-1} \frac{1}{T}\sum_t \widehat{F}_t \widehat{F}_t' \widehat{\epsilon}_{t+1}^2 K_{t,T} \left(\frac{1}{T}\sum_t \widehat{F}_t \widehat{F}_t' K_{t,T}\right)^{-1} \\ \widehat{\mathrm{Var}}_F &= \frac{1}{N}\sum_i \widehat{\eta}_{i,T}^2 \widehat{\lambda}_{i,t-1}^{\mathrm{IV}} \widehat{\lambda}_{i,t-1}^{\mathrm{IV}\,\prime}. \end{aligned} \tag{4.4}$$

Here $\widehat{F}_t = (1, \widehat{g}_T')'$, and $\widehat{\epsilon}_t$ and $\widehat{\eta}_{i,T}$ are the estimated residuals. To gain some intuition behind the standard error formula (4.3), we can write: (with a rotation matrix H_g)

$$\widehat{y}_{T+1|T} - y_{T+1|T} = (\widehat{\rho}_g - H_g^{-1}\rho_{g,T})' H_g F_T + \widehat{\rho}_g'(\widehat{F}_T - H_g F_T). \tag{4.5}$$

The forecast standard error takes into account two sources of forecast uncertainty: the first component $\frac{1}{Th}\widehat{F}_T'\widehat{\text{Var}}_\rho \widehat{F}_T$ arises from the uncertainty of $(\widehat{\rho} - H_g^{-1}\rho)$. The second component $\frac{1}{N}\widehat{\rho}_g'\widehat{\text{Var}}_F\widehat{\rho}_g$ arises from the uncertainty of the estimated factors $(\widehat{F}_T - H_g F_T)$. Analyzing its effect is non-standard as it combines several sources of uncertainty from the following steps: (i) period-by-period neural networks regression to obtain expected returns; (ii) apply local PCA to estimate stock-factors; (iii) use stock-beta as instrumental variables.

The standard error for neural networks is vastly different from standard errors in parametric forecast models. For example, in a forecast model based on PCA the forecast confidence interval is developed by Bai and Ng (2006). The estimation error of our method and that of PCA based forecasts differs strongly due to the different errors in estimating factors, the term $(\widehat{F}_T - H_g F_T)$ in equation (4.5). For our DNN-based factors, we have the following rates of convergence:

$$\widehat{F}_T - H_g F_T = O_P\left(\frac{1}{\sqrt{N}}\right) + o_P\left(\frac{1}{\sqrt{N}}\right)\text{Var}(\eta_{i,t}),$$

where the second term on the right hand side, $o_P\left(\frac{1}{\sqrt{N}}\right)\text{Var}(\eta_{i,t})$, only depends on the number of cross-sectional units and the complexity of the neural networks, but does not depend on the length of the time series (T). In contrast, let $\widehat{F}_{T,\text{PCA}}$ denote the PCA-based factors. Then Bai and Ng (2006) show that

$$\widehat{F}_{T,\text{PCA}} - H_g F_T = O_P\left(\frac{1}{\sqrt{N}}\right) + O_P\left(\frac{1}{T}\right)\text{Var}(\eta_{i,t})$$

where the second term on the right hand side depends on the variance of the idiosyncratic shocks and the length of the time series. Therefore, the parametric confidence interval can be severely affected by the idiosyncratic variance in short-horizon forecasts, While both methods depend on the idiosyncratic variance $\text{Var}(\eta_{i,t})$, the DNN-based method is robust to this variance as long as N is large.

4.2 Economic Explainability

To explain the economic impacts of the economy's stability on the forecast quality, we introduce a new measurement of the economy's stability, which we call "Creative Destruction Index" (CDI). It is defined as cross-sectional correlation of rankings between firms' sales and market capitalization:

$$\mathrm{CDI}_t = \mathrm{Corr}_t(S_t, M_t)$$

where Corr_t takes the sample correlation, at time t, between cross-sectional elements of $S_t = (S_{1,t}, ..., S_{N,t})$ and $M_t = (M_{1,t}, ..., M_{N,t})$. Here $S_{i,t}$ and $M_{i,t}$ respectively denote the cross-sectional rankings of the sales and market capitalization for firm i. The CDI is motivated from the intuition that the creative structure changes the exposure of firms to different shocks. When the economy is stable without much innovative technological shocks, the ranking of sales share and ranking of market value shares should be highly correlated. Meanwhile, in periods when the economy undergoes technological innovation, emerging firms will have lower sales in ranking but higher market values, whereas declining firms will have higher sales but lower market values in ranking.

The CDI may or may not coincide with peaks and drops of business cycles, but it affects the predictability through the "volatility of forecast uncertainty". In their empirical study, Pesaran and Timmermann (1995) found that a decrease of forecastability often coincides with a significant increase in the standard errors of the forecasting equations, suggesting that the predictability can be explained by the forecast standard error. One of the appealing features of our CML model is that the forecast standard error can be derived, which is a good measure of the forecast uncertainty. The variations of the forecast standard error with CDI therefore can be used to explain the success and failure of forecasts over time.

As shown in Proposition 1, the forecast standard error depends on $\frac{1}{N}\rho_g' \overline{\mathrm{Var}}_F \rho_g + \frac{1}{Th} F_T' \overline{\mathrm{Var}}_\rho F_T$: The first term arises from the uncertainty of estimating the factors, which depends on the idiosyncratic volatility

$$\frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}(\eta_{i,T}).$$

The predictability is relatively high/low at periods when the idiosyncratic volatility is large/small. In addition, the second term $F_T' \overline{\mathrm{Var}}_\rho F_T$, is essentially determined by g_T^2, the squared realized factor in the forecast period. Large values of g_T^2 would also harm the predictability. Therefore, the evolution of both the idiosyncratic volatility and squared realized factors can explain the dynamic performance of the CML-based forecast over time.

Moving on to the pooled ML forecast, whose explainability can be explained by examining the prediction error. [11] As derived in our theory, one of the main sources of its forecasting error is

$$\text{factor realization} = \rho_g' \left[g_T - \mathbb{E}(g_T | \mathcal{F}_{z,T}) \right],$$

whose magnitude is determined by the time series variance $\mathrm{Var}(\rho_g g_T)$. We refer to this term as the "factor-impact volatility". It is important to note that the factor-impact volatility is related, however different, from the factor volatility defined as $\mathrm{Var}(g_T)$. The former is the volatility of the interaction of the forecast coefficient and the factors, so reflecting the effect of lagged factors on predictions, while the latter is determined by the factor itself. Above all, the evolution of the predictability can be explained by its magnitude over time: small values of this volatility should explain the superior of the pooled ML, and vice versa, and should vary over time.

4.3 The OOS R^2 Process

The above explainability stems from the assumption that the forecasting model is correctly specified, which uses our theoretical framework of forecast uncertainty and statistical errors. Meanwhile, we can also explain the impact of potential misspecifications, through the out-of-sample R^2 analysis (OOSR2).

The most common measure of forecast accuracy is the out-of-sample R^2 (OOSR2), defined as:

$$R_t^2 := 1 - \frac{\sum_{s \in \mathcal{S}_t} (y_{s+1} - \widehat{y}_{s+1|s})^2}{\sum_{s \in \mathcal{S}_t} (y_{s+1} - \bar{y}_s)^2}$$

where $\widehat{y}_{s+1|s}$ denotes the one-step-ahead forecast using data up to period s and $\mathcal{S}_t$ is a set of out-of-sample observations. The forecast is then compared with $\bar{y}_s$, the in-sample average. While many papers report this quantity to assess the success of predictive procedures, often only a *single number* is reported. It is therefore a concern, that "t" maybe chosen

[11] To date, studying the forecast standard error of the pooled ML is still an open question, which we shall leave for future research.

somewhat arbitrarily. While this is certainly a useful statistic for the full sample, it does not capture the variation of model specification over time. In this subsection, we provide a novel decomposition of the OOSR2, which explains the impact of model specification on the forecastability.

4.3.1 R^2 Decomposition

From the definition of R_t^2, it is clear that it is also a time series and we can study its evolution over time. Intuitively, a robust forecast $\widehat{y}_{s+1|s}$ should have a stable R_t^2 process, whereas a turbulent trajectory of R_t^2 indicates that the forecast may not be robust. We formalize this intuition below. Suppose a researcher chooses a model M for forecasting y_{s+1} based on the information set $\mathcal{F}_s$, and produces a conditional forecast $\widehat{y}_{s+1|s}$. We can then decompose the forecast error into:

$$\begin{aligned} y_{s+1} - \widehat{y}_{s+1|s} &= \epsilon_{s+1} + \zeta_{s+1} \\ & \quad \text{where} \\ \epsilon_{s+1} &= y_{s+1} - \mathbb{E}(y_{s+1}|\mathcal{F}_s) = \text{innovation shocks} \\ \zeta_{s+1} &= \mathbb{E}(y_{s+1}|\mathcal{F}_s) - \widehat{y}_{s+1|s} = \text{misspecification of the forecast model.} \end{aligned} \tag{4.6}$$

The first term is the difference between the outcome and the true conditional expectation, $\mathbb{E}(y_{s+1}|\mathcal{F}_s)$. This is akin to the irreducible error. The second term, ζ_{s+1}, arises because a researcher typically does not know the true conditional expectation and is using a model, M, to approximate it. Since any model will typically be misspecified it constitutes a second sources of error.[12] Because a particular model, M, is a deliberate choice by the researcher, regardless of whether $\widehat{y}_{s+1|s}$ is consistent for the true conditional mean, it is often the case as the out-of-sample period becomes very long, we have a limit:

$$\frac{1}{|\mathcal{S}_t|}\sum_{s\in\mathcal{S}_t}\mathbb{E}(\zeta_{s+1}^2) \to \sigma_\zeta^2(t, \mathrm{M})$$

[12] To give a simple example for these terms, suppose the return is generated by $y_{s+1} = x_s\beta + \epsilon_{s+1}$ for some linear regressor x_s. Then $\mathbb{E}(y_{s+1}|\mathcal{F}_s) = x_s\beta$. Suppose we use a "wrong" beta, denoted by β_1, to forecast the return for $s+1$, then $\widehat{y}_{s+1|s} = x_s\beta_1$ in this case, and $\zeta_{s+1} = x_s(\beta - \beta_1)$ is the misspecification of the forecast. As a result $\sigma_\zeta^2(t, \mathrm{M}) = \frac{1}{|\mathcal{S}_t|}\sum_{s\in\mathcal{S}_t}(x_s\beta - x_s\beta_1)^2$.

where the limit depends on the model M being used:

$$\sigma_\zeta^2(t, \mathrm{M}) = \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} \left(\mathbb{E}(y_{s+1}|\mathcal{F}_s) - \mathbb{E}(y_{s+1}|\mathcal{F}_s, \mathrm{M})\right)^2.$$

Here $\mathbb{E}(y_{s+1}|\mathcal{F}_s, \mathrm{M})$ is the probability limit of $\widehat{y}_{s+1|s}$, the forecast outcome produced using model M; $|\mathcal{S}_t|$ denotes the number of elements in $\mathcal{S}_t$. If the model were correctly specified, the limit would be zero. But if the true model is very complicated and we have to confine ourselves to some class of functions, then there will always be some specification error. Therefore, depending on whether the model specification is correct, $\sigma_\zeta^2(t, \mathrm{M})$ can vary from zero to a large quantity over time.

Define the variance of the innovation in y as:

$$\sigma_\epsilon^2(t) := \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}} \mathrm{Var}(\epsilon_{s+1}).$$

Note that it does not depend on the forecast model, M. Moreover, because the two components in decomposition (4.6) are uncorrelated, we have

$$\begin{aligned} \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} (y_{s+1} - \widehat{y}_{s+1|s})^2 &= \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}} \mathrm{Var}(\epsilon_{s+1}) + \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} \mathbb{E}(\zeta_{s+1}^2) + o_P(1) \\ &\to \sigma_\epsilon^2(t) + \sigma_\zeta^2(t, M). \end{aligned}$$

Meanwhile, we can express the limit of the unconditional forecast error as:

$$\frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_s} (y_{s+1} - \bar{y}_s)^2 \to^P \sigma_{un}^2(t).$$

The conditional forecast using the true $\mathbb{E}(y_{s+1}|\mathcal{F}_s)$ is almost always better than the unconditional mean forecast, so we almost always have $\sigma_{un}^2(t) > \sigma_\epsilon^2(t)$. We can then define the *benefits from conditional forecasts*:

$$\delta_t := \sigma_{un}^2(t) - \sigma_\epsilon^2(t) > 0.$$

Thus we have proved that for some negligible error term $\Delta_t = o_P(1)$,

$$R_t^2 = \theta_t + \Delta_t,$$

where

$$\theta_t = \frac{1}{\sigma_{un}^2(t)} \left[\delta_t - \sigma_\zeta^2(t, \mathrm{M})\right]. \tag{4.7}$$

From (4.7) we can see that both the sign and magnitude of R_t^2 depend the sign of $\delta_t - \sigma_\zeta^2(t, \mathrm{M})$. Therefore, the practice of examining OOSR2 to assess forecasts is essentially about comparing the quality of the specified model $\sigma_\zeta^2(t, \mathrm{M})$ with the benefit of conditional forecasts δ_t. If the specified forecasting model, M, is doing a good out-of-sample job so that $\sigma_\zeta^2(t, \mathrm{M})$ is smaller than the benefits from conditional forecasts, OOSR2 will be positive, otherwise OOSR2 will be negative: the model is so badly mis-specified that the cost exceeds the benefits from conditional forecasts. Such a comparison apparently may vary across t, so is a dynamic comparison. We shall study this dynamic nature more carefully in the next subsection.

Pesaran and Timmermann (1995) empirically documented that the change of return volatility can also affect the predictive power. One of the interesting empirical observations they found is that both the volatility in the U.S. stock market and the predictability of stock returns increased around 1974, and conjectured that this is related to the economic "regime switches". Our analysis in (4.7) provides a theoretical ground of their empirical finding. Specifically, the predictability, measured by the OOSR2 sequence, depends on the volatility through the term:

$$\frac{\delta_t}{\sigma_{un}^2(t)} = \frac{\sigma_{un}^2(t) - \sigma_\epsilon^2(t)}{\sigma_{un}^2(t)}.$$

Hence the impact of economic regime switches on the predictability is pronounced through the ratio of the *benefits from conditional forecast* to the *unconditional volatility*, instead of either the $\sigma_{un}^2(t)$ or $\sigma_\epsilon^2(t)$ by itself.

4.3.2 Connection to standard errors: impacts of misspecification

Recall that the OOSR2 decomposition explicitly depends on the impact of misspecification through $\sigma_\zeta^2(t, \mathrm{M})$, which measures the discrepancy between the true conditional expected return $\mathbb{E}(y_{t+1}|\mathcal{I}_t)$ and the model-specified expected return $\mathbb{E}(y_{t+1}|\mathcal{I}_t, \mathrm{M})$. While it is challenging to directly evaluate $\sigma_{\zeta}^2(t, \mathrm{M})$ without the knowledge of the true expected return $\mathbb{E}(y_{s+1}|\mathcal{I}_s)$, it is however possible to indirectly assess its impact by examining the evolution of the forecast coefficient ρ_t. The degree of evolution of ρ_t can be measured by its time series variance, which can be computed using the moving window variance of $\widehat{\rho}_t$.

$$\mathrm{Var}(\rho_t) = \frac{1}{L}\sum_{s=t-L}^{t}(\widehat{\rho}_s - \bar{\rho}_t)^2, \quad \bar{\rho}_t = \frac{1}{L}\sum_{s=t-L}^{t}\widehat{\rho}_s$$

where $\widehat{\rho}_s$ is the estimated forecast coefficient in the s th forecast rolling window, and L is a predetermined number of rolling estimators. Large/low values of $\mathrm{Var}(\rho_t)$ is a good measure of assessing the degree of misspecifications of our forecasting model.

4.3.3 R^2 Structural Break Tests

The previous subsection shows that the out-of-sample R_t^2 depends on weighing the benefits of conditional forecasts vs. the drawbacks of misspecfication, i.e. comparing δ_t and $\sigma_{\zeta}^2(t, \mathrm{M})$. Naturally both components vary over time, but if a forecasting model is good and robust, it should not vary too strongly over time, otherwise, at periods when the forecasting method is not robust to the change of forecasting environment, R_t^2 may possess *structural breaks.*

We re-write (4.7) as

$$R_t^2 = a_t - b_t(\mathrm{M}) + \Delta_t, \text{ where } \quad a_t = \frac{\delta_t}{\sigma_{un}^2(t)}, \quad b_t(\mathrm{M}) = \frac{\sigma_{\zeta}^2(t, M)}{\sigma_{un}^2(t)}, \tag{4.8}$$

and Δ_t is the statistical error, which is negligible in the discussion below. The decomposition (4.8) identifies two sources of structural breaks on the time series of R_t^2:

"common-break": breaks on a_t, the benefits from conditional forecasts
"model-break": breaks on $b_t(\mathrm{M})$, the specific error in the forecast model.

While both terms $a_t, b_t(\mathrm{M})$ may possess structural breaks over time, the former is shared by all models, and the latter is model-specific. The robustness is mainly assessed by comparing the structural breaks on the second component among different forecast models.

On one hand, if the forecasting model, M, is robust to market instabilities, so that $\widehat{y}_{s+1|s}$ is close to the true conditional expected mean $\mathbb{E}(y_{s+1}|\mathcal{F}_s)$ during most of periods, then $b_t(\mathrm{M})$ should be close to zero. This would lead to $R_t^2 \approx a_t$, so the OOSR2 time series may only possess the common-break. On the other hand, if the model is not robust to market instabilities/breaks, the large discrepancy between in-sample and out-of-sample $\mathbb{E}(y_{t+1}|\mathcal{F}_t)$ would make $b_t(\mathrm{M})$ be a quite unstable process, then in addition to the common-break, the OOSR2 series would also contain model specific breaks. [13] Consequently, models that lead to more robust forecasts would contain less structural breaks on the OOSR2 series on $b_t(\mathrm{M})$.

The following proposition formalizes the above discussions. Let $h(\frac{t}{T}) = (1, \frac{t}{T}, ..., (\frac{t}{T})^k)$ be a polynomial trending function up to some order $k \geq 0$.

Proposition 2. *Suppose Assumption 6 in the appendix holds. Then*

(i) $R_t^2 = a_t - b_t(\mathrm{M}) + o_P(1)$.

(ii) Suppose both a_t and $b_t(\mathrm{M})$ have time trends meaning that they can be written as $a_t = c'h(\frac{t}{T})$ and $b_t(\mathrm{M}) = d'h(\frac{t}{T})$, where c and d are vectors of trending coefficients, and either may subject to multiple structural breaks at periods $(\tau_1, ..., \tau_q)$. Then

$$R_t^2 = \mu' h\left(\frac{t}{T}\right) + o_P(1),$$

where μ is a vector of trending coefficients that have multiple structural breaks at periods $(\tau_1, ..., \tau_q)$.

In Proposition 2, we say a *trending coefficient* vector c associated with a time series X_t can have multiple *structural breaks* at periods $(\tau_1, ..., \tau_q)$ if it can be represented as

$$X_t = c'h\left(\frac{t}{T}\right) \quad \text{where } c = \begin{cases} c_1 & t \leq \tau_1 \\ c_2 & \tau_1 < t \leq \tau_2 \\ \vdots & \\ c_{q+1} & t > \tau_q \end{cases}.$$

[13] There is a third case, where the model is *constantly bad* so that $|b_t(\mathrm{M})|$ is stably large over time. This model is robust, but uninteresting and we do not analyze this case further.

These vectors $c_1, ..., c_{q+1}$ are different vectors on the $q+1$ regimes $(1, \tau_1], (\tau_1, \tau_2], ..., (\tau_q, T]$. As an intuitive example, suppose we use the linear trend $h(\frac{t}{T}) = (1, \frac{t}{T})'$, and $a_t = a$ is a constant. In addition, suppose $b_t(\mathrm{M})$ has a linear trend with a single structural break occurring at time τ_1:

$$b_t(\mathrm{M}) = \begin{cases} \frac{t}{T} & t \leq \tau_1 \\ 2 - \frac{t}{T} & t > \tau_1. \end{cases}$$

Then the R_t^2 series also has a structural break at τ_1, because (ignoring the $o_P(1)$ term),

$$R_t^2 \approx \begin{cases} a - \frac{t}{T} & t \leq \tau_1 \\ a - 2 + \frac{t}{T} & t > \tau_1. \end{cases}$$

Our OOSR2 decomposition is similar in spirit to results in Giacomini and Rossi (2009), Paye and Timmermann (2006); Pettenuzzo and Timmermann (2011), who identified the parameter instabilities as one of the major sources of forecast breakdowns in parametric models. In a nonparametric setting, our analysis reaches qualitatively similar conclusion that the forecast breakdown may occur through the term $b_t(\mathrm{M})$. [14]

Because of the trending structure of R_t^2, testing for structural breaks in $a_t - b_t(\mathrm{M})$ gives us a way of evaluating the robustness of the forecasting model. However, for technical reasons, it is *not* convenient to directly implement structural break tests in the R_t^2 sequence itself. The results of West (1996); McCracken (2007) show that for regular forecasting models, the error term in the R_t^2 process satisfies $\Delta_t = \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} \varepsilon_s (1 + o_P(1))$ for some zero-mean random variable ε_s. Hence

$$\Delta_t = \Delta_{t-1} + \frac{1}{|\mathcal{S}_t|} \varepsilon_{t-1} + o_P(|\mathcal{S}_t|^{-1}). \tag{4.9}$$

So Δ_t is nearly a unit-root process which creates obstacles of applying the test directly on R_t^2. We resolve this issue by taking the first difference, $\Delta R_t^2 := R_t^2 - R_{t-1}^2$. Then (4.7) and

[14] Similar ideas for assessing forecast stability also appeared in the forecast literature. For instance, Rossi (2021) argue that forecast instabilities can refer to the forecast performance rather than the forecasts themselves. Even if forecasts may be stable, yet the forecast performance may display instabilities because predictability varies over time.

(4.9) imply

$$\Delta R_t^2 = \vartheta_t + \frac{1}{|\mathcal{S}_t|}\varepsilon_{t-1} + o_P(|\mathcal{S}_t|^{-1}), \quad \text{where } \vartheta_t := \theta_t - \theta_{t-1}.$$

The change process now has a stationary error term $\frac{1}{|\mathcal{S}_t|}\varepsilon_{t-1}$, and we can test for *outliers* in ϑ_t. In practice, it is also meaningful to test for breaks in ΔR_t^2 to see if the change of R^2 dramatically differs at certain periods. Another motivation of testing for breaks in the ΔR_t^2 process rather than in R_t^2 itself is that differencing can eliminate serial correlations.[15]

For implementation, we assume that ϑ_t has a linear trend, i.e.

$$\Delta R_t^2 = c_1 + c_2\left(\frac{t}{T}\right) + \text{noise}_t \tag{4.10}$$

where the coefficients $c = (c_1, c_2)$ may be subject to multiple breaks at unknown locations. We can apply the techniques originally developed by Bai and Perron (1998, 2003), to detect breaks/outliers in (4.10), which treat both the number of breaks and the locations of breaks as unknown parameters.[16] Also note that while all forecasting models fail to capture $\mathbb{E}(y_{t+1}|\mathcal{F}_t)$ during periods of volatile markets such as financial crises, we should favor models with fewer estimated breaks.

5 Comparison with other ML based predictions

5.1 A high level summary of the comparison

There are two more commonly used ML approaches that readers may wonder how they perform. One of them, which we call "naive" CML approach, is to simply apply the cross-sectional DNN on the last period:

$$\widehat{m}_T(z) = \arg\min_{m\in\text{DNN}} \sum_{i=1}^{N} (x_{i,T} - m(z_{i,T-1}))^2 \tag{5.1}$$

[15] Bai and Perron (2003) has conducted extensive simulations to verify the performance of their tests, and found that serial correlation in the data series can induce significant size distortions.

[16] As discussed by Bai and Perron (2003), the same method can be applied to estimating both structural breaks and outliers, which is based on the least squares treating the breaking regimes and outlier periods as unknown parameters in the least squares problem. Also see discussions in Perron and Rodríguez (2003), where first-order differencing for $I(1)$ processes produces higher powers for detecting outliers.

and substitute with the "new" $z_{i,T}$ and forecast using:

$$\widehat{y}_{T+1,\text{naive CML}} := \sum_{i=1}^{N} w_i \widehat{m}_T(z_{i,T}).$$

The other approach is more often referred to as "pooled ML" (Gu et al., 2020): First, estimate a conditional expectation function as:

$$\widehat{m}(z) = \arg\min_{m \in \text{ML}} \sum_{t=1}^{T} \sum_{i=1}^{N} (x_{i,t} - m(z_{i,t-1}))^2 \tag{5.2}$$

where "ML" denotes a machine learning space. This function $\widehat{m}(\cdot)$ is trained using data $(x_{i,t}, z_{i,t-1})$ pooled over all time periods and cross-sections. Then plug-in the "new" characteristic, $z_{i,T}$, and construct value-weighted market predictor:

$$\widehat{y}_{T+1,\text{pooled ML}} := \sum_{i=1}^{N} w_i \widehat{x}_{i,T+1,\text{ML}}, \quad \text{where } \widehat{x}_{i,T+1,\text{ML}} := \widehat{m}(z_{i,T}). \tag{5.3}$$

In this section we shall analyze the structure of these two machine learning predictions and compare with our proposed method. Recall that the true out-of-sample return is

$$y_{T+1} = \widetilde{\rho}_{f,T}' f_{T+1} + \widetilde{\epsilon}_{T+1} = \rho_{0,T} + \rho_{g,T}' g_T + \epsilon_{T+1}$$

which has two representations: either via contemporaneous stock factors, or via lagged BM-factors. Let

$$y_{T+1|T} = \rho_{0,T} + \rho_{g,T}' g_T.$$

We show:

$$\begin{aligned} \widehat{y}_{T+1|T} &\to^P \quad y_{T+1|T}, && \text{our proposed method} \\ \widehat{y}_{T+1,\text{naive CML}} &\to^P \quad \widetilde{\rho}_{f,T}' f_T, && \text{naive CML} \\ \widehat{y}_{T+1,\text{pooled ML}} &\to^P \quad \widetilde{\rho}_{f,T}' \mathbb{E}(f_{T+1} | \mathcal{F}_{z,T}) = \mathbb{E}(y_{T+1|T} | \mathcal{F}_{z,T}) && \end{aligned} \tag{5.4}$$

where $\mathcal{F}_{z,T}$ denotes the filtration generated by characteristics $z_{i,t}$ up to T. Result (5.4) provides clear econometric insights of these three ML-based predictors, where the first is the proposed CML prediction, and the other two have been popularly used in asset pricing. They show that: (1) under correct specification, the CML correctly captures the conditional expected return. (2) The naive CML is *wrong* as it only captures the lagged stock factors, yet there is a huge gap between the contemporaneous and lagged stock factors. (3) the pooled ML captures the unconditional expected return $E(y_{T+1|T}|\mathcal{F}_{z,T})$, but is missing the factor realization $y_{T+|T} - \mathbb{E}(y_{T+|T}|\mathcal{F}_{z,T})$.

5.2 Compare with "naive" CML

Unlike the proposed CML, this method uses only the last period to conduct DNN but does not estimate local PCA or any factors. As the last period $x_{i,T} = h_{\beta,T}(z_{i,T-1})'f_T + u_{i,T}$, the cross-sectional DNN removes the idiosyncratc error but retains the risk factor f_T, so

$$\widehat{m}_T(z) \to^P h_{\beta,T}(z)'f_T.$$

Note that the factor realization f_T on the right hand side is latent and part of the estimated neural network function $\widehat{m}_T(z)$. Now substitute to $z = z_{i,T}$,

$$\widehat{m}_T(z_{i,T}) \approx h_{\beta,T}(z_{i,T})'f_T = \beta_{i,T}'f_T. \tag{5.5}$$

As a result, it forecasts:

$$\widehat{y}_{T+1,\text{naive CML}} \to^P \sum_{i=1}^{N} w_i \beta_{i,T}' f_T, \tag{5.6}$$

whereas the true asset return should be

$$y_{T+1} = \sum_i w_i x_{i,t+1} \approx \sum_i w_i \beta_{i,T}' f_{T+1}. \tag{5.7}$$

Comparing the last two displayed equalities, we see that there is a substantial gap between the forecaster and the target: the former depends on the lagged factor f_T whereas the latter is driven by the contemporaneous factor f_{T+1}. The gap between the two could lead to substantial misleading forecast results because the lagged stock return factors f_T may carry little information of the factor innovation f_{T+1}.

5.3 Compare with pooled ML

We now analyze the pooled ML. According to (2.3), the true future market index can be represented as:

$$y_{T+1} \quad = \quad y_{T+1|T} + \epsilon_{T+1}, \quad \text{where } y_{T+1|T} = \rho_{0,T} + \rho_{g,T}' g_T. \tag{5.8}$$

We can further decompose the conditional expected return into the sum of a unconditional expected return, plut the factor realization:

$$y_{T+1|T} = \underbrace{\mathbb{E}(y_{T+1|T}|\mathcal{F}_{z,T})}_{\text{unconditional ER}} + \underbrace{\rho_g' \left[g_T - \mathbb{E}(g_T|\mathcal{F}_{z,T})\right]}_{\text{factor realization}} \tag{5.9}$$

Proposition 3 below shows that the pooled machine learning predictor captures the unconditional expected return, but not the factor realization:

$$\widehat{y}_{T+1,\text{pooled ML}} \rightarrow^P \mathbb{E}(y_{T+1|T}|\mathcal{F}_{z,T}) = \rho_0 + \rho_g' \mathbb{E}(g_T|\mathcal{F}_{z,T}).$$

So the pooled machine learning *does not* predict $y_{T+1|T}$, but only $\mathbb{E}(y_{T+1|T}|\mathcal{F}_{z,T})$. The key difference between the two is that $\mathbb{E}(y_{T+1|T}|\mathcal{F}_{z,T})$ does not preserve the BM-factor realization g_T, which may be the main source of variation driving the conditional expected return (CER).

In contrast, our approach can forecast the CER:

$$\widehat{y}_{T+1|T} \rightarrow^P y_{T+1|T}. \tag{5.10}$$

Combining (5.9)-(5.10), we can express the true future market return using these estimates:

$$y_{T+1} = \underbrace{\widehat{y}_{T+1,\text{pooled ML}} + \text{factor realizations}}_{\widehat{y}_{T+1|T}} + \epsilon_{T+1} + o_P(1),$$

which shows:

$$\begin{aligned} \text{forecast error of } \widehat{y}_{T+1,\text{pooled ML}} &= \text{factor realization} + \epsilon_{T+1} \\ \text{forecast error of } \widehat{y}_{T+1|T} &= \epsilon_{T+1}. \end{aligned}$$

Therefore, one can infer that $\widehat{y}_{T+1|T}$, once consistently estimating $y_{T+1|T}$, has strictly less forecast error than the pooled ML approach asymptotically.

To see the consequence of such a comparison on a multiple-forecast context, we conduct one-step-ahead forecast y_{t+1} for multiple periods $t \in \mathcal{S}$ respectively using the proposed method $\widehat{y}_{t+1|t}$ and the pooled machine learning $\widehat{y}_{T+1,\text{pooled ML}}$, and obtain their corresponding OOSR2. The comparison as being discussed well relates to the OOSR analysis in Section 4.3. For model M_{pooled} being the pooled ML, its specification error is

$$\mathbb{E}(y_{T+1}|\mathcal{F}_T) - \mathbb{E}(y_{T+1}|\mathcal{F}_T, \mathrm{M}_{pooled}) = y_{T+1|T} - \mathbb{E}(y_{T+1|T}|\mathcal{F}_{z,T}),$$

which gives rise to

$$b_T(\mathrm{M}_{pooled}) = \frac{1}{\sigma^2_{un}(T)} \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \left(y_{s+1|s} - \mathbb{E}(y_{s+1|s}|\mathcal{F}_{z,s})\right)^2 .$$

Such specification error impacts on the OOSR2 of the pooled machine learning, leading to suboptimal forecasts. The following proposition formalizes this in the neural network context.

Proposition 3 (Comparison with Pooled ML)**.** *Consider a factor model for stock returns* $x_{i,t} = g_\alpha(z_{i,t-1}) + h_{\beta,t}(z_{i,t-1})' f_t + u_{it}$ *where* g_α *and* g_β *are respectively mapping characteristics to alphas and betas. In addition, consider the pooled machine learning method (5.2)-(5.3) with neural network methods, and suppose Assumption 5 in the appendix hold. Then*

$$\widehat{y}_{T+1,\text{pooled ML}} \to^P \mathbb{E}(y_{T+1|T}|\mathcal{F}_{z,T}). \tag{5.11}$$

In addition, suppose model (2.5)-(2.2) is correctly specified. Then when the number of predictions in $\mathcal{S}$ *grows, with probability approaching one,*

$$R^2(\widehat{y}_{T+1|T}) > R^2(\widehat{y}_{T+1,\text{pooled ML}}), \tag{5.12}$$

where the left and right hand side are respectively the OOSR2 of the proposed method and the pooled ML.

5.4 Model Ensemble Using Forecast Standard Error

The CML model requires the factor structure assumption outlined in Section . Namely, both the market return and the contemporaneous stock factors depend on the lagged BM-factor g_t with stable coefficients:

$$\begin{aligned} y_{t+1} &= \rho_{0,t} + \rho_{g,t}' g_t + \epsilon_{t+1}, \\ f_{t+1} &= \Phi_0 + \Phi_g g_t + e_{t+1}. \end{aligned} \tag{5.13}$$

It is crucial to note that the validity of these assumptions may exhibit temporal variability. Moreover, as elucidated in the preceding subsection, the CML dominates the pooled ML by capturing the factor realization $\rho_g'[g_T - \mathbb{E}(g_T|\mathcal{F}_{z,T})]$. However, this advantage diminishes when the contemporaneous return depends less on the lagged g_T. Consequently, it is reasonable to anticipate that neither the CML nor the pooled ML uniformly dominates each other.

This is a natural concern as good forecasting models may be successful only for a brief period of time as market conditions change and investors adapt their behavior. This phenomenon is also supported by the efficient market hypothesis – when a trading rule helps forecast future returns and makes near arbitrage profits, other money follows and the profits vanish and the trading strategy "stops working." An important progress in the literature that addresses this issue is Farmer et al. (2022), who develop a "pocket of forecast" framework to incorporate multivariate information into the forecast. Each forecasting algorithm is successful only for a brief period of time, so the overall performance can be improved by ensembling multiple models. Bianchi et al. (2023) investigated the idea of model averaging in more depth, and and showed that it can compete with economically motivated predictive regressions.

We develop their idea by ensembling the two machine learning methods: CML and pooled ML. Ensembling both ML-based forecasts could take advantage of the dynamic performance of either methods. Instead of developing a "pockets forecast" we adopt the procedure of Black and Litterman (1990), which is essentially a Bayesian updating rule for computing the conditional mean $y_{t+1|t}$. Specifically, suppose in period t the investor has a *prior distribution* for the expected return:

$$y_{t+1|t} \sim \mathcal{N}(\pi_t, \Sigma_t) \tag{5.14}$$

Then with the predicted index $\widehat{y}_{T+1|T}$, the investor wants to update the expected index return. The Black-Litterman model treats $\widehat{y}_{T+1|T}$ as a noisy observation generated from a distribution centered at $y_{T+1|T}$. We can apply Proposition 1 and set this distribution as the predictive distribution:

$$\widehat{y}_{t+1|t} \sim \mathcal{N}(y_{t+1|t}, \mathrm{SE}(\widehat{y}_{t+1|t})^2). \tag{5.15}$$

Combining (5.14) and (5.15), the posterior mean of $y_{t+1|t}$ is, by Bayes' theorem,

$$\widehat{y}^{en}_{t+1|t} := \mathbb{E}(y_{t+1|t}|\widehat{y}_{t+1|t}) = w_t\widehat{y}_{t+1|t} + (1-w_t)\pi_t, \quad w_t = \frac{\Sigma_t}{\mathrm{SE}(\widehat{y}_{t+1|t})^2 + \Sigma_t}. \tag{5.16}$$

which ensembles the proposed predictor $\widehat{y}_{t+1|t}$ with the benchmark predictor π_t. As for the "prior", we specify $\pi_t = \widetilde{y}_{t+1,\mathrm{ML}}$, the pooled-ML forecast so that $\widehat{y}^{en}_{t+1|t}$ is an ensemble forecast of the conditional-ML and the pooled-ML. In addition, we use the (backward-looking) average of the squared standard error:

$$\Sigma_t = \text{average of } \mathrm{SE}(\widehat{y}_{s+1|s})^2, \quad s = 1...t$$

This yields the ensemble forecast:

$$\widehat{y}^{en}_{t+1|t} = w_t\widehat{y}_{t+1|t} + (1-w_t)\widetilde{y}_{t+1,\mathrm{ML}}. \tag{5.17}$$

We choose the pooled-ML as the prior predictor because as the prior, it requires much less conditions on the model specification than the CML does. That is, its validity does not require conditions (5.13). In contrast, due to the extra conditions of factor structures, the CML is not suitable as the prior. Meanwhile, on one hand the predictive distribution requires the availability of forecast standard error, which is not readily available for the pooled-ML. On the other hand, our proposed CML admits a nice forecast standard error $\mathrm{SE}(\widehat{y}_{t+1|t})$, making it suitable to be specified in as a complete predictive distribution as in (5.15).

The forecast weight w_t is decreasing function of $\mathrm{SE}(\widehat{y}_{t+1|t})^2$. In periods when the forecast standard error, and thus the uncertainty, is smaller, $\widehat{y}^{en}_{t+1|t}$ is weighted more by the CML. The intuition is that a large standard error of the CML method would indicate a large value of idiosyncratic volatility, suggesting that the factor-based forecast would not help at this period. It is then better off to put higher weights on the pooled ML, which does not depend on any realized factors in the prediction.

6 Empirical Analysis

6.1 Data and the moving window

We use 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.

Our main dataset is obtained from Chen and Zimmermann (2021) and consists of 87 firm characteristics that are available from 1955 - 2021. The firm characteristics feature a combination of accounting information as well as versions of momentum and functions of trading volume. Table 4 provides an overview of the characteristics we use in our main empirical analysis. We use the imputation approach of Freyberger et al. (2021) to impute missing characteristics. It should be noted that the predictors are not a randomly selected set of features, but have been found to be successful cross-sectional predictors in the literature. We detail the construction of the data set used for the main analysis in Appendix A.

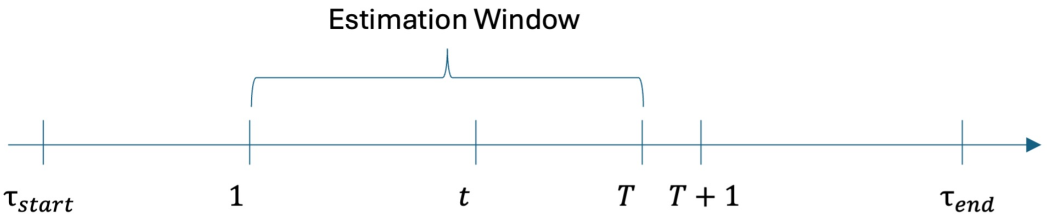

Figure 6.1: The estimation window and notation adopted in this paper.

We use moving windows for estimation and predicting the market excess returns, in which we fix the window size at $T = 60$ for estimation, and predict the market at month $T + 1$, then slide forward by one month. The first prediction occurs for December 1964, and the last prediction is in December 2021.

Our main forecast model uses only the BM-factor $\widehat{g}_t$, with a single estimated factor ($K_g = 1$):

$$\widehat{y}_{T+1|T} := \widehat{\rho}_0 + \widehat{\rho}_g \widehat{g}_T.$$

We refer this model to "CML". As in many machine learning procedures, our approach also contains tuning parameters. In many instances, there is little theoretical guidance on how to choose these tuning parameters. For the purpose of equity premium forecasting, it is desirable to adopt a tuning method that does not select models that display extremely good in-sample fit, but are often prone to breaking down out-of-sample.

There have been mainly two tuning approaches in the forecasting literature. In the first approach, one fixes a window of "tuning period" on which data are being used to tune the model, and then fix the tuned model for subsequent forecasts in the "post-tuning" periods. The other approach is to use time series cross-validation. It turns out neither approach is suitable in our context for the concern of robust forecasts. First, the dynamics of the equity risk premium are very subtle and even, the in-sample relationship is subject to change. For example, Kelly and Pruitt (2013) document that dividend growth is predictable over some periods, but not over others. Therefore, any models that is fixed for a long period of time or predicates on stable predictive relationships is likely going to disappoint at some point. Secondly, the standard statistical approaches for determining tuning parameters aim to find the best model, by maximizing some measure of fit such as R^2 or relatively mean-squared error. These statistics are however, very sensitive to extreme observations, so that they can be "fooled" by models that happen to work particularly well (or poorly) in 1-2 periods.

We therefore adopt a new approach to tuning to achieve more robust predictions. Campbell and Thompson (2008) argue that it is sensible from an economic point of view to constrain models for forecasting the equity risk premium to always yield a positive forecast and that it also leads to better out-of-sample performance. We incorporate this insight in our tuning procedure and use the model for out-of-sample prediction that most consistently produces a positive forecast of over the tuning period. More formally, we find the best model to achieve

$$\text{at tuning period } \tau : \mathrm{M}^*_\tau := \arg \max_{\mathrm{M}_\tau \in \mathcal{M}} \sum_{t \in \text{tuning period}_\tau} 1\{\widehat{y}_{t+1|t}(\mathrm{M}_\tau) > 0\}$$

where $\widehat{y}_{t+1|t}(\mathrm{M})$ is the forecast index using model M during the tuning period.[17] This corresponds to a specific choice of tuning parameters in the tuning set $\mathcal{M}$.

We repeat this tuning procedure every twelve months using the past 60 months as a tuning period. Note that as the tuning period rolls forward, we could in principle choose a different model every twelve months. We view this as a desirable feature since the equity premium dynamics are likely very subtle and complicated so that choosing a different model more frequently may yield a better approximation.

We evaluate the out-of-sample R^2_t (OOSR2) to assess the prediction performance on two aspects: the prediction accuracy and the prediction robustness. Let y_{s+1} denote the true return from s to $s+1$, $\widehat{y}_{s+1|s}$ denotes the predicted return, and $\bar{y}_s$ denote the in-sample time series average of the return up to time s.

We consider two versions of OOSR2. In both definitions, the realized future return y_{s+1} never enters into model estimation or tuning, i.e. all measures of fit are strictly out-of-sample.

I. The $[t : \text{end}]$-R^2_t:

$$[t : \text{end}] \quad R^2_t := 1 - \frac{\sum_{s \geq t} (y_{s+1} - \widehat{y}_{s+1|s})^2}{\sum_{s \geq t} (y_{s+1} - \bar{y}_s)^2}. \tag{6.1}$$

This version of OOSR2 varies the starting point, i.e. it evaluates the cumulative predictive performance, starting in month t through the end of the sample.

II. The $[t_0 : t]$-R^2_t:

$$[1 : t] \quad R^2_t := 1 - \frac{\sum_{s=t_0}^{t} (y_{s+1} - \widehat{y}_{s+1|s})^2}{\sum_{s=t_0}^{t} (y_{s+1} - \bar{y}_s)^2}. \tag{6.2}$$

For this version of OOSR2, we fix the starting point and vary the end point, i.e. this measure evaluates the cumulative predictive performance up to month t. This definition of R^2 is used e.g. in Goyal et al. (2021).

[17] If there are ties, we use the median forecast over the tied models.

6.2 Forecast Accuracy

We assess the prediction accuracy using the time series average of OOSR2 and compare six predictive methods:[18]

(i) (CML) The proposed conditional forecast. We estimate a single BM-factor and use three neural networks at the firm level (Section A.2). The bandwidth for kernel smoothing is tuned from 0.6 to 1 in 0.025 increments. In total, we choose from 54 models to apply the adaptive tuning.

(ii) (PCA) Regular PCA applied to individual stock returns. We estimate three stock-factors.

(iii) (PCA-ker) Kernel-based PCA applied to individual stock returns. We apply the adaptive tuning method to tune the number of stock-factors in $\{1, .., 4\}$ and the bandwidth for kernel smoothing in the range from 0.6 to 1 in 0.025 increments.

(iv) (GW-linear) Linear forecast using 16 predictors from Goyal et al. (2021).

(v) (GW-Fourier) Random Fourier transformation of 16 predictors from Goyal et al. (2021).

(vi) (Pooled ML) The pooled machine learning method.

We divide the forecast period into three parts: pre-millennium (1964-1999), pre-2008-crisis (2000-2007) and post-2008-crisis (2008-2021), and respectively calculate the time series average of R_t^2, with respect to each of the three definitions over each period. The results are given in Table 1 and time series plots are given in Figure 6.2.

Table 1 shows that the proposed method, CML, compares favorably relative to the other methods, closely followed by the GW-Fourier predictors. Almost all measures of out-of-sample R^2 are positive for both methods, whereas they are mainly negative for the other methods.

Figure 6.2 illustrates that the time series of $[t : \text{end}]$-R_t^2 is much more stable at beginning than toward the end, whereas the trend is opposite for the $[1 : t]$-R_t^2. This observation is also reasonable because toward the end of the forecast periods, less months of forecast errors

[18] We refer to Section A.2 for more detailed explanation of PCA-ker, GW-linear and GW-Fourier.

Table 1: Time series averages of R_t^2 (in percentage)

This table shows the time series average of OOSR2 during the period given in the first column. Here R_t^2 is evaluated using two definitions: $[t : \text{end}]$ and $[1 : t]$. We respectively trim off the the last 20 months for the $[t : \text{end}]$-R_t^2, and the first 20 months for the $[1 : t]$-R_t^2 so that there are at least 20 months of forecasts in computing R_t^2 at each t. The estimation uses 60-month moving window as the in-sample period, and conducts one-month ahead forecast, then slides forward by one month. Both PCA and PCA-ker used three factors. All R^2 measure are in percentage. The sample period is December 1964 - December 2021.

		Forecast periods			
		1964-1999	2000-2007	2008-2021	full period
$[t : \text{end}]$-R_t^2	CML	1.733	1.384	-0.185	1.258
	PCA	-8.928	-2.824	-8.654	-7.967
	PCA-ker	-5.244	-6.578	-14.259	-7.443
	GW-linear	-25.17	-47.69	-39.85	-31.45
	GW-Fourier	0.291	0.026	0.013	0.196
	Pooled-ML	0.160	0.459	1.663	0.536
$[1 : t]$-R_t^2	CML	0.657	1.606	1.741	1.069
	PCA	-3.108	-10.002	-7.841	-5.221
	PCA-ker	-2.135	-3.241	-3.149	-2.551
	GW-linear	-6.601	-0.899	-11.56	-6.898
	GW-Fourier	0.457	0.373	0.319	0.413
	Pooled-ML	-0.629	-0.703	-0.599	-0.632

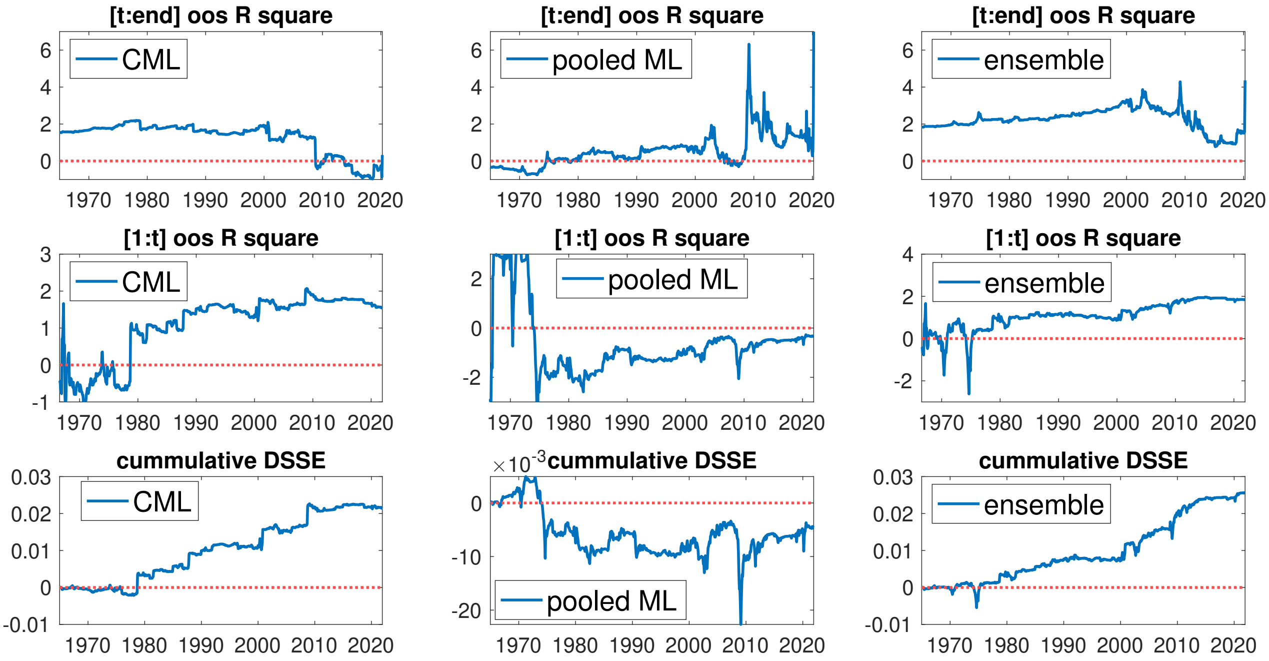

Figure 6.2: The out-of-sample predictive R_t^2 (in percentage) and cumulative DSSE. The "conditional ML" refers to the proposed conditional deep learning based forecast; "conditional ML" refers to the pooled deep learning forecast. "ensemble" refers the weighted average of conditional ML and pooled ML, using the forecast standard error as the weight. The estimation uses five-year moving windows ($T = 60$) to forecast market index for $T + 1$. The computed R_t^2 correspond to one of the definitions in Section 6.1. The model is tuned using the adaptive tuning described in Section 6.1.

are being computed in the $[t : \text{end}]$-R_t^2, and this trend is opposite for the $[1 : t]$-R_t^2. These plots provide a more complete picture of the overall predictive performance than reporting a single number out-of-sample R_t^2 at an arbitrarily chosen t. In addition, the bottom panels of Figure 6.2 plot the cumulative difference-sum-squared-errors, defined as

$$\text{DSSE}(t) = \sum_{s \le t} (y_{s+1} - \bar{y}_s)^2 - \sum_{s \le t} (y_{s+1} - \widehat{y}_{s+1})^2$$

where $\widehat{y}_{s+1} = \widehat{y}_{s+1|s}$ for our predictor, and $\widehat{y}_{s+1} = \widehat{y}_{s+1,\text{ML}}$ for the pooled ML. An interesting feature of the DSSE(t) plot is that an increase signifies improved predictive performance of the model for period $t+1$, while a decrease indicates better performance of the sample-mean. So for good predictors, the DSSE should be steadily rising when there is consistently good predictability, using the terminology of Farmer et al. (2022), "pockets of predictability", with fewer crashes.

The bottom left panel of the cumulative DSSE in Figure 6.2 shows that prior to 1978, the proposed conditional ML forecast does worse than the sample average, and then has an overall increasing trend. There are two noticeable "jumps" where it significantly outperforms the sample average: July 1987 and June 2008. In contrast, the pooled ML does better with a steady increasing cumulative DSSE prior to 1975. However, it decreases after May 1975 and then switches between increases and decreases afterwards until Jun 2008. The steady increasing DSSE of the pooled ML forecast at the end of the forecasting period shows that it outperforms the sample average during most of the time after June 2008.

The three right-most panels of Figure 6.2 plot the R_t^2 sequences and the cumulative DSSE of the ensemble forecast. As introduced in Section 5.4, the ensemble forecast takes a weighted average of the CML and the pooled ML, where the weight is determined by the contemporaneous forecast standard error of the $\widehat{y}_{t+1|t}$, reflecting the forecast uncertainty of the CML. Overall, incorporating the uncertainty information delivered by the forecast standard error leads to good forecasts overall. From a pure accuracy perspective, the prediction performance is comparable with that of $\widehat{y}_{t+1|t}$ before 2000, and is noticeably improved afterwards. The DSSE plot steadily increases during most of the time after 2008, except for a period of drop during 2004, showing that the ensemble method outperforms the benchmark sample mean during most of the forecast periods.

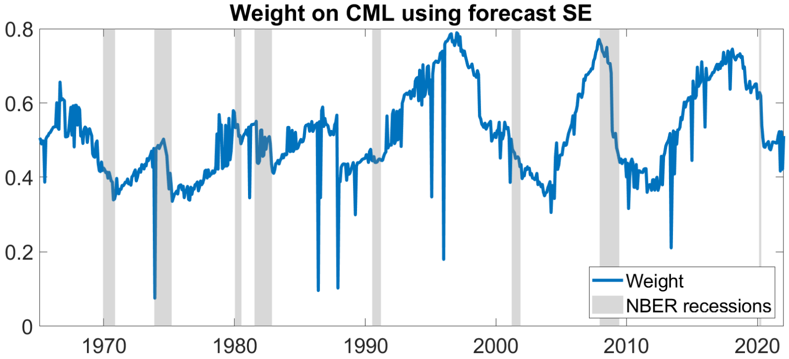

Figure 6.3: The weight w_t on the CML of the ensemble forecast (5.16) and (5.17), overlayed by shaded US recession bands as reported by the National Bureau of Economic Research (NBER).

Finally, one important observation from Figure 6.2 is the sudden change in June 2008, which is during the 2008 financial crises. The top left panel shows a large "drop" in the [t:end]-R_t^2 of the conditional ML. At first sight, this might suggest that our method performance poorly during this time and onward. It appears counter intuitive, but our method actually performs well at this period and thereafter. At this point, t_0, we have the following ordering of squared deviations:

$$\underbrace{(y_{t_0} - \widehat{y}_{t_0|t_0-1})^2}_{\text{squared deviation our method}} < \underbrace{(y_{t_0} - \bar{y}_{t_0-1})^2}_{\text{squared deviation mean}} < \underbrace{(y_{t_0} - \widetilde{y}_{t_0,\text{ML}})^2}_{\text{squared deviation pooled ML}} .$$

This is corroborated by the first DSSE plots in the two bottom panels of Figure 6.2. The DSSE(t_0) of the conditional ML jumps, but that of the pooled ML drops at June 2008, showing that the former outperforms whereas the latter underperforms relative to the sample mean.

To understand this phenomenon, recall that the [t : end] - R_t^2, which by definition, summarizes the aggregated forecast performance *after* period t. Hence for all periods prior to t_0, the computed [t : end] - R_t^2 benefits from the superior performance of our method at t_0. Before the financial crises, the closer t is to June 2008, the larger outperformance is observed. This explains the increase of R_t^2 on the left panel before the sudden drop. Afterwards, the outperformance at t_0 is no longer part of the sample which explains the drop in [t : end] - R_t^2 for our method. The phenomenon can also be seen from the middle panel of Figure 6.2, which plots the [1 : t] - R_t^2. This version R_t^2 does not benefit from the outperformance until t_0 is included, which explains the jump in June 2008. In contrast, pooled ML, uses a long history of data and reacts slowly to market instability and thus forecasts poorly in June 2008 compared to the sample mean.

6.3 Explainability

To explain the dynamic predictability of the ML-based forecasts, we now analyze the relevant volatilities constituting to the forecast standard error and prediction error. We plot the evolution of moving average of squared factor realizations

$$g_T^2 \sim \text{average of past 24 months estimated } \widehat{g}_T^2,$$

the idiosyncratic volatility

$$\mathrm{Var}(\eta_{i,T}) \sim \text{average of past 24 months estimated } \frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}(\eta_{i,T}),$$

the factor-effect volatility

$$\mathrm{Var}(\rho' g_T) \sim \text{average of past 24 months estimated } \mathrm{Var}(\rho' g_T),$$

and change of forecast coefficient:

$$\mathrm{Var}(\rho_t) \sim \text{time series variance of past 24 months estimated } \widehat{\rho}_s.$$

Figure 6.4 plots the evolution of these volatilities.

6.3.1 The conditional ML

Figure 6.4 shows thas the evolution of both g_T^2 and $\mathrm{Var}(\eta_{i,T})$ increase in the early sample until mid-1980s, which explains the underperformance of the conditional ML forecast relative to the sample mean in the early period as indicated by the R^2 plot in Figure 6.2. Indeed, as explained by our theory, the increasing volatilities amplify the forecast uncertainty and the predictive error. In addition, the cumulative DSSE also decreases during 1995-2000, which is also explainable by the evolution of the idiosyncratic volatility $\mathrm{Var}(\eta_{i,T})$, which maintains a high level in this period.

Meanwhile, the superior prediction during 1985-1995 of the CML, evidenced by the monotone increasing trend of the DSSE plot, is also explainable by the decreasing trend of g_T^2 in this period. In particular, the g_T^2 plot in Figure 6.4 shows a clear decrease during 1990-1995, and reaches a minimum value around June 1995. Also, both the idiosyncratic volatility and squared factor realizations noticeably decrease during 2010-2015, explaining the superior predictability of the CML in this period.

The two noticeable “jumps” on the DSSE plot, occurring at May 1998 and June 2008 are both well explainable by the change of volatilities: in Figure 6.4, the evolution plots of g_T^2 and $\mathrm{Var}(\eta_{i,T})$ decrease and maintain at relatively small levels respectively in the neighboring months of May 1998 and June 2008. In addition, the noticeable peek in both volatilities in

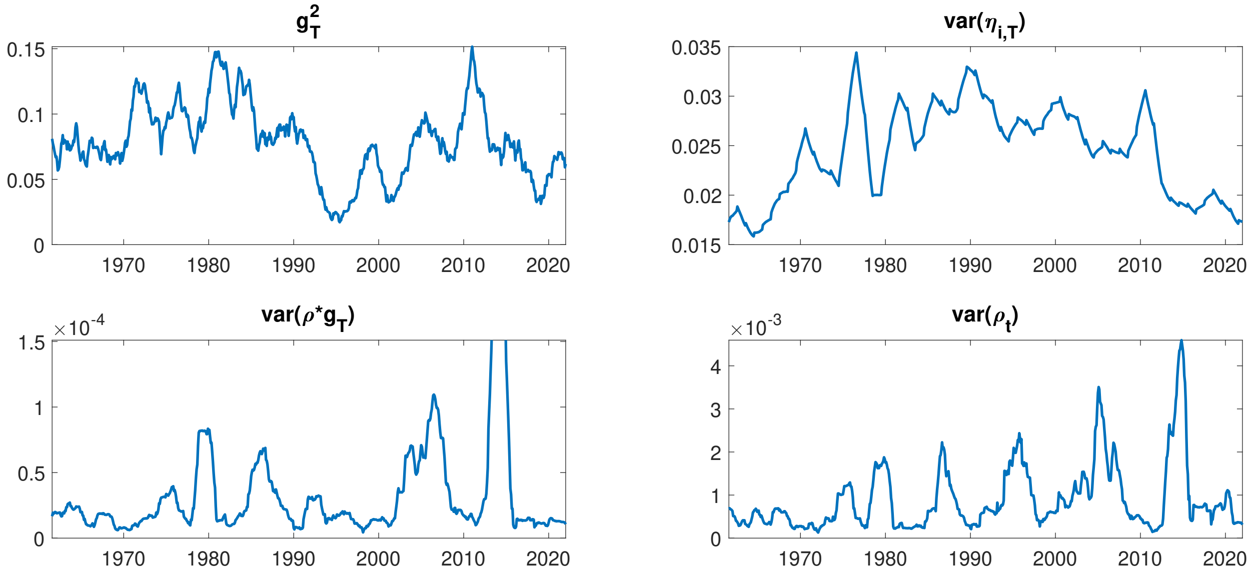

Figure 6.4: The evolution plots of estimated g_T^2, Var$(\eta_{i,T})$, Var$(\rho_T' g_T)$ and Var(ρ_T). They are first estimated using the conditional ML method. We then plot their moving average using the 24-month moving window.

early 2010s explain the underperformance of the conditional ML with a decreasing DSSE shortly after the 2008 jump.

6.3.2 Creative Destruction Index

To see a more pronounced economic relation between the market stability and the forecast uncertainty, let us consider the "volatility of uncertainty", defined as

$$\text{std}_t(w_t) = 24 \text{ months time series standard deviation of } w_t,$$

which measures the variation of the forecast standard errors in a moving window up to period t. This measures the volatility of forecast uncertainty at each time. As we discussed for the explainability of the CML forecast, the forecast uncertainty depends on the volatilities idiosyncratic noise and factors, which explains the success/failure of CML forecast. If one of these volatilities varies dramatically for a period of time, $\text{std}_t(w_t)$ would significantly increase.

Figure 6.5 plots the creative destruction index (CDI), defined as the correlation between firm sales and market capitalization rankings, along with $\text{std}_t(w_t)$. There is a clear pattern that features the relationship of the CDI and $\text{std}_t(w_t)$. In periods when CDI drops and reaches local valleys, $\text{std}_t(w_t)$ increases and peaks. This relationship can be confirmed by runnning the regression:

$$\text{CDI}_t = 0.81 - 0.23 \times \text{std}_t(w_t) + \text{residual}.$$

The slope is significantly negative (p value less than 6×10^{-4}). In summary, the correlation between sales and market capitalization rankings is economically related to the forecast uncertainty, and the relation is more pronounced to the volatility of the forecast standard error. In periods when the sales and market capitalization rankings are less correlated, the forecast standard error is less stable, indicating less robust market equity forecasts. In addition, in periods when the correlation is low, the ML-based forecast can be remedied by assigning less weights to the CML method, while giving more weights to the pooled ML method.

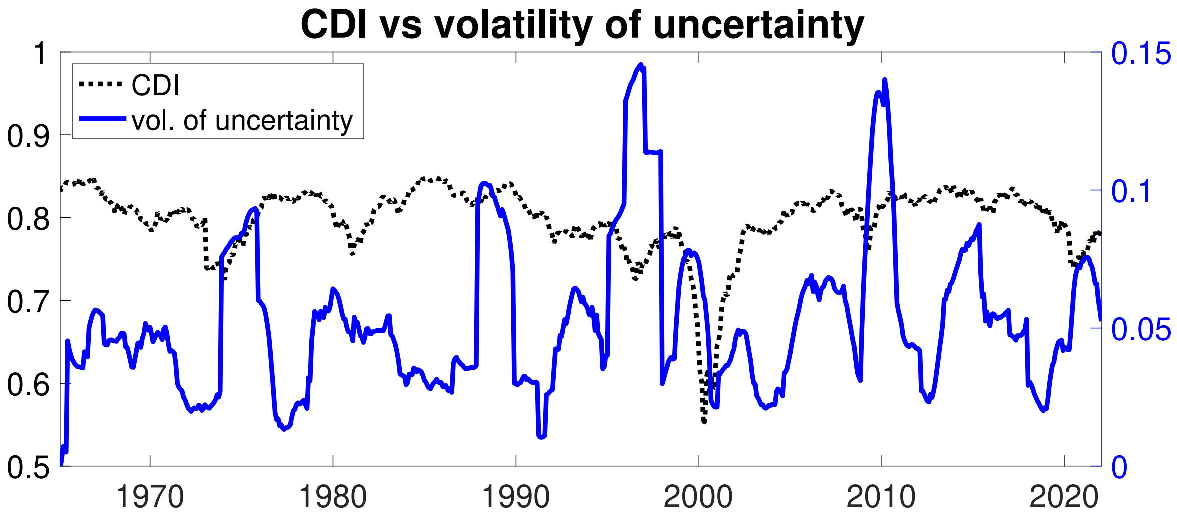

Figure 6.5: Creative Destruction Index (CDI) v.s. the 24-month moving standard deviation of w_t, the latter measure the volatility of the forecast uncertainty.

6.3.3 The pooled ML

Moving on to the pooled ML, its predictability is also well explainable under this framework. Theories show that the major source of forecast error of the pooled ML is the factor shocks $\rho'(g_T - \mathbb{E}(g_T|\mathcal{F}_{z,T}))$, so the predictability is negatively impacted by the evolution of the volatility $\mathrm{Var}(\rho' g_t)$, which we also plot for its 24-month moving averages.

There are two noticeable periods in which the DSSE of pooled ML steadily increases: before 1970 and after 2015. Notice that these are the periods where $\mathrm{Var}(\rho' g_t)$ takes a relatively low value. Especially, $\mathrm{Var}(\rho' g_t)$ decays after 2010 and reaches a local minimum around early 2012. This explains the superior of pooled ML over CML in this period. Except for these periods, neither the cumulative DSSE nor the evolution of $\mathrm{Var}(\rho' g_t)$ have monotonic trends, explaining that the pooled ML does not universally outperform the in-sample average during most of the forecasting period.

6.3.4 The impact of potential misspecifications

Finally, we also explain, to some extent, the impact of potential misspecifications. While our model allows the forecast coefficient to vary over time, it is our assumption that is should not vary too fast, and the faster it changes over time, the more severe the misspecification problem is. The degree of evolution can be measured by its time series variance of $\mathrm{Var}(\rho_t)$ in each forecast rolling window.

The last panel of Figure 6.4 shows that while the evolution of $\mathrm{Var}(\rho_t)$ periodically varies throughout the forecasting periods, it peaks around early 2015, which also explains the underperformance of the CML shortly after the jump in 2008 on the DSSE plot. Also, there are periodical "local peaks" on the $\mathrm{Var}(\rho_t)$, which shows that the forecast coefficients are not stable in these periods. These peaks suggest possible instabilities/strucural breaks for the predicting performance, which are well detected by the structural break test, to be studied in the next subsection.

6.4 Prediction Robustness

We evaluate the robustness of OOSR2 using two measures for the time series $\{R_t^2 : t \geq 1\}$: (i) time series variance and (ii) number of structural breaks. These measures capture distinct aspects of forecast stability and robustness, and are not interchangeable. The time series

Table 2: Time series standard deviation of R_t^2 (in percentage)

This table shows the time series standard deviation of OOSR2 during the period given in the first column. Here R_t^2 is evaluated using three different definitions: $[t : \text{end}]$, $[1 : t]$, and 5-year moving. We respectively trim off the the last 20 months for the $[t : \text{end}]$-R_t^2, and the first 20 months for the $[1 : t]$-R_t^2 so that there are at least 20 months of forecasts in computing R_t^2 at each t. The estimation uses 60-month moving window as the in-sample period, and conducts one-month ahead forecast, then slides forward by one month. Both PCA and PCA-ker used three factors. All R^2 measure are in percentage. The entire sample period is December 1959 - December 2020, with the first forecast occurs in December 1964.

		Forecast periods			
		1964-1999	2000-2007	2008-2020	full period
$[t : \text{end}]$-R_t^2	CML	0.186	0.235	0.570	0.839
	PCA	3.581	0.231	2.367	3.755
	PCA-ker	0.845	0.753	4.238	4.247
	GW-linear	10.41	2.747	5.991	12.56
	GW-Fourier	0.061	0.166	0.266	0.193
	Pooled-ML	0.463	0.585	1.115	0.911
$[1 : t]$-R_t^2	CML	0.919	0.121	0.106	0.891
	PCA	4.831	0.460	0.447	4.725
	PCA-ker	0.916	0.197	0.406	0.904
	GW-linear	19.66	3.059	2.531	15.91
	GW-Fourier	0.644	0.062	0.059	0.512
	Pooled-ML	1.746	0.291	0.275	1.368

variance represents the average squared deviations from the mean and is influenced by the magnitude range of the time series. Conversely, the number of structural breaks reflects the sensitivity of predictions to sudden market changes. Both are useful indicators to measure the prediction robustness. In the subsequent analysis, we examine each measure separately.

6.4.1 Time series variance of R_t^2

Table 2 presents the time series variance of the R_t^2 sequence across the three periods of interest. In most scenarios, CML and GW-Fouriers exhibit comparable and generally smaller variances, followed by PCA-ker applied to individual stock returns.

Interestingly, the variance of $[t : \text{end}]$-R_t^2 tends to be larger towards the end of the prediction period compared to the beginning period. Conversely, the variance of $[t_0 : t]$-R_t^2 tends to be smaller towards the end of the prediction period. This pattern aligns with intuition since

$[t : \mathrm{end}]$-R_t^2 is computed using fewer samples towards the end, while $[t_0 : t]$-R_t^2 is computed using fewer samples at the beginning. Moreover, variance is generally sensitive to outliers when the sample size is small.

Consequently, to primarily capture the nature of predictions while mitigating the influence of sample size, the variance of $[t : \mathrm{end}]$-R_t^2 is more appropriate for assessing prediction robustness during the beginning through mid-periods. Similarly, the variance of $[t_0 : t]$-R_t^2 is more suitable for the mid- through the end period.

6.4.2 Structural break tests

An alternative method to evaluate the prediction robustness is to test for *structural breaks* of the R_t^2 series for each method. We test for structural breaks using:

$$\Delta R_t^2 \quad = \quad \theta_t + \vartheta_t \cdot \left(\frac{t}{T}\right) + e_t. \tag{6.3}$$

Figure 6.6 plots the R_t^2 series with the detected structural breaks associated with each series. For the $[t : end]$-R_t^2, six breaks are detected on the CML prediction; seven breaks are detected on the pooled-ML prediction, and no break is found on the ensemble prediction. The detected breaks of the two ML-based predictions are fairly uniformly distributed across the entire forecast periods. Meanwhile, for the $[1 : t]$-R_t^2 series, five breaks are detected on CML, one dected on pooled-ML, and six are detected on the ensemble prediction.

As discussed in Section 4.3, we can roughly classify the detected breaks by “common breaks” – due to changes in volatility, and by “model breaks”– due to changes of the model misspecification. While it might be challenging to quantitatively classify the two types of breaks, we can intuitively classify those “shared” by multiple forecast models as common breaks, and those only appear to specific models as model specific breaks. We therefore list the detected breaks in Table 3. Here “Common breaks” are the breaks that are detected on the OOSR2 of at least two predictors among CML, pooled-ML and the ensemble prediction, while the “Model breaks” are the breaks that are detected on model specific methods.

Important crises defined by NBER-recessions are detected as common breaks by the structural break test. For instance, the common breaks detected by the $[1 : t]$-R_t^2 series mainly occurred during early periods, respectively in September 1978 and October 1987, matching with the Oil Shock in the 1970s. The break in October 1987, while not defined by

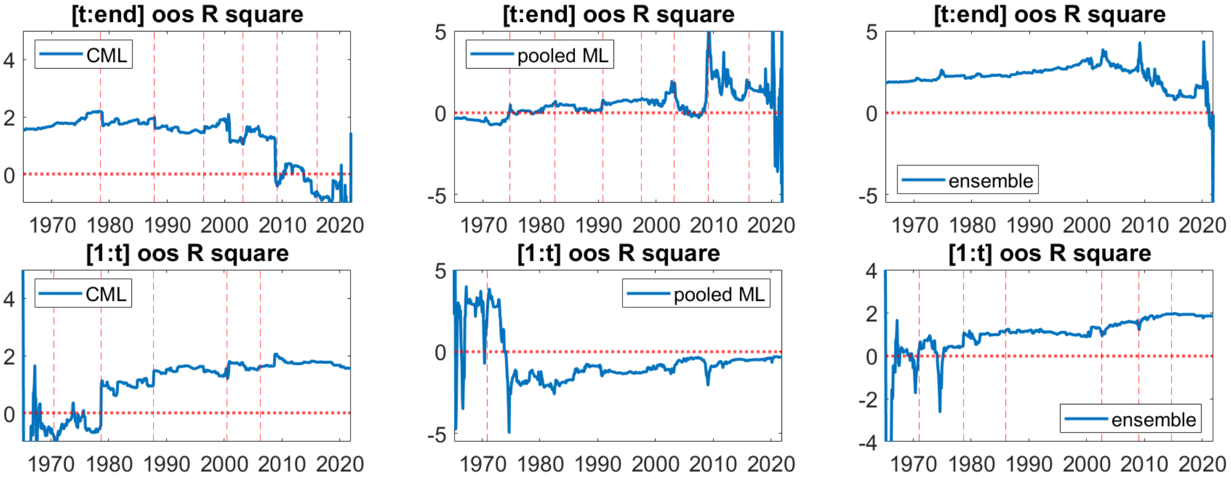

Figure 6.6: Structural breaks of R_t^2 (top panel: the proposed method; the bottom panel: the pooled ML). Vertical lines indicates where structural breaks are detected.

NBER recessions, is likely due to a sudden change on the true expected index return, causing the sample-mean benchmark forecasts badly for this month. In addition, the common breaks detected by the $[t : end]$-R_t^2 series occurred mainly towards the end of the forecasting period: February 2003, January 2009 and February 2016. The first two are also defined by the NBER recessions, respectively corresponding to the end of dot-bubble and the global financial crisis.

6.5 Economic Relevance

6.5.1 Interpretations of BM-factors

To relate the estimated BM-factors to economic state variables, Figure 6.8 plots the estimated BM-factors alongside with Shiller CAPE ratio over the entire prediction period December-1954 through December 2021. The plot shows that the two time series are positively correlated during most of the forecasting periods. The correlation during 1980 through 2000 is particularly high, which equals 0.80, and the overall correlation is 0.42 in this entire period. But PE itself does not forecast the market index as well, with the overall OORS2 being negative.

The close correlation illustrates that our BM-factor is closely related to discount rates and discount rate variation. But unlike the usual aggregated indices that utilizes firms'

Table 3: Detected structural breaks on the R_t^2 time series

This table shows the structural breaks detected by the structural break test (Bai and Perron, 2003) respectively on the $[t : end]$-R_t^2 and $[1 : t]$-R_t^2 time series, corresponding to the OOSR2 of three predictors: conditional machine learning (CML), pooled machine learning (pooled-ML) and the ensemble method of the two based on the ML- forecast standard error. The "Common breaks" are the breaks that are detected on the OOSR2 of at least two predictors, while the "Model breaks" are the breaks that are detected on model specific methods.

	$[t : end] - R_t^2$		
Common breaks	Feb-2003, Jan-2009, Feb-2016		
Model breaks	CML	Pooled-ML	Ensemble
	Jun-1978	Aug-1974	
	Sep-1987	Jun-1982	
	Apr-1996	Sep-1990	
		Jun-1997	
	$[1 : t] - R_t^2$		
Common breaks	Sep-1978, Oct-1987		
Model breaks	CML	Pooled-ML	Ensemble
	Jul-1970	Nov-1970	Jan-1971
	Jun-2000		Jul-2002
	Mar-2006		Dec-2008
			Aug-2014

market capitalization as the weights, our BM-factor is weighted by the stock-betas. As the stock-betas are much less noisy than the market capitalization, the BM-factor is relatively stable except during a few periods of US recessions reported by the NBER. These are the periods for which forecasting is very challenging due to the presence of structural breaks on the conditional mean index returns.

6.5.2 Forecast Distribution for Portfolio Allocation

We apply the distribution derived in Proposition 1 to construct portfolios that incorporate investors' forecasts and the corresponding uncertainty around these forecasts into a portfolio decision. At period T, an investor wants to allocate her investment into two assets: one risky asset whose return is y_{T+1} and one risk-free asset r_f. In addition, she has also obtained

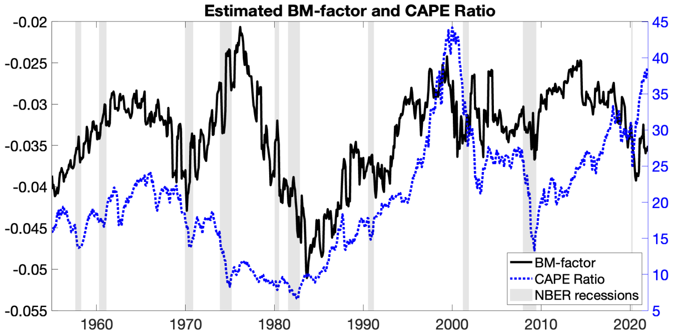

Figure 6.7: Plots of estimated BM-factors and Shiller cape ratio, overlayed by shaded US recession bands, as reported by the National Bureau of Economic Research (NBER).

a forecast index $\widehat{y}_{T+1|T}$ using the proposed method.

Let ω be the investor's allocation in y_{T+1} at the end of period T. Also let W_T denote the investor's wealth at the end of period T. Then her next period wealth is

$$W_{T+1} = W_T \left[\omega(1 + y_{T+1}) + (1 - \omega)(1 + r_f)\right].$$

Taking a utility function $\mathcal{U}(W) = \log(W)$ the investor chooses ω to maximize expected utility,

$$\max_{\omega} \mathbb{E}\mathcal{U}(W_{T+1}) = \max_{\omega} \int \log W_{T+1} p_T(y) dy \tag{6.4}$$

where $p_T(\cdot)$ is the predictive distribution of the future index return. We assume the predictive distribution is normal, and write

$$p_T(y) \sim \mathcal{N}(\mu_T, V_T) \tag{6.5}$$

with conditional mean μ_T and variance V_T. Here V_T is set as the estimated residual variance in the forecast model $y_{t+1} = \widehat{y}_{t+1|t} + e_t$. The key question is how to compute μ_T such that it incorporates the investor's knowledge of the predicted index return.

The Black-Litterman's model, as adopted for the model ensembling, is also useful to compute μ_T. As we discussed earlier, it is essentially a Bayesian updating rule to reflect the investor's "view" of the market index into the investment. Importantly, applying the Black-Litterman's updating requires the distribution of the machine learning predictor $\widehat{y}_{T+1|T}$ to serve as the "likelihood" function. Meanwhile, the asymptotic theory we derived for our robust machine learning predictor naturally yields such a distribution:

$$\widehat{y}_{t+1|t} \sim \mathcal{N}(y_{t+1|t}, \mathrm{SE}(\widehat{y}_{t+1|t})^2).$$

We use a prior distribution $y_{t+1|t} \sim \mathcal{N}(\pi, \Sigma)$. Then set μ_T as the *posterior mean* for $y_{T+1|T}$:

$$\mathbb{E}(y_{t+1|t}|\widehat{y}_{t+1|t}) = w\widehat{y}_{t+1|t} + (1-w)\pi, \quad w = \frac{\Sigma}{\mathrm{SE}(\widehat{y}_{t+1|t})^2 + \Sigma}.$$

Therefore, we set $\mu_T = \mathbb{E}(y_{T+1|T}|\widehat{y}_{T+1|T})$ for the predictive distribution $p_T(y)$ in (6.5). An appealing feature of such choice of μ_T is taking the forecast uncertainty, $\mathrm{SE}(\widehat{y}_{t+1|t})$, into forecast account. For the prior distribution, we set $\pi = \bar{y}_T$ as the index's historical mean. This prior can be viewed as an *empirical Bayes* approach that specifies prior distributions that also depend on the data. For the prior variance, we apply Zellner's g-prior (Zellner, 1986): $\Sigma = g\sigma^2$, where σ^2 is the sample variance of the market index return in the estimation period, and we fix $g = 1$. Then μ_T incorporates both the historical mean in the usual mean-variance analysis and the investor's view about the index prediction.

The optimal portfolio weight, ω, is then obtained by maximizing expected utility of terminal wealth:

$$\int \log(W_T r_{T+1}) p(y_{T+1}) dy_{T+1}, \quad r_{T+1} = \omega(1 + y_{T+1}) + (1-\omega)(1 + r_f),$$

and $y_{T+1} \sim \mathcal{N}(\mu_T, V_T)$. We solve this problem numerically, varying ω on a grid of $[0, 1.3]$, where the upper limit 1.3 is chosen so that the the average ω over time is roughly equal to one (average investment weight on the market). We label this procedure as "Robust-FD" approach which constructs the portfolio that utilizes our robust forecast distribution. We also implement this procedure by simply using the in-sample mean and variance of the market index return for μ_T and V_T, and construct the optimal portfolio without considering the uncertainty around the input parameters. This serves as a natural benchmark, which we label "regular".

We assess the portfolio performance using the cumulative realized wealth $W_{T+1} = W_T r_{T+1}$

and annualized Sharpe ratios:

$$\mathrm{SR}_t = \frac{\bar{r}_{[t_0:t]}}{\mathrm{SE}(r_{[t_0:t]})} \times \sqrt{12}$$

where $\bar{r}_{[t_0:t]}$ and $\mathrm{SE}(r_{[t_0:t]})$ respectively denote the out-of-sample mean and standard deviations of r_{T+1} during the periods of the expanding window $[t_0 : t]$. The period t_0 of the first investment is fixed to December 1959. Figure 6.8 plots the realized log wealth and Sharpe ratio for a sequence of out-of-sample periods, where the realized wealth is defined using the true return y_{T+1}:

$$W_T[\widehat{\omega}(1 + y_{T+1}) + (1 - \widehat{\omega})(1 + r_{f,T+1})]$$

and $\widehat{\omega}$ is the portfolio that numerically optimizes the expected utility. For a comparison, we also plot the sequences of the market index return of the same period. The realized wealth for the Robust-FD is significantly higher than the market prior to 2010, and is significantly higher than the regular method after 2005. The plot of realized wealth for all the three models shows two substantial drops: one occurs during the early 2000s recession and the other one occurs during the 2007-2008 financial crisis. For Sharpe ratios, we also find that the Robust-FD sequence is constantly higher than the other two methods.

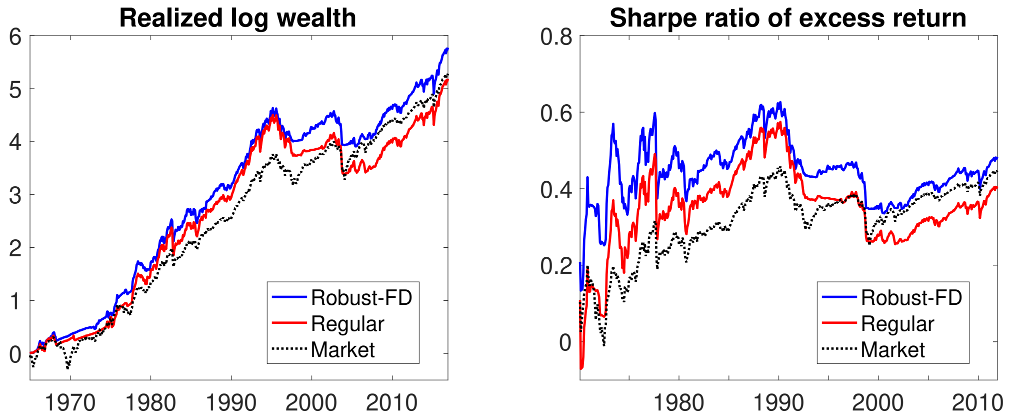

Figure 6.8: The Sharpe ratio is plotted based on expanding window ranging from January 1970 through December 2021. "Robust-FD" refers to the portfolio constructed using the forecast distribution of our proposed robust machine learning $\widehat{y}_{t+1|t} \sim \mathcal{N}(y_{t+1|t}, \mathrm{SE}(\widehat{y}_{t+1|t})^2)$, coupled with the Black-Litterman model; "Regular" refers to the portfolio constructed using the sample average and variance for (μ_T, V_T); "Market" refers to the market index return.

7 Conclusion

In this paper we develop a conditional machine learning approach to predict an aggregate market index leveraging the rich information from many predictors in a large cross section. We focus on the robustness of predictions which does not predicate on stable long term predictive relationships and does not require an expanding time series dimension to achieve good properties. We also provide a rigorous theoretical analysis of the formal properties of our estimation procedure and establish consistency and also derive confidence intervals of the neural network predictions to quantify the uncertainty associated with the forecasts. In addition, we show that the prediction robustness can be assessed by testing for structural breaks on the out-of-sample R_t^2 time series.

Empirically, we find that our method compares favorably with leading competitors in predicting a notoriously difficult target, the monthly US equity premium. Importantly, we find that our method's favorable performance does not stem from only a small number of periods, but is distributed throughout the sample, and has fewer structural breaks than some of the commonly used prediction models. We expect that our method will be useful in many different forecast applications such as international equity and bond markets, in which robustness might be an even larger concern due to limited time series availability. Our method can also be applied to forecast macroeconomic quantities, which might also be subject to instability as the structure of the economy is changing.

A Data & Implementation Details

A.1 Main Data Set

We take the data set of Chen and Zimmermann (2021) as our starting point. The dataset contains monthly data of a set of firm specific characteristics, as listed in Table 4. We delete all cases for which book-to-market is not observed, and additionally delete the first 24 months of observations for each firm to avoid possible forward looking biases following Banz and Breen (1986). This leaves us with a data set of 2,343,844 firm months observation starting in January 1955. The last column in Table 4 shows what fraction of the data is missing (after we have imposed the Banz and Breen (1986) filter and requiring that the book-to-market ratio is observed). For all analyses, we rank-transform the characteristics to $[0, 1]$ as in Freyberger et al. (2020) and Kozak et al. (2020).

We also follow Freyberger et al. (2021) to impute the missing observations which uses the cross-sectional and temporal relationship between and within a characteristic to obtain an estimate of the characteristic value for the missing case. In order to simplify the procedure, we only rely on the cross-sectional relationship between the characteristics for imputation and use a linear model for imputation. We refer to Freyberger et al. (2021) for the full procedure and more details, but for convenience, we outline the approach for two characteristics $(X_{i,1},\ X_{i,2})$. For some firms, we observe X_1 and X_2, whereas X_2 may be missing for some firms, from the complete case, we can estimate the model $X_{i,2} = \gamma_0 + \gamma_1 X_{i,1} + u_i$. We can then obtain an imputation for the missing characteristic X_2 (for the firms for which it is not observed) as $\hat{X}_{i,2} = \hat{\gamma}_0 + \hat{\gamma}_1 X_{i,1}$. We estimate the imputation models each month and complete the missing observations this way.

Table 4: Overview of the Characteristics

This Table gives an overview of the characteristic used in the empirical analysis. They are obtained from Chen and Zimmermann (2021). We refer to their paper and the companion website for the precise construction and reference to the original paper that proposed these predictors.

Acronym	Description	Publication Year	% missing
AM	Total assets to market	1992	7.92
Accruals	Accruals	1996	0.69

Table 4: Overview of the Characteristics *(continued)*

AssetGrowth	Asset growth	2008	0.05
BMdec	Book to market using December ME	1992	0.00
BMf	Book to market using most recent ME	1985	0.00
Beta	CAPM beta	1973	0.00
BetaFP	Frazzini-Pedersen Beta	2014	0.00
BetaTailRisk	Tail risk beta	2014	23.49
BidAskSpread	Bid-ask spread	1986	5.90
BookLeverage	Book leverage (annual)	1992	0.00
CF	Cash flow to market	1994	7.92
CashProd	Cash Productivity	2009	8.89
ChAssetTurnover	Change in Asset Turnover	2008	15.74
ChInv	Inventory Growth	2002	0.05
ChInvIA	Change in capital inv (ind adj)	1998	9.39
ChNNCOA	Change in Net Noncurrent Op Assets	2008	0.98
ChNWC	Change in Net Working Capital	2008	0.69
ConvDebt	Convertible debt indicator	2016	0.00
Coskewness	Coskewness	2000	0.00
DelCOA	Change in current operating assets	2005	0.05
DelCOL	Change in current operating liabilities	2005	0.69
DelFINL	Change in financial liabilities	2005	0.96
DelLTI	Change in long-term investment	2005	0.05
DelNetFin	Change in net financial assets	2005	0.96
DivInit	Dividend Initiation	1995	7.92
DivOmit	Dividend Omission	1995	7.92
DivSeason	Dividend seasonality	2013	53.74
DolVol	Past trading volume	1998	3.43
EP	Earnings-to-Price Ratio	1977	30.14
EntMult	Enterprise Multiple	2011	22.77
ExchSwitch	Exchange Switch	1995	7.92
FirmAge	Firm age based on CRSP	1984	62.62
GrLTNOA	Growth in long term operating assets	2003	2.40
GrSaleToGrInv	Sales growth over inventory growth	1998	20.03
Herf	Industry concentration (sales)	2006	15.81
HerfAsset	Industry concentration (assets)	2006	15.81
HerfBE	Industry concentration (equity)	2006	15.81
IdioRisk	Idiosyncratic risk	2006	0.00
IdioVol3F	Idiosyncratic risk (3 factor)	2006	0.00
IdioVolAHT	Idiosyncratic risk (AHT)	2003	0.00
Illiquidity	Amihud's illiquidity	2002	4.30
IndIPO	Initial Public Offerings	1991	7.92

Table 4: Overview of the Characteristics *(continued)*

IndMom	Industry Momentum	1999	7.92
IntMom	Intermediate Momentum	2012	8.02
InvGrowth	Inventory Growth	2012	37.05
InvestPPEInv	change in ppe and inv/assets	2008	10.75
Investment	Investment to revenue	2004	18.35
LRreversal	Long-run reversal	1985	8.57
Leverage	Market leverage	1988	8.16
MRreversal	Medium-run reversal	1985	8.08
MaxRet	Maximum return over month	2010	0.00
Mom12m	Momentum (12 month)	1993	8.05
Mom12mOffSeason	Momentum without the seasonal part	2008	7.95
Mom6m	Momentum (6 month)	1993	7.98
MomOffSeason	Off season long-term reversal	2008	8.03
MomOffSeason06YrPlus	Off season reversal years 6 to 10	2008	20.44
MomSeason	Return seasonality years 2 to 5	2008	8.03
MomSeason06YrPlus	Return seasonality years 6 to 10	2008	20.34
MomSeasonShort	Return seasonality last year	2008	7.98
NetEquityFinance	Net equity financing	2006	0.72
NetPayoutYield	Net Payout Yield	2007	32.70
PayoutYield	Payout Yield	2007	46.87
Price	Price	1972	0.00
PriceDelayRsq	Price delay r square	2005	2.81
PriceDelaySlope	Price delay coeff	2005	2.81
PriceDelayTstat	Price delay SE adjusted	2005	2.83
RDIPO	IPO and no R&D spending	2006	0.00
ResidualMomentum	Momentum based on FF3 residuals	2011	1.37
ReturnSkew	Return skewness	2016	0.45
ReturnSkew3F	Idiosyncratic skewness (3F model)	2016	0.00
SP	Sales-to-price	1996	8.08
STreversal	Short term reversal	1989	0.00
ShareIss1Y	Share issuance (1 year)	2008	8.09
ShareIss5Y	Share issuance (5 year)	2006	15.37
ShareRepurchase	Share repurchases	1995	0.00
Size	Size	1981	0.00
Spinoff	Spinoffs	1993	7.92
Tax	Taxable income to income	2004	11.50
TotalAccruals	Total accruals	2005	3.98
VarCF	Cash-flow to price variance	1996	8.12
VolMkt	Volume to market equity	1996	3.79
VolSD	Volume Variance	2001	4.79

Table 4: Overview of the Characteristics *(continued)*

VolumeTrend	Volume Trend	1996	5.90
grcapx	Change in capex (two years)	2006	10.26
zerotrade	Days with zero trades	2006	3.76
zerotradeAlt1	Days with zero trades	2006	3.34
zerotradeAlt12	Days with zero trades	2006	4.37

A.2 Implementations

In all our applications, we use feed-forward neural networks for estimation of expected returns. As customary in neural network modeling, model fitting is carried out by stochastic gradient descent (SGD). We use the adaptive moment estimation algorithm (Adam) introduced by Kingma and Ba (2014). We use the default parameters for Adam in the Julia package Flux. For all networks, we use ReLu activation functions, a learning rate of 0.001 and 2000 epochs. To ameliorate the sensitivity to starting values and similar to Gu et al. (2020) we create an ensemble of five forecasts (each month) and then use the ensemble average in the subsequent analysis. We use between one and three hidden layers with 32 nodes on the first hidden layer, 16 nodes on the second hidden layer (if it exists) and 8 nodes on the third hidden layer (if it exists). This results in three different versions for estimates of expected returns.

As one of the competing forecast models, we implemented a linear forecast using the 16 predictors from Goyal et al. (2021), estimated using ridge regression (GW-linear). In addition, we also implemented the random Fourier transformations of 16 predictors (GW-Fourier). This method, originated from Rahimi and Recht (2007), uses d random Fourier bases as $g_{i,t} := (\sin(\pi\omega_i' w_t/2), \cos(\pi\omega_i' w_t/2))$ for $i = 1, ..., d$, where w_t contains the 16 predictors, and $\omega_i \sim \mathcal{N}(0,1)$. Then run the prediction regression:

$$y_{t+1} = \rho_0 + \sum_{i=1}^{d} \rho_i' g_{i,t} + \text{noise}. \tag{A.1}$$

We take $d = 300$ and use ridge regression for (A.1). In addition, as the randomness of the transformations would produce unstable forecasts, we repeat this process for 10 times, and take the final forecaster $\widehat{y}_{t+1|t}$ as the average of them. Kelly et al. (2021) apply random Fourier features to the Welch and Goyal (2008) predictors and find these nonlinear transformations greatly improve predictability at the aggregate index level, whereas a linear model using the same 16 variables does not lead to predictive gains relative to the historical mean.

Finally, we also compared with the PCA-ker method, which is to allow time-varying betas using kernel based PCA on individual stocks. In the econometric literature, this method was first studied by Su and Wang (2017). Specifically, we estimate the stock-betas using $\widehat{\beta}_{t-1}$ which equals $\sqrt{N}$ times the $N \times K_f$ eigenvector matrix of S_t, corresponding to the top eigenvalues of kernel-weighted stocks returns: $\frac{1}{T}\sum_{s=1}^{T} x_s x_s' K_{s,t}$. then apply a linear forecast model using estimated stock-factors:

$$\widehat{f}_t = \frac{1}{N}\sum_{i=1}^{N} \widehat{\beta}_{i,t-1} x_{i,t}.$$

B Technical Appendix

B.1 Technical Assumptions

B.1.1 Assumptions for Proposition 1: forecast confidence intervals

Our theoretical analysis requires on the forecast uncertainty relies on a set of regularity conditions. Distribution theories for deep neural networks and time-varying conditional factor models are rather sophisticated, hence needless to say, some technical conditions are required, although we do not attempt to pursuing the minimum number of necessary conditions. The conditions below can be classified into four categories:

I. Regulates the dependences of the data generating process (DGP).

II. Impose some moment-bounds on various quantities.

III. Quantifies the degree of time-varyingness on the betas and some second moments.

IV. Complexity and approximation theories of the deep neural network, as well as various statistical convergences.

Assumption 1 (DGP). *(i)* $\mathbb{E}(u_t|\mathcal{I}_{t-1}, f_t, g_t, z_{t-1}) = 0$ *and* $\mathbb{E}(\eta_t|\mathcal{I}_{t-1}, f_t, g_t, z_{t-1}) = 0$. *Conditioning on* $z_{i,t}$ *and* f_t, $\eta_{i,s}$ *are independent over* i *for all* s, t. *In addition,* $(u_{it}, z_{i,t})$ *is independent over* i.

(ii) $u_{i,t}$ *and* η_{it} *are sub-Gaussian with uniformly bounded sub-Gaussian norm.*

The first condition assumes the dependence structure of the DGP. In particular, we allow the stock-idiosyncratic and BM- idiosyncratic η_t and u_t to be possibly conditionally correlated. We suppose that these noises are sub-Gaussian.

The following conditions requies some moment - bounds of various quantities. These conditions extend those for the usual PCA to the time-varying case.

Assumption 2 (Bounds). *(i)* $\max_t \frac{1}{T}\sum_s (1+\|f_s\|)|K_{s,t}| < C$, $\mathrm{Var}(\eta_{it}|z_{i,t}, f_t) < C$, $\mathbb{E}\|f_s\|^6 + \mathbb{E}\|f_s\|^4\|g_t\|^2 + \mathbb{E}\|f_s\|^4\epsilon_{t+1}^2 < C$, $\max_t \|\beta_t\| + \max_t \|\lambda_t\| < C\sqrt{N}$.

(ii) For all s, t, $\mathbb{E}\|\frac{1}{\sqrt{N}}\beta_{s-1}' u_t\|^2 < C$, $\|\mathbb{E}(u_t u_t'|f_t, g_t, \epsilon_{t+1})\| < C$, $\|\mathbb{E}(u_t u_t'|f_t, \mathcal{I}_{t-1})\| < C$.

(iii) All singular values of $\frac{1}{N}\beta_t'\beta_t$, $S_{f,t} = \frac{1}{Th}\sum_s f_s f_s' K_{s,t}$ *are bounded away from zero; the top* K_g *singular values of* $\frac{1}{N}\beta_t'\lambda_t$ *are bounded away from zero, uniformly over* t.

(iv) $S_{f,t}^{1/2}\Sigma_{\beta,t}S_{f,t}^{1/2}$ *has distinct eigenvalues. By "distinct", we mean there is* $c > 0$, *so that* $|\nu_j(t) - \nu_i(t)| > c$ *for all* $j \leq K_f$ *and* $t \leq T$. *Here* $\nu_j(t)$ *denotes the* j *th largest eigenvalue.*

Next, we quantify the "slow-varying" betas, to adopt the local-PCA. By "slow-varying" betas, we assume that their trajectories are twice differentiable functions of time. This is the standard treatment for conditional asset pricing, as in Ang and Kristensen (2012).

Assumption 3 (Slow-varying). *There are functions $\beta_i(\cdot)$, $\lambda_i(\cdot)$, $s_f(\cdot)$ and $J(\cdot)$ so that $\beta_{i,t-1} = \beta_i(\frac{t}{T})$, $\lambda_{i,t-1} = \lambda_i(\frac{t}{T})$, $\mathbb{E} f_t f_t' = s_f(\frac{t}{T})$ and $v_t = J(\frac{t}{T})$, where $\Sigma_{\beta,t} = plim \frac{1}{N}\beta_{t-1}'\beta_{t-1}$ and v_t denotes the matrix of eigenvector of $\Sigma_{f,t}^{1/2}\Sigma_{\beta,t}\Sigma_{f,t}^{1/2}$. In addition, $\beta_i(\cdot)$, $s_f(\cdot)$ and $J(\cdot)$ are twice continuously differentiable, with bounded second derivatives.*

Next, we impose three types of convergence properties as in the following assumption.

Assumption 4 (Convergence). *(i) There is a diagonal matrix $\bar{A}_t$ so that*

$$\max_t \|\bar{A}_t - \Sigma_{\beta,t}^{1/2} S_{f,t} \Sigma_{\beta,t}^{1/2}\| = o_P((Th)^{-1/2}).$$

(ii) $\max_t \frac{1}{N}\beta_{t-1}' \frac{1}{TN}\sum_s u_s f_s' K_{s,t} = o_P((Th)^{-1/2})$ *and*
$\max_t \frac{1}{N}\beta_{t-1}' \frac{1}{T}\sum_s (\beta_{s-1} - \beta_{t-1}) \frac{1}{N}(f_s f_s' - \mathbb{E} f_s f_s')\beta_{s-1}'\widehat{\beta}_{t-1} V_t^{-1} K_{s,t} = o_P((Th)^{-1/2})$.
$\max_{it} \max_{r_1,r_2} \|\frac{1}{T}\sum_s \beta_{i,s-1,r_1}\beta_{i,s-1,r_2}(f_s f_s' - \Sigma_f)K_{s,t}\| = o_P(1)$.
Furthermore, conditioning on $\{f_T, g_T\}$, for some covariance matrix $W = diag\{W_1, W_2\}$,

$$\Big(\frac{\sqrt{Th}}{T}\sum_t F_t \epsilon_{t+1} K_{t,T}, \frac{1}{\sqrt{N}} \mathrm{vec}(\bar{\Upsilon}_T' \beta_{T-1}' \xi_T)\Big) \to^d \mathcal{N}(0, W)$$

where $\bar{\Upsilon}_t$ is defined as the probability limit of $\Upsilon_t = \frac{1}{T}\sum_s \frac{1}{N} f_s f_s' \beta_{s-1}' \widehat{\beta}_{t-1} V_t^{-1} K_{s,t}$.

(iii) $Th^3 = o(1)$ and $Th \to \infty$. Also, $r_T^2 + r_T\varphi_N + r_T\delta_N = o(N^{-1})$ and

$$\|\pi m_1 - \pi m_2\|_\infty \leq C\|m_1 - m_2\|_\infty,$$

where $\pi()$ denotes the projection onto the DNN space under the $\|h\|_\infty = \sup_x |h(x)|$ norm; $p(\mathrm{DNN})$ denotes the pseudo-dimension of the DNN space, see the definition in Bartlett et al. (2019), and

$$\begin{aligned} \delta_N^2 &= \frac{p(\mathrm{DNN})\log(NT)}{N}, \\ \varphi_N^2 &= \max_t \inf_{h\in \mathrm{DNN}} \sup_z |h(z) - m_t(z)|^2 \\ r_T^2 &= \sup_{h\in\{\tau_T g_t + \mathrm{DNN}\}} \inf_{m\in \mathrm{DNN}} \|m-h\|_\infty^2. \end{aligned} \qquad \text{(B.1)}$$

Condition (i) is an identifiable condition. As it is well known that PCA- based methods estimate factors and betas up to a rotation. Condition (i) in the assumption identifies the probability limit of the rotation matrix, and is spiritually similar to the conditions imposed by Bai and Ng (2013), but extends to the conditional factor model context.

Condition (ii) is a set of statistical convergence conditions. They are imposed in a relatively high-level to allow some temporal dependences. In simpler cases with purely independent data, both uniform convergence and the central limit theorem can be directly verified.

Condition (iii) regulates the relative rates of the bandwidth for local-PCA, as well as the approximation rates and complexity of the deep neural network space.

B.1.2 Assumptions for Propositions 2, 3

Assumption 5 (pooled ML). *Let $\mathcal{F}_{z,t}$ denote the filtration of characteristics up to time t. Define $v_{t+1} := f_{t+1} - \mathbb{E}(f_{t+1}|\mathcal{F}_{z,t})$. Define $m(z) = g_\alpha(z) + h_{\beta,t}(z)'\mathbb{E}(f_{t+1}|\mathcal{F}_{z,t})$. Let πm denote the projection of m in the DNN space. Then:*

(i) Conditioning on $z_{i,t-1}$, the sequence of v_t is alpha-mixing and sub-Gaussian and $z_{i,t}$ is independent across i. (see various papers for the precise definition of alpha-mixing and subexponential (Fan et al., 2013).)

(ii) $\mathbb{E}(\epsilon_{T+1}|\mathcal{F}_{z,T}) = \mathbb{E}(u_{i,T+1}|\mathcal{F}_{z,T}) = 0$, and $\mathbb{E}(f_{t+1}|\mathcal{F}_{z,t})$ does not vary across t.

(iii) $\|g_\alpha\|_\infty + \|g_\beta\|_\infty + \sup_{\mathrm{DNN}} \|h\|_\infty < C$ for some constant $C > 0$.

(iv) $\sup_z |\pi m(z) - m(z)| = o_P(1)$ and $p(\mathrm{DNN}) \log N = o(T)$.

Assumption 6 (structural breaks). *Let $\mathcal{F}_t$ denote the usual information set up to time t for conditional forecasting, which includes both characteristic filtration $\mathcal{F}_{z,t}$ and latent factors f_t.*

(i) For the forecasting model M, suppose $\widehat{y}_{s+1|s}$ "converges" to some quantity, denoted by $\mathbb{E}(y_{s+1}|\mathcal{F}_s, M)$, in the sense that $\frac{1}{|\mathcal{S}_t|}\sum_{s\geq t}[\mathbb{E}(y_{s+1}|\mathcal{F}_s, M) - \widehat{y}_{s+1|s}]^2 = o_P(1)$.

(ii) $\mathbb{E}(y_{s+1}|\mathcal{F}_s, M)$ may not be equal to the true conditional mean $\mathbb{E}(y_{s+1}|\mathcal{F}_s)$ due to possible model specification errors/structural breaks. But the discrepancy has a limit $\sigma^2_\zeta(t, M)$ in the sense that $\frac{1}{|\mathcal{S}_t|}\sum_{s\geq t}[\mathbb{E}(y_{s+1}|\mathcal{F}_s) - \mathbb{E}(y_{s+1}|\mathcal{F}_s, M)]^2 \to^P \sigma^2_\zeta(t, M)$.

(iii) $\mathbb{E}(\epsilon_{t+1}|\mathcal{F}_t) = 0$.

(iv) There are $\sigma^2_{un}(t)$ and $\sigma^2_\epsilon(t)$ so that $\frac{1}{|\mathcal{S}_t|}\sum_{s\geq t}(y_{s+1} - \bar{y}_s)^2 \to^P \sigma^2_{un}(t)$ and $\frac{1}{|\mathcal{S}_t|}\sum_{s\geq t}\epsilon^2_{s+1} \to^P \sigma^2_\epsilon(t)$ at the periof of interest t.

References

Ang, A. and G. Bekaert (2007). Stock return predictability: Is it there? *The Review of Financial Studies 20*(3), 651–707.

Ang, A. and D. Kristensen (2012). Testing conditional factor models. *Journal of Financial Economics 106*(1), 132–156.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*, 135–171.

Bai, J. and S. Ng (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica 74*(4), 1133–1150.

Bai, J. and S. Ng (2013). Principal components estimation and identification of static factors. *Journal of econometrics 176*(1), 18–29.

Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 47–78.

Bai, J. and P. Perron (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics 18*(1), 1–22.

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.

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

Bianchi, D., A. Rubesam, and A. Tamoni (2023). It takes two to tango: Economic theory and model uncertainty for equity premium prediction. *Available at SSRN 4513241*.

Black, F. and R. Litterman (1990). Asset allocation: combining investor views with market equilibrium. *Goldman Sachs Fixed Income Research 115*(1), 7–18.

Campbell, J. Y. and S. B. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies 21*(4), 1509–1531.

Carriero, A., T. E. Clark, and M. Marcellino (2018). Measuring uncertainty and its impact on the economy. *Review of Economics and Statistics 100*(5), 799–815.

Chen, A. Y. and T. Zimmermann (2021). Open source cross sectional asset pricing. *Critical Finance Review, Forthcoming*.

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.

Clark, T. E., M. W. McCracken, and E. Mertens (2020). Modeling time-varying uncertainty of multiple-horizon forecast errors. *Review of Economics and Statistics 102*(1), 17–33.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance 66*(4), 1047–1108.

Connor, G., M. Hagmann, and O. Linton (2012). Efficient semiparametric estimation of the fama–french model and extensions. *Econometrica 80*(2), 713–754.

Dangl, T. and M. Halling (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics 106*(1), 157–181.

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

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.

Fan, J., Y. Liao, and W. Wang (2016). Projected principal component analysis in factor models. *Annals of Statistics 44*(1), 219–254.

Farmer, L., L. Schmidt, and A. Timmermann (2022). Pockets of predictability. *Journal of Finance, forthcoming*.

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., B. Höppner, A. Neuhierl, and M. Weber (2021). Missing data in asset pricing panels. *Available at SSRN*.

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

Giacomini, R. and B. Rossi (2009). Detecting and predicting forecast breakdowns. *The Review of Economic Studies 76*(2), 669–705.

Goyal, A. and I. Welch (2003). Predicting the equity premium with dividend ratios. *Management Science 49*(5), 639–654.

Goyal, A., I. Welch, and A. Zafirov (2021). A comprehensive look at the empirical performance of equity premium prediction ii. *Available at SSRN 3929119*.

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

Gu, S., B. T. Kelly, and D. Xiu (2019). Autoencoder asset pricing models.

Hansen, L. P. and S. F. Richard (1987). The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. *Econometrica 55*, 587–613.

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

Kelly, B. and S. Pruitt (2013). Market expectations in the cross-section of present values. *The Journal of Finance 68*(5), 1721–1756.

Kelly, B. and S. Pruitt (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics 186*(2), 294–316.

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.

Kelly, B. T., S. Pruitt, and Y. Su (2020). Instrumented principal component analysis. *Available at SSRN 2983919*.

Kim, S., R. A. Korajczyk, and A. Neuhierl (2021). Arbitrage portfolios. *The Review of Financial Studies 34*(6), 2813–2856.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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

Lettau, M. and S. Ludvigson (2010). Measuring and modeling variation in the risk-return trade-off. *Handbook of Financial Econometrics 1*, 617–690.

Lettau, M. and S. Van Nieuwerburgh (2008). Reconciling the return predictability evidence. *The Review of Financial Studies 21*(4), 1607–1652.

Li, Q. and J. S. Racine (2007). *Nonparametric econometrics: theory and practice.* Princeton University Press.

McCracken, M. W. (2007). Asymptotics for out of sample tests of granger causality. *Journal of econometrics 140*(2), 719–752.

Merton, R. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science 4*, 141–183.

Paye, B. S. and A. Timmermann (2006). Instability of return prediction models. *Journal of Empirical Finance 13*(3), 274–315.

Perron, P. and G. Rodríguez (2003). Searching for additive outliers in nonstationary time series. *Journal of Time Series Analysis 24*(2), 193–220.

Pesaran, M. H. and A. Timmermann (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance 50*(4), 1201–1228.

Pettenuzzo, D. and A. Timmermann (2011). Predictability of stock returns and asset allocation under structural breaks. *Journal of Econometrics 164*(1), 60–78.

Polk, C., S. Thompson, and T. Vuolteenaho (2006). Cross-sectional forecasts of the equity premium. *Journal of Financial Economics 81*(1), 101–141.

Rahimi, A. and B. Recht (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems 20*.

Rapach, D. E. and M. E. Wohar (2006). Structural breaks and predictive regression models of aggregate us stock returns. *Journal of Financial Econometrics 4*(2), 238–274.

Rosenberg, B. and W. McKibben (1973). The prediction of systematic and specific risk in common stocks. *Journal of Financial and Quantitative Analysis 8*(2), 317–333.

Rossi, B. (2021). Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *Journal of Economic Literature 59*(4), 1135–90.

Rossi, B. and T. Sekhposyan (2015). Macroeconomic uncertainty indices based on nowcast and forecast error distributions. *American Economic Review 105*(5), 650–655.

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

Shanken, J. (1990). Intertemporal asset pricing: An empirical investigation. *Journal of Econometrics 45*(1-2), 99–120.

Stock, J. and M. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Su, L. and X. Wang (2017). On time-varying factor models: Estimation and testing. *Journal of Econometrics 198*(1), 84–101.

Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting 24*(1), 1–18.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies 21*(4), 1455–1508.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica: Journal of the Econometric Society*, 1067–1084.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.