

The nonstationarity-complexity tradeoff in return prediction

Agostino Capponi* Chengpiao Huang[†] J. Antonio Sidaoui[‡] Kaizheng Wang[§]
Jiacheng Zou[¶]

This version: December 29, 2025

Abstract

We investigate machine learning models for stock return prediction in non-stationary environments, revealing a fundamental *nonstationarity-complexity tradeoff*: complex models reduce misspecification error but require longer training windows that introduce stronger non-stationarity. We resolve this tension with a novel model selection method that jointly optimizes model class and training window size using a tournament procedure that adaptively evaluates candidates on non-stationary validation data. Our theoretical analysis demonstrates that this approach balances misspecification error, estimation variance, and non-stationarity, performing close to the best model in hindsight.

Applying our method to 17 industry portfolio returns, we consistently outperform standard rolling-window benchmarks, improving out-of-sample R^2 by 14–23% on average. During NBER-designated recessions, improvements are substantial: our method achieves positive R^2 during the Gulf War recession while benchmarks are negative, and improves R^2 in absolute terms by at least 80bps during the 2001 recession as well as superior performance during the 2008 Financial Crisis. Economically, a trading strategy based on our selected model generates 31% higher cumulative returns averaged across the industries.

Keywords: Non-stationarity, Model complexity, Return prediction, Model selection, Adaptive window selection

1 Introduction

Machine learning (ML) models have emerged as powerful tools for return prediction in financial markets. Recent studies demonstrate that ML models can effectively approximate stochastic discount factor (SDF) by capturing complex nonlinear relationships between risk information carried by characteristics and asset returns (Gu et al., 2020; Freyberger et al., 2020; Kelly and Xiu, 2023;

Authors are listed in alphabetical order.

*Department of IEOR and Columbia Business School, Columbia University. Email: ac3827@columbia.edu.

[†]Department of IEOR, Columbia University. Email: chengpiao.huang@columbia.edu.

[‡]Department of IEOR, Columbia University. Email: j.sidaoui@columbia.edu.

[§]Department of IEOR and Data Science Institute, Columbia University. Email: kaizheng.wang@columbia.edu.

[¶]Email: jiachengzou@alumni.stanford.edu.

Kelly et al., 2024). While these studies have shown impressive predictive performance, they implicitly admit a degree of freedom vis-a-vis how historical data is utilized for estimation. The literature typically adopts one of two regimes: either an expanding window that uses all historical data, or a fixed-length rolling window which uses only the most recent observations in a fixed look-back horizon. Both conventions follow the same operating logic: to firstly consider the choice of model class, such as linear regression, random forest, or any other ML model; secondly and conditioned on the model choice, pick a training window. Since financial markets are subject to shocks and exhibit cycles, this separation of model class and training window is ad-hoc rather than built upon rigorous statistical design.

We show that in a non-stationary environment, the complexity of an approximation model of SDF and the estimation window length are deeply intertwined and cannot be optimized independently. They are linked by a fundamental *nonstationarity-complexity tradeoff*. While complex models are effective at reducing model misspecification error, they simultaneously require a larger volume of training data to mitigate their inherent estimation variance. Extending the training window to meet this data requirement increases the risk of incorporating outdated economic regimes, thereby introducing non-stationarity bias that can degrade a SDF estimator’s predictive performance. This tension creates a “less can be more” dilemma, where a complex model trained on a long window of past data may be outperformed by a simpler model trained on a shorter, more recent window of data. Consequently, the optimal model complexity depends on the training window size, and vice versa.

A rapidly growing branch of literature has recently established the “virtue of complexity” in return prediction, demonstrating that complex, high-dimensional models can significantly outperform simpler, parsimonious benchmarks (Kelly and Malamud, 2025; Kelly et al., 2022, 2024). Drawing on the universal approximation property of neural networks, these studies prove that models where the number of parameters exceeds the number of observations can better leverage the information content of predictive signals by accurately approximating unknown nonlinear functions that govern asset returns. Our work complements this literature by introducing a new dimension to this framework: the role of *non-stationarity* in the training environment for return prediction.

The “virtue of complexity” literature demonstrates that complex ML models can effectively capture nonlinear SDF relationships. These studies typically employ expanding windows with all available historical data. They show that approximation gains from model flexibility outweigh the statistical costs of heavy parameterization in the classical bias-variance tradeoff. Our work complements this literature by examining how non-stationarity in financial markets, arising from structural breaks and economic cycles, affects the optimal choice of model complexity and training window. When the data generating process shifts over time, realizing the virtue of complexity requires carefully selecting how much historical data to include in training.

Our findings suggest that model complexity and training window size cannot be optimized independently as fixed hyperparameters; rather, they must be chosen jointly to balance misspecification error, statistical uncertainty, and environmental drift. This is the central problem studied in our

paper:

How to jointly choose the model complexity and training window size?

We complement the insights of the machine learning asset pricing literature, including the surprising dominance of large factor models (Didisheim et al., 2024) and the success of deep learning and complexity in return prediction (Kelly et al. (2022) and their extensions), by considering the case of unknown temporal distribution shifts. By proposing a data-driven framework that adaptively selects the optimal model class and training window size simultaneously, we offer a method for navigating the complex relationships between predictors and returns as they evolve over time. Our method adaptively selects validation data tailored to the local non-stationarity, allowing for a near-optimal estimation of a model’s future performance. Our framework is general: it can compare any candidate models from different model classes trained on different horizons in any manner.

Our main contributions are three-fold. First, we provide empirical and theoretical investigations of a fundamental nonstationarity-complexity tradeoff in return prediction under non-stationarity. In an empirical study on industry portfolio return prediction, we show that models with greater expressive power or longer training windows may underperform when the environment changes over time. We then formalize this phenomenon through a finite-sample bound that characterizes the prediction error of a model f in terms of its model class \mathcal{F} and training window size k :

$$\text{Prediction Error}(f) \lesssim \text{Misspecification}(\mathcal{F}) + \text{Uncertainty}(\mathcal{F}, n_k) + \text{Non-stationarity}(k). \quad (1.1)$$

The bound decomposes the prediction error into three sources: the model misspecification error of the model class \mathcal{F} , the statistical uncertainty associated with learning the model using n_k samples in the training window, and the non-stationarity within the last k periods. This characterization quantifies how model complexity and training window length jointly influence the model’s predictive performance.

Second, motivated by this tradeoff, we develop an adaptive model selection approach for jointly choosing the model class and training window length. Our method is a sequential elimination tournament procedure, and uses a pairwise model comparison subroutine that adaptively selects non-stationary validation data to compare two given models. We prove that our algorithm jointly chooses a model class and training window that near-optimally balance the nonstationarity-complexity tradeoff (1.1), up to logarithmic factors. Furthermore, we develop a variant tailored to the out-of-sample R^2 metric commonly used in asset pricing.

Third, we demonstrate the empirical efficacy of this framework on daily returns of 17 industry portfolios, and show that it adapts to the local non-stationarity and significantly improves the out-of-sample (OOS) R^2 compared to non-adaptive fixed-window baselines. Over the 1990–2016 OOS period, our method delivers an average R^2 of 0.049 across all industries, representing a 14% improvement over fixed-horizon training with long-horizon validation, and more than doubling the performance of short-horizon validation.

Our method’s advantages are most pronounced during recessions, when non-stationarity is most

evident. We examine the three recessions identified by National Bureau of Economic Research (NBER) in its [NBER Business Cycle Dating](#) that fall within our OOS period. During the 1990 Gulf War recession, our framework achieves a positive R^2 of 0.027 while all benchmarks produce negative R^2 , demonstrating the critical importance of handling non-stationarity properly. In the 2001 recession, our method attains an R^2 of 0.125, outperforming the cross-validation benchmark which attains 0.071, a 540 basis point improvement, and the long-window validation benchmark which achieves 0.117, an 80 basis point improvement. During the 2008 Financial Crisis, our method again delivers the strongest performance. These gains are robust across all benchmark methods (Table 2) and persistent across industries (Figure 5), confirming that our adaptive approach effectively navigates the nonstationarity-complexity tradeoff.

Economically, our predictive gains translate to meaningful value: a simple trading strategy based on our selected models generates 31% higher returns than the best-performing validation benchmark, averaged across the 17 industries. This confirms that jointly optimizing model complexity and training window size to address non-stationarity yields substantial benefits for investors.

1.1 Related Literature

The integration of machine learning into asset pricing was initially driven by the “multidimensional challenge”, that is, the need to identify which of the hundreds of proposed firm characteristics provide independent information for expected returns.

Early influential work by [Gu et al. \(2020\)](#) demonstrated that nonlinear interactions missed by traditional regressions are a primary source of predictive gains, identifying trees and neural networks as superior methods. [Freyberger et al. \(2020\)](#) used adaptive group LASSO to show that only a small subset of characteristics provides incremental information when nonlinearities are properly accounted for. [Choi et al. \(2025\)](#) applied machine learning to 32 international markets, concluding that market-specific neural networks achieve stronger results than global models by capturing local return-characteristic relationships. We refer to [Kelly and Xiu \(2023\)](#) for an excellent survey on financial machine learning.

Building on these empirical successes, a series of theoretical papers have formalized the virtue of complexity, proving that out-of-sample forecast accuracy and portfolio Sharpe ratios can be strictly increasing in model complexity. This phenomenon occurs because high complexity induces “implicit shrinkage”, which reduces prediction variance without the heavy bias costs associated with explicit shrinkage. This line of research advocates for the largest approximating model one can compute, because the gains from better approximation of the unknown truth dominate the statistical costs of heavy parameterization. Foundational works in this stream of literature include [Kelly et al. \(2022\)](#); [Kelly and Malamud \(2025\)](#); [Kelly et al. \(2024\)](#). They focus on time-series return prediction and market timing, resolving the “double limit” problem of growing parameters and observations to show that complexity captures unknown nonlinearities that improve Sharpe ratios. A recent study by [Didisheim et al. \(2024\)](#) extends these insights to the cross-section of returns, tackling a “three infinities” problem involving a simultaneously large number of assets, parameters,

and observations. Their work proposes using random Fourier features to generate vast numbers of nonlinear factors, shifting the statistical objective from pure return prediction to minimizing pricing errors and constructing a high-complexity stochastic discount factor that reflects the true drivers of investors’ marginal rates of substitution.

The above surveyed works provide compelling empirical evidence and theoretical justifications for the superiority of complex machine learning models over simple linear models. However, they typically treat the training window as a fixed hyperparameter, often setting it to an expanding window that includes all available historical data. Furthermore, their theoretical analysis typically assumes that the training data is i.i.d. Our work points out that, as the financial market is in constant motion due to structural breaks, shifting risk regimes, and economic cycles, accounting for non-stationarity beyond complexity may lead to even further improvements.

A rich literature has developed statistical frameworks for detecting structural breaks and change points (Banerjee and Urga, 2005). Foundational works such as Chow (1960); Andrews (1993); Bai and Perron (1998); Chib (1998) established rigorous methods to identify structural changes, which have been applied to financial time series including realized volatility (Liu and Maheu, 2007) and speculative bubbles (Homm and Breitung, 2012). While these studies focus on identifying when breaks occur, our work addresses how to optimally use non-stationary data for estimation. Rather than pinpointing change points, we determine the optimal training window to minimize prediction error in the presence of non-stationarity.

A complementary literature examines optimal training window selection under non-stationarity. Pesaran and Timmermann (2007) showed that under structural breaks, optimal window selection should balance the bias from including pre-break data against the variance from using only post-break data. Subsequent work explored various selection criteria, including minimizing estimation loss functions (Pesaran and Timmermann, 2007; Inoue et al., 2017) and aggregating predictions across multiple windows (Pesaran and Pick, 2011). However, these approaches typically assume linear models and specific non-stationary structures, such as single breaks or random walks. Our contribution differs in two key ways. First, we take a model-free approach that does not impose parametric assumptions on either the prediction function or the non-stationary dynamics, allowing us to handle more general patterns. Second, we extend the bias-variance tradeoff identified by Pesaran and Timmermann (2007) to the machine learning context by jointly optimizing model complexity and training window size, whereas prior work typically selects windows for a pre-specified model. This joint selection accounts for the fact that more complex models introduce additional misspecification-variance tradeoffs that interact with the choice of training window.

The rest of the paper is organized as follows. Section 2 describes the problem setup. Section 3 investigates the nonstationarity-complexity tradeoff. Section 4 presents the adaptive model selection algorithm. Section 5 illustrates our algorithm on real datasets. Section 6 concludes the paper and discusses future directions. Mathematical proofs are deferred to the supplemental materials.

Notation and Terminology. We introduce the mathematical notation used throughout the paper. Let $\mathbb{Z}_+ = \{1, 2, \dots\}$ be the set of positive integers. For $n \in \mathbb{Z}_+$, define $[n] = \{1, 2, \dots, n\}$. For $a, b \in \mathbb{R}$, define $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For $x \in \mathbb{R}$, let $x_+ = x \vee 0$. The sign of a real number $x \in \mathbb{R}$ is defined by $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = 0$ if $x = 0$, and $\text{sign}(x) = -1$ if $x < 0$. For non-negative sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we write $a_n = O(b_n)$ if there exists $C > 0$ such that for all $n \in \mathbb{Z}_+$, $a_n \leq Cb_n$. We write $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. Unless otherwise stated, $a_n \lesssim b_n$ also represents $a_n = O(b_n)$. For a finite set S , we use $|S|$ to denote its cardinality.

2 Problem Setup

We consider the problem of predicting a response variable $y \in \mathbb{R}$, such as an asset return, using a vector of covariates \mathbf{x} in a space $\mathcal{X} \subseteq \mathbb{R}^d$. A key feature in our setting is *non-stationarity*: in each time period $t = 1, \dots, T$, the covariates and response (\mathbf{x}, y) follow a time-varying joint distribution P_t . At the beginning of period t , we have access to historical data $\{\mathcal{D}_j\}_{j=1}^{t-1}$, where $\mathcal{D}_j = \{(\mathbf{x}_{j,i}, y_{j,i})\}_{i=1}^{B_j}$ is a set of i.i.d. samples collected from P_j at time j . Throughout our paper, we will assume that the samples are independent across time.

Assumption 2.1 (Independent data). *For each $j \in \mathbb{Z}_+$, the dataset $\{(\mathbf{x}_{j,i}, y_{j,i})\}_{i=1}^{B_j}$ consists of i.i.d. samples. The datasets $\{\mathcal{D}_j\}_{j=1}^\infty$ are independent.*

While financial time series inherently exhibit temporal dependence, Assumption 2.1 is a standard simplification in the theoretical analysis of machine learning for return prediction and asset pricing (Kelly et al., 2024; Didisheim et al., 2024). Adopting this independence assumption allows us to isolate the effect of non-stationarity, without introducing additional technicalities from temporal dependence.

Our goal is to use the historical data $\{\mathcal{D}_j\}_{j=1}^{t-1}$ to construct a prediction model $f_t : \mathcal{X} \rightarrow \mathbb{R}$ that performs well on the current, unobserved distribution P_t . The performance of a model $f : \mathcal{X} \rightarrow \mathbb{R}$ with respect to the data distribution P_t is measured by the mean squared error (MSE):

$$L_t(f) = \mathbb{E}_{(\mathbf{x}, y) \sim P_t} \left[(f(\mathbf{x}) - y)^2 \right]. \quad (2.1)$$

In line with the empirical finance literature, we also use the R^2 metric to evaluate the performance of a given modeling procedure or algorithm `alg` that produces a prediction model f_t^{alg} at each time

t . The out-of-sample R^2 for the algorithm `alg` over an evaluation period $[t_1, t_2]$ is computed as¹

$$R^2_{[t_1, t_2]}(\text{alg}) = 1 - \frac{\sum_{t=t_1}^{t_2} \sum_{i=1}^{B_t} \left(f_t^{\text{alg}}(x_{t,i}) - y_{t,i} \right)^2}{\sum_{t=t_1}^{t_2} \sum_{i=1}^{B_t} y_{t,i}^2}. \quad (2.2)$$

For completeness, in the appendices we also present results using the statistical R^2 metric

$$R^2_{[t_1, t_2], \text{s}}(\text{alg}) = 1 - \frac{\sum_{t=t_1}^{t_2} \sum_{i=1}^{B_t} \left(f_t^{\text{alg}}(x_{t,i}) - y_{t,i} \right)^2}{\sum_{t=t_1}^{t_2} \sum_{i=1}^{B_t} (y_{t,i} - \bar{y})^2}, \quad (2.3)$$

where \bar{y} is the mean of the samples $\{y_{t,i} : t \in [t_1, t_2], i \in [B_t]\}$.

In a stationary environment where $P_t = P$ for all $t \in [T]$, the standard approach for learning a model f consists in choosing a model class \mathcal{F} (e.g., linear model, random forest) and then finding a model $\hat{f} \in \mathcal{F}$ by minimizing the empirical loss over the training data. The choice of the model class \mathcal{F} involves a classic bias-variance trade-off. A simple class may exhibit high bias due to model misspecification, while a complex class may suffer from high estimation variance.

When the environment is non-stationary (that is, $P_i \neq P_j$ for $i \neq j$), the problem becomes significantly more complicated. One must now make two critical choices simultaneously: the model class \mathcal{F} and the amount of historical data used for training. Data from the distant past may no longer be representative of the current environment, and can be misleading for model training. This creates the core tension of our paper: complex models require more data to reduce estimation variance, but using more data may introduce stronger non-stationarity that increases bias. Thus, it is possible for simple models with less training data to outperform complex models trained on more data. Our goal is to develop an approach to jointly choose the model class and training window size.

3 The Nonstationarity-Complexity Tradeoff

3.1 Empirical Evidence

We begin with an empirical illustration that highlights the challenges of jointly choosing a model class and a training window under non-stationarity. The task is to forecast the excess returns of 17 industry portfolios from Kenneth French’s data library using a set of covariates, with training data starting from September 1987 and ending in October 2016.² We highlight that our data spans several recessions documented in [NBER Business Cycle Dating](#): the 1990 Gulf War recession, the 2001 dot-com bubble bust and the 9/11 attack, and the 2007-2009 Financial Crisis. To show that

¹We note that in the R^2 metric (2.2), the denominator is the sum of the squared responses $y_{t,i}^2$ *without demeaning*. In other words, we are benchmarking against a forecast of zero rather than the historical mean as in the statistical R^2 metric. As noted by [Gu et al. \(2020\)](#), predicting future excess stock returns with historical averages can be problematic and is not assumption-free, because the historical mean is estimated with significant noise, often performing worse than a forecast of zero.

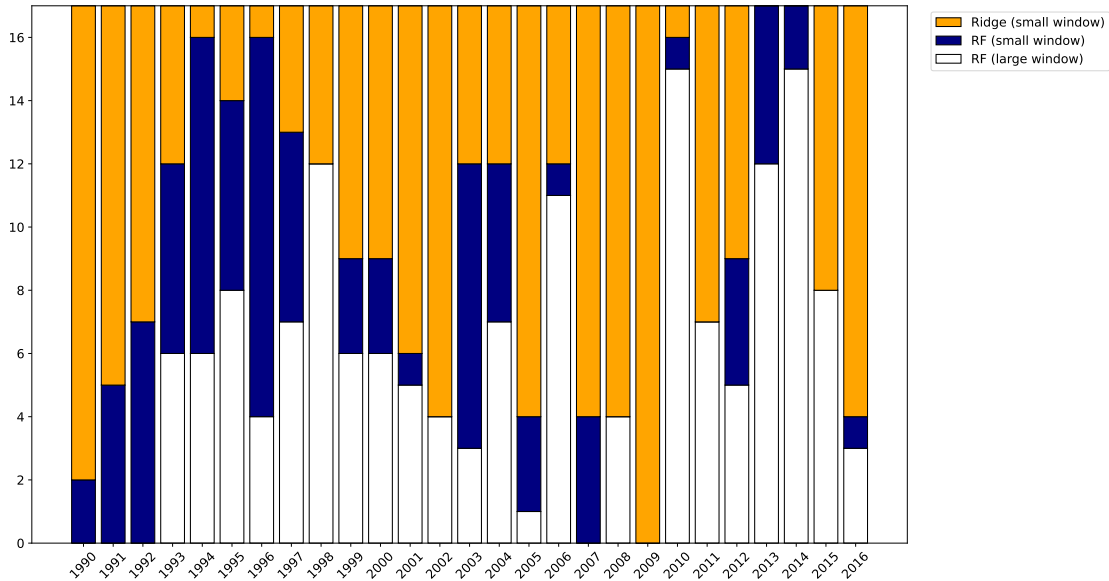
²More details about the dataset are provided in Section 5.

model performance is fundamentally linked to non-stationarity, we document a simple ranking of linear and nonlinear models in each period, across the different industries.

In each month t , for each industry, we fit three prediction models: (1) a linear model trained by ridge regression using the most recent 64 months of data, (2) a random forest trained on the most recent 64 months of data, and (3) a random forest trained on all historical data. More details of the experiments are given in Appendix F.2.

We compute each model’s annual out-of-sample R^2 for every industry. To visualize the models’ relative performance across industries, we count the number of industries in which each model achieves the highest out-of-sample R^2 for a given year. Figure 1 summarizes the result. To provide a more granular understanding of the models’ performance, Figure 2 further plots the annual out-of-sample R^2 of the models for the 17 industries.

Figure 1: Number of Industries where Each Model Attains the Highest Annual Out-of-Sample R^2 .

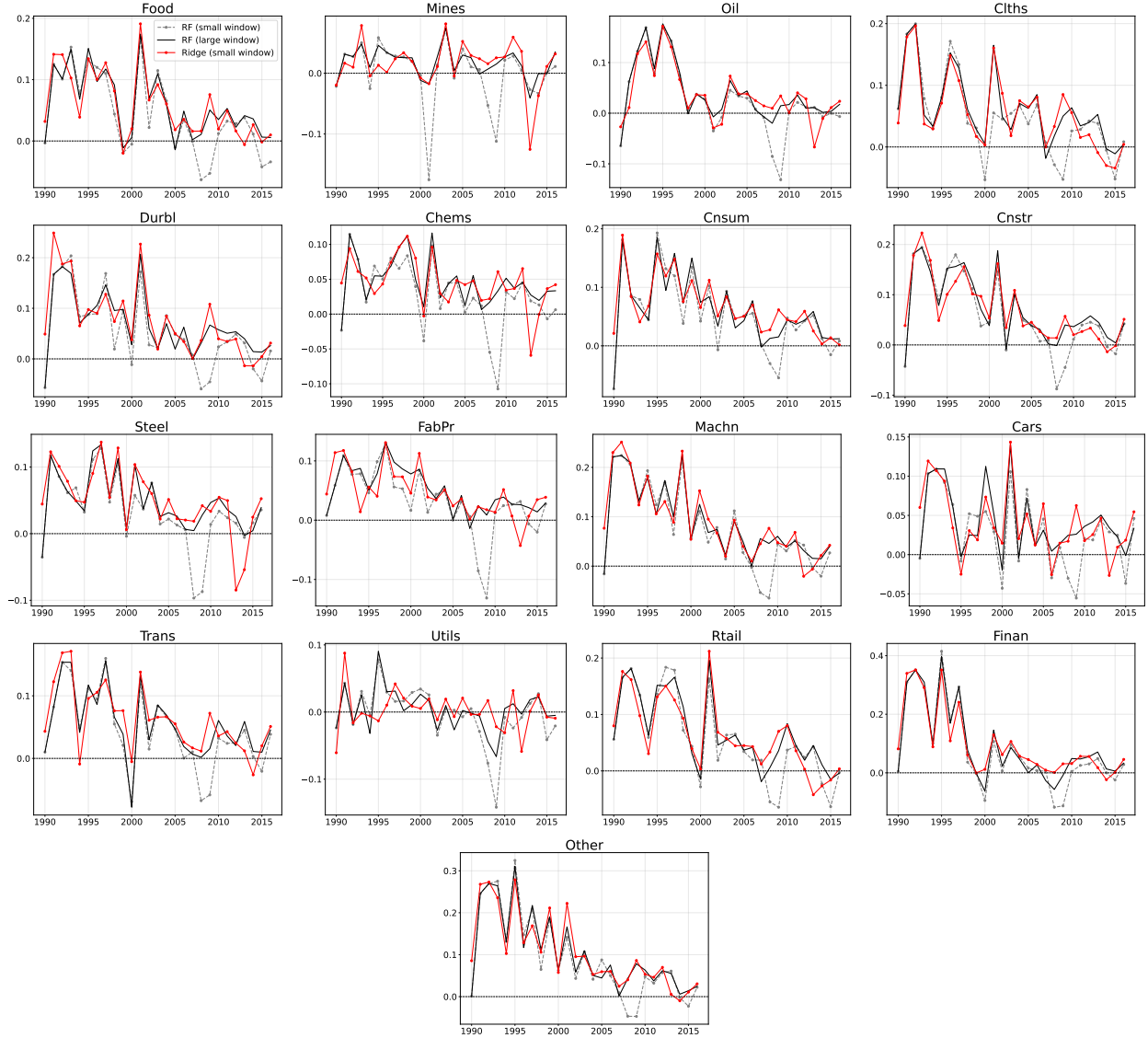


This figure reports the relative performance of three models in predicting the excess returns of the 17 industry portfolios. The three models are: (1) a linear model trained by ridge regression on the most recent 64 months of data (orange), (2) a random forest trained on the most recent 64 months of data (blue), and (3) a random forest trained on all historical data (white). For each year from 1990 to 2016, we compute the annual out-of-sample R^2 of the models for each of the 17 industry portfolios, and then count, for each model, the number of industries in which it outperforms the other two models in terms of the annual out-of-sample R^2 .

We make two key observations. First, within the same model class, using less training data may lead to better performance. For example, in several years including 1994, 1996 and 2003, the random forest trained on the most recent 64 months of data outperforms the random forest trained on all historical data for at least half of the industries.

Second, and more strikingly, a simple model trained on a short window can outperform a complex model trained on a long window. In particular, during the three NBER-designated recessions, the simple linear model trained on 64 months of data outperforms the more complex random forest model trained on all historical data for over half of the industries. This consistent pattern shows that the

Figure 2: Annual Out-of-Sample R^2 of Three Models for 17 Industry Portfolios.



This figure reports, for each of the 17 industry portfolios, the annual out-of-sample R^2 from 1990 to 2016 for three models: (1) a linear model trained by ridge regression on the most recent 64 months of data (red), (2) a random forest trained using the most recent 64 months of data (gray), and (3) a random forest trained using all available historical data up to that year (black). In periods of strong non-stationarity, such as 1990-1991, 2001-2002 and 2008-2009, the linear model trained on a small window constantly outperforms the more complex random forest trained on a large window. The labels in each figure is the Kenneth French acronym for the industries. For full names of these industries, please refer to Table 4.

advantage of a more expressive model class can be completely negated by the non-stationarity in the training data. In Section 4, we propose data-driven approaches to select the best-performing model during such unusual economic regimes.

These empirical findings highlight that in a non-stationary environment, the model complexity and training data size are intricately linked with each other. We call this phenomenon the *nonstationarity-complexity tradeoff*. Crucially, the optimal choice of model class and training win-

dow size is not fixed; instead, it generally varies with the degree of the non-stationarity.

3.2 Theoretical Characterization

We provide theoretical support for the nonstationarity-complexity tradeoff, by deriving a finite-sample bound on a model's prediction error under non-stationarity. The bound decomposes the prediction error into three key components: model misspecification error, statistical uncertainty, and non-stationarity, and shows how they interact with the choice of model class and training window length.

Consider a model \hat{f} trained from a model class \mathcal{F} by minimizing the empirical loss over training data from the last k periods, denoted by $\{\mathcal{D}_j^{\text{tr}}\}_{j=t-k}^{t-1}$, where $\mathcal{D}_j^{\text{tr}} = \{(\mathbf{x}_{j,i}^{\text{tr}}, y_{j,i}^{\text{tr}})\}_{i=1}^{m_j}$ is the training data in period j . That is,

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} \sum_{i=1}^{m_j} [f(\mathbf{x}_{j,i}^{\text{tr}}) - y_{j,i}^{\text{tr}}]^2, \quad (3.1)$$

where $m_{t,k} = \sum_{j=t-k}^{t-1} m_j$ is the number of training data points in $\{\mathcal{D}_j^{\text{tr}}\}_{j=t-k}^{t-1}$.

Define the Bayes optimal least squares estimator $f_t^*(\cdot) = \mathbb{E}_{(\mathbf{x}, y) \sim P_t}[y \mid \mathbf{x} = \cdot]$, which minimizes the MSE $L_t(f)$ over all possible prediction models $f : \mathcal{X} \rightarrow \mathbb{R}$. Our bound will be stated in terms of the *excess risk*

$$\mathcal{E}_t(f) = L_t(f) - L_t(f_t^*),$$

which compares the prediction error of a model f against that of f_t^* . To facilitate analysis, we make the following boundedness assumption.

Assumption 3.1 (Boundedness). *There exists a constant $M > 0$ such that for all models f in the class \mathcal{F} , $(\mathbf{x}, y) \sim P_j$ and $j \in \mathbb{Z}_+$, we have $|f(\mathbf{x})| \leq M$, and $|y| \leq M$. Without loss of generality we assume $M \geq 1$.*

To quantify the effective complexity of the model class \mathcal{F} relative to the training window size k , we employ a measure $r_{t,k}(\mathcal{F})$ derived from the theory of *local Rademacher complexity* (Bartlett et al., 2005). Given the technical nature of this measure, we defer its formal definition to Appendix C. The local Rademacher complexity measures the ability of the near-optimal models in \mathcal{F} to fit random noise using data within the training window k . A higher complexity indicates a richer model class that is capable of approximating complex patterns, but also signals a higher estimation variance and thus a higher risk of overfitting. As an illustration, we now present the complexity measure $r_{t,k}(\mathcal{F})$ for several common model classes. The results are proved in Appendix C.5.

Example 3.1 (Finite class). *If $|\mathcal{F}| < \infty$, then $r_{t,k}(\mathcal{F}) \leq (4M \log |\mathcal{F}|)/m_{t,k}$.*

Example 3.2 (Linear class). *Recall $\mathcal{X} \subseteq \mathbb{R}^d$. For every $\boldsymbol{\theta} \in \mathbb{R}^d$, define $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathbb{R}$ by $f_{\boldsymbol{\theta}}(\mathbf{x}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle$. Suppose that $\mathcal{F} \subseteq \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^d\}$. Then, $r_{t,k}(\mathcal{F}) \leq cd/m_{t,k}$ holds with some constant c .*

Example 3.3 (Kernel class). Let \mathbb{H} be a reproducing kernel Hilbert space (Wahba, 1990) with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|_{\mathbb{H}}$, and $\phi : \mathcal{X} \rightarrow \mathbb{H}$ be a feature mapping. For any $\boldsymbol{\theta} \in \mathbb{H}$, define $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow \mathbb{R}$ by $f_{\boldsymbol{\theta}}(\mathbf{x}) = \langle \boldsymbol{\theta}, \phi(\mathbf{x}) \rangle$. Consider the function class $\mathcal{F} = \{f_{\boldsymbol{\theta}} : \|\boldsymbol{\theta}\|_{\mathbb{H}} \leq R\}$ for some constant $R > 0$. Model fitting in this class can be efficiently implemented through kernel ridge regression, which is a finite-dimensional convex program even if \mathbb{H} and \mathcal{F} are infinite-dimensional.

Suppose there exists a trace-class operator $\mathbf{S} : \mathbb{H} \rightarrow \mathbb{H}$ such that for any $j \in \mathbb{Z}_+$ and $\mathbf{v} \in \mathbb{H}$, we have $\mathbb{E}_{(\mathbf{x}, y) \sim P_j} |\langle \phi(\mathbf{x}), \mathbf{v} \rangle|^2 \leq \langle \mathbf{v}, \mathbf{S} \mathbf{v} \rangle$. Let $\{\mu_k\}_{k=1}^{\infty}$ be the eigenvalues of \mathbf{S} sorted in descending order. We have the following results:

- (Exponential decay) If there are constants $c_1, c_2 > 0$ such that $\mu_k \leq c_1 e^{-c_2 k}$ holds for all k , then $r_{t,k}(\mathcal{F}) \leq (C \log m_{t,k})/m_{t,k}$ holds with some constant C .
- (Polynomial decay) If there are constants $c > 0$ and $\alpha \geq 1$ such that $\mu_k \leq c k^{-2\alpha}$ holds for all k , then $r_{t,k}(\mathcal{F}) \leq C m_{t,k}^{-\frac{2\alpha}{2\alpha+1}}$ holds with some constant C .

Examples of the above two cases include function spaces induced by the Gaussian kernel and Sobolev spaces, respectively (Wainwright, 2019).

In the classical setting where the training data $\{\mathcal{D}_j\}_{j=t-k}^{t-1}$ is i.i.d., the complexity measure $r_{t,k}(\mathcal{F})$ is a key component in bounding the excess risk of \hat{f} : with high probability,

$$\mathcal{E}_t(\hat{f}) \lesssim \min_{f \in \mathcal{F}} \mathcal{E}_t(f) + \left(r_{t,k}(\mathcal{F}) + \frac{1}{m_{t,k}} \right). \quad (3.2)$$

In particular, the prediction error is decomposed into two terms:

1. *Model misspecification error*

$$\min_{f \in \mathcal{F}} \mathcal{E}_t(f) = \min_{f \in \mathcal{F}} L_t(f) - L_t(f_t^*),$$

which describes how well \mathcal{F} can approximate the Bayes optimal least squares estimator f_t^* at time t . A more complex model class tends to reduce the model misspecification error.

2. *Statistical uncertainty*

$$r_{t,k}(\mathcal{F}) + \frac{1}{m_{t,k}},$$

which quantifies the estimation variance of the model \hat{f} . As is discussed above, using a more complex model class increases the statistical uncertainty of the fitted model. Consequently, a more complex model typically requires a longer training window k to mitigate its estimation variance.

The classical error bound (3.2) shows that in the i.i.d. case, increasing the training window size k always reduces the statistical uncertainty, thereby lowering the total prediction error. However, we now present our theory to show that under non-stationarity, this logic is incomplete. As we increase

the window size k to reduce estimation variance, we inadvertently include older data distributions that differ from the target, introducing a third error component. We formalize this in the following theorem.

Theorem 3.1 (Prediction error bound). *Let Assumptions 2.1 and 3.1 hold, and fix $\delta \in (0, 1)$. With probability at least $1 - \delta$, the model \hat{f} defined by (3.1) satisfies*

$$\mathcal{E}_t(\hat{f}) \lesssim \min_{f \in \mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{\log(1/\delta)}{m_{t,k}} \right) + M^2 \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t).$$

Here \lesssim hides a universal constant, and $\text{TV}(P_j, P_t) = \max_A |P_j(A) - P_t(A)|$ is the total variation distance.

Proof of Theorem 3.1. See Appendix C. □

Theorem 3.1 reveals that the non-stationarity adds a third dimension to the classical prediction error bound (3.2), namely, a *non-stationarity* term

$$\max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t),$$

which quantifies the distribution drift in the environment within the last k periods. Unlike the statistical uncertainty, this error component increases with the window size k . In Table 1, we summarize how the model complexity and the training window size k impact the three sources of error.

Table 1: Impacts of Model Complexity and Training Window Size on Prediction Error

	Misspecification Error	Statistical Uncertainty	Non-Stationarity
Model Complexity ↗	↘	↗	-
Training Window k ↗	-	↘	↗

The error decomposition in Theorem 3.1 formalizes the empirical observations in Section 3.1:

(i) Using a more expressive model class reduces misspecification error but increases the risk of overfitting, and (ii) using a longer training window reduces statistical uncertainty but increases non-stationarity. As a result, neither greater model complexity nor more training data is uniformly beneficial under non-stationarity. Below we illustrate this phenomenon through a simple example.

Example 3.4 (Selection of model class and window under non-stationarity). *Let $\eta, \gamma \in [0, 1]$ be two small constants. Suppose that at each time t , the covariate and response $(x, y) \sim P_t$ satisfy $x \sim \text{Uniform}[0, 1]$, $y|x \sim N(f_t^*(x), 1)$, and*

$$f_t^*(x) = c_t x + \gamma \sin(2\pi x),$$

where $\{c_t\}_{t=1}^\infty$ is a deterministic sequence in $[0, 1]$ satisfying $|c_{t+1} - c_t| = \eta$. We observe a single sample per period. Consider two model classes: linear class and kernel class with a first-order Sobolev kernel (see, e.g., Example 12.16 in Wainwright (2019)).

- If we train a linear model with a training window k , then the three components of the prediction error bound in Theorem 3.1 satisfy

$$\min_{f \in \mathcal{F}} \mathcal{E}_t(f) \asymp \gamma^2, \quad r_{t,k}(\mathcal{F}) \asymp k^{-1}, \quad \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t) \asymp k\eta.$$

Optimizing their sum over k yields the optimal window size $k^* \asymp \eta^{-1/2}$, which leads to an $O(\gamma^2 + \eta^{1/2})$ bound on the prediction error.

- If we use the kernel class, then f_t^* is well-specified. For a training window k , we have

$$\min_{f \in \mathcal{F}} \mathcal{E}_t(f) = 0, \quad r_{t,k}(\mathcal{F}) \asymp k^{-2/3}, \quad \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t) \asymp k\eta.$$

The optimal training window is $k^* \asymp \eta^{-3/5}$, which results in a prediction error of $O(\eta^{2/5})$.

We observe that for both classes, the optimal window size depends on the severity of the drift η , and is in general not the full window size. If one naïvely uses the kernel class with a large window size, then the resulting error scales as $O(k\eta)$, which is linear in k and can be much worse than the above bounds.

As expected, the preferable model class depends on the interplay between misspecification γ and drift η . The kernel class is more expressive but more sensitive to drift. When $\eta = O(\gamma^5)$, drift is relatively mild and the kernel is optimal, consistent with the “virtue of complexity” (Kelly and Malamud, 2025; Kelly et al., 2022, 2024). However, when $\eta \gg \gamma^5$, severe non-stationarity requires shorter training windows under which sample sizes are too limited for the kernel estimator to fully exploit its flexibility advantage. In this high-drift regime, the linear class achieves better performance with its shorter optimal window, explaining the “less can be more” phenomenon observed in our experiments.

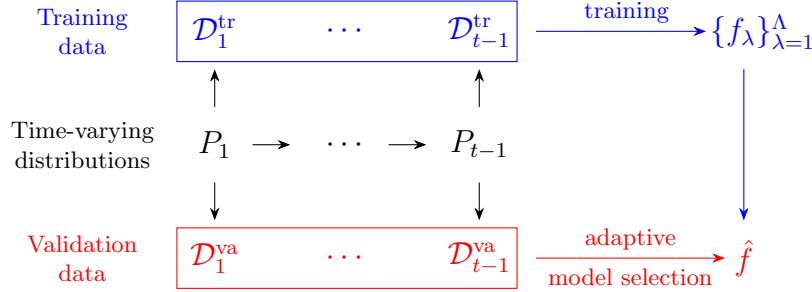
4 Adaptive Model and Data Selection under Non-Stationarity

Section 3 shows that the predictive performance of a model depends jointly on its complexity and the size of the training data, and that the optimal choice often varies over time with the non-stationarity. As the non-stationarity is generally unknown *a priori*, the selection of the model class and training window calls for a data-driven approach.

In this section, we develop a novel method that uses historical validation data to select the best model from a set of candidates. These candidate models can come from different model classes, be trained on different time windows, or use different hyperparameters. The main challenge is that the same non-stationarity that complicates model training also incapacitates standard model selection techniques such as holdout and cross validation. Specifically, in a non-stationary environment, a model that performs well on a validation set from the distant past may not perform as well in the future. Our solution is to adaptively select the relevant validation data that best reflects the current environment, allowing for a more accurate comparison of the candidate models’ future performance.

We now formally set up the framework, illustrated in Figure 3. In each period t , we split the available data $\{\mathcal{D}_j\}_{j=1}^{t-1}$ into a training dataset $\{\mathcal{D}_j^{\text{tr}}\}_{j=1}^{t-1}$ and a validation dataset $\{\mathcal{D}_j^{\text{va}}\}_{j=1}^{t-1}$. We use the training data $\{\mathcal{D}_j^{\text{tr}}\}_{j=1}^{t-1}$ to produce a finite set of candidate models $\{f_\lambda\}_{\lambda=1}^\Lambda$. These candidates can come from different model classes, be trained on different data horizons, or use different hyperparameters. We will use the validation data $\{\mathcal{D}_j^{\text{va}}\}_{j=1}^{t-1}$ to select a good model $\hat{f} = f_{\hat{\lambda}}$ that performs best at time t .

Figure 3: Our Framework for Model Training and Selection under Non-stationarity.



4.1 Adaptive Tournament Model Selection

In this section, we describe our model selection approach, which uses a sequential elimination tournament. The procedure relies on a pairwise comparison subroutine \mathcal{A} which is designed to compare two given models f and f' , and output the better model, denoted by $\mathcal{A}(f, f')$. In each round, we choose one remaining model f as a pivot model and compare it against each remaining model f' using \mathcal{A} . If the pivot f wins all pairwise comparisons, it is declared the winner; otherwise, the models that defeated f advance to the next round. The procedure is formally described in Algorithm 1.

Algorithm 1 Adaptive Tournament Model Selection (ATOMS)

Input: Candidate models $\{f_\lambda\}_{\lambda=1}^\Lambda$, validation data $\{\mathcal{D}_j^{\text{va}}\}_{j=1}^t$, pairwise model comparison subroutine \mathcal{A} .
Initialize $S = \{f_\lambda\}_{\lambda=1}^\Lambda$. // collection of remaining models
while $|S| > 1$
 Choose a pivot model $f \in S$ uniformly at random.
 Initialize $S' \leftarrow \emptyset$. // collection of models in S that outperform f
 for $f' \in S \setminus \{f\}$
 Run \mathcal{A} to compare $\{f, f'\}$ to obtain $\mathcal{A}(f, f')$.
 If $\mathcal{A}(f, f') = f'$, set $S' \leftarrow S' \cup \{f'\}$.
 if $S' = \emptyset$
 return $\hat{f} = f$. // if no model outperforms f , output f
 else
 Set $S \leftarrow S'$.
return the only model $\hat{f} \in S$.

Algorithm 1 has two attractive properties. First, in terms of computational efficiency, the expected number of pairwise comparisons scales linearly with the number of models Λ .

Lemma 4.1 (Computational complexity). *Algorithm 1 calls the subroutine \mathcal{A} for $\Theta(\Lambda)$ times in expectation.*

Proof of Lemma 4.1. See Appendix D.1. □

Second, regarding the statistical accuracy of model selection, we show later that Algorithm 1 preserves the performance guarantee of any pairwise comparison subroutine \mathcal{A} , incurring only a logarithmic factor overhead in the number of models Λ .

Pairwise comparison subroutine. We now detail the model comparison subroutine \mathcal{A} . As we mentioned before, directly comparing the models on the non-stationary validation data $\{\mathcal{D}_j^{\text{va}}\}_{j=1}^{t-1}$ may lead to significantly biased estimates of the model performance. To address this problem, we take an approach based on the adaptive rolling window framework developed by Han et al. (2024).

The main idea is as follows. To choose between two models, it suffices to determine the sign of their *performance gap*

$$\Delta_t = L_t(f_1) - L_t(f_2).$$

Indeed, f_2 is better than f_1 if and only if $\Delta_t > 0$. To estimate Δ_t from the non-stationary validation data, a natural idea is to take a *look-back window* $\ell \in [t-1]$, and use the validation data from the last ℓ periods, $\mathcal{D}_j^{\text{va}} = \{(\mathbf{x}_{j,i}^{\text{va}}, y_{j,i}^{\text{va}})\}_{i=1}^{n_j}, j = t-\ell, \dots, t-1$, to form a rolling window estimate

$$\hat{\Delta}_{t,\ell} = \frac{1}{n_{t,\ell}} \sum_{j=t-\ell}^{t-1} \sum_{i=1}^{n_j} u_{j,i}, \quad \text{where} \quad u_{j,i} = [f_1(\mathbf{x}_{j,i}^{\text{va}}) - y_{j,i}^{\text{va}}]^2 - [f_2(\mathbf{x}_{j,i}^{\text{va}}) - y_{j,i}^{\text{va}}]^2, \quad (4.1)$$

and $n_{t,\ell} = \sum_{j=t-\ell}^{t-1} n_j$. The accuracy of model comparison depends on the estimation accuracy of $\hat{\Delta}_{t,\ell}$. The critical challenge is choosing a validation window size ℓ such that the estimation error $|\hat{\Delta}_{t,\ell} - \Delta_t|$ is small.

The choice of the validation window ℓ involves a bias-variance tradeoff: with probability at least $1 - \delta$,

$$|\hat{\Delta}_{t,\ell} - \Delta_t| \leq \phi(t, \ell) + \psi(t, \ell, \delta). \quad (4.2)$$

Here,

$$\phi(t, \ell) = \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|$$

is the bias term that measures the non-stationarity of Δ_j in the last ℓ periods, and

$$\psi(t, \ell, \delta) = \begin{cases} 8M^2, & \text{if } n_{t,\ell} = 1 \\ \sigma_{t,\ell} \sqrt{\frac{2 \log(2/\delta)}{n_{t,\ell}}} + \frac{16M^2 \log(2/\delta)}{3n_{t,\ell}}, & \text{if } n_{t,\ell} \geq 2 \end{cases}, \quad \text{with} \quad \sigma_{t,\ell}^2 = \frac{1}{n_{t,\ell}} \sum_{j=t-\ell}^{t-1} n_j \text{var}(u_{j,1}),$$

is the variance term that quantifies the statistical uncertainty associated with the estimate $\widehat{\Delta}_{t,\ell}$ via a Bernstein concentration inequality. In general, as the window ℓ increases, we expect the bias $\phi(t, \ell)$ to increase and the variance term $\psi(t, \ell)$ to decrease. The ideal validation window size ℓ^* should strike a balance between the bias and variance:

$$\ell^* = \operatorname{argmin}_{\ell \in [t-1]} \{ \phi(t, \ell) + \psi(t, \ell, \delta) \}.$$

However, as both $\phi(t, \ell)$ and $\psi(t, \ell)$ depend on the unknown non-stationarity, ℓ^* cannot be directly computed.

To tackle this problem, we construct proxies $\widehat{\psi}$ and $\widehat{\phi}$ for ψ and ϕ , respectively. The proxy for $\psi(t, \ell, \delta)$ is constructed by replacing the unknown variance $\sigma_{t,\ell}^2$ by the sample variance

$$\widehat{v}_{t,\ell}^2 = \frac{1}{n_{t,\ell} - 1} \sum_{j=t-\ell}^{t-1} \sum_{i=1}^{n_j} (u_{j,i} - \widehat{\Delta}_{t,\ell})^2, \quad (4.3)$$

which gives

$$\widehat{\psi}(t, \ell, \delta) = \begin{cases} 8M^2, & \text{if } n_{t,\ell} = 1 \\ \widehat{v}_{t,\ell} \sqrt{\frac{2 \log(2/\delta)}{n_{t,\ell}}} + \frac{64M^2 \log(2/\delta)}{3(n_{t,\ell} - 1)}, & \text{if } n_{t,\ell} \geq 2 \end{cases}. \quad (4.4)$$

The proxy for the bias term is inspired by the Goldenshluger-Lepski method for adaptive non-parametric estimation (Goldenshluger and Lepski, 2008):

$$\widehat{\phi}(t, \ell, \delta) = \max_{i \in [\ell]} \left(|\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}| - [\widehat{\psi}(t, \ell, \delta) + \widehat{\psi}(t, i, \delta)] \right)_+. \quad (4.5)$$

To interpret $\widehat{\phi}$, in light of the bias-variance decomposition in (4.2), the quantity

$$\left(|\widehat{\Delta}_{t,\ell} - \widehat{\Delta}_{t,i}| - [\widehat{\psi}(t, \ell, \delta) + \widehat{\psi}(t, i, \delta)] \right)_+ \quad (4.6)$$

can be viewed as a measure of the bias between the window ℓ and a smaller window $i \leq \ell$, where subtracting $\widehat{\psi}(t, \ell, \delta)$ and $\widehat{\psi}(t, i, \delta)$ eliminates the stochastic error and teases out the bias. The term $\widehat{\phi}(t, \ell, \delta)$ is then formed by taking the maximum of (4.6) over all smaller windows $i \in [\ell]$.

After constructing the bias and variance proxies, one chooses a window size $\widehat{\ell}$ that minimizes their sum:

$$\widehat{\ell} = \operatorname{argmin}_{\ell \in [t-1]} \left\{ \widehat{\phi}(t, \ell) + \widehat{\psi}(t, \ell, \delta) \right\}. \quad (4.7)$$

We then use $\widehat{\Delta}_{t,\widehat{\ell}}$ as our estimate of Δ_t for model comparison. In particular, the subroutine selects f_1 if and only if $\widehat{\Delta}_{t,\widehat{\ell}} \leq 0$. The procedure is summarized in Algorithm 2.

By using Algorithm 2 as the model comparison subroutine in Algorithm 1, we obtain an algorithm that adaptively chooses non-stationary data to perform model selection. We call the

Algorithm 2 Adaptive Rolling Window for Model Comparison

Input: Candidate models $\{f_1, f_2\}$, validation data $\{\mathcal{D}_j^{\text{va}}\}_{j=1}^{t-1}$, hyperparameters δ' and M .

for $\ell = 1, \dots, t-1$

 Compute $\hat{\Delta}_{t,\ell}$, $\hat{\psi}(t, \ell, \delta')$ and $\hat{\phi}(t, \ell, \delta')$ according to (4.1), (4.4) and (4.5).

 Choose window size

$$\hat{\ell} \in \operatorname{argmin}_{\ell \in [t-1]} \left\{ \hat{\phi}(t, \ell, \delta') + \hat{\psi}(t, \ell, \delta') \right\}. \quad (4.8)$$

 Select $\hat{\lambda} = 1$ if $\hat{\Delta}_{t,\hat{\ell}} \leq 0$, and $\hat{\lambda} = 2$ otherwise.

return $\hat{f} = f_{\hat{\lambda}}$.

algorithm Addaptive Tournament Model Selection, or **ATOMS** in short.

4.2 Theoretical Guarantees

We now present the theoretical guarantees for our model selection framework. Theorem 4.1 below establishes a performance bound for the pairwise model comparison subroutine (Algorithm 2).

Theorem 4.1 (Near-optimal model comparison). *Let Assumptions 2.1 and 3.1 hold. Choose $\delta \in (0, 1)$ and take $\delta' = \delta/(3t)$ in Algorithm 2. With probability at least $1 - \delta$, the output \hat{f} of Algorithm 2 satisfies*

$$\mathcal{E}_t(\hat{f}) \lesssim \min\{\mathcal{E}_t(f_1), \mathcal{E}_t(f_2)\} + M^2 \log(t/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}} \right\}. \quad (4.9)$$

Here \lesssim hides a universal constant.

Proof of Theorem 4.1. See Appendix D.2. □

Theorem 4.1 gives a finite-sample oracle inequality (4.9). It states that the excess risk of the \hat{f} does not exceed that of the better model between f_1 and f_2 , plus an additional error term that reflects the difficulty of using the non-stationary data to make the comparison. Inside this additional term, the quantity

$$\max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}}$$

represents the two sources of errors that arise when using a validation window ℓ to compare models, namely, the non-stationarity $\max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t)$ and the statistical uncertainty $1/n_{t,\ell}$ associated with the $n_{t,\ell}$ validation samples. The bound takes the minimum over all validation window sizes ℓ , meaning that Algorithm 2 performs almost as well as an oracle that knows in hindsight which validation window size ℓ would lead to the most accurate comparison. This shows that Algorithm 2 adaptively chooses a near-optimal validation window tailored to the local non-stationarity.

Building on this pairwise guarantee, Theorem 4.2 below shows that our model selection algorithm **ATOMS** inherits the same oracle property when selecting from multiple candidate models $\{f_\lambda\}_{\lambda=1}^\Lambda$.

Theorem 4.2 (Near-optimal model selection). *Let Assumptions 2.1 and 3.1 hold. Choose $\delta \in (0, 1)$ and take $\delta' = \delta/(3\Lambda^2 t)$ in **ATOMS**. With probability at least $1 - \delta$, **ATOMS** outputs a model \hat{f} satisfying*

$$\mathcal{E}_t(\hat{f}) \lesssim \min_{\lambda \in [\Lambda]} \mathcal{E}_t(f_\lambda) + M^2 \log(\Lambda t / \delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}} \right\}.$$

Here \lesssim hides a universal constant.

Proof of Theorem 4.2. See Appendix D.3. □

Theorem 4.2 states that the excess risk of the model \hat{f} chosen by the tournament is at most the excess risk of the best model in $\{f_\lambda\}_{\lambda=1}^\Lambda$, up to an additional term that has the same form as in the pairwise comparison bound, with an extra $O(\log \Lambda)$ multiplicative factor. In other words, **ATOMS** identifies a model whose performance is nearly as good as the best candidate one could have selected in hindsight using the non-stationary validation data.

We remark that our model selection framework (Algorithm 1) is general and can be combined with any model comparison subroutine \mathcal{A} . In particular, in Appendix D.3, we prove a general reduction lemma (Lemma D.1) that converts any theoretical guarantee of the subroutine \mathcal{A} to a guarantee of Algorithm 1. In Appendix B, we further develop a R^2 -based pairwise comparison subroutine that targets the R^2 metric. When equipped with this R^2 -based subroutine, Algorithm 1 enjoys a guarantee with respect to the R^2 metric.

Remark 1 (Comparison with prior work). Our model selection framework builds upon the model comparison method of Han et al. (2024). Below we briefly discuss the main differences between our work and theirs. First, their analysis of the model comparison procedure (Algorithm 2) assumes that the distribution of the covariates \mathbf{x} remains fixed across time. Our theory removes this assumption entirely, and covers the general non-stationary setting where the joint data distribution (\mathbf{x}, y) can change arbitrarily. Second, for model selection, they propose a single-elimination procedure which performs $\Lambda - 1$ model comparisons, but incurs additional factors $(\log \Lambda)^2$ in the performance bound. In contrast, our approach maintains a linear complexity in Λ in expectation while achieving a sharper bound.

4.3 Application to Joint Model Class and Training Window Size Selection

Finally, we apply Theorem 4.2 to the joint selection of model class and training sample size. Let \mathcal{F} be a finite collection of model classes, e.g., $\mathcal{F} = \{\text{linear model, random forest of a certain size}\}$. For each model class $\mathcal{F} \in \mathcal{F}$, we train models on different windows $k \in [t-1]$ of the training data $\{\mathcal{D}_j^{\text{tr}}\}_{j=1}^{t-1}$. Let $\hat{h}(\mathcal{F}, k)$ denote the model from \mathcal{F} trained on $\{\mathcal{D}_j^{\text{tr}}\}_{j=t-k}^{t-1}$. Then, the set of candidate models is given by

$$\{f_\lambda\}_{\lambda=1}^\Lambda = \left\{ \hat{h}(\mathcal{F}, k) : \mathcal{F} \in \mathcal{F}, k \in [t-1] \right\}. \quad (4.10)$$

Applying Theorem 4.2 to this set of candidate models yields the following guarantee. For simplicity, we assume that training-validation data splitting ratio is fixed across time.

Assumption 4.1 (Balanced training-validation split). *There exists $c > 0$ such that $|\mathcal{D}_j^{\text{tr}}|/|\mathcal{D}_j^{\text{va}}| = c$ for all $j \in \mathbb{Z}_+$.*

Theorem 4.3 (Near-optimal model-and-data selection). *Let Assumptions 2.1, 3.1 and 4.1 hold. Suppose the set of candidate models is given by (4.10). Choose $\delta \in (0, 1)$ and take $\delta' = \delta/(6|\mathcal{F}|^2 t^3)$ in **ATOMS**. Then, with probability at least $1 - \delta$, the output \hat{f} of **ATOMS** satisfies*

$$\mathcal{E}_t(\hat{f}) \lesssim \min_{\mathcal{F} \in \mathcal{F}, k \in [t-1]} \left\{ \min_{f \in \mathcal{F}} \mathcal{E}_t(f) + \left(r_{t,k}(\mathcal{F}) + \frac{1}{B_{t,k}} \right) + \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t) \right\}, \quad (4.11)$$

where $B_{t,k} = \sum_{j=t-k}^{t-1} (|\mathcal{D}_j^{\text{tr}}| + |\mathcal{D}_j^{\text{va}}|)$ is the total sample size, and \lesssim hides the constants M and c and logarithmic factors of t , δ^{-1} and $|\mathcal{F}|$.

Proof of Theorem 4.3. See Appendix D.4. □

We note that the term

$$\min_{f \in \mathcal{F}} \mathcal{E}_t(f) + \left(r_{t,k}(\mathcal{F}) + \frac{1}{B_{t,k}} \right) + \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t)$$

on the right hand side of (4.11) is exactly the model performance bound of the model $h(\mathcal{F}, k)$ in Theorem 3.1. Thus, our algorithm selects a near-optimal pair of model class and training window size, up to logarithmic factors.

5 Explaining the Cross-Section of Industry Portfolio Returns

In this section, we investigate whether our algorithm helps explain the cross-section of stock returns using industry portfolios as test assets. Rather than simply building predictive models, we approximate time-varying stochastic discount factors that capture the evolving relationship between risk and return. First, we describe our comprehensive dataset of firm-characteristic managed portfolios and industry portfolios. Then, we report our finding that our adaptive algorithm **ATOMS** achieves superior out-of-sample performance in explaining expected returns compared to fixed window and expanding window approaches across different economic regimes.

5.1 Data

We examine the pricing of 17 industry portfolio returns from Kenneth French’s data library, covering the period from September 1987 to November 2016. Our predictor set combines macroeconomic factors, risk premia from characteristic-sorted portfolios, and lagged returns to capture the complex dynamics driving industry returns, sourced for widely cited public datasets.³ We provide full details

³More specifically, our data combines daily and monthly sources to construct a comprehensive time series of covariates combining macroeconomic and cross-sectional signals. The final sample spans the time period from September 1987 to November 2016. We merge daily CRSP excess returns with monthly characteristics from Gu et al. (2020), which provides 94 standardized characteristics for U.S. equities encompassing valuation ratios, profitability measures,

of the dataset construction in Appendix A.1, including data preprocessing and long-short portfolio constructions.

Macroeconomic and Systematic Factors. We incorporate 15 factors from Chen et al. (2024), who estimate a SDF using deep learning while imposing no-arbitrage restrictions. These factors include: (i) the estimated SDF representing the aggregate price of risk; (ii) ten beta-sorted decile portfolios based on firms’ SDF exposure; and (iii) four macroeconomic hidden states extracted from 178 macro time series via a generative adversarial network. These monthly observations are assigned to all trading days within each month. We also include the daily Fama-French three factors (market, size, and value) from Fama and French (1993) as benchmark risk factors.

Characteristic-Sorted Portfolios. Following Gu et al. (2020), we construct 94 long-short portfolios sorted on firm characteristics that capture price trends, liquidity, size, and risk measures. For each characteristic, we form decile portfolios using all CRSP-listed stocks and create a long-short strategy that buys the top decile and shorts the bottom decile. This approach transforms firm-level characteristics into interpretable factor returns that isolate the pricing implications of each characteristic.

Predictor Set. Our final predictor set comprises: (i) 15 macroeconomic factors from Chen et al. (2024); (ii) 3 Fama-French factors; (iii) 94 characteristic-sorted long-short portfolio returns; and (iv) the 17 lagged industry returns. This comprehensive set combines traditional risk factors with modern high-dimensional representations, allowing us to test whether our adaptive asset pricing framework can effectively navigate the complex, time-varying relationships between these predictors and industry returns.

5.2 Return model

We evaluate our adaptive algorithm **ATOMS** using candidate models from different specifications with varying parameters and estimation windows. We take one month as a period, where the data $\mathcal{D}_t = \{(\mathbf{x}_{t,i}, y_{t,i})\}_{i=1}^{B_t}$ in month t consists of daily covariate-return pairs within that month.

Model Specifications. We consider the following specifications that approximate stochastic discount factors. For a vector of covariates $\mathbf{x} \in \mathbb{R}^d$, we write $\tilde{\mathbf{x}} = (\mathbf{x}^\top, 1)^\top \in \mathbb{R}^{d+1}$.

1. Non-linear specification using random forests (RF). Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, and two parameters, namely the number of trees n_{tree} and the maximum tree depth d_{max} , RF estimates a random forest model.

investment activity, liquidity, and past return dynamics. These characteristics have become the canonical set of firm-level predictors in modern empirical asset pricing. We construct daily long-short portfolios by sorting firms into deciles based on each characteristic and taking the difference between the top and bottom decile returns, following the methodology of Gu et al. (2020).

2. Linear specification estimated with ridge regularization (Ridge). Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and regularization parameter $\alpha > 0$, Ridge estimates a linear model $f(\mathbf{x}) = \langle \hat{\boldsymbol{\theta}}, \tilde{\mathbf{x}} \rangle$ by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n (\langle \boldsymbol{\theta}, \tilde{\mathbf{x}}_i \rangle - y_i)^2 + \alpha \|\boldsymbol{\theta}\|_2^2 \right\}.$$

3. Linear specification with LASSO regularization (LASSO). Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and regularization parameter $\alpha > 0$, LASSO estimates a linear model $f(\mathbf{x}) = \langle \hat{\boldsymbol{\theta}}, \tilde{\mathbf{x}} \rangle$ by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ \frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^n (\langle \boldsymbol{\theta}, \tilde{\mathbf{x}}_i \rangle - y_i)^2 + \alpha \|\boldsymbol{\theta}\|_1 \right\}.$$

4. Linear specification with elastic net regularization (E-Net). Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and regularization parameters $\alpha > 0$ and $r \in (0, 1)$, E-Net estimates a linear model $f(\mathbf{x}) = \langle \hat{\boldsymbol{\theta}}, \tilde{\mathbf{x}} \rangle$ by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ \frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^n (\langle \boldsymbol{\theta}, \tilde{\mathbf{x}}_i \rangle - y_i)^2 + \alpha r \|\boldsymbol{\theta}\|_1 + \frac{\alpha}{2} (1-r) \|\boldsymbol{\theta}\|_2^2 \right\}.$$

In each month t , we estimate models from these specifications on estimation windows of $4^k \wedge (t-1)$ months, where $0 \leq k \leq 5$. We detail the parameter choices for the specifications in Appendix A.3.

Benchmark Approaches. To verify the adaptivity of our framework, we compare it with two non-adaptive benchmarks that use a fixed window for estimation and/or validation.

1. Fixed validation window for specification selection (**Fixed-val**(ℓ)). This is the non-adaptive fixed-window counterpart of ATOMS. In each month t , we estimate the same candidate specifications above, then use validation data from the last ℓ periods $\{\mathcal{D}_j^{\text{va}}\}_{j=t-\ell}^{t-1}$ to perform specification selection. The detailed description of **Fixed-val**(ℓ) is given in Algorithm 3. We consider validation window sizes $\ell = 32, 128, 512$ months, where $\ell = 512$ corresponds to using all historical validation data at all times.
2. Fixed-window cross-validation (**Fixed-CV**). In each month t , we use data from the last 36 months $\{\mathcal{D}_j\}_{j=t-36}^{t-1}$ to perform 5-fold cross-validation to estimate and select a specification out of the candidate specifications with the same sets of parameters.

We run each of these approaches over 20 random splits of estimation and validation data. More details can be found in Appendix A.3.

Performance Metrics. We measure the performance of each approach using the out-of-sample R^2 metric (2.2) that benchmarks against a zero forecast. We compute both the overall out-of-sample R^2 from January 1990 to November 2016, and the annual out-of-sample R^2 . The latter provides a more granular understanding of the approaches' performance over time. In Appendix G.1, we also report results for the standard R^2 metric (2.3).

Algorithm 3 Fixed Validation Window for Specification Selection (**Fixed-val**(ℓ))

Input: Candidate specifications $\{f_\lambda\}_{\lambda=1}^\Lambda$, validation data $\{\mathcal{D}_j^{\text{va}}\}_{j=1}^{t-1}$, validation window size ℓ .
Select

$$\hat{\lambda} = \underset{\lambda \in [\Lambda]}{\operatorname{argmin}} \sum_{j=(t-\ell) \vee 1}^{t-1} \sum_{i=1}^{n_j} [f_\lambda(\mathbf{x}_{j,i}^{\text{va}}) - y_{j,i}^{\text{va}}]^2.$$

return $\hat{f} = f_{\hat{\lambda}}$.

5.3 Empirical Results: 17 Industry Portfolios

We next turn to the empirical analysis of pricing the 17 industry portfolios. The fundamental premise of our adaptive framework is that asset pricing relationships exhibit time-varying dynamics rather than remaining stationary across economic conditions. This non-stationarity is particularly pronounced during economic recessions, when structural breaks in risk premia, sudden shifts in investor risk aversion, and disruptions to market liquidity mechanisms create environments where long-term historical pricing relationships provide poor guidance for future returns. Recessions therefore serve as a natural laboratory for testing the adaptivity of our framework: if our approach can successfully navigate these turbulent periods when non-stationarity is most severe, it provides compelling evidence for the value of adaptive model selection in asset pricing more generally.

Recession Performance Analysis. Our most striking empirical finding relates to the differential performance of **ATOMS** during economic downturns. Table 2 presents out-of-sample R^2 values across distinct economic regimes, revealing that **ATOMS** exhibits particular strength during recessionary periods when market dynamics are most volatile and traditional models typically fail. The adaptive framework achieves an out-of-sample R^2 of 0.049 across the full sample period, representing a 14.0% improvement over the best fixed-window benchmark **Fixed-val**(512) which has $R^2 = 0.043$.

During the 2001 recession, characterized by the dot-com bubble collapse and the September 11 terrorist attacks, **ATOMS** achieves an impressive R^2 of 0.125, outperforming **Fixed-val**(512) by 6.8% (0.117) and substantially exceeding **Fixed-CV**'s 0.071. This superior performance suggests that our adaptive framework effectively captures the rapid regime shifts that occurred during this period, when technology-related stocks experienced dramatic revaluation and risk premia underwent fundamental restructuring.

The 1990 Gulf War recession provides particularly compelling evidence of our framework's adaptability. While **ATOMS** maintains a positive R^2 of 0.027 during this sharp but brief contraction, the fixed-window benchmark **Fixed-val**(512) produces a negative R^2 of -0.031 , indicating worse performance than a simple forecast of zero. This divergence highlights the critical importance of adaptivity during periods of sudden market stress, when historical relationships between risk factors and returns break down most severely. The adaptive framework's ability to rapidly adjust its validation window allows it to recognize and respond to the changing market dynamics that fixed-window models miss entirely.

Table 2: Out-of-Sample R^2 Averages Across Industries by Time Period

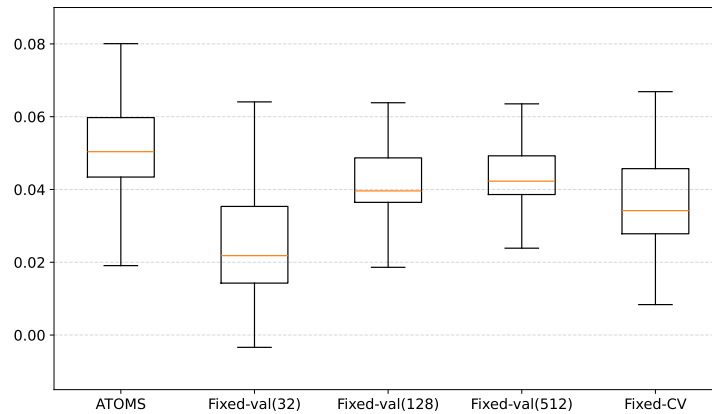
Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	0.049	0.027	0.125	0.041
Fixed-val(32)	0.022	0.009	0.096	−0.001
Fixed-val(512)	0.043	−0.031	0.117	0.039
Fixed-CV	0.035	−0.007	0.071	0.014

This table reports out-of-sample (OOS) R^2 averages for return prediction models across all 17 industry portfolios. Full OOS Period refers our largest available OOS period covering 01/1990~11/2016. Columns report OOS R^2 averages across all industries and highlight this metric during three recessions, as documented in [NBER Business Cycle Dating](#):

- the 1990 Gulf War recession (06/1990~10/1990);
- the 2001 Recession of dot-com bubble burst and the 9/11 attack (05/2001~10/2001);
- the Financial Crisis led by defaults of subprime mortgages (11/2007~06/2009).

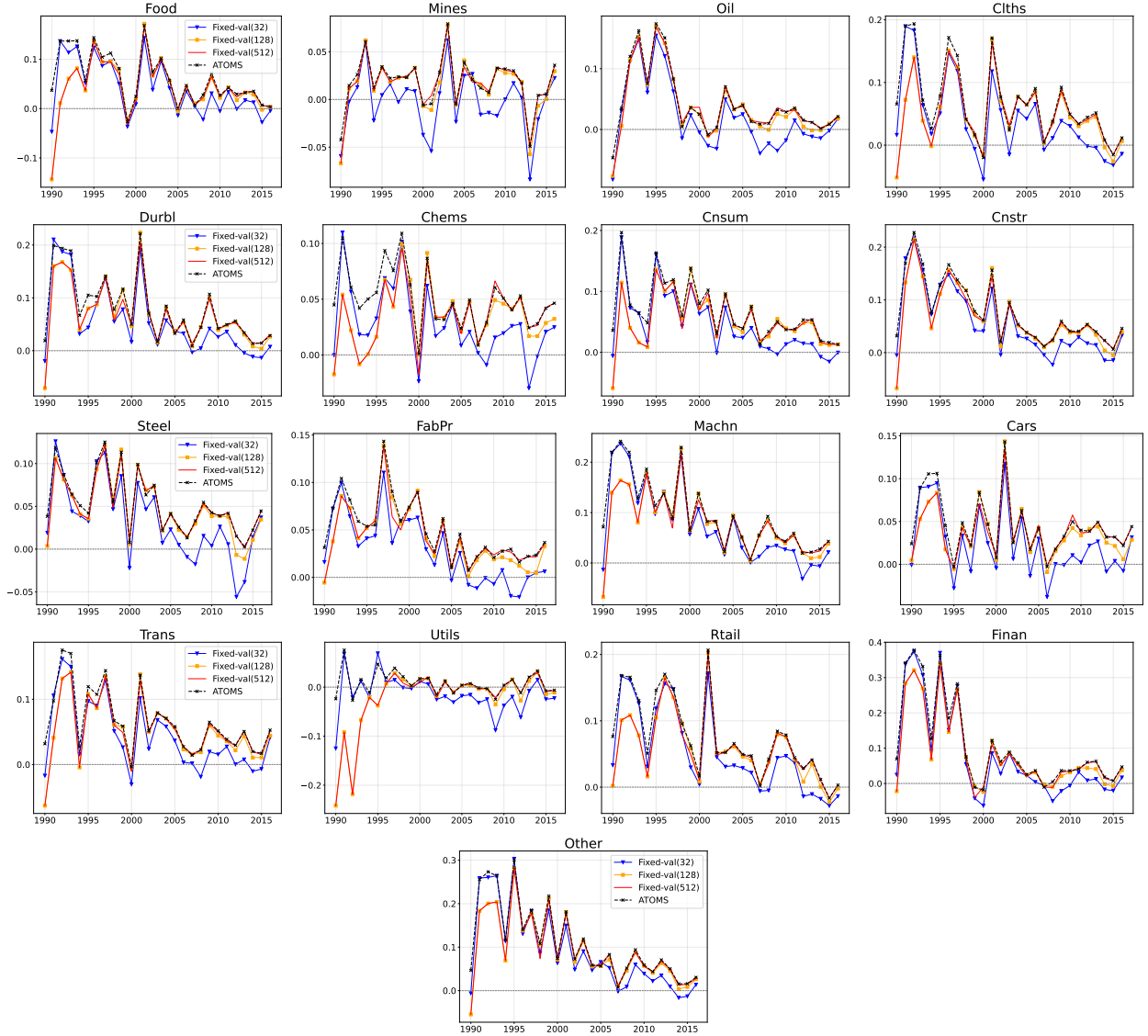
That is, the OOS performance in Gulf War column focuses on model performance comparisons exclusively in the out-of-sample period of 06/1990~10/1990. All values are calculated using monthly return data.

During the Global Financial Crisis of 2007-2009, **ATOMS** achieves an R^2 of 0.041, marginally outperforming **Fixed-val(512)** (0.039) and substantially exceeding **Fixed-CV** (0.014). The relatively smaller performance gap during this period is primarily driven by the fact that NBER defines the Financial Crisis-related recession with a much longer period that spans from 2007 to 2009, where eventually new data observed in the recession itself could be factored into training and validation for our benchmarked model that uses fixed look-back horizon. In other words, the long duration of this recession attenuates the advantages of our method. Nevertheless, **ATOMS** maintained its advantage throughout this long period, demonstrating robustness across different types of economic contractions.

Figure 4: Box Plot of Out-of-Sample R^2 of **ATOMS** and Fixed-Window Baselines for 17 Industry Portfolios.

This figure describes the distribution of each method’s OOS R^2 . Each box corresponds to all industries and all years in our OOS horizon.

Figure 5: Annual Out-of-Sample R^2 of ATOMS and Fixed-Window Baselines for 17 Industry Portfolios.



This figure reports the annual out-of-sample R^2 of our adaptive model selection algorithm ATOMS (black dashed line with \times 's), as well as the fixed-window baselines Fixed-val(32) (blue ∇ 's), Fixed-val(128) (orange \blacksquare 's), and Fixed-val(512) (red), which use the last 32, last 128 and all months of validation data. The title in each subfigure is Kenneth French's acronym for each industry. For the full names of these industries, please refer to Table 4.

Economic Interpretation. The superior recession performance of ATOMS has important implications for asset pricing theory and practice. Traditional asset pricing models assume stationary risk-return relationships, an assumption that becomes particularly problematic during economic downturns when risk aversion typically increases and market liquidity conditions deteriorate. Our adaptive framework explicitly recognizes this non-stationarity by allowing the validation window to expand or contract based on recent predictive performance.

The empirical evidence suggests that during recessions, the optimal window for model selection shrinks significantly, reflecting the rapid evolution of risk premia. The 1990 Gulf War recession

provides the clearest example: its sudden onset and brief duration created an environment where only models with very recent validation data could accurately capture the new pricing dynamics. Conversely, during more prolonged downturns like the Global Financial Crisis, the optimal window likely expanded gradually as new market conditions became established.

From a theoretical perspective, these findings support the view that stochastic discount factors exhibit time-varying dynamics that are particularly pronounced during economic stress. The adaptive framework’s ability to track these dynamics more effectively than fixed-window approaches suggests that the non-stationarity of asset pricing relationships is not merely a statistical artifact but reflects fundamental economic mechanisms that vary with the business cycle.

Robustness Across Industries. We conduct industry-by-industry robustness check to confirm that the recession outperformance of **ATOMS** is not driven by a subset of industries but represents a broad-based phenomenon. Firstly, we report in Figure 4 that **ATOMS** overall has better R^2 across all years, as its median, and level position of the box is higher than those of the other methods.

In more details, Figure 5 plots the annual out-of-sample R^2 of **ATOMS** and the baselines in each industry. We observe that **ATOMS** maintains its advantage across diverse sectors including cyclical industries (Durbl, Cars, Trans) and defensive sectors (Food, Utils, Cnsum). This cross-sectional consistency strengthens our confidence that the observed performance reflects genuine adaptivity to changing market conditions rather than industry-specific anomalies.

Notably, the adaptive algorithm shows particular strength in industries most sensitive to business cycle fluctuations, such as durable goods (Durbl), consumer discretionary (Rtail), and financial services (Finan). This pattern aligns with economic intuition, as these sectors experience the most dramatic shifts in risk premia during economic transitions. Our method’s ability to capture these dynamics more effectively than benchmarks suggests it successfully identifies the changing risk-return tradeoffs that characterize different phases of the business cycle.

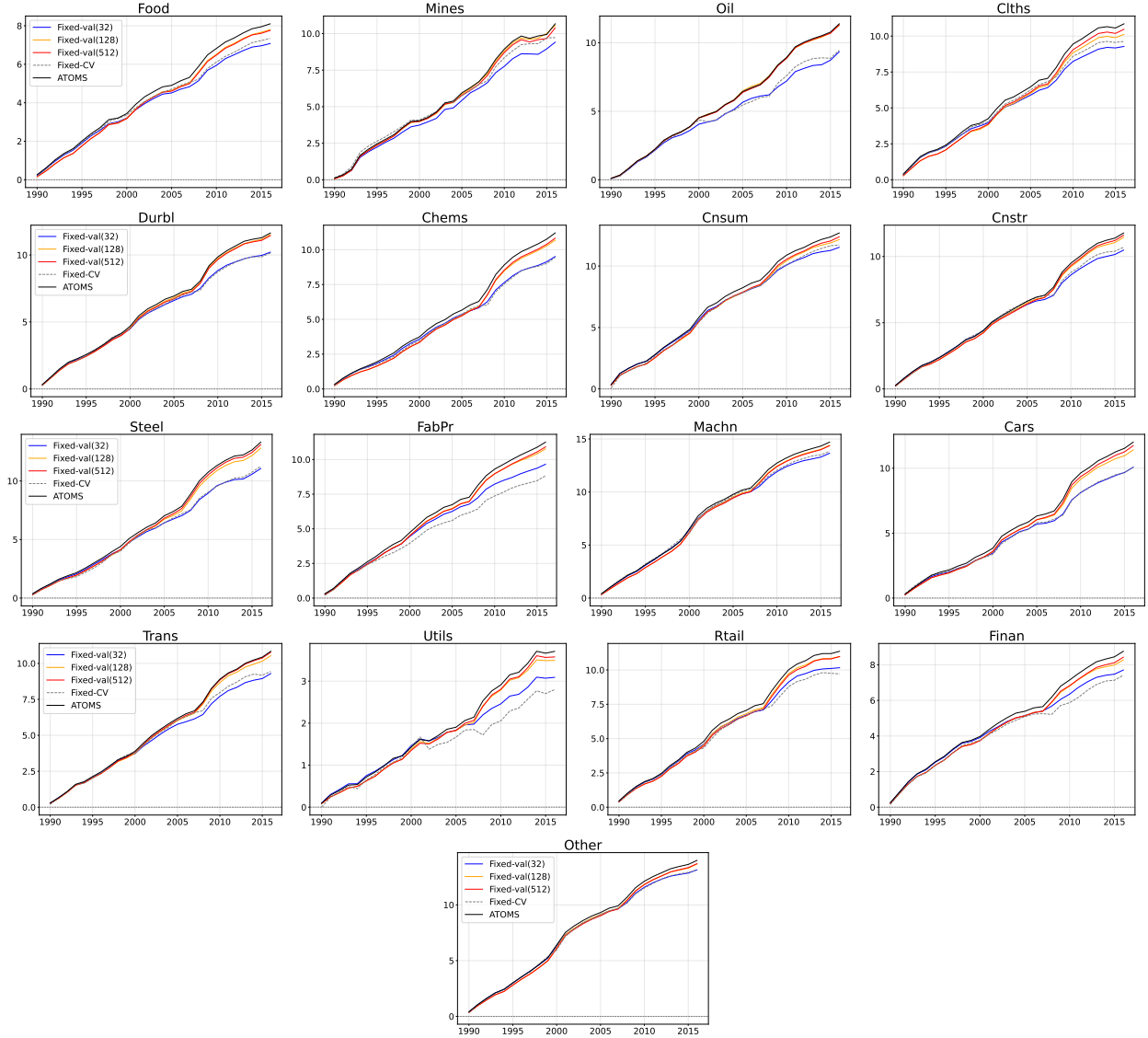
5.4 Trading Strategies

To assess the economic significance of our asset pricing framework, we implement trading strategies based on its return predictions and evaluate wealth accumulation—this tests whether the documented predictive power translates into economic value for investors.

Our trading protocol is standard: we start with initial wealth $W_0 = 1$. At the beginning of each month t , the model yields a predicted return \hat{f}_t . We trade based on the sign of this predicted return. That is, on each trading day $i \in [B_t]$ of month t , we trade according to the sign of the predicted return $\hat{f}_t(\mathbf{x}_{t,i})$: we take a long position if $\hat{f}_t(\mathbf{x}_{t,i}) > 0$, and a short position otherwise. Consequently, the portfolio wealth evolves according to the realized daily returns $y_{t,i}$, adjusted by the position direction. The cumulative wealth W_t at the end of month t is given by:

$$W_t = W_{t-1} \prod_{i=1}^{B_t} \left(1 + y_{t,i} \cdot \text{sign}(\hat{f}_t(\mathbf{x}_{t,i})) \right).$$

Figure 6: Cumulative Wealth Curve of ATOMS and Baselines for 17 Industry Portfolios.



This figure reports the cumulative wealth (in log scale) of trading strategies based on predictions from our adaptive algorithm **ATOMS** (black line), as well as the fixed-window baselines **Fixed-val(32)** (blue), **Fixed-val(128)** (orange), **Fixed-val(512)** (red), and **Fixed-CV** (gray dashed line). For most industries, our algorithm **ATOMS** consistently attains more cumulative wealth than the fixed-window baselines. The title in each subfigure is Kenneth French's acronym for each industry. For the full names of these industries, please refer to Table 4.

Iterating this process, the final wealth W_T at time T is

$$W_T = W_0 \prod_{t=1}^T \prod_{i=1}^{B_t} \left(1 + y_{t,i} \cdot \text{sign}(\hat{f}_t(\mathbf{x}_{t,i})) \right).$$

Our investment starts in January 1990 as the first month $t = 1$, and ends in November 2016. We generate wealth trajectories $\{W_t^{\text{alg}}\}_{t=1}^T$ for our algorithm and the fixed-window baselines $\text{alg} \in \{\text{ATOMS}, \text{Fixed-val}(\ell), \text{Fixed-CV}\}$.

Table 3: Average Cumulative Wealth Relative Excess of **ATOMS** over Baselines.

Baseline Model	Fixed-val (32)	Fixed-val (128)	Fixed-val (512)	Fixed-CV
Excess Ratio	3.38	0.48	0.31	3.54

This table reports the Excess Ratio. For each industry, we compute an Excess Ratio, then report average across the 17 computed industries. This metric can also be considered as the Excess Ratio if we were to invest in an equal-weighted portfolio that allocates evenly among all 17 industries in the initial period.

Figure 6 depicts the evolution of the log cumulative wealth $\log W_t^{\text{alg}}$ for each method at the end of each year from 1990 to 2016. We observe that for most industries, the adaptive algorithm **ATOMS** consistently yields higher cumulative wealth than the fixed-window baselines, surpassing both the short and long validation windows. This superior performance highlights the algorithm’s capacity to generate substantial excess returns by balancing the trade-off between non-stationarity and model complexity.

To further quantify the performance gain across the 17 industries, we compute an Excess Ratio of $W_T^{\text{ATOMS}}/W_T^{\text{alg}} - 1$ against any baseline $\text{alg} \in \{\text{Fixed-val}(\ell), \text{Fixed-CV}\}$, and take its average across all 17 industries. This simple arithmetic average can be considered as an equal-weighted portfolio that invests 1/17th initial wealth in each of the industries. A value greater than zero indicates that our adaptive method **ATOMS** accumulates higher terminal wealth than the benchmark alg . In Table 3, we report the average Excess Ratio over the 17 industries. The positive values of Excess Ratio indicate that **ATOMS** consistently generates superior wealth accumulation compared to the fixed-window benchmarks for the equal-weighted portfolio of industries. Compared with the best-performing benchmark of long horizon validation **Fixed-val**(512), our method yields 31% higher return for its investor by the end of our OOS period. With weaker benchmarked methods such as model picked by **Fixed-CV** that uses cross-validation, our method obtains 3.54 times more wealth over the investment horizon of 1990 to 2016.

6 Conclusions

Our empirical results demonstrate the practical value of this framework across multiple dimensions. Most notably, during periods of heightened economic stress, our adaptive method **ATOMS** exhibits superior performance compared to fixed-window approaches. Our approach is well motivated by the documented facts in Section 3.1 that, during recession periods including the 1990 Gulf War recession, the 2001 dot-com bubble burst and 9/11 attack, and the 2007-2009 Financial Crisis, simpler models trained on shorter windows consistently outperformed more complex models trained on longer windows. This empirical evidence validates our theory of the nonstationarity-complexity tradeoff.

The adaptive algorithm’s performance during economic downturns is worth pointing out. As shown in Table 2, **ATOMS** achieves an out-of-sample R^2 of 0.027 during the brief but severe 1990 Gulf War recession, while the best fixed-window benchmark **Fixed-val**(512) produces a negative R^2 of -0.031 . During the 2001 recession, **ATOMS** attains an impressive R^2 of 0.125, outperforming

Fixed-val(512) by 6.8% (0.117). Even during the prolonged Global Financial Crisis of 2007-2009, **ATOMS** maintains its advantage with an R^2 of 0.041 compared to **Fixed-val**(512)’s 0.039. Beyond statistical performance metrics, the economic significance of our approach is demonstrated through trading strategy analysis. Averaged across the industries, our model yields 31% higher return than the best performing benchmark in the OOS period.

Several future directions are worth exploring. First, our adaptive model selection framework relies on the assumption that data is independent across time even though the distribution can change arbitrarily. While numerical experiments show that our method is robust against temporal dependence in real-world financial time series, it would be interesting and important to extend the framework in a principled way. Second, our framework of joint model and training window selection requires training a large number of candidate models, which can be computationally intensive. A valuable future direction is to reduce these training costs. For example, a heuristic approach is to utilize the optimized parameters from previous periods as “warm starts” for subsequent training.

Acknowledgement

Chengpiao Huang and Kaizheng Wang’s research is supported by National Science Foundation grants DMS-2210907 and DMS-2515679. Agostino Capponi’s research was supported in part through grants of the Global Risk Institute and of Fi-Tek.

References

- ANDREWS, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* **61** 821–856.
URL <http://www.jstor.org/stable/2951764>
- BAI, J. and PERRON, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78.
URL <http://www.jstor.org/stable/2998540>
- BANERJEE, A. and URGA, G. (2005). Modelling structural breaks, long memory and stock market volatility: an overview. *Journal of Econometrics* **129** 1–34. Modelling structural breaks.
URL <https://www.sciencedirect.com/science/article/pii/S0304407604001630>
- BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *The Annals of Statistics* **33** 1497 – 1537.
URL <https://doi.org/10.1214/009053605000000282>
- BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
URL <https://doi.org/10.1093/acprof:oso/9780199535255.001.0001>

- CHEN, L., PELGER, M. and ZHU, J. (2024). Deep learning in asset pricing. *Management Science* **70** 714–750.
URL <https://doi.org/10.1287/mnsc.2023.4695>
- CHIB, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics* **86** 221–241.
URL <https://www.sciencedirect.com/science/article/pii/S0304407697001152>
- CHOI, D., JIANG, W. and ZHANG, C. (2025). Alpha go everywhere: Machine learning and international stock returns. *The Review of Asset Pricing Studies* **15** 288–331.
URL <https://academic.oup.com/raps/article-abstract/15/3-4/288/8172522>
- CHOW, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica* **28** 591–605.
URL <http://www.jstor.org/stable/1910133>
- CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L. and STEIN, C. (2022). *Introduction to algorithms*. MIT press.
- DIDISHEIM, A., KE, S. B., KELLY, B. T. and MALAMUD, S. (2024). APT or “AIPT”? The surprising dominance of large factor models. Tech. rep., National Bureau of Economic Research.
- DUAN, Y., JIN, C. and LI, Z. (2021). Risk bounds and rademacher complexity in batch reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*. PMLR.
URL <https://proceedings.mlr.press/v139/duan21a.html>
- FAMA, E. F. and FRENCH, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33** 3–56.
URL <https://www.sciencedirect.com/science/article/pii/0304405X93900235>
- FREYBERGER, J., NEUHIERL, A. and WEBER, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies* **33** 2326–2377.
- GOLDENSHLUGER, A. and LEPSKI, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli* **14** 1150 – 1190.
URL <https://doi.org/10.3150/08-BEJ144>
- GU, S., KELLY, B. and XIU, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies* **33** 2223–2273.
URL <https://doi.org/10.1093/rfs/hhaa009>
- HAN, E., HUANG, C. and WANG, K. (2024). Model assessment and selection under temporal distribution shift. In *Proceedings of the 41st International Conference on Machine Learning* (R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett and F. Berkenkamp,

- eds.), vol. 235 of *Proceedings of Machine Learning Research*. PMLR.
URL <https://proceedings.mlr.press/v235/han24b.html>
- HOMM, U. and BREITUNG, J. (2012). Testing for speculative bubbles in stock markets: A comparison of alternative methods. *Journal of Financial Econometrics* **10** 198–231.
URL <https://doi.org/10.1093/jjfinec/nbr009>
- INOUE, A., JIN, L. and ROSSI, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics* **196** 55–67.
URL <https://www.sciencedirect.com/science/article/pii/S0304407616301713>
- KELLY, B., MALAMUD, S. and ZHOU, K. (2024). The virtue of complexity in return prediction. *The Journal of Finance* **79** 459–503.
URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13298>
- KELLY, B. and XIU, D. (2023). Financial machine learning. *Foundations and Trends® in Finance* **13** 205–363.
- KELLY, B. T. and MALAMUD, S. (2025). Understanding the virtue of complexity. Tech. Rep. 25-96, Swiss Finance Institute.
URL <https://ssrn.com/abstract=5346842>
- KELLY, B. T., MALAMUD, S. and ZHOU, K. (2022). The virtue of complexity everywhere. Tech. Rep. 22-57, Swiss Finance Institute.
URL <https://ssrn.com/abstract=4166368>
- LIU, C. and MAHEU, J. M. (2007). Are there structural breaks in realized volatility? *Journal of Financial Econometrics* **6** 326–360.
URL <https://doi.org/10.1093/jjfinec/nbn006>
- MASSART, P. (2000). About the constants in Talagrand’s concentration inequalities for empirical processes. *The Annals of Probability* **28** 863 – 884.
URL <https://doi.org/10.1214/aop/1019160263>
- MENDELSON, S. (2002). Geometric parameters of kernel machines. In *Computational Learning Theory* (J. Kivinen and R. H. Sloan, eds.). Springer Berlin Heidelberg, Berlin, Heidelberg.
- PESARAN, M. H. and PICK, A. (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics* **29** 307–318.
URL <https://doi.org/10.1198/jbes.2010.09018>
- PESARAN, M. H. and TIMMERMAN, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics* **137** 134–161.
URL <https://www.sciencedirect.com/science/article/pii/S0304407606000418>

WAHBA, G. (1990). *Spline models for observational data*. SIAM.

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Appendix

A Additional details for empirical asset pricing

In this section, we provide additional details for the numerical experiments in Section 5.

A.1 Data

We aim to explain the cross-section of 17 industry portfolio returns from Kenneth French’s website⁴. We first describe our rich set of covariates, which combines daily and monthly sources to construct a comprehensive time series combining macroeconomic and cross-sectional signals. The final daily-frequency sample spans the time period from September 1987 to November 2016.

Common Factors. To capture the underlying macroeconomic and systematic risk structure in our response variables, we include 15 common factors from the monthly dataset of Chen et al. (2024), who develop a deep learning framework for estimating the stochastic discount factor (SDF) consistent with the no-arbitrage condition.⁵

The Chen et al. (2024) dataset provides three distinct elements relevant to our analysis: (i) the estimated stochastic discount factor (SDF), capturing the aggregate price of risk implied by their no-arbitrage model; (ii) the returns of ten equal-weighted decile portfolios sorted by firms’ exposure to the SDF (beta-sorted), which serve as cross-sectional test assets; and (iii) four macroeconomic hidden state variables, derived from the GAN model, that summarize the joint dynamics of 178 macroeconomic time series into a small set of nonlinear factors reflecting business cycle conditions and systemic risk.⁶

In addition to these 15 factors, we incorporate the daily Fama-French three factors (FF3) from Fama and French (1993), which include the excess market return (MKT), the size factor (SMB), and the value factor (HML). These factors provide a benchmark set of linear risk exposures that have been shown to explain broad cross-sectional patterns in stock returns. Each factor is lagged appropriately to avoid look-ahead bias.

This unified set of common factors allows us to incorporate both the traditional linear risk structure captured by FF3 and the nonlinear, macroeconomically conditioned dynamics extracted from Chen et al. (2024).

⁴https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_17_ind_port.html

⁵Their model integrates three neural network components: a feedforward network to approximate the nonlinear functional form of the SDF, a recurrent Long-Short Term Memory (LSTM) network to extract macroeconomic state variables, and a generative adversarial network (GAN) that constructs the most informative test assets for pricing. This architecture yields an empirically estimated SDF that represents the conditional mean-variance efficient portfolio of all U.S. equities.

⁶These components jointly span the main drivers of systematic variation in expected returns. The SDF and beta-sorted deciles capture the cross-sectional risk-return trade-offs, while the macroeconomic hidden states track time variation in the pricing kernel associated with expansions, recessions, and crisis periods. The dataset covers the period 1967-2016; to align frequencies with our daily response data, each monthly observation is assigned to all trading days within the corresponding month.

Firm-Level Returns and Characteristics. We augment the dataset with daily firm-level stock returns obtained from the CRSP database. Excess returns are computed relative to the daily risk-free rate from Kenneth French’s data library.⁷ To characterize the cross-sectional heterogeneity in firm fundamentals and trading behavior, we incorporate the comprehensive panel of firm-level characteristics from Gu et al. (2020). Their dataset provides 94 standardized monthly characteristics for U.S. equities, constructed to ensure comparability across firms and over time.

Specifically, Gu et al. (2020) identify three dominant groups of predictive signals: (i) **Price trend variables:** measures of short- and long-term momentum, industry momentum, and short-term reversal that capture the persistence of returns and investor underreaction; (ii) **Liquidity and size variables:** market capitalization, trading volume, turnover, and bid-ask spread, which reflect trading frictions and the limits to arbitrage; (iii) **Volatility and risk variables:** total and idiosyncratic volatility, market beta, and higher-order terms such as beta squared, which proxy for systematic and residual risk exposure.⁸

A.1.1 Size Filter

We apply a size-based filter to focus the analysis on the subset of firms that are most representative of the small-cap universe. Specifically, within each day t , we retain only the bottom 25% of firms by market equity.⁹

Formally, let $S_{i,m}$ denote the size of firm i in month m . For all trading days t belonging to month m , we retain only those firms satisfying

$$S_{i,m} \leq Q_{0.25}(\{S_{j,m}\}_{j \in \mathcal{F}_m}),$$

where $Q_{0.25}(\cdot)$ denotes the 25th percentile of the cross-sectional size distribution, and \mathcal{F}_m is the set of all firms available in month m . Restricting the sample in this way allows us to study the predictive role of firm characteristics and macroeconomic factors within a homogeneous segment of the equity market, mitigating heterogeneity arising from scale effects in firm size.

A.1.2 Long-Short Decile Portfolios from Characteristics

To transform the firm-level characteristics into tradable, interpretable covariates, we construct daily long-short portfolios sorted on each of the Gu et al. (2020) characteristics. For each of the 94 characteristics, denoted by $c \in \{1, \dots, 94\}$, and for each day t , we sort firms into ten equal-weighted portfolios based on the value of characteristic c .

⁷To mitigate the impact of outliers that can distort cross-sectional averages, we Winsorize the daily cross-section of excess returns by removing the top and bottom 1% of firms within each day based on their excess return values. This procedure ensures that the resulting portfolio and factor constructions are not unduly influenced by extreme realizations.

⁸Because the Gu et al. (2020) characteristics are available only at a monthly frequency, we assign each monthly vector of characteristics to all trading days within that month for the corresponding firm. This alignment ensures consistency between the daily CRSP return data and the lower-frequency fundamental information.

⁹We measure market equity with the (monthly) “mvel1” size covariate from Gu et al. (2020), this variable represents the firm’s market capitalization at the end of month m , standardized for comparability across firms and time.

Formally, let $X_{i,c,m}$ denote the value of characteristic c for firm i in month m , which is assigned to all trading days within that month, and let $R_{i,t}^{\text{ex}}$ denote the daily excess return of firm i on day t . We compute monthly breakpoints for characteristic c using cross-sectional quantiles:

$$b_{c,k,m} = Q_{k/10}(\{X_{i,c,m}\}_{i \in \mathcal{F}_t}), \quad k = 1, \dots, 10,$$

where $Q_p(\cdot)$ denotes the p -th empirical quantile, and \mathcal{F}_t is the set of all firms observed in day t .

We then compute the daily return of the k -th decile portfolio for characteristic c as:

$$D_{c,k,t} = \frac{1}{N_{c,k,t}} \sum_{i: b_{c,k-1,m} < X_{i,c,m} \leq b_{c,k,m}} R_{i,t}^{\text{ex}},$$

where $N_{c,k,t}$ is the number of firms assigned to the k -th decile at time t . Each decile is equal-weighted to ensure that portfolio performance reflects cross-sectional variation in firm characteristics rather than differences in market capitalization.

The corresponding long-short portfolio return for characteristic c is defined as:

$$\text{LS}_{c,t} = D_{c,10,t} - D_{c,1,t},$$

representing the daily return to a strategy that is long the highest-decile firms (those with the largest values of $X_{i,c,m}$) and short the lowest-decile firms (those with the smallest values of $X_{i,c,m}$).

This construction yields a balanced set of long-short factor returns that isolate the pricing implications of each firm characteristic. Repeating this procedure for all $c \in \{1, \dots, 94\}$ produces a time series matrix of 94 characteristic-sorted long-short portfolio returns at the daily frequency.

A.1.3 Final Covariate Set

Our final set of predictors integrates macroeconomic, factor-based, and cross-sectional sources of variation to form a unified time series of covariates for forecasting the Russell 2000 index and the French 17 industry portfolio returns. The resulting dataset combines information from three complementary dimensions of the asset pricing literature:

1. **Macroeconomic and systematic factors:** the 10 equal-weighted beta-sorted decile portfolios, the 4 macroeconomic hidden states, and the estimated stochastic discount factor (SDF) from [Chen et al. \(2024\)](#).
2. **Benchmark risk factors:** the daily Fama-French 3 factors (FF3) from [Fama and French \(1993\)](#), consisting of the market excess return (MKT), the size factor (SMB), and the value factor (HML).
3. **Cross-sectional characteristic factors:** the 94 daily long-short characteristic-sorted portfolio returns, from the [Gu et al. \(2020\)](#) characteristics.

4. **Lagging features:** we augment the feature set with one-day lagged returns. For the prediction of the Russell 2000 index return, we use the index’s own lag. For the prediction of the 17 industry portfolio returns, we use the full vector of the 17 industry portfolio lagged returns.

By combining the first three components, our dataset links the macroeconomic and cross-sectional perspectives on asset pricing. The [Chen et al. \(2024\)](#) factors embed the nonlinear, time-varying structure of the stochastic discount factor, while the [Gu et al. \(2020\)](#) characteristic-sorted long-short portfolios summarize the cross-sectional distribution of risk premia across firms. Incorporating the Fama-French 3 factors provides a benchmark for evaluating whether these modern, high-dimensional representations offer predictive power beyond the traditional linear framework.

A.2 Figures and tables

Table 4: Name Mapping for the 17 Industries

Industry Acronym	Full Industry Name
Food	Food
Mines	Mining and Minerals
Oil	Oil and Petroleum Products
Clths	Textiles, Apparel & Footwear
Durbl	Consumer Durables
Chems	Chemicals
Cnsum	Drugs, Soap, Perfumes, Tobacco
Cnstr	Construction and Construction Materials
Steel	Steel Works Etc
FabPr	Fabricated Products
Machn	Machinery and Business Equipment
Cars	Automobiles
Trans	Transportation
Utils	Utilities
Rtail	Retail Stores
Finan	Banks, Insurance Companies, and Other Financials
Other	Other

A.3 Machine learning model implementation details

In this appendix, we provide detailed specifications for the hyperparameter tuning procedures and training configurations used in our empirical analysis. These technical details complement the main text by offering comprehensive information about the model selection process and computational implementation.

A.3.1 Model Hyperparameter Grids

For each model class considered in our analysis, we systematically explore a comprehensive grid of hyperparameter values for model selection. The hyperparameter grids are designed to balance computational efficiency with thorough exploration of the model space.

Linear Models with Regularization. For the ridge regression (Ridge), LASSO (LASSO), and elastic net (E-Net) models, we consider the following hyperparameter specifications:

1. For ridge regression, we consider values of the regularization parameter α on a logarithmic scale:

$$\alpha \in \{10^{-3}, 10^{-1.5}, 1, 10^{1.5}, 10^3\}.$$

This range allows for both strong regularization (small α) and weak regularization (large α), accommodating different levels of multicollinearity in our high-dimensional covariate space.

2. For LASSO, we consider values of the regularization parameter α on a logarithmic scale:

$$\alpha \in \{10^{-5}, 10^{-3.5}, 10^{-2}, 10^{-0.5}, 10\}.$$

The ℓ_1 penalty in LASSO facilitates feature selection, which is particularly valuable given our large set of covariates.

3. For the elastic net, we consider the following combinations of the regularization parameter α and the mixing parameter r :

$$\alpha \in \{10^{-3}, 1, 10^3\}, \quad r \in \{0.01, 0.05, 0.1\}.$$

This grid explores the balance between feature selection and coefficient shrinkage.

Random Forest. For the random forest models, we consider the following combinations of the number of trees n_{tree} and the maximum tree depth d_{max} :

$$n_{\text{tree}} \in \{10, 100, 200\}, \quad d_{\text{max}} \in \{3, 5, 10\}.$$

Increasing the number of trees generally improves model stability and reduces variance, though with diminishing returns beyond a certain point. Trees with shallower depths provide stronger regularization, while deeper trees can capture more complex nonlinear relationships.

A.3.2 Training Window Configurations

To assess the performance of our adaptive model selection algorithm across different data regimes, we train models on estimation windows of varying lengths. For each month t , we consider training

windows of $4^k \wedge (t - 1)$ months with $0 \leq k \leq 5$. In particular, since $t \in \{1, \dots, 350\}$, then $k = 5$ corresponds to a full training window of $(t - 1)$ months. This yields the following window lengths:

Table 5: Training Window Lengths by Value of k

k	Window Length (months)	Approximate Years
0	1	0.08
1	4	0.33
2	16	1.33
3	64	5.33
4	256	21.33
5	$t - 1$	$(t - 1)/12$

This exponential scaling allows us to examine how model performance varies with the amount of historical data available for training. Shorter windows capture recent market dynamics but may be susceptible to noise, while longer windows provide more stable parameter estimates but may miss structural changes in the data-generating process.

A.3.3 Hyperparameter for ATOMS

In our implementation of ATOMS, we set $\delta' = 0.1$ and $M^2 = 5 \times 10^{-4}$.

A.3.4 Computational Implementation

All models are implemented using Python 3.9 with the following libraries:

- Linear models: `scikit-learn` version 1.0.2, specifically `Ridge`, `Lasso`, and `ElasticNet`.
- Random forest: `scikit-learn`'s `RandomForestRegressor` with `random_state=0`.
- Data manipulation: `pandas` version 1.4.2 and `numpy` version 1.21.5.

B Extension: Model Selection with the R^2 Metric

In this section, we propose a variant of our model selection method that is tailored to the R^2 metric. We consider the following population form:

$$\tilde{R}_t^2(f) = 1 - \frac{\mathbb{E}_{(\mathbf{x}, y) \sim P_t} \left[(f(\mathbf{x}) - y)^2 \right]}{\mathbb{E}_{(\mathbf{x}, y) \sim P_t} [y^2]}, \quad (\text{B.1})$$

Define $V_t = \mathbb{E}_{(\mathbf{x}, y) \sim P_t} [y^2]$, then $\tilde{R}_t^2(f) = 1 - L_t(f)/V_t$. For simplicity, we assume that at the beginning of each period $t \in \mathbb{Z}_+$, we have access to $\{V_j\}_{j=1}^{t-1}$. In our numerical experiments, we will approximate V_j by its empirical counterpart computed from the validation data $\mathcal{D}_j^{\text{va}}$. For the population R^2 metric to be well defined, we assume for simplicity that $\{V_t\}_{t=1}^\infty$ are bounded away from zero.

Assumption B.1 (Uniformly lower bounded second moments). *There exists $v > 0$ such that $V_t \geq v$ for all $t \in \mathbb{Z}_+$.*

We first define the model comparison subroutine, which aims to output the better of two given candidate models f_1 and f_2 . Define the R^2 performance gap

$$\Delta_t^R = \tilde{R}_t^2(f_2) - \tilde{R}_t^2(f_1) = \frac{L_t(f_1) - L_t(f_2)}{V_t}.$$

Then f_2 outperforms f_1 if and only if $\Delta_t^R > 0$. For each window size $\ell \in [t-1]$, we can form a rolling-window estimator of Δ_t^R , given by

$$\hat{\Delta}_{t,\ell}^R = \frac{\hat{\Delta}_{t,\ell}}{V_{t,\ell}} \quad (\text{B.2})$$

where $\hat{\Delta}_{t,\ell}$ is defined by (4.1), and $V_{t,\ell} = \frac{1}{n_{t,\ell}} \sum_{j=t-\ell}^{t-1} n_j V_j$. We establish a bias-variance decomposition for the estimation error of $\hat{\Delta}_{t,\ell}^R$.

Lemma B.1 (Bias-variance decomposition). *Let Assumptions 2.1, 3.1 and B.1 hold. Let $\sigma_{t,\ell}^R = \sigma_{t,\ell}/V_{t,\ell}$, where $\sigma_{t,\ell}$ is defined by (4.4). For $\delta \in (0, 1)$, define*

$$\begin{aligned} \phi_R(t, \ell) &= \max_{t-\ell \leq j \leq t-1} |\Delta_j^R - \Delta_t^R|, \\ \psi_R(t, \ell, \delta) &= \begin{cases} 8M^2/v, & \text{if } n_{t,\ell} = 1 \\ \sigma_{t,\ell}^R \sqrt{\frac{2 \log(2/\delta)}{n_{t,\ell}}} + \frac{16(M^2/v) \log(2/\delta)}{3n_{t,\ell}}, & \text{if } n_{t,\ell} \geq 2 \end{cases}. \end{aligned}$$

With probability at least $1 - \delta$,

$$|\hat{\Delta}_{t,\ell}^R - \Delta_t^R| \leq \phi_R(t, \ell) + \psi_R(t, \ell, \delta).$$

Proof. See Appendix D.5. □

Based on Lemma B.1 and following the same idea as (4.4), we form a data-driven proxy for $\psi_R(t, \ell, \delta)$, given by

$$\hat{\psi}_R(t, \ell, \delta) = \begin{cases} 8M^2/v, & \text{if } n_{t,\ell} = 1 \\ \hat{v}_{t,\ell}^R \sqrt{\frac{2 \log(2/\delta)}{n_{t,\ell}}} + \frac{64(M^2/v) \log(2/\delta)}{3(n_{t,\ell} - 1)}, & \text{if } n_{t,\ell} \geq 2 \end{cases}, \quad (\text{B.3})$$

where

$$\hat{v}_{t,\ell}^R = \frac{\hat{v}_{t,\ell}}{V_{t,\ell}}, \quad (\text{B.4})$$

and $\widehat{v}_{t,\ell}$ is defined in (4.3). Following the same idea as (4.5), we also form a data-driven proxy for $\phi_R(t, \ell, \delta)$:

$$\widehat{\phi}_R(t, \ell, \delta) = \max_{i \in [\ell]} \left(|\widehat{\Delta}_{t,\ell}^R - \widehat{\Delta}_{t,i}^R| - [\widehat{\psi}_R(t, \ell, \delta) + \widehat{\psi}_R(t, i, \delta)] \right)_+. \quad (\text{B.5})$$

This yields Algorithm 4 as the model comparison routine.

Algorithm 4 Adaptive Rolling Window for Model Comparison (R^2 Metric)

Input: Candidate models $\{f_1, f_2\}$, validation data $\{\mathcal{D}_j^{\text{va}}\}_{j=1}^{t-1}$, variances $\{V_j\}_{j=1}^{t-1}$, hyperparameters δ', M, v .

for $\ell = 1, \dots, t-1$

 Compute $\widehat{\Delta}_{t,\ell}^R, \widehat{v}_{t,\ell}^R, \widehat{\psi}_R(t, \ell, \delta')$ and $\widehat{\phi}_R(t, \ell, \delta')$ according to (B.2), (B.4), (B.3) and (B.5).

 Choose window size

$$\widehat{\ell} \in \operatorname{argmin}_{\ell \in [t-1]} \left\{ \widehat{\phi}_R(t, \ell, \delta') + \widehat{\psi}_R(t, \ell, \delta') \right\}. \quad (\text{B.6})$$

 Select $\widehat{\lambda} = 1$ if $\widehat{\Delta}_{t,\widehat{\ell}}^R \leq 0$, and $\widehat{\lambda} = 2$ otherwise.

return $\widehat{f} = f_{\widehat{\lambda}}$.

By using Algorithm 4 as the pairwise comparison subroutine \mathcal{A} in Algorithm 1, we obtain an R^2 -based model selection algorithm, which we call **ATOMS-R2**. We establish the following guarantee in terms of the R^2 metric.

Theorem B.1 (Near-optimal model selection with R^2). *Let Assumptions 2.1 and 3.1 hold. Choose $\delta \in (0, 1)$ and set $\delta' = 1/(3\Lambda^2 t)$ in Algorithm 1. With probability at least $1 - \delta$, the output \widehat{f} of **ATOMS-R2** satisfies*

$$\max_{\lambda \in [\Lambda]} \widetilde{R}_t^2(f_\lambda) - \widetilde{R}_t^2(\widehat{f}) \lesssim \log(\Lambda t / \delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} \max_{\lambda \in [\Lambda]} |\widetilde{R}_j^2(f_\lambda) - \widetilde{R}_t^2(f_\lambda)| + \frac{M^2/v}{\sqrt{n_{t,\ell}}} \right\}.$$

Here \lesssim hides a universal constant.

Proof of Theorem B.1. See Appendix D.6. □

Theorem B.1 shares a similar interpretation as Theorem 4.2. The term

$$\max_{t-\ell \leq j \leq t-1} \max_{\lambda \in [\Lambda]} |\widetilde{R}_j^2(f_\lambda) - \widetilde{R}_t^2(f_\lambda)|$$

quantifies the shift in the R^2 metric of the models within the last ℓ periods, and $(M^2/v)/\sqrt{n_{t,\ell}}$ represents the statistical uncertainty associated with the $n_{t,\ell}$ validation data points. Together, they represent the errors that arise when using a fixed validation window ℓ to select models. Theorem B.1 shows that our R^2 -based model selection algorithm **ATOMS-R2** is comparable to an oracle that uses the optimal validation window in hindsight to attain the highest R^2 .

In Appendix G.2, we present numerical experiment results for **ATOMS-R2**, which are similar to those for **ATOMS**.

C Theoretical Analysis of Nonstationarity-Complexity Tradeoff

In this section, we provide more details on Theorem 3.1 and its proof.

We first set up some mathematical notation. Recall that $f_t^*(\cdot) = \mathbb{E}_{(x,y) \sim P_t}[y \mid x = \cdot]$ is the Bayes optimal least squares estimator, which minimizes $L_t(f)$ over all measurable $f : \mathcal{X} \rightarrow \mathbb{R}$. Let

$$\bar{f}_t = \operatorname{argmin}_{f \in \mathcal{F}} L_t(f),$$

which minimizes $L_t(f)$ over all $f \in \mathcal{F}$. For each t , let $P_{t,\mathbf{x}}$ be the marginal distribution of P_t with respect to the covariates \mathbf{x} . The distribution $P_{t,\mathbf{x}}$ induces a norm $\|\cdot\|_t$, given by

$$\|f\|_t = \sqrt{\mathbb{E}_{\mathbf{x} \sim P_{t,\mathbf{x}}} [f(\mathbf{x})^2]}.$$

It can be shown that $\mathcal{E}_t(f) = \|f - f_t^*\|_t^2$ for all $f : \mathcal{X} \rightarrow \mathbb{R}$.

C.1 A Non-Stationary Local Rademacher complexity

We now formally define $r_{t,k}(\mathcal{F})$ through a non-stationary version of the *local Rademacher complexity* (Bartlett et al., 2005). We first define the *Rademacher complexity*, which reflects the richness of a function class with respect to certain data samples.

Definition C.1 (Rademacher complexity). *Let $\{z_i\}_{i=1}^n$ be independent random variables in \mathcal{X} . Let \mathcal{G} be a class of functions from \mathcal{X} to \mathbb{R} . The **Rademacher complexity** of \mathcal{G} associated with $\{z_i\}_{i=1}^n$ is defined by*

$$\mathfrak{R}(\mathcal{G}; \{z_i\}_{i=1}^n) = \mathbb{E} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(z_i) \right],$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. random variables following the Rademacher distribution $\mathbb{P}(\varepsilon_1 = 1) = \mathbb{P}(\varepsilon_1 = -1) = 1/2$, and are independent of $\{z_i\}_{i=1}^n$.

The local Rademacher complexity is the Rademacher complexity of some local function class centered at \bar{f}_t .

Definition C.2 (Local function class). *For every $r \geq 0$, define the local function class*

$$\mathcal{F}_{t,k}(r) = \left\{ f \in \mathcal{F} : \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \|f - \bar{f}_t\|_j^2 \leq r \right\}.$$

In words, $\mathcal{F}_{t,k}(r)$ consists of functions $f \in \mathcal{F}$ that are close to \bar{f}_t with respect to the distributions $\{P_{j,\mathbf{x}}\}_{j=t-k}^{t-1}$ on average. We are now ready to define the quantity $r_{t,k}(\mathcal{F})$, which is more formally known as the *critical radius* of the function

$$\mathfrak{R}_{t,k}(r; \mathcal{F}) = \mathfrak{R}(\mathcal{F}_{t,k}(r); \mathcal{D}_{t,k}^{\text{tr}}).$$

Definition C.3 (Subroot function). A function $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is **subroot** if it is increasing and $r \mapsto \psi(r)/\sqrt{r}$ is decreasing on $(0, \infty)$.

Definition C.4 (Critical radius). The **critical radius** of $\mathfrak{R}_{t,k}(r; \mathcal{F})$ is defined by

$$r_{t,k}(\mathcal{F}) = \inf \{r \geq 0 : \exists \text{ subroot } \psi \text{ such that } \psi(r) = r, \text{ and } \psi(s) \geq \mathfrak{R}_{t,k}(s; \mathcal{F}) \forall s \geq r\}.$$

C.2 Proof of Theorem 3.1

We now prove Theorem 3.1. In the proof, we will write $\hat{f} = \hat{f}_{t,k}$ to emphasize its dependence on the time t and the training window size k . The key of the proof is the following lemma, which is a non-stationary version of Theorem 3.3 in Bartlett et al. (2005).

Lemma C.1 (Localized uniform concentration). Let $\mathbf{z}_1, \dots, \mathbf{z}_n$ be independent random variables taking values in a space $\mathcal{Z} \subseteq \mathbb{R}^{d+1}$. Let \mathcal{G} be a collection of functions from \mathcal{Z} to $[a, b]$. Suppose that there exist $T : \mathcal{G} \rightarrow \mathbb{R}_+$ and $C, \eta_1, \dots, \eta_n \geq 0$ such that the following noise condition holds:

$$\frac{1}{n} \sum_{i=1}^n \text{var}[g(\mathbf{z}_i)] \leq T(g) \leq C \cdot \frac{1}{n} \sum_{i=1}^n (\mathbb{E}[g(\mathbf{z}_i)] + \eta_i), \quad \forall g \in \mathcal{G}. \quad (\text{C.1})$$

Let ψ be a sub-root function with a fixed point r^* satisfying

$$\psi(r) \geq C \cdot \mathfrak{R}(\{g \in \mathcal{G} : T(g) \leq r\}; \{\mathbf{z}_i\}_{i=1}^n), \quad \forall r \geq r^*.$$

Take $\delta \in (0, 1)$. With probability at least $1 - \delta$, for all $g \in \mathcal{G}$ and $K > 1$,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[g(\mathbf{z}_i)] \leq \frac{K}{K-1} \cdot \frac{1}{n} \sum_{i=1}^n g(\mathbf{z}_i) + c \left[\frac{K}{C} r^* + ((b-a) + CK) \frac{\log(1/\delta)}{n} \right] + \frac{1}{K-1} \cdot \frac{1}{n} \sum_{i=1}^n \eta_i.$$

Here $c > 0$ is a universal constant.

Proof of Lemma C.1. See Appendix C.3. □

For every $f \in \mathcal{F}$, define $\ell_f : \mathcal{X} \times \mathbb{R} \rightarrow \mathbb{R}$ by $\ell_f((\mathbf{x}, y)) = [f(\mathbf{x}) - y]^2$. We also denote (\mathbf{x}, y) by \mathbf{z} . In Lemma C.1, take $\{\mathbf{z}_i\}_{i=1}^n = \mathcal{B}_{t,k}^{\text{tr}}$ and $\mathcal{G} = \{\ell_f - \ell_{\bar{f}} : f \in \mathcal{F}\}$. The following lemma suggests a choice of g_i and T for which (C.1) holds.

Lemma C.2 (Noise condition). Let Assumption 3.1 hold. For all $f, \bar{f} \in \mathcal{F}$,

$$\begin{aligned} \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \text{var}_{\mathbf{z} \sim P_j} [\ell_f(\mathbf{z}) - \ell_{\bar{f}}(\mathbf{z})] &\leq 16M^2 \cdot \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \|f - \bar{f}\|_j^2 \\ &\leq 32M^2 \cdot \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \{ \mathbb{E}_{\mathbf{z} \sim P_j} [\ell_f(\mathbf{z}) - \ell_{\bar{f}}(\mathbf{z})] + 2\mathcal{E}_j(\bar{f}) \}. \end{aligned}$$

Proof of Lemma C.2. See Appendix C.4. □

Define

$$T(\ell_f - \ell_{\bar{f}_t}) = 16M^2 \cdot \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \|f - \bar{f}_t\|_j^2, \quad \forall f \in \mathcal{F}.$$

By Lemma C.2, the noise condition (C.1) holds with $C = 32M^2$, and Lemma C.1 is applicable. Moreover, for all $r \geq 0$,

$$\begin{aligned} & \mathfrak{R}(\{\ell_f - \ell_{\bar{f}_t} : f \in \mathcal{F}, T(\ell_f - \ell_{\bar{f}_t}) \leq r\}; \mathcal{D}_{t,k}^{\text{tr}}) \\ &= \mathbb{E} \sup \left\{ \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} \sum_{i=1}^{m_j} \varepsilon_{j,i} [f(\mathbf{x}_{j,i}) - y_{j,i}]^2 : f \in \mathcal{F}, T(\ell_f - \ell_{\bar{f}_t}) \leq r \right\} \\ &\leq 4M \cdot \mathbb{E} \sup \left\{ \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} \sum_{i=1}^{m_j} \varepsilon_{j,i} [f(\mathbf{x}_{j,i}) - y_{j,i}] : f \in \mathcal{F}, T(\ell_f - \ell_{\bar{f}_t}) \leq r \right\} \\ &= 4M \cdot \mathbb{E} \sup \left\{ \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} \sum_{i=1}^{m_j} \varepsilon_{j,i} f(\mathbf{x}_{j,i}) : f \in \mathcal{F}, T(\ell_f - \ell_{\bar{f}_t}) \leq r \right\} \\ &= 4M \cdot \mathfrak{R}(\{f \in \mathcal{F} : T(\ell_f - \ell_{\bar{f}_t}) \leq r\}; \mathcal{D}_{t,k}^{\text{tr}}), \end{aligned}$$

where $\{\varepsilon_{j,i}\}$ are i.i.d. Rademacher random variables, and the inequality is due to Theorem A.6 in Bartlett et al. (2005).

Define

$$L_{t,k}^{\text{tr}}(f) = \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j L_j(f).$$

Applying Lemma C.1 with $g = \hat{f}_{t,k} - \bar{f}_t$, we obtain that if $\tilde{\psi} : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a sub-root function with fixed point \tilde{r} satisfying

$$\tilde{\psi}(r) \geq 128M^3 \cdot \mathfrak{R}(\{f \in \mathcal{F} : T(\ell_f - \ell_{\bar{f}_t}) \leq r\}; \mathcal{D}_{t,k}^{\text{tr}}), \quad \forall r \geq \tilde{r}, \quad (\text{C.2})$$

then with probability at least $1 - \delta$,

$$\begin{aligned} L_{t,k}^{\text{tr}}(\hat{f}_{t,k}) - L_{t,k}^{\text{tr}}(\bar{f}_t) &\lesssim \left(\hat{L}_{t,k}^{\text{tr}}(\hat{f}_{t,k}) - \hat{L}_{t,k}^{\text{tr}}(\bar{f}_t) \right) + \left[\frac{\tilde{r}}{M^2} + \frac{M^2 \log(1/\delta)}{m_{t,k}} \right] + \max_{t-k \leq j \leq t-1} \mathcal{E}_j(\bar{f}_t) \\ &\lesssim \left[\frac{\tilde{r}}{M^2} + \frac{M^2 \log(1/\delta)}{m_{t,k}} \right] + \max_{t-k \leq j \leq t-1} \mathcal{E}_j(\bar{f}_t), \end{aligned} \quad (\text{C.3})$$

where \lesssim hides a universal constant, and the second inequality is due to $\hat{L}_{t,k}^{\text{tr}}(\hat{f}_{t,k}) \leq \hat{L}_{t,k}^{\text{tr}}(\bar{f}_t)$.

It remains to express \tilde{r} in terms of $r_{t,k}$, and convert (C.3) into a bound for $L_t(\hat{f}_{t,k})$. We work

on \tilde{r} first. Take a subroot function $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with fixed point $r_{t,k}$ such that

$$\psi(r) \geq \mathfrak{R} \left(\left\{ f \in \mathcal{F} : \frac{1}{n_{t,k}} \sum_{j=t-k}^{t-1} n_j \|f - \bar{f}_t\|_j^2 \leq r \right\} \right), \quad \forall r \geq r_{t,k}.$$

We now show that $\tilde{\psi}(r) = 128M^3 \psi\left(\frac{r}{16M^2}\right)$ satisfies (C.2). By Lemma E.1, the fixed point \tilde{r} of $\tilde{\psi}$ satisfies

$$\min\{1, 8M\}^2 16M^2 r_{t,k} \leq \tilde{r} \leq \max\{1, 8M\}^2 16M^2 r_{t,k}. \quad (\text{C.4})$$

For all $r \geq \tilde{r}$, since $r \geq r_{t,k}$, then

$$\tilde{\psi}(r) \geq 128M^3 \cdot \mathfrak{R} \left(\left\{ f \in \mathcal{F} : 16M^2 \cdot \frac{1}{n_{t,k}} \sum_{j=t-k}^{t-1} n_j \|f - \bar{f}_t\|_j^2 \leq r \right\} \right),$$

so (C.2) holds for this choice of $\tilde{\psi}$, and (C.3) becomes

$$L_{t,k}^{\text{tr}}(\hat{f}_{t,k}) - L_{t,k}^{\text{tr}}(\bar{f}_t) \lesssim M^2 \left[r_{t,k} + \frac{\log(1 + \delta^{-1})}{m_{t,k}} \right] + \max_{t-k \leq j \leq t-1} \mathcal{E}_j(\bar{f}_t). \quad (\text{C.5})$$

Finally, we will convert (C.5) into a bound for $L_t(\hat{f}_{t,k})$. We invoke the following lemma.

Lemma C.3. *For all $f \in \mathcal{F}$ and $j, t \in \mathbb{Z}_+$,*

$$\begin{aligned} |L_j(f) - L_t(f)| &\leq 4M^2 \text{TV}(P_j, P_t), \\ |\mathcal{E}_j(f) - \mathcal{E}_t(f)| &\leq 4M^2 \text{TV}(P_j, P_t). \end{aligned}$$

Proof of Lemma C.3. For every $f \in \mathcal{F}$,

$$|L_j(f) - L_t(f)| = \left| \mathbb{E}_{(x,y) \sim P_j} \left\{ [f(x) - y]^2 \right\} - \mathbb{E}_{(x,y) \sim P_t} \left\{ [f(x) - y]^2 \right\} \right| \leq 4M^2 \cdot \text{TV}(P_j, P_t).$$

To prove the second inequality, since $L_t(f_t^*) \leq L_t(f_j^*)$, then

$$\begin{aligned} \mathcal{E}_j(f) - \mathcal{E}_t(f) &= [L_j(f) - L_j(f_j^*)] - [L_t(f) - L_t(f_t^*)] \\ &\leq [L_j(f) - L_j(f_j^*)] - [L_t(f) - L_t(f_j^*)] \\ &= \mathbb{E}_{(x,y) \sim P_j} \left\{ [f(x) - y]^2 - [f_j^*(x) - y]^2 \right\} - \mathbb{E}_{(x,y) \sim P_t} \left\{ [f(x) - y]^2 - [f_j^*(x) - y]^2 \right\} \\ &\leq 4M^2 \cdot \text{TV}(P_j, P_t). \end{aligned}$$

By symmetry, $\mathcal{E}_t(f) - \mathcal{E}_j(f) \leq M^2 \cdot \text{TV}(P_j, P_t)$, so

$$\max_{t-k \leq j \leq t-1} |\mathcal{E}_j(f) - \mathcal{E}_t(f)| \leq 4M^2 \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t).$$

This finishes the proof. □

Since

$$|L_{t,k}^{\text{tr}}(f) - L_t(f)| \leq \max_{t-k \leq j \leq t-1} |L_j(f) - L_t(f)|,$$

then substituting Lemma C.3 into (C.5) yields

$$L_t(\hat{f}_{t,k}) - L_t(\bar{f}_t) \lesssim M^2 \left[r_{t,k} + \frac{\log(1 + \delta^{-1})}{m_{t,k}} \right] + \mathcal{E}_t(\bar{f}_t) + M^2 \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t).$$

Since $\mathcal{E}_t(f) = L_t(f) - L_t(f_t^*)$, then

$$\mathcal{E}(\hat{f}_{t,k}) \lesssim \mathcal{E}_t(\bar{f}_t) + M^2 \left[r_{t,k} + \frac{\log(1 + \delta^{-1})}{m_{t,k}} \right] + M^2 \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t).$$

This completes the proof.

C.3 Proof of Lemma C.1

The core techniques are the same as those of Theorem 3.3 in Bartlett et al. (2005), with small changes in the quantities to bound. For $r, \lambda > 0$, let

$$w(g) = \min\{r\lambda^k : k \in \{0\} \cup \mathbb{Z}_+, r\lambda^k \geq T(g)\}, \quad \mathcal{G}_r = \left\{ \frac{r}{w(g)}g : g \in \mathcal{G} \right\},$$

and

$$V_r^+ = \sup_{g \in \mathcal{G}} \left\{ \frac{r}{w(g)} \cdot \frac{1}{n} \sum_{i=1}^n (\mathbb{E}[g(\mathbf{z}_i)] - g(\mathbf{z}_i)) \right\}.$$

Similar to Lemma 3.8 of Bartlett et al. (2005), for every $K > 1$ and $g \in \mathcal{G}$,

$$V_r^+ \leq \frac{r}{\lambda CK} \quad \text{implies} \quad \frac{1}{n} \sum_{i=1}^n \mathbb{E}[g(\mathbf{z}_i)] \leq \frac{K}{K-1} \cdot \frac{1}{n} \sum_{i=1}^n g(\mathbf{z}_i) + \frac{r}{\lambda CK} + \frac{1}{K-1} \cdot \frac{1}{n} \sum_{i=1}^n \eta_i.$$

We now invoke a uniform convergence result to give a bound for V_r^+ . It is a non-stationary version of Theorem 2.1 in Bartlett et al. (2005).

Lemma C.4 (Uniform concentration). *Consider the setting of Lemma C.1. Define*

$$v = \frac{1}{n} \sup_{g \in \mathcal{G}} \sum_{i=1}^n \text{var}[g(\mathbf{z}_i)].$$

Let $\delta \in (0, 1)$. With probability at least $1 - \delta$,

$$\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n (\mathbb{E}[g(\mathbf{z}_i)] - g(\mathbf{z}_i))$$

$$\lesssim \inf_{\alpha>0} \left\{ (1+\alpha)\mathfrak{R}(\mathcal{G}; \{\mathbf{z}_i\}_{i=1}^n) + \sqrt{\frac{v \log(1/\delta)}{n}} + (1+\alpha^{-1}) \frac{(b-a) \log(1/\delta)}{n} \right\},$$

where \lesssim hides a universal constant.

Proof of Lemma C.4. Let $Z = \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n [g(\mathbf{z}_i) - \mathbb{E}g(\mathbf{z}_i)]$. Adapting the proof of Theorem 4 in Massart (2000), with probability at least $1 - \delta$,

$$Z - \mathbb{E}Z \lesssim \sqrt{\frac{\log(1/\delta)}{n}} (v + (b-a)\mathbb{E}Z) + \frac{(b-a) \log(1/\delta)}{n},$$

where \lesssim hides a universal constant. By Young's inequality, for every $\alpha > 0$,

$$\sqrt{\frac{\log(1/\delta)}{n}} (b-a)\mathbb{E}Z \leq \alpha \mathbb{E}Z + \alpha^{-1} \frac{(b-a) \log(1/\delta)}{n}.$$

By Rademacher symmetrization (e.g., Lemma A.5 in Bartlett et al. (2005)), for i.i.d. Rademacher random variables $\varepsilon_1, \dots, \varepsilon_n$ independent of $\mathbf{z}_1, \dots, \mathbf{z}_n$, it holds that

$$\mathbb{E}Z \leq 2\mathbb{E} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(\mathbf{z}_i) \right] = 2\mathfrak{R}(\mathcal{G}; \{\mathbf{z}_i\}_{i=1}^n).$$

This completes the proof. \square

By Lemma C.4, each of the following inequalities holds with probability at least $1 - \delta$:

$$V_r^+ \lesssim \inf_{\alpha>0} \left\{ (1+\alpha)\mathfrak{R}(\mathcal{G}_r; \{\mathbf{z}_i\}_{i=1}^n) + \sqrt{\frac{v \log(1/\delta)}{n}} + (1+\alpha^{-1}) \frac{(b-a) \log(1/\delta)}{n} \right\},$$

The rest of the proof follows that in Section 3.2 of Bartlett et al. (2005).

C.4 Proof of Lemma C.2

Recall that for $\mathbf{z} = (\mathbf{x}, y)$, we define $\ell_f(\mathbf{z}) = [f(\mathbf{x}) - y]^2$. Then

$$\begin{aligned} & \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \text{var}_{\mathbf{z} \sim P_j} [\ell_f(\mathbf{z}) - \ell_{\bar{f}}(\mathbf{z})] \\ &= \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \text{var}_{(\mathbf{x}, y) \sim P_j} \left[(f(\mathbf{x}) - y)^2 - (\bar{f}(\mathbf{x}) - y)^2 \right] \\ &\leq \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \mathbb{E}_{(\mathbf{x}, y) \sim P_j} \left\{ \left[(f(\mathbf{x}) - y)^2 - (\bar{f}(\mathbf{x}) - y)^2 \right]^2 \right\} \\ &= \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \mathbb{E}_{(\mathbf{x}, y) \sim P_j} \left[(f(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 (f(\mathbf{x}) + \bar{f}(\mathbf{x}) - 2y)^2 \right] \end{aligned}$$

$$\begin{aligned}
&\leq 16M^2 \cdot \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \mathbb{E}_{(\mathbf{x}, y) \sim P_j} \left[(f(\mathbf{x}) - \bar{f}(\mathbf{x}))^2 \right] = 16M^2 \cdot \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \|f - \bar{f}\|_j^2 \\
&\leq 32M^2 \cdot \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j (\|f - f_j^*\|_j^2 + \|\bar{f} - f_j^*\|_j^2) \\
&= 32M^2 \cdot \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} m_j \left\{ \mathbb{E}_{\mathbf{z} \sim P_j} [\ell_f(\mathbf{z}) - \ell_{\bar{f}}(\mathbf{z})] + 2\mathcal{E}_j(\bar{f}) \right\}.
\end{aligned}$$

This finishes the proof.

C.5 Proofs for Example 3.1, Example 3.2 and Example 3.3

Example 3.1 follows from the result below. It is an immediate extension of Proposition 6.1 and Lemma D.1 in Duan et al. (2021) to independent samples with non-identical distributions. The proof is omitted.

Lemma C.5. *Let $\mathbf{z}_1, \dots, \mathbf{z}_n$ be independent random variables taking values in \mathcal{X} , and let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. Rademacher random variables independent of $\{\mathbf{z}_i\}_{i=1}^n$. Let \mathcal{F} be a finite class of functions from \mathcal{X} to $[-M, M]$. Take an arbitrary function $\bar{f} : \mathcal{X} \rightarrow [-M, M]$. For $r \geq 0$, define*

$$\mathcal{R}(r) = \mathbb{E} \max \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(\mathbf{z}_i) : f \in \mathcal{F}, \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[(f(\mathbf{z}_i) - \bar{f}(\mathbf{z}_i))^2 \right] \leq r \right\}.$$

Then for every $r \geq 0$,

$$\mathcal{R}(r) \leq 2 \max \left\{ \sqrt{\frac{r \log |\mathcal{F}|}{n}}, \frac{2M \log |\mathcal{F}|}{n} \right\}.$$

Moreover, the function on the right hand side is subroot, and has a unique fixed point

$$r = \frac{4 \max\{M, 1\} \log |\mathcal{F}|}{n}.$$

To obtain the results in Example 3.2 and Example 3.3, we invoke a useful lemma. It is an extension of Theorem 41 in Mendelson (2002) to the non-i.i.d. case. The proof is omitted.

Lemma C.6. *Take $\bar{\boldsymbol{\theta}} \in B(\mathbf{0}, \sqrt{M}) \subseteq \mathbb{R}^d$. Suppose $\mathbf{x}_i \sim Q_i$, $i \in [n]$ are independent random vectors in \mathbb{R}^d . Let $\varepsilon_1, \dots, \varepsilon_n$ be i.i.d. Rademacher random variables independent of $\{\mathbf{x}_i\}_{i=1}^n$. For $r \geq 0$, define*

$$\mathcal{R}(r) = \mathbb{E} \sup \left\{ \frac{1}{n} \sum_{i=1}^n \varepsilon_i \mathbf{x}_i^\top \boldsymbol{\theta} : \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[(\mathbf{x}_i^\top (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}))^2 \right] \leq r, \|\boldsymbol{\theta}\|_2^2 \leq M \right\}.$$

Let $\Sigma = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(\mathbf{x}_i \mathbf{x}_i^\top)$, and denote by $\{\lambda_i\}_{i=1}^d$ its eigenvalues sorted in descending order. Then

$$\mathcal{R}(r) \leq \sqrt{\frac{c}{n} \sum_{i=1}^d \min\{r, M\lambda_i\}}.$$

The above lemma leads to $\mathcal{R}(r) \leq \sqrt{crd/n}$. The right-hand side is subroot and has a unique fixed point $r = cd/n$. This verifies Example 3.2.

To get Example 3.3, we apply Lemma C.6 to the transformed features $\phi(\mathbf{x}_{j,i})$ rather than the raw ones. Correspondingly, the feature space becomes \mathbb{H} , and the matrix Σ in Example 3.3 becomes $\frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} \sum_{i=1}^{m_j} \mathbb{E}[\phi(\mathbf{x}_{j,i}) \otimes \phi(\mathbf{x}_{j,i})]$. Here \otimes denotes the tensor product.

In Example 3.3, the assumption regarding the trace-class operator \mathbf{S} forces $\Sigma \preceq \mathbf{S}$. Then, Lemma C.6 yields

$$\mathcal{R}(r) \lesssim \sqrt{\frac{1}{m_{t,k}} \sum_{i=1}^{\infty} \min\{r, \mu_i\}} \leq \frac{1}{\sqrt{m_{t,k}}} \min_{s \geq 1} \sqrt{sr + \sum_{i=s+1}^{\infty} \mu_i},$$

where \lesssim hides a constant factor.

- If there are constants $c_1, c_2 > 0$ such that $\mu_k \leq c_1 e^{-c_2 k}$ holds for all k , then $\sum_{i=s+1}^{\infty} \mu_i \lesssim e^{-c_2 s}$. Taking $s = \lceil c_2^{-1} \log(1/r) \rceil$, we get

$$\mathcal{R}(r) \lesssim \frac{1}{\sqrt{m_{t,k}}} \sqrt{r \lceil c_2^{-1} \log(1/r) + 1 \rceil}.$$

The right-hand side is sub-root. Elementary calculation yields $r_{t,k}(\mathcal{F}) \lesssim (\log m_{t,k})/m_{t,k}$.

- If there are constants $c > 0$ and $\alpha \geq 1$ such that $\mu_k \leq ck^{-2\alpha}$ holds for all k , then $\sum_{i=s+1}^{\infty} \mu_i \lesssim s^{1-2\alpha}$. Taking $s = \lceil r^{-1/(2\alpha)} \rceil$, we get

$$\mathcal{R}(r) \lesssim \sqrt{\frac{r^{1-1/(2\alpha)}}{m_{t,k}}}.$$

The right-hand side is sub-root. Then, we can easily get $r_{t,k}(\mathcal{F}) \lesssim m_{t,k}^{-\frac{2\alpha}{2\alpha+1}}$.

D Proofs for Section 4 and Appendix B

D.1 Proof of Lemma 4.1

Given two models $f, f' \in \mathcal{F}$, denote the output of \mathcal{A} by $\mathcal{A}(f, f') \in \{f, f'\}$. Let $T(\Lambda)$ be the maximum expected number of times Algorithm 1 can call \mathcal{A} after there are Λ remaining models $\{f_\lambda\}_{\lambda=1}^\Lambda$ at the end of a while loop, where the maximum is taken over all possible choices of $\{f_\lambda\}_{\lambda=1}^\Lambda$. Then $T(\Lambda)$ is increasing in Λ , and $T(\Lambda) \leq \Lambda^2/2$. Let N denote the number of remaining models at

the end of the next while loop. Since that while loop calls \mathcal{A} at most $\Lambda - 1$ times, then

$$T(\Lambda) \leq (\Lambda - 1) + \mathbb{E}[T(N)]. \quad (\text{D.1})$$

For each $\lambda \in [\Lambda]$, let $n_\lambda = |\{\lambda' \in [\Lambda] \setminus \{\lambda\} : \mathcal{A}(f_\lambda, f_{\lambda'}) = f_{\lambda'}\}|$ be the number of remaining models that would beat f_λ if they were paired. Since Algorithm 1 chooses each f_λ as the pivot model uniformly at random, then

$$\mathbb{E}[T(N)] = \mathbb{E}\left[\frac{1}{\Lambda} \sum_{\lambda=1}^{\Lambda} T(n_\lambda)\right]. \quad (\text{D.2})$$

Since $\sum_{\lambda=1}^{\Lambda} n_\lambda$ counts exactly one of $(f_\lambda, f_{\lambda'})$ and $(f_{\lambda'}, f_\lambda)$ for all $\lambda \neq \lambda'$, then $\sum_{\lambda=1}^{\Lambda} n_\lambda = \Lambda(\Lambda-1)/2$. Let $n_{(1)} \leq \dots \leq n_{(\Lambda)}$ be the order statistics of n_1, \dots, n_Λ . Then for all $i = 1, \dots, \lceil \Lambda/3 \rceil$,

$$n_{(i)} \leq n_{(\lceil \Lambda/3 \rceil)} \leq \frac{1}{\Lambda - \lceil \Lambda/3 \rceil + 1} \sum_{i=\lceil \Lambda/3 \rceil}^{\Lambda} n_i \leq \frac{3}{2\Lambda} \cdot \frac{\Lambda(\Lambda-1)}{2} = \frac{3(\Lambda-1)}{4}.$$

By the monotonicity of T ,

$$\begin{aligned} \frac{1}{\Lambda} \sum_{\lambda=1}^{\Lambda} T(n_\lambda) &= \frac{1}{\Lambda} \sum_{i=1}^{\Lambda} T(n_{(i)}) \leq \frac{1}{\Lambda} \left(\lfloor \Lambda/3 \rfloor T(\lceil 3\Lambda/4 \rceil) + (\Lambda - \lfloor \Lambda/3 \rfloor) T(\Lambda-1) \right) \\ &\leq \frac{1}{3} T(\lceil 3\Lambda/4 \rceil) + \left(\frac{2}{3} + \frac{1}{\Lambda} \right) T(\Lambda) \\ &\leq \frac{1}{3} T(\lceil 3\Lambda/4 \rceil) + \frac{2}{3} T(\Lambda) + \Lambda. \end{aligned}$$

Substituting this into (D.2) and (D.1) gives

$$T(\Lambda) \leq 6\Lambda + T(\lceil 3\Lambda/4 \rceil).$$

By the Master Theorem (Theorem 4.1) in [Cormen et al. \(2022\)](#), we conclude that $T(\Lambda) = \Theta(\Lambda)$.

D.2 Proof of Theorem 4.1

We will prove the following stronger result: with probability at least $1 - \delta$,

$$\sqrt{\mathcal{E}_t(\hat{f})} - \sqrt{\mathcal{E}_t(f_2)} \lesssim M \sqrt{\log(t/\delta)} \cdot \left(\max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}} \right)^{1/2}. \quad (\text{D.3})$$

The bound in Theorem 4.1 is obtained by squaring both sides of (D.3).

Without loss of generality, assume $L_t(f_1) \geq L_t(f_2)$, so $\min_{\lambda \in [2]} L_t(f_\lambda) = L_t(f_2)$. By Theorem 4.2 in [Han et al. \(2024\)](#), with probability at least $1 - \delta$, Algorithm 2 outputs a model \hat{f} satisfying

$$L_t(\hat{f}) - L_t(f_2) \leq |\hat{\Delta}_{t,\ell} - \Delta_t|$$

$$\lesssim \min_{\ell \in [t-1]} \left\{ \sqrt{\log(t/\delta)} \cdot \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t| + \sigma_{t,\ell} \sqrt{\frac{\log(t/\delta)}{n_{t,\ell}}} + \frac{M^2 \log(t/\delta)}{n_{t,\ell}} \right\},$$

where \lesssim hides a universal constant. When this event happens, it holds for all $\ell \in [t-1]$ that

$$L_t(\hat{f}) - L_t(f_2) \lesssim \sqrt{\log(t/\delta)} \cdot \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t| + \sigma_{t,\ell} \sqrt{\frac{\log(t/\delta)}{n_{t,\ell}}} + \frac{M^2 \log(t/\delta)}{n_{t,\ell}}.$$

When $\hat{f} = f_2$, (D.3) automatically holds and there is nothing to prove.

Now consider the case when $\hat{f} = f_1$. Fix $\ell \in [t-1]$. Define

$$\mathcal{E}_{t,\ell}^{\text{va}}(f) = \frac{1}{n_{t,\ell}} \sum_{j=t-\ell}^{t-1} n_j [L_j(f) - L_j(f_j^*)].$$

Since

$$|[\mathcal{E}_{t,\ell}^{\text{va}}(f_1) - \mathcal{E}_{t,\ell}^{\text{va}}(f_2)] - [L_t(f_1) - L_t(f_2)]| \leq \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t|,$$

then

$$\begin{aligned} \mathcal{E}_{t,\ell}^{\text{va}}(f_1) - \mathcal{E}_{t,\ell}^{\text{va}}(f_2) &\leq L_t(f_1) - L_t(f_2) + \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t| \\ &\leq C \left[\sqrt{\log(t/\delta)} \cdot \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t| + \sigma_{t,\ell} \sqrt{\frac{\log(t/\delta)}{n_{t,\ell}}} + \frac{M^2 \log(t/\delta)}{n_{t,\ell}} \right] \end{aligned} \quad (\text{D.4})$$

for some universal constant $C \geq 1$. The variance term $\sigma_{t,\ell}^2$ can be bounded by

$$\begin{aligned} \sigma_{t,\ell}^2 &= \frac{1}{n_{t,\ell}} \sum_{j=t-\ell}^{t-1} n_j \text{var}_{(\mathbf{x},y) \sim P_j} \left[(f_1(\mathbf{x}) - y)^2 - (f_2(\mathbf{x}) - y)^2 \right] \\ &\leq 2 \sum_{\lambda=1}^2 \left[\frac{1}{n_{t,\ell}} \sum_{j=t-\ell}^{t-1} n_j \text{var}_{(\mathbf{x},y) \sim P_j} \left[(f_\lambda(\mathbf{x}) - y)^2 - (f_j^*(\mathbf{x}) - y)^2 \right] \right]. \end{aligned}$$

For each $\lambda \in [2]$ and $j \in \mathbb{Z}_+$,

$$\begin{aligned} \text{var}_{(\mathbf{x},y) \sim P_j} \left[(f_\lambda(\mathbf{x}) - y)^2 - (f_j^*(\mathbf{x}) - y)^2 \right] &\leq \mathbb{E}_{(\mathbf{x},y) \sim P_j} \left\{ \left[(f_\lambda(\mathbf{x}) - y)^2 - (f_j^*(\mathbf{x}) - y)^2 \right]^2 \right\} \\ &= \mathbb{E}_{(\mathbf{x},y) \sim P_j} \left[(f_\lambda(\mathbf{x}) - f_j^*(\mathbf{x}))^2 (f_\lambda + f_j^*(\mathbf{x}) - 2y)^2 \right] \\ &\leq 16M^2 \mathbb{E}_{(\mathbf{x},y) \sim P_j} \left[(f_\lambda(\mathbf{x}) - f_j^*(\mathbf{x}))^2 \right] \\ &= 16M^2 \mathcal{E}_j(f_\lambda). \end{aligned}$$

Thus,

$$\sigma_{t,\ell}^2 \leq 32M^2 \cdot \sum_{\lambda=1}^2 \left[\frac{1}{n_{t,\ell}} \sum_{j=t-\ell}^{t-1} n_j \mathcal{E}_j(f_\lambda) \right] = 32M^2 [\mathcal{E}_{t,\ell}^{\text{va}}(f_1) + \mathcal{E}_{t,\ell}^{\text{va}}(f_2)]. \quad (\text{D.5})$$

Substituting (D.5) into (D.4) yields

$$\mathcal{E}_{t,\ell}^{\text{va}}(f_1) - \mathcal{E}_{t,\ell}^{\text{va}}(f_2) \leq 2A \left(\sqrt{\mathcal{E}_{t,\ell}^{\text{va}}(f_1)} + \sqrt{\mathcal{E}_{t,\ell}^{\text{va}}(f_2)} \right) + D,$$

where

$$A = 2\sqrt{2}CM \sqrt{\frac{\log(t/\delta)}{n_{t,\ell}}} \quad \text{and} \quad D = C \left[\sqrt{\log(t/\delta)} \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t| + \frac{M^2 \log(t/\delta)}{n_{t,\ell}} \right]$$

Completing the squares gives

$$\left(\sqrt{\mathcal{E}_{t,\ell}^{\text{va}}(f_1)} - A \right)^2 \leq \left(\sqrt{\mathcal{E}_{t,\ell}^{\text{va}}(f_2)} + A \right)^2 + D,$$

which implies

$$\begin{aligned} \sqrt{\mathcal{E}_{t,\ell}^{\text{va}}(f_1)} - \sqrt{\mathcal{E}_{t,\ell}^{\text{va}}(f_2)} &\leq 2A + \sqrt{D} \\ &\lesssim M \sqrt{\frac{\log(t/\delta)}{n_{t,\ell}}} + \left[\sqrt{\log(t/\delta)} \max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t| + \frac{M^2 \log(t/\delta)}{n_{t,\ell}} \right]^{1/2} \\ &\lesssim \sqrt{\log(t/\delta)} \cdot \left(\max_{t-\ell \leq j \leq t-1} |\Delta_j - \Delta_t| + \frac{M^2}{n_{t,\ell}} \right)^{1/2} \\ &\lesssim M \sqrt{\log(t/\delta)} \cdot \left(\max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}} \right)^{1/2}, \end{aligned}$$

where the last inequality is due to

$$|\Delta_j - \Delta_t| \lesssim M^2 \cdot \text{TV}(P_j, P_t).$$

Finally, by Lemma C.3, for every $f \in \{f_1, f_2\}$,

$$\left| \sqrt{\mathcal{E}_{t,\ell}^{\text{va}}(f)} - \sqrt{\mathcal{E}_t(f)} \right| \leq \sqrt{|\mathcal{E}_{t,\ell}^{\text{va}}(f) - \mathcal{E}_t(f)|} \leq \max_{t-\ell \leq j \leq t-1} \sqrt{|\mathcal{E}_j(f) - \mathcal{E}_t(f)|} \leq 2M \max_{t-\ell \leq j \leq t-1} \sqrt{\text{TV}(P_j, P_t)},$$

so

$$\begin{aligned} \sqrt{\mathcal{E}_t(f_1)} - \sqrt{\mathcal{E}_t(f_2)} &\leq \sqrt{\mathcal{E}_{t,\ell}^{\text{va}}(f_1)} - \sqrt{\mathcal{E}_{t,\ell}^{\text{va}}(f_2)} + 4M \max_{t-\ell \leq j \leq t-1} \sqrt{\text{TV}(P_j, P_t)} \\ &\lesssim M \sqrt{\log(t/\delta)} \cdot \left(\max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}} \right)^{1/2}. \end{aligned}$$

As $\hat{f} = f_1$, this finishes the proof.

D.3 Proof of Theorem 4.2

We first prove the following lemma, which converts any performance guarantee of the subroutine \mathcal{A} to that of Algorithm 1.

Lemma D.1 (From comparison to selection). *Take a performance metric $\mathcal{L} : \{f_\lambda\}_{\lambda=1}^\Lambda \rightarrow \mathbb{R}$, and $U : (0, 1) \rightarrow \mathbb{R}_+$. Fix $\delta \in (0, 1)$. Suppose that the model comparison subroutine \mathcal{A} in Algorithm 1 satisfies the following property: given two models $h_1, h_2 \in \{f_\lambda\}_{\lambda=1}^\Lambda$, it outputs a model $\hat{h} \in \{h_1, h_2\}$ satisfying*

$$\mathbb{P}\left(\mathcal{L}(\hat{h}) - \min\{\mathcal{L}(h_1), \mathcal{L}(h_2)\} \leq U(\delta)\right) \geq 1 - \delta.$$

Then the output \hat{f} of Algorithm 1 satisfies

$$\mathbb{P}\left(\mathcal{L}(\hat{f}) - \min_{\lambda \in [\Lambda]} \mathcal{L}(f_\lambda) \leq 2U(\delta)\right) \geq 1 - \Lambda^2\delta.$$

Proof of Lemma D.1. Given two models $f, f' \in \mathcal{F}$, denote the output of \mathcal{A} by $\mathcal{A}(f, f') \in \{f, f'\}$. For notational convenience we also set $\mathcal{A}(f, f) = f$ for every $f \in \mathcal{F}$. By a union bound, with probability at least $1 - \Lambda^2\delta$,

$$\mathcal{L}(\mathcal{A}(f_{\lambda'}, f_{\lambda''})) - \min_{\lambda \in \{\lambda', \lambda''\}} \mathcal{L}(f_\lambda) \leq U(\delta), \quad \forall \lambda', \lambda'' \in [\Lambda]. \quad (\text{D.6})$$

From now on suppose that (D.6) holds. Take $\bar{f} \in \{f_\lambda\}_{\lambda=1}^\Lambda$ such that $\mathcal{L}(\bar{f}) = \min_{\lambda \in [\Lambda]} \mathcal{L}(f_\lambda)$.

If Algorithm 1 outputs $\hat{f} = \bar{f}$, then there is nothing to prove. Now assume that $\hat{f} \neq \bar{f}$. Then, there exists $\ell \in \mathbb{Z}_+$ such that at the end of the K -th while loop, \bar{f} is not in S' . Take the smallest such K . Let g_K denote the pivot model f during the K -th while loop, and let S'_K denote the set S' at the end of the K -th while loop. There are two cases.

1. If at the end of the K -th while loop, $S'_K = \emptyset$ and Algorithm 1 outputs \hat{f} , then during this while loop, a call of Algorithm 2 has yielded $\mathcal{A}(\{\hat{f}, \bar{f}\}) = \hat{f}$, so by (D.6), $\mathcal{L}(\hat{f}) - \mathcal{L}(\bar{f}) \leq U(\delta)$.
2. Otherwise, at the end of the K -th while loop, $S'_K \neq \emptyset$. There are two cases.
 - (a) If $g_K = \bar{f}$, then every $f \in S'_K$, a call of \mathcal{A} has yielded $\mathcal{A}(\{\bar{f}, f\}) = f$, so by (D.6), $\mathcal{L}(f) - \mathcal{L}(\bar{f}) \leq U(\delta)$. Since the output \hat{f} must come from S'_K , then automatically $\mathcal{L}(\hat{f}) - \mathcal{L}(\bar{f}) \leq U(\delta)$.
 - (b) If $g_K \neq \bar{f}$, then for every $f \in S'_K$, a call of \mathcal{A} has yielded $\mathcal{A}(\{g_K, f\}) = f$, so by (D.6),

$$\mathcal{L}(f) - \mathcal{L}(g_K) \leq U(\delta). \quad (\text{D.7})$$

Since $\bar{f} \notin S'_K$, then a call of Algorithm 2 has yielded $\mathcal{A}(\{g_K, \bar{f}\}) = g_K$, so by (D.6),

$$\mathcal{L}(g_K) - \mathcal{L}(\bar{f}) \leq U(\delta). \quad (\text{D.8})$$

Putting together (D.7) and (D.8) yields that for all $f \in S'_K$,

$$\mathcal{L}(f) \leq \mathcal{L}(g_K) + U(\delta) \leq \mathcal{L}(\bar{f}) + 2U(\delta).$$

Since the output \hat{f} must come from S'_K , then automatically $\mathcal{L}(\hat{f}) - \mathcal{L}(\bar{f}) \leq 2U(\delta)$.

In all the cases above, we have $\mathcal{L}(\hat{f}) - \mathcal{L}(\bar{f}) \leq 2U(\delta)$. \square

We now prove Theorem 4.2. By the stronger bound (D.3) in the proof of Theorem 4.1, we can set in Lemma D.1 $\mathcal{L}(f) = \sqrt{\mathcal{E}_t(f)}$ and

$$U(\delta) = CM\sqrt{\log(t/\delta)} \cdot \min_{\ell \in [t-1]} \left(\max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}} \right)^{1/2},$$

for some universal constant $C > 0$. Then, by Lemma D.1 and the choice of δ' , with probability $1 - \delta$, the output \hat{f} of Algorithm 1 satisfies

$$\sqrt{\mathcal{E}_t(\hat{f})} - \min_{\lambda \in [\Lambda]} \sqrt{\mathcal{E}_t(f_\lambda)} \lesssim U(\delta/\Lambda^2) \lesssim M\sqrt{\log(\Lambda t/\delta)} \cdot \min_{\ell \in [t-1]} \left(\max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}} \right)^{1/2}.$$

Here \lesssim only hides universal constants. This finishes the proof.

D.4 Proof of Theorem 4.3

Recall that $|\mathcal{D}_j^{\text{tr}}| = m_j$, $|\mathcal{D}_j^{\text{va}}| = n_j$, $|\mathcal{D}_j| = m_j + n_j$, $m_{t,k} = \sum_{j=t-k}^{t-1} m_j$, $n_{t,\ell} = \sum_{j=t-\ell}^{t-1} n_j$, and $B_{t,k} = m_{t,k} + n_{t,k}$. Since there are $(t-1)|\mathcal{F}|$ candidate models, then by Theorem 4.2, with probability at least $1 - \delta/2$, the output \hat{f} of Algorithm 1 satisfies

$$\mathcal{E}_t(\hat{f}) \lesssim \min_{\mathcal{F} \in \mathcal{F}, k \in [t-1]} \mathcal{E}_t(\hat{h}(\mathcal{F}, k)) + M^2 \log(t^2|\mathcal{F}|/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}} \right\}. \quad (\text{D.9})$$

By Theorem 3.1 and a union bound, with probability at least $1 - \delta/2$,

$$\mathcal{E}_t(\hat{h}(\mathcal{F}, k)) \lesssim \min_{f \in \mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{\log(t|\mathcal{F}|/\delta)}{m_{t,k}} \right) + M^2 \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t). \quad (\text{D.10})$$

Combining (D.9) and (D.10) yields that, with probability at least $1 - \delta$,

$$\begin{aligned} \mathcal{E}_t(\hat{f}) &\lesssim \min_{\mathcal{F} \in \mathcal{F}, k \in [t-1]} \left\{ \min_{f \in \mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{\log(t|\mathcal{F}|/\delta)}{m_{t,k}} \right) + M^2 \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t) \right\} \\ &\quad + M^2 \log(t^2|\mathcal{F}|/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{n_{t,\ell}} \right\} \\ &\lesssim \min_{\mathcal{F} \in \mathcal{F}, k \in [t-1]} \left\{ \min_{f \in \mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{\log(t|\mathcal{F}|/\delta)}{B_{t,k}} \right) + M^2 \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t) \right\} \\ &\quad + M^2 \log(t^2|\mathcal{F}|/\delta) \cdot \min_{k \in [t-1]} \left\{ \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t) + \frac{1}{B_{t,k}} \right\} \quad (\text{by Assumption 4.1}) \end{aligned}$$

$$\lesssim \log(t|\mathcal{F}|/\delta) \cdot \min_{\mathcal{F} \in \mathcal{F}, k \in [t-1]} \left\{ \min_{f \in \mathcal{F}} \mathcal{E}_t(f) + M^2 \left(r_{t,k}(\mathcal{F}) + \frac{1}{B_{t,k}} \right) + M^2 \max_{t-k \leq j \leq t-1} \text{TV}(P_j, P_t) \right\}.$$

Here the second inequality hides the constant c in Assumption 4.1. This completes the proof.

D.5 Proof of Lemma B.1

By the triangle inequality,

$$|\hat{\Delta}_{t,\ell}^R - \Delta_t^R| \leq |\hat{\Delta}_{t,\ell}^R - \Delta_{t,\ell}^R| + |\Delta_{t,\ell}^R - \Delta_t^R|, \quad (\text{D.11})$$

where

$$\Delta_{t,\ell}^R = \mathbb{E}[\hat{\Delta}_{t,\ell}^R] = \frac{\Delta_{t,\ell}}{V_{t,\ell}}.$$

By Lemma E.3,

$$|\Delta_{t,\ell}^R - \Delta_t^R| = \left| \frac{\Delta_{t,\ell}}{V_{t,\ell}} - \frac{\Delta_t}{V_t} \right| \leq \max_{t-\ell \leq j \leq t-1} |\Delta_j^R - \Delta_t^R|. \quad (\text{D.12})$$

We have

$$\hat{\Delta}_{t,\ell}^R = \frac{1}{n_{t,\ell}} \sum_{j=t-\ell}^{t-1} \sum_{i=1}^{n_j} \frac{u_{j,i}}{V_{t,\ell}},$$

where $u_{j,i} = [f_1(\mathbf{x}_{j,i}^{\text{va}}) - y_{j,i}^{\text{va}}]^2 - [f_2(\mathbf{x}_{j,i}^{\text{va}}) - y_{j,i}^{\text{va}}]^2$. By Assumptions 3.1 and B.1, $|u_{j,i}/V_{t,\ell}| \leq 8M^2/v$ for all j and i . By Bernstein's concentration inequality (Lemma E.2), with probability at least $1 - \delta$,

$$|\hat{\Delta}_{t,\ell}^R - \Delta_{t,\ell}^R| \leq \sigma_{t,\ell}^R \sqrt{\frac{2 \log(2/\delta)}{n_{t,\ell}}} + \frac{16(M^2/v) \log(2/\delta)}{3n_{t,\ell}}. \quad (\text{D.13})$$

Substituting (D.12) and (D.13) into (D.11) completes the proof.

D.6 Proof of Theorem B.1

We first prove the following theoretical guarantee for the R^2 -based comparison subroutine Algorithm 4.

Theorem D.1 (Near-optimal model comparison with R^2). *Let Assumptions 3.1 and B.1 hold. Choose $\delta \in (0, 1)$ and set $\delta' = 1/(3t)$ in Algorithm 1. With probability at least $1 - \delta$, the output \hat{f} of Algorithm 4 satisfies*

$$\max_{\lambda \in [2]} \tilde{R}_t^2(f_\lambda) - \tilde{R}_t^2(\hat{f}) \lesssim \log(t/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} \max_{\lambda \in [2]} |\tilde{R}_j^2(f_\lambda) - \tilde{R}_t^2(f_\lambda)| + \frac{M^2/v}{\sqrt{n_{t,\ell}}} \right\}. \quad (\text{D.14})$$

Here \lesssim hides a universal constant.

Proof of Theorem D.1. Following the same argument as Theorem 4.2 in Han et al. (2024), we can

show that with probability at least $1 - \delta$,

$$\begin{aligned} \max_{\lambda \in [2]} \tilde{R}_t^2(f_\lambda) - \tilde{R}_t^2(\hat{f}) &\leq |\hat{\Delta}_{t,\hat{\ell}}^R - \Delta_t^R| \\ &\lesssim \log(t/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} |\Delta_j^R - \Delta_t^R| + \frac{\hat{v}_{t,\ell}^R}{\sqrt{n_{t,\ell}}} + \frac{M^2/v}{n_{t,\ell}} \right\}. \end{aligned}$$

We finish the proof by noting that $\sigma_{t,\ell}^R \lesssim M^2/v$ and

$$|\Delta_j^R - \Delta_t^R| = \left| \left[\tilde{R}_j^2(f_1) - \tilde{R}_j^2(f_2) \right] - \left[\tilde{R}_t^2(f_1) - \tilde{R}_t^2(f_2) \right] \right| \leq 2 \max_{\lambda \in [2]} \left| \tilde{R}_j^2(f_\lambda) - \tilde{R}_t^2(f_\lambda) \right|.$$

□

We can now use Lemma D.1 to translate Theorem 4.1 to a theoretical guarantee for general model selection. Set $\mathcal{L}(f) = 1 - \tilde{R}_t^2(f)$ and

$$U(\delta) = C \log(t/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} \max_{\lambda \in [\Lambda]} \left| \tilde{R}_j^2(f_\lambda) - \tilde{R}_t^2(f_\lambda) \right| + \frac{M^2/v}{\sqrt{n_{t,\ell}}} \right\},$$

for a sufficiently large universal constant $C > 0$. Then for any f ,

$$\mathcal{L}(f) - \min_{\lambda \in [\Lambda]} \mathcal{L}(f_\lambda) = \max_{\lambda \in [\Lambda]} \tilde{R}_t^2(f_\lambda) - \tilde{R}_t^2(f).$$

By Lemma D.1 and the choice of δ' , the output \hat{f} of Algorithm 1 satisfies that with probability at least $1 - \delta$,

$$\begin{aligned} \max_{\lambda \in [\Lambda]} \tilde{R}_t^2(f_\lambda) - \tilde{R}_t^2(\hat{f}) &\lesssim U(\delta/\Lambda^2) \\ &\lesssim \log(\Lambda t/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} \max_{\lambda \in [\Lambda]} \left| \tilde{R}_j^2(f_\lambda) - \tilde{R}_t^2(f_\lambda) \right| + \frac{M^2/v}{\sqrt{n_{t,\ell}}} \right\}. \end{aligned}$$

Here \lesssim only hides universal constants.

E Technical Lemmas

Lemma E.1. *Let $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a sub-root function with fixed point $r^* > 0$. For all $A, a > 0$, the function $\tilde{\psi}(r) = A\psi(ar)$ is sub-root and its fixed point \tilde{r} satisfies*

$$\frac{\min\{1, Aa\}^2}{a} r^* \leq \tilde{r} \leq \frac{\max\{1, Aa\}^2}{a} r^*.$$

Proof of Lemma E.1. It is easy to verify that $\tilde{\psi}$ is subroot. We now study \tilde{r} . First consider the case $a = 1$. Since $A\psi(\tilde{r}) = \tilde{r}$, then $\psi(\tilde{r})/\sqrt{\tilde{r}} = \sqrt{\tilde{r}}/A$. There are two cases.

- If $\tilde{r} \geq r^*$, then $\psi(\tilde{r})/\sqrt{\tilde{r}} \leq \psi(r^*)/\sqrt{r^*} = \sqrt{r^*}$, so $\tilde{r} \leq A^2 r^*$.
- If $\tilde{r} \leq r^*$, then $\psi(\tilde{r})/\sqrt{\tilde{r}} \geq \psi(r^*)/\sqrt{r^*} = \sqrt{r^*}$, so $\tilde{r} \geq A^2 r^*$.

Therefore, if $A < 1$, then $A^2 r^* \leq \tilde{r} < r^*$. If $A > 1$, then $r^* < \tilde{r} \leq A^2 r^*$. This shows that

$$\min\{1, A\}^2 r^* \leq \tilde{r} \leq \max\{1, A\}^2 r^*.$$

In the general case of $a > 0$, the function $r \mapsto a^{-1}\psi(ar)$ is sub-root and has fixed point $a^{-1}r^*$. The proof is finished by noting $\tilde{\psi}(r) = (Aa) \cdot a^{-1}\psi(ar)$. \square

Lemma E.2 (Bernstein's concentration inequality). *Let $\{x_i\}_{i=1}^n$ be independent random variables taking values in $[a, b]$ almost surely. Define the average variance $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \text{var}(x_i)$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\left| \frac{1}{n} \sum_{i=1}^n (x_i - \mathbb{E}x_i) \right| \leq \sigma \sqrt{\frac{2 \log(2/\delta)}{n}} + \frac{2(b-a) \log(2/\delta)}{3n}.$$

Proof of Lemma E.2. Inequality (2.10) in [Boucheron et al. \(2013\)](#) implies that for any $t \geq 0$,

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (x_i - \mathbb{E}x_i) > t\right) \leq \exp\left(-\frac{nt^2/2}{\sigma^2 + (b-a)t/3}\right).$$

Fix $\delta \in (0, 1)$. Then,

$$\begin{aligned} & \exp\left(-\frac{nt^2/2}{\sigma^2 + (b-a)t/3}\right) \leq \delta \\ \Leftrightarrow & \frac{nt^2}{2} \geq \sigma^2 \log(1/\delta) + \frac{t(b-a) \log(1/\delta)}{3} \\ \Leftrightarrow & \frac{n}{2} \left(t - \frac{(b-a) \log(1/\delta)}{3n}\right)^2 \geq \sigma^2 \log(1/\delta) + \frac{n}{2} \left(\frac{(b-a) \log(1/\delta)}{3n}\right)^2 \\ \Leftrightarrow & \left(t - \frac{(b-a) \log(1/\delta)}{3n}\right)^2 \geq \left(\sigma \sqrt{\frac{2 \log(1/\delta)}{n}} + \frac{(b-a) \log(1/\delta)}{3n}\right)^2 \\ \Leftrightarrow & t \geq \sigma \sqrt{\frac{2 \log(1/\delta)}{n}} + \frac{2(b-a) \log(1/\delta)}{3n}. \end{aligned}$$

Hence,

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (x_i - \mathbb{E}x_i) > \sigma \sqrt{\frac{2 \log(1/\delta)}{n}} + \frac{2(b-a) \log(1/\delta)}{3n}\right) \leq \delta.$$

Replacing each x_i by $-x_i$ gives bounds on the lower tail and the absolute deviation. \square

Lemma E.3. For all $a, a_1, \dots, a_n \geq 0$ and $b, b_1, \dots, b_n > 0$, it holds that

$$\left| \frac{a}{b} - \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \right| \leq \max_{i \in [n]} \left| \frac{a}{b} - \frac{a_i}{b_i} \right|.$$

Proof of Lemma E.3. This is due to

$$\begin{aligned} \left| \frac{a}{b} - \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \right| &= \left| \frac{a}{b} - \sum_{i=1}^n \frac{b_i}{\sum_{j=1}^n b_j} \cdot \frac{a_i}{b_i} \right| = \left| \sum_{i=1}^n \frac{b_i}{\sum_{j=1}^n b_j} \cdot \left(\frac{a}{b} - \frac{a_i}{b_i} \right) \right| \\ &\leq \sum_{i=1}^n \frac{b_i}{\sum_{j=1}^n b_j} \cdot \left| \frac{a}{b} - \frac{a_i}{b_i} \right| \leq \max_{i \in [n]} \left| \frac{a}{b} - \frac{a_i}{b_i} \right|. \end{aligned}$$

This finishes the proof. □

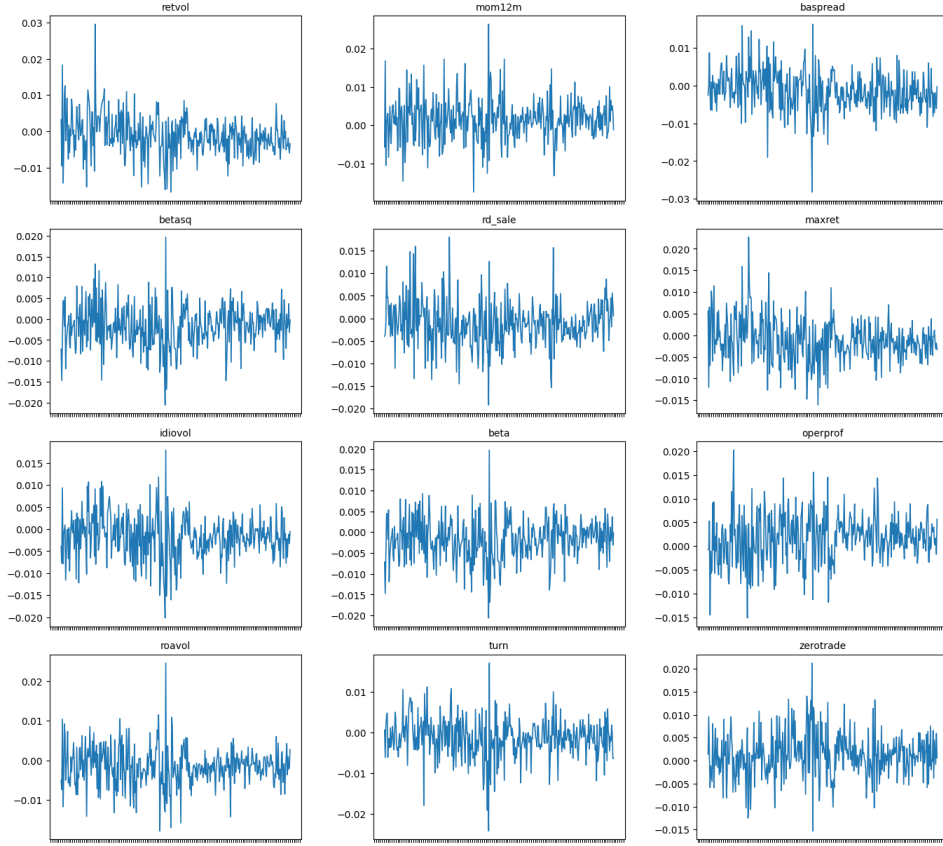
F Additional Experiment Details

F.1 Summary Statistics of the Dataset

We now provide an overview of the long-short firm characteristic covariates used in the analysis, their time-series behavior, and cross-sectional dependence as well as a brief summary of the stochastic discount factor (SDF) and decile portfolios from [Chen et al. \(2024\)](#). Recall that all of the long-short characteristic portfolios are computed at the daily frequency, for the subsequent summary plots, we have aggregated them into the monthly frequency using within-month averages, in line with the standard practice of aligning signals with monthly returns. The monthly aggregation smooths out day-to-day noise and highlights the economically relevant medium-horizon variations.

Monthly Evolution of Covariates. Figure 7 displays the time series of monthly mean values for the twelve most volatile covariates, ranked by their total-sample standard deviation. The figure highlights that variables such as `retvol`, `mom12m`, and `baspread` exhibit pronounced month-to-month fluctuations, while others such as `turn` and `operprof` remain relatively stable. These series reveal persistent heteroskedasticity and regime shifts over time, particularly during market dislocations such as the early 2000s and 2008 crises.

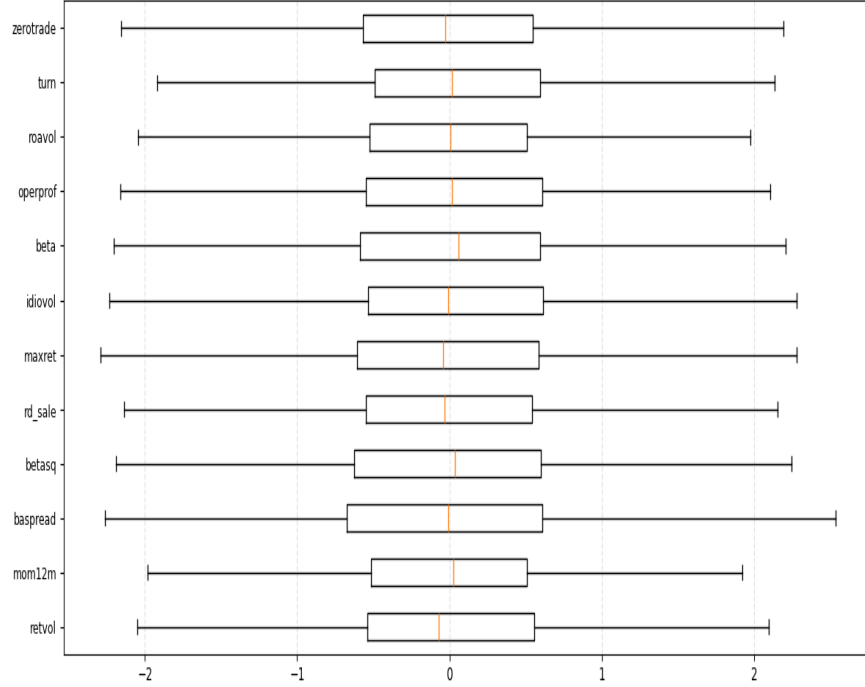
Figure 7: Monthly Means of the 12 Most Volatile Covariates.



Each panel shows the monthly mean of a long–short firm characteristic. Covariates are ranked by total-sample volatility. The time series reveal which characteristics exhibit the greatest month-to-month variation and long-run persistence.

Distributional and Correlation Structure. Figure 8 summarizes the time-series distributions of the same twelve covariates using standardized (z-scored) monthly values. The median, interquartile range, and whiskers capture the magnitude and symmetry of fluctuations across time. Most variables display near-zero median values but differ in dispersion and tail behavior, consistent with heterogeneous economic mechanisms underlying each characteristic.

Figure 8: Distributions (Boxplots) of Standardized Monthly Covariates.



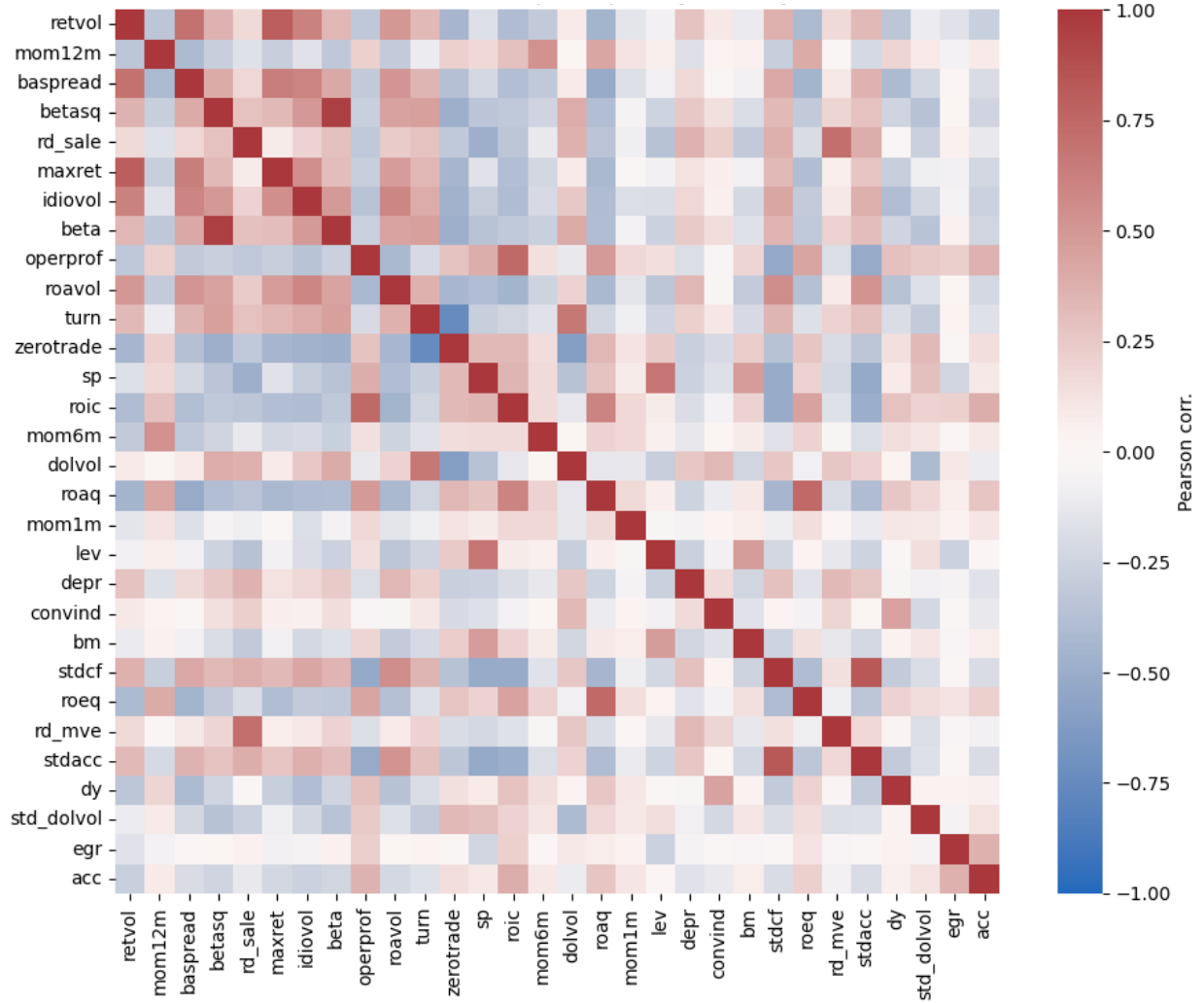
Z-scored monthly series for the twelve most volatile covariates. The figure compares dispersion and tail behavior across characteristics, highlighting differences in amplitude and symmetry.

For example, variables such as **retvol**, **baspread**, and **mom12m** exhibit wide interquartile ranges and thick tails, suggesting that these signals experience substantial time variation and occasional extreme realizations. In contrast, variables such as **turn** and **operprof** have narrower boxes, implying greater stability through time.

Each distribution is constructed by pooling monthly observations over the entire sample period for that specific covariate. This provides a concise view of the temporal heterogeneity and persistence of each characteristic after accounting for scale differences. The figure thus complements the time-series plots in Figure 7 by providing a scale-free summary of long-run variability and skewness in the underlying long-short characteristics.

The pairwise dependence structure among the top thirty covariates (in terms of volatility) is visualized in Figure 9. The heatmap reveals clusters of strongly correlated signals, such as volatility-related measures (**retvol**, **idiovol**, **roavol**) and liquidity-related variables (**baspread**, **zerotrade**, **turn**). The presence of such correlation blocks indicates there could be shared economic channels.

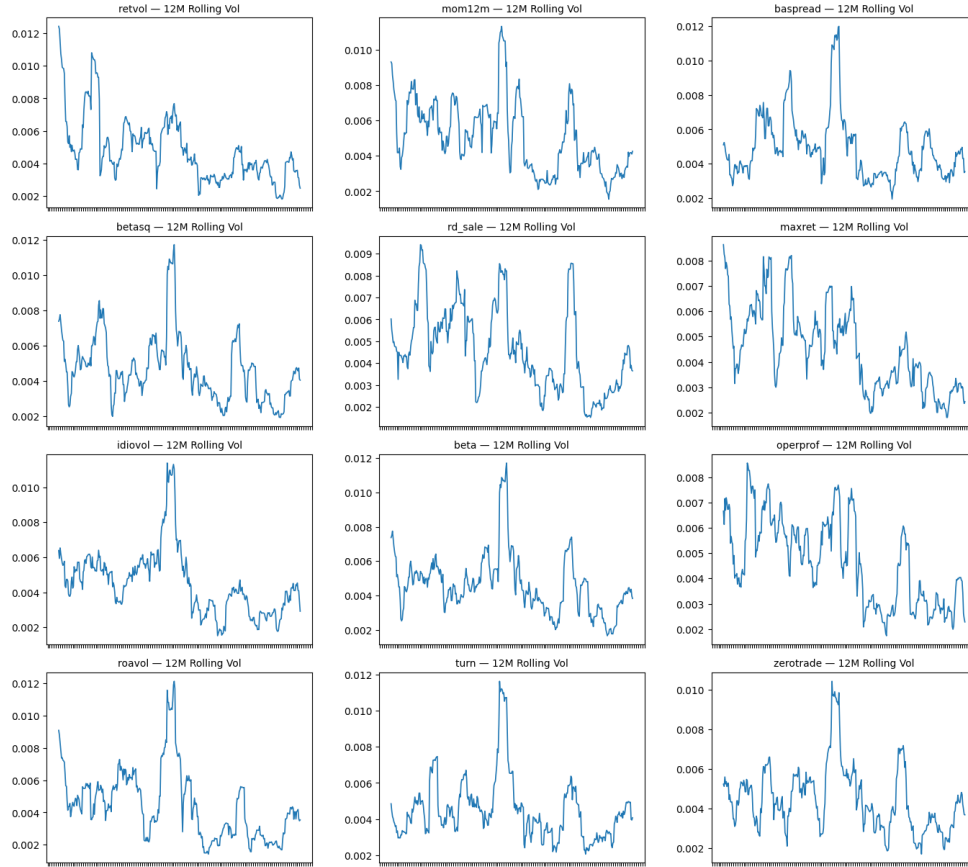
Figure 9: Correlation Heatmap of the 30 Most Volatile Covariates.



The matrix shows pairwise Pearson correlations between the thirty most volatile monthly covariates. Red indicates positive correlation and blue indicates negative correlation. Distinct blocks suggest clusters of related characteristics.

Time-Varying Volatility of Covariates. Figure 10 plots the 12-month rolling standard deviation of the twelve most volatile covariates. Unlike Figure 7, which ranks variables by overall volatility, the rolling volatility tracks how the variability of each covariate evolves through time. Periods such as the dot-com bubble and the global financial crisis correspond to distinct spikes in volatility across multiple signals, indicating that the informational strength and instability of certain factors are regime-dependent.

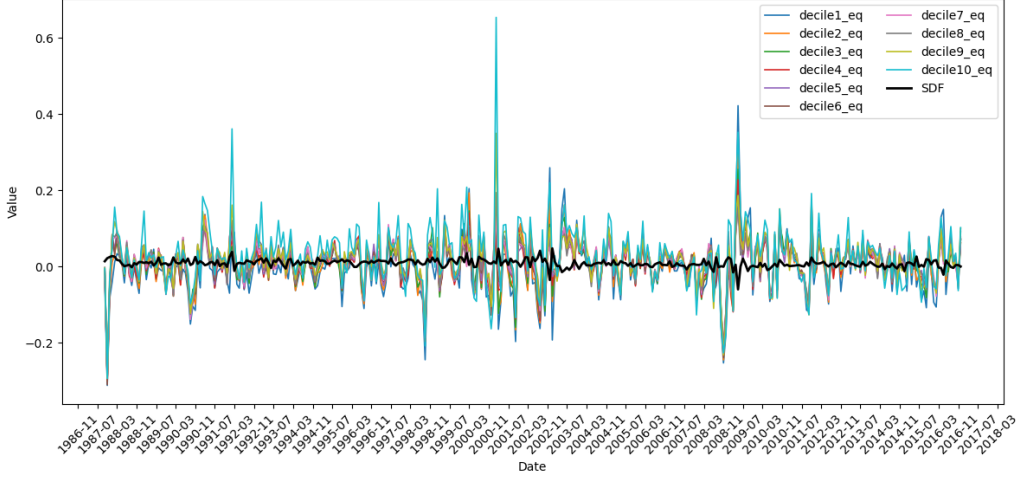
Figure 10: Twelve-Month Rolling Volatility of the Most Volatile Covariates.



The panels show rolling standard deviations computed using a 12-month moving window for each of the twelve most volatile covariates. This highlights temporal variation in the stability and amplitude of the long-short signals.

SDF and Decile Portfolios. Finally, Figure 11 plots the monthly time series of the stochastic discount factor (SDF) alongside the ten equal-weighted decile portfolios sorted by the underlying characteristic. The decile portfolios exhibit substantial comovement, with the SDF (shown in black) fluctuating more smoothly. This figure provides a benchmark for comparing the magnitude and temporal alignment of the SDF with characteristic-sorted portfolio returns, and serves as a diagnostic for whether the constructed SDF captures systematic components of asset pricing variation.

Figure 11: Monthly SDF and Decile Portfolios.



The black line represents the stochastic discount factor (SDF), while the colored lines correspond to ten equal-weighted decile portfolios sorted on firm characteristics. The comovement between the SDF and the characteristic-sorted portfolios provides an initial indication of factor relevance.

F.2 Experiment Details for Section 3.1

In this section, we give more details on our empirical investigations of the nonstationarity-complexity tradeoff in Section 3.1.

The three prediction models, along with their hyperparameters, are: (1) a linear model trained by Ridge regression using the most recent 64 months of data, with $\alpha = 1$, (2) a random forest trained on the most recent 64 months of data, with $n_{\text{tree}} = 200$ and $d_{\text{max}} = 5$, and (3) a random forest trained on all historical data, with $n_{\text{tree}} = 200$ and $d_{\text{max}} = 5$.

In each month t , we construct training data by randomly subsampling 4/5 of the observations \mathcal{D}_j in each previous month $j \in [t - 1]$. The process is repeated 20 times with independent random seeds. We then average the out-of-sample R^2 over the 20 random seeds, which are then used to produce the figures in Section 3.1.

G Additional Experiment Results

G.1 Experiment Results with the Standard R^2 Metric

In Section 5, we evaluated predictive performance using the zero-benchmark R^2 to avoid the noise inherent in historical mean estimation. For completeness, this section reports the corresponding results using the standard out-of-sample R^2 metric, which benchmarks model performance against the historical sample mean. Qualitatively, the relative performance among models remains consistent with the observations in Section 5: our adaptive algorithm **ATOMS** continues to outperform fixed-window benchmarks. Quantitatively, we observe that the standard R^2 values are generally lower than their zero-benchmark counterparts.

Table 6 presents out-of-sample standard R^2 values of ATOMS and baselines across distinct economic regimes, serving as the counterpart to Table 2 in the main text.

Table 6: OOS Standard R^2 Averages Across Industries by Time Period.

Method	Full OOS Period	Recessions		
		Gulf War	2001 Recession	Financial Crisis
ATOMS	0.041	−0.019	0.115	0.039
Fixed-val(32)	0.013	−0.038	0.085	−0.003
Fixed-val(512)	0.034	−0.080	0.107	0.037
Fixed-CV	0.026	−0.056	0.060	0.012

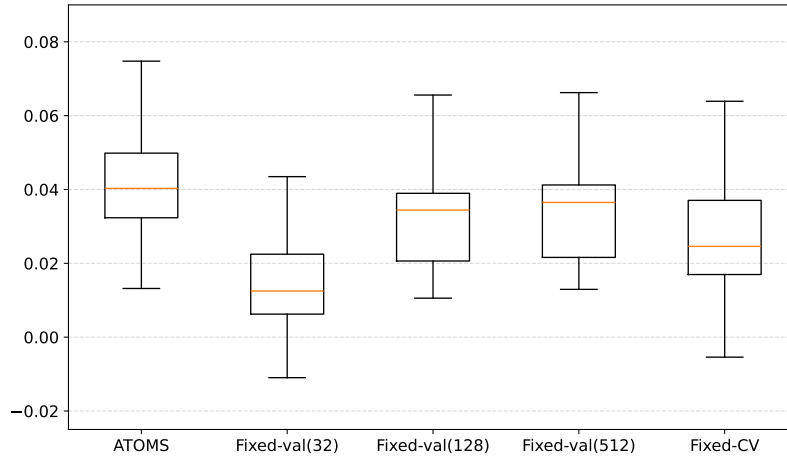
This table reports OOS standard R^2 averages for return prediction models across all 17 industry portfolios. Full OOS Period refers to OOS period covering 01/1990~11/2016. Columns report OOS R^2 averages across all industries and highlight this metric during three recessions, as documented in [NBER Business Cycle Dating](#):

- the 1990 Gulf War recession (06/1990~10/1990);
- the 2001 Recession of dot-com bubble burst and the 9/11 attack (05/2001~10/2001);
- the Financial Crisis led by defaults of subprime mortgages (11/2007~06/2009).

That is, the OOS performance in Gulf War column focuses on model performance comparisons exclusively in the out-of-sample period of 06/1990~10/1990. All values are calculated using monthly return data.

Figure 12 gives a box plot of the OOS standard R^2 of ATOMS and the fixed-window baselines over the 17 industry portfolios, mirroring Figure 4.

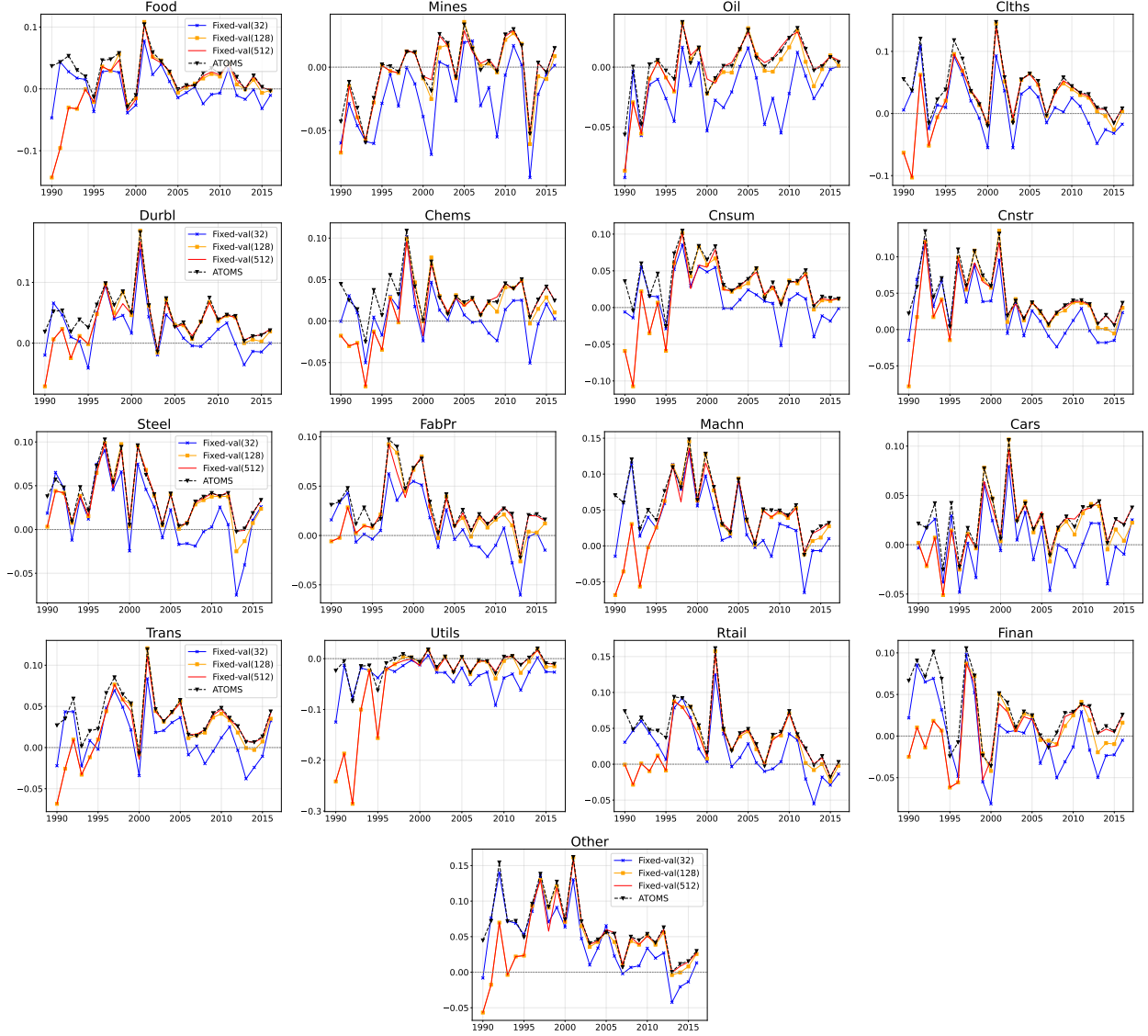
Figure 12: Box Plot of OOS Standard R^2 of ATOMS and Baselines for 17 Industry Portfolios.



This figure describes the distribution of each method’s OOS R^2 . Each box corresponds to all industries and all years in our OOS horizon.

Finally, Figure 13 plots the annual out-of-sample R^2 for the 17 industry portfolios, paralleling Figure 5.

Figure 13: Annual OOS Standard R^2 of Different Approaches for 17 Industry Portfolios.



This figure reports the annual OOS standard R^2 of our adaptive model selection algorithm **ATOMS** (black dashed line with \times 's), as well as the fixed-window baselines **Fixed-val(32)** (blue ∇ 's), **Fixed-val(128)** (orange \blacksquare 's), and **Fixed-val(512)** (red), which use the last 32, last 128 and all months of validation data. The title in each subfigure is Kenneth French's acronym for each industry. For the full names of these industries, please refer to Table 4.

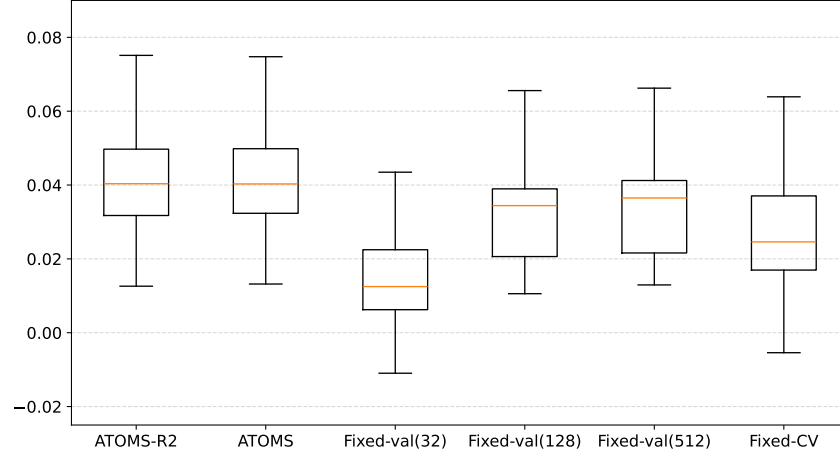
G.2 Experiment Results for ATOMS-R2

In this section, we present experiment results for the R^2 -based model selection method **ATOMS-R2** developed in Appendix B. We set its hyperparameters $\delta' = 0.1$ and $M^2 = 5$. Its performance is similar to the MSE-based approach **ATOMS**. We will report results for the R^2 metric that benchmarks against a zero forecast.

In Figure 14, we give a box plot of the overall out-of-sample R^2 of **ATOMS-R2** along with **ATOMS** and the fixed-window benchmarks across the 17 industries. In Figure 15, we compare the annual

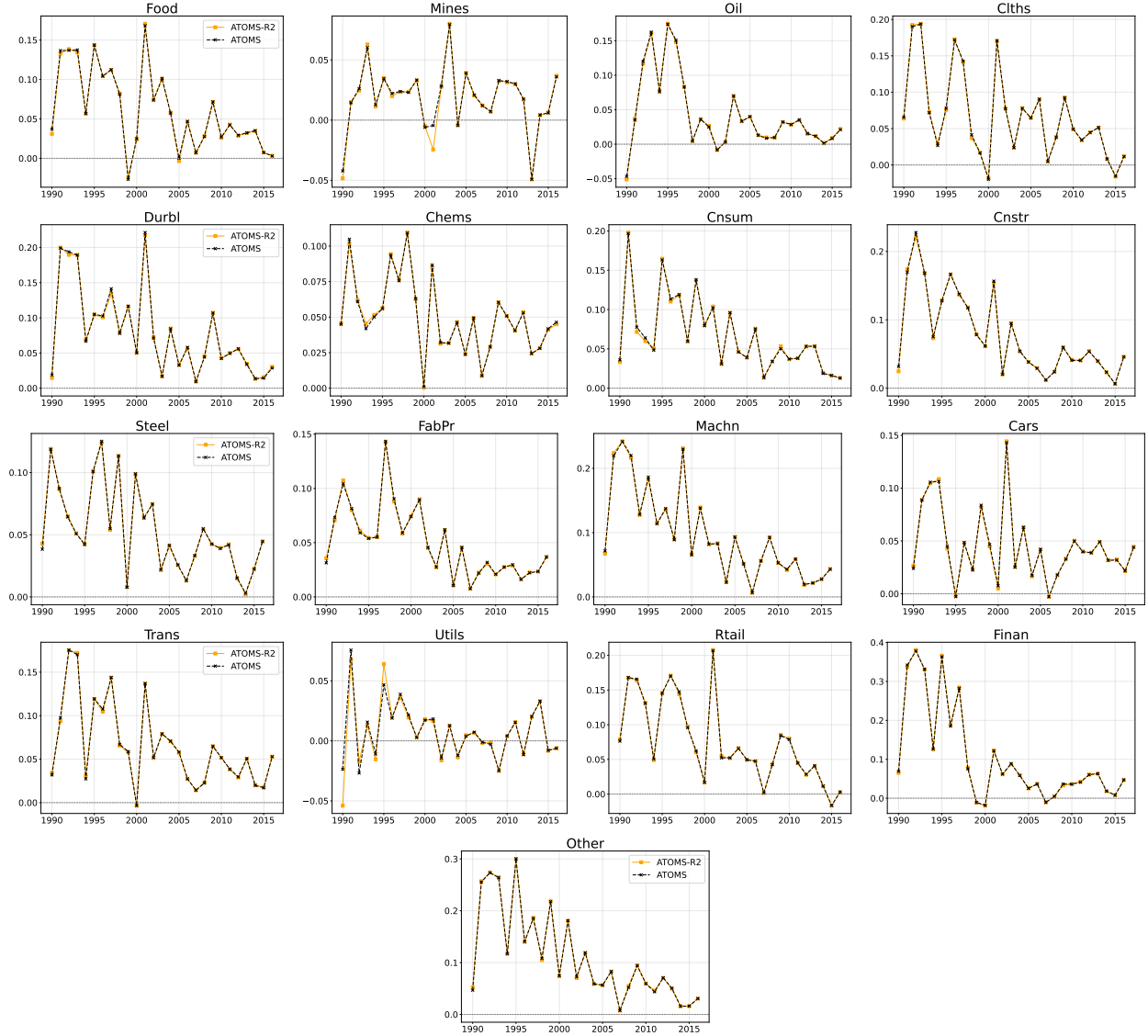
out-of-sample R^2 of **ATOMS-R2** and **ATOMS** for the 17 industries. Again, **ATOMS-R2** and **ATOMS** have similar performance.

Figure 14: Box Plot of Out-of-Sample R^2 of **ATOMS-R2**, **ATOMS** and Baselines for 17 Industry Portfolios.



This figure describes the distribution of each method's OOS R^2 . Each box corresponds to all industries and all years in our OOS horizon.

Figure 15: Annual Out-of-Sample R^2 of ATOMS-R2 and ATOMS for 17 Industry Portfolios.



This figure reports the annual OOS R^2 of our adaptive model selection algorithms **ATOMS-R2** (orange line with ■'s) and **ATOMS** (black dashed line with ×'s). The title in each subfigure is Kenneth French's acronym for each industry. For the full names of these industries, please refer to Table 4.