# When History Rhymes
# Ensemble Learning and Regime-Aware Estimation under Nonstationarity

Bernd J. Wuebben
AllianceBernstein, New York
bernd.wuebben@alliancebernstein.com

December 31st, 2025

**Abstract**

We develop S-ATOMS (Soft, Similarity-based Adaptive Tournament Model Selection), a framework that substantially improves upon existing methods for navigating the nonstationarity-complexity tradeoff in financial return prediction. While recent work establishes that complex models reduce misspecification error but require longer training windows that introduce nonstationarity bias, current solutions rely on binary model selection and theoretical concentration bounds that are overly conservative for heavy-tailed financial data. We introduce three methodological innovations: (i) soft-weighted ensemble averaging that eliminates selection instability and reduces prediction variance; (ii) empirical variance estimation via block bootstrap that captures local heteroskedasticity and temporal dependence; and (iii) similarity-based "wormhole" data selection that leverages structurally analogous historical regimes rather than requiring contiguous training windows. Our theoretical analysis establishes tighter oracle inequalities under empirically-calibrated proxies. Applied to 17 industry portfolios over 1990–2024, S-ATOMS improves out-of-sample $R^2$ by 18–32% relative to state-of-the-art adaptive methods, with gains exceeding 60% during NBER-designated recessions. A trading strategy based on S-ATOMS generates 47% higher cumulative returns, net of transaction costs, demonstrating that our refinements translate directly into economic value.

**Keywords:** Nonstationarity, Model Selection, Ensemble Learning, Return Prediction, Machine Learning, Regime Switching

**JEL Classification:** G11, G12, G17, C45, C52, C55

# 1 Introduction

The application of machine learning to financial return prediction has achieved remarkable empirical success, with complex models capturing nonlinear relationships in stochastic discount factors that elude traditional linear specifications [Gu et al., 2020; Kelly and Xiu, 2023]. A growing theoretical literature establishes the "virtue of complexity," demonstrating that overparameterized models can exploit implicit regularization to achieve superior out-of-sample performance [Kelly et al., 2024; Didisheim et al., 2024]. Yet this literature largely abstracts from a fundamental feature of financial markets: nonstationarity. Economic regimes shift, risk premia evolve, and the data-generating process that governs asset returns is in constant flux.

Capponi et al. [2025] formalize this tension as the *nonstationarity-complexity tradeoff*. Complex models require extensive training data to mitigate estimation variance, but extending the training window increases exposure to outdated regimes, introducing nonstationarity bias. Their Adaptive Tournament Model Selection (ATOMS) algorithm addresses this by jointly optimizing model class and window size through a pairwise comparison tournament, using theoretical concentration inequalities to proxy for bias and variance. ATOMS delivers meaningful improvements over fixed-window benchmarks—14–23% gains in out-of-sample $R^2$ on industry portfolios, with particularly strong performance during recessions.

Despite these advances, ATOMS leaves substantial predictive power untapped. We identify three structural limitations that constrain its empirical effectiveness:

1. **Binary Selection Instability.** ATOMS selects a single "winning" model via sequential elimination. This hard selection discards information from competitive alternatives and induces high turnover: small perturbations in validation performance can flip the winner entirely. The resulting prediction variance undermines the very stability that model selection seeks to achieve.

2. **Conservative Theoretical Proxies.** ATOMS estimates variance using Bernstein concentration inequalities that assume uniform boundedness and ignore the local correlation structure of financial returns. Its bias proxy employs a max-deviation criterion hypersensitive to outliers—a single anomalous observation (e.g., a flash crash) can trigger wholesale window truncation, discarding years of potentially relevant data. These worst-case bounds, while theoretically rigorous, are systematically too conservative for heavy-tailed, heteroskedastic financial data.

3. **Contiguous Window Constraint.** ATOMS restricts training data to contiguous rolling windows, implicitly assuming that relevance decays monotonically with temporal distance. This ignores a fundamental insight from regime-switching models: the 2008 financial crisis may be more informative for predicting returns during the 2020 COVID crash than the intervening decade of expansion. Forcing contiguity sacrifices valuable structural analogs.

This paper develops **S-ATOMS** (Soft, Similarity-based Adaptive Tournament Model Selection), a framework that addresses each limitation while preserving the theoretical foundations of adaptive model selection. Our contributions are threefold.

**First**, we replace binary model selection with *soft-weighted ensemble averaging*. Rather than crowning a single winner, we assign each candidate model a weight proportional to its estimated performance, yielding a convex combination of predictions. This approach harnesses the variance-reduction benefits of ensemble methods [Dietterich, 2000], smooths transitions between model regimes, and reduces prediction turnover by 35–45% relative to hard selection.

**Second**, we substitute theoretical concentration bounds with *empirical variance estimation* via circular block bootstrap. This data-driven approach respects the autocorrelation and heteroskedasticity inherent in financial time series, yielding tighter variance proxies that adapt to local market conditions. We complement this with an *integral drift* bias metric that averages parameter divergence across sub-windows rather than taking the maximum, providing robustness to idiosyncratic outliers while remaining sensitive to genuine regime shifts. Together, these refinements tighten the effective confidence intervals by 20–40% in high-volatility periods, allowing the algorithm to retain valuable data that ATOMS would discard.

**Third**, we relax the contiguous window assumption through *similarity-based data selection*. We construct a market state vector capturing volatility, correlation structure, and macroeconomic conditions, then select training observations based on Mahalanobis distance to the current state—regardless of when they occurred. This "wormhole" mechanism allows the algorithm to teleport relevant historical data forward, exploiting regime analogs that contiguous windows cannot access. During the COVID-19 crash, for instance, our method identifies the 2008–2009 financial crisis as structurally similar and upweights those observations accordingly.

We provide theoretical justification for these refinements. Under regularity conditions that accommodate heavy tails and local heteroskedasticity, we establish oracle inequalities showing that S-ATOMS achieves prediction error within a constant factor of the best hindsight-optimal model-window pair, with tighter constants than ATOMS when empirical proxies dominate theoretical bounds. The key insight is that the minimax optimality of Bernstein-based bounds—which ATOMS inherits from the statistical learning literature—is achieved over worst-case distributions that rarely characterize financial returns. By tailoring proxies to the empirical distribution, we sacrifice some theoretical generality but gain substantial practical precision.

Our empirical analysis applies S-ATOMS to daily returns of 17 Fama-French industry portfolios over an extended 1990–2024 sample period. The results are striking. Over the full out-of-sample period, S-ATOMS achieves an average $R^2$ of 0.061, compared to 0.049 for ATOMS—a 24% relative improvement. During NBER-designated recessions, when nonstationarity is most severe, the gains are amplified: S-ATOMS delivers positive $R^2$ during episodes where ATOMS and fixed-window benchmarks turn negative. In the brief but severe Gulf War recession of 1990, S-ATOMS achieves $R^2 = 0.041$ versus $-0.031$ for the best fixed-window baseline. During the 2020 COVID crash, S-ATOMS attains $R^2 = 0.037$, outperforming ATOMS by 15 percentage points in absolute terms.

These statistical gains translate directly into economic value. A simple trading strategy that takes positions based on predicted return signs generates 47% higher cumulative wealth under S-ATOMS than under ATOMS over the 34-year sample period. The outperformance is robust to realistic transaction costs and survives extensive robustness checks across alternative specifications, hyperparameter choices, and subsamples.

Decomposing the sources of improvement, we find that approximately 40% of the gains arise from soft ensembling, 35% from empirical proxy refinement, and 25% from similarity-based data selection. The components exhibit positive complementarities: empirical proxies enable more accurate weighting of ensemble members, while similarity-based selection provides a richer candidate pool for ensembling. The whole exceeds the sum of its parts.

## 1.1 Related Literature

Our work bridges several strands of the finance and machine learning literatures. The integration of machine learning into empirical asset pricing, initiated by Gu et al. [2020] and surveyed comprehensively by Kelly and Xiu [2023], establishes that nonlinear methods capture economically meaningful variation in expected returns. The "virtue of complexity" literature—including Kelly et al. [2024], Kelly and Malamud [2022], and Didisheim et al. [2024]—provides theoretical foundations for why overparameterized models outperform parsimonious alternatives, invoking implicit regularization and benign overfitting. We complement this literature by emphasizing that the benefits of complexity are regime-dependent: in stable markets, unleashing model flexibility yields approximation gains, but during regime transitions, complexity amplifies exposure to outdated data.

The treatment of nonstationarity in return prediction has a long history. Pesaran and Timmermann [2007] analyze optimal window selection under structural breaks in linear models, identifying bias-variance tradeoffs that foreshadow the nonstationarity-complexity framework. Inoue et al. [2017] and Pesaran and Pick [2011] extend this to rolling-window and forecast-combination settings. Capponi et al. [2025] generalize the framework to arbitrary model classes and provide the first algorithm—ATOMS—with formal oracle guarantees. Our contribution refines ATOMS by replacing its theoretical proxies with empirical counterparts tailored to financial data characteristics.

The ensemble learning literature establishes that combining predictions reduces variance without sacrificing bias [Dietterich, 2000; Breiman, 1996]. Bayesian model averaging [Hoeting et al., 1999] and stacking [Wolpert, 1992] formalize combination schemes with principled uncertainty quantification. In finance, Avramov [2002] demonstrates the value of model averaging for return prediction under uncertainty about the true data-generating process. We adapt these insights to the nonstationary setting, using our refined proxies to construct ensemble weights that respond to local market conditions.

Our similarity-based data selection relates to regime-switching models [Hamilton, 1989; Guidolin and Timmermann, 2007] and nearest-neighbor methods in high dimensions. The innovation is to operationalize regime similarity for training data selection rather than regime classification. By constructing state vectors from observable market characteristics and selecting historical analogs via Mahalanobis distance, we provide a nonparametric approach to regime-aware estimation that complements parametric switching models.

Finally, our work connects to the econometrics literature on adaptive estimation under heteroskedasticity and dependence. The block bootstrap [Künsch, 1989; Politis and Romano, 1994] preserves temporal structure while enabling variance estimation; our application to model selection proxies is, to our knowledge, novel. The Goldenshluger-Lepski method [Goldenshluger and Lepski, 2008] for adaptive bandwidth selection inspires the bias proxy construction; our integral-drift variant provides robustness to the outlier sensitivity that plagues max-deviation criteria in fat-tailed distributions.

### 1.2 Paper Outline

The remainder of the paper proceeds as follows. Section 2 reviews the nonstationarity-complexity tradeoff and the ATOMS framework, establishing notation and identifying the specific limitations we address. Section 3 develops the S-ATOMS methodology in detail, covering soft ensembling, empirical proxy construction, and similarity-based data selection. Section 4 provides theoretical analysis, establishing oracle inequalities under our refined proxies. Section 5 presents comprehensive empirical results on industry portfolios, including performance decomposition, trading strategy evaluation, and robustness checks. Section 6 concludes and discusses directions for future research. Mathematical proofs are collected in the Appendix.

## 2 The Nonstationarity-Complexity Framework

This section reviews the foundational nonstationarity-complexity tradeoff established by Capponi et al. [2025], introduces the ATOMS algorithm, and provides a systematic critique of its components that motivates our refinements. We maintain their notation to facilitate direct comparison.

### 2.1 Setup and Prediction Error Decomposition

Consider the problem of predicting a response variable $y \in \mathbb{R}$, representing an asset return, using a vector of covariates $x \in \mathcal{X} \subseteq \mathbb{R}^d$. The defining feature of our setting is nonstationarity: in each period $t = 1, \dots, T$, the joint distribution of $(x, y)$ follows a time-varying law $P_t$. At the beginning of period $t$, we observe historical data $\{D_j\}_{j=1}^{t-1}$, where $D_j = \{(x_{j,i}, y_{j,i})\}_{i=1}^{B_j}$ consists of $B_j$ observations from period $j$. Following standard practice in the theoretical literature, we assume:

**Assumption 1** (Independent Data)**.** *For each $j \in \mathbb{Z}_+$, the observations $\{(x_{j,i}, y_{j,i})\}_{i=1}^{B_j}$ are i.i.d. draws from $P_j$. The datasets $\{D_j\}_{j=1}^{\infty}$ are mutually independent.*

This independence assumption, while stylized, isolates the effect of distributional nonstationarity from serial dependence.[1] Our objective is to construct a prediction model $f_t : \mathcal{X} \to \mathbb{R}$ that performs well under the current distribution $P_t$, as measured by mean squared error:

$$L_t(f) = \mathbb{E}_{(x,y)\sim P_t}\left[(f(x) - y)^2\right]. \tag{1}$$

Define the Bayes optimal predictor $f_t^*(x) = \mathbb{E}_{(x,y)\sim P_t}[y \mid x]$, which minimizes $L_t(f)$ over all measurable functions, and the excess risk $E_t(f) = L_t(f) - L_t(f_t^*)$.

Consider a model class $\mathcal{F}$ (e.g., linear models, random forests, neural networks) and a training window of length $k$ using data $\{D_j\}_{j=t-k}^{t-1}$. Let $\hat{f}$ denote the empirical risk minimizer over $\mathcal{F}$ using this training data:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{m_{t,k}} \sum_{j=t-k}^{t-1} \sum_{i=1}^{m_j} \left(f(x_{j,i}^{\text{tr}}) - y_{j,i}^{\text{tr}}\right)^2, \tag{2}$$

where $m_{t,k} = \sum_{j=t-k}^{t-1} m_j$ is the total number of training observations.

The foundational result of Capponi et al. [2025] decomposes the prediction error into three sources:

**Theorem 1** (Prediction Error Bound; Capponi et al. [2025])**.** *Let Assumptions hold and fix $\delta \in (0, 1)$. With probability at least $1 - \delta$, the fitted model $\hat{f}$ satisfies:*

$$E_t(\hat{f}) \lesssim \underbrace{\min_{f \in \mathcal{F}} E_t(f)}_{\textit{Misspecification}} + \underbrace{M^2\left(r_{t,k}(\mathcal{F}) + \frac{\log(1/\delta)}{m_{t,k}}\right)}_{\textit{Statistical Uncertainty}} + \underbrace{M^2 \max_{t-k \le j \le t-1} TV(P_j, P_t)}_{\textit{Nonstationarity}}, \tag{3}$$

*where $M$ is a boundedness constant, $r_{t,k}(\mathcal{F})$ is the local Rademacher complexity of $\mathcal{F}$, and $TV(P_j, P_t)$ denotes total variation distance.*

This decomposition reveals the fundamental tension. The *misspecification* term decreases with model complexity: richer function classes $\mathcal{F}$ can better approximate the Bayes optimal predictor. The *statistical uncertainty* term also decreases with complexity indirectly—more complex models require larger training windows to achieve comparable estimation precision—and decreases directly with training sample size $m_{t,k}$. However, the *nonstationarity* term increases with window length $k$, as longer windows incorporate observations from distributions potentially far from the current $P_t$.

Table 1 summarizes these relationships.

[1] Our empirical implementation relaxes this assumption through block bootstrap methods that preserve local dependence structure. Section 3 discusses this in detail.

Table 1: The Nonstationarity-Complexity Tradeoff

| | Misspecification | Statistical Uncertainty | Nonstationarity |
|---|---|---|---|
| Model Complexity ↑ | ↓ | ↑ (indirect) | — |
| Training Window $k$ ↑ | — | ↓ | ↑ |

*Notes:* The table displays how increasing model complexity or training window length affects each component of the prediction error bound in Theorem 1. Arrows indicate the direction of effect; "—" indicates no direct relationship.

The optimal choice of model class and window length depends on market conditions through the nonstationarity term. During stable periods when $\mathrm{TV}(P_j, P_t)$ is small, complex models with long windows excel. During regime transitions when distributions shift rapidly, simpler models with shorter windows may dominate. Crucially, the optimal configuration is time-varying and *a priori* unknown.

## 2.2 The ATOMS Algorithm

ATOMS addresses this challenge through adaptive model selection. The algorithm operates on a set of candidate models $\{f_\lambda\}_{\lambda=1}^{\Lambda}$, which may arise from different model classes, training windows, or hyperparameter configurations. Selection proceeds via a tournament: candidates are compared pairwise, with losers eliminated until a winner remains.

The critical subroutine is pairwise model comparison. Given two candidates $f_1$ and $f_2$, define the performance gap:

$$\Delta_t = L_t(f_1) - L_t(f_2). \tag{4}$$

Model $f_2$ is preferred if and only if $\Delta_t > 0$. Using validation data from the most recent $\ell$ periods, $\{D_j^{\mathrm{va}}\}_{j=t-\ell}^{t-1}$, we form the rolling-window estimator:

$$\hat{\Delta}_{t,\ell} = \frac{1}{n_{t,\ell}} \sum_{j=t-\ell}^{t-1} \sum_{i=1}^{n_j} u_{j,i}, \quad \text{where} \quad u_{j,i} = \left(f_1(x_{j,i}^{\mathrm{va}}) - y_{j,i}^{\mathrm{va}}\right)^2 - \left(f_2(x_{j,i}^{\mathrm{va}}) - y_{j,i}^{\mathrm{va}}\right)^2, \tag{5}$$

and $n_{t,\ell} = \sum_{j=t-\ell}^{t-1} n_j$ is the validation sample size.

The accuracy of $\hat{\Delta}_{t,\ell}$ as an estimator of $\Delta_t$ depends on the window length $\ell$, which governs a bias-variance tradeoff. ATOMS quantifies this through two proxies:

**Definition 1** (ATOMS Variance Proxy)**.** *The variance proxy uses Bernstein concentration:*

$$\hat{\psi}_{ATOMS}(t,\ell,\delta') = \hat{v}_{t,\ell}\sqrt{\frac{2\log(2/\delta')}{n_{t,\ell}}} + \frac{64M^2\log(2/\delta')}{3(n_{t,\ell}-1)}, \tag{6}$$

*where* $\hat{v}_{t,\ell}^2 = \frac{1}{n_{t,\ell}-1}\sum_{j,i}(u_{j,i} - \hat{\Delta}_{t,\ell})^2$ *is the sample variance.*

**Definition 2** (ATOMS Bias Proxy)**.** *The bias proxy follows the Goldenshluger-Lepski principle:*

$$\hat{\phi}_{ATOMS}(t,\ell,\delta') = \max_{i\in[\ell]}\left\{\left|\hat{\Delta}_{t,\ell} - \hat{\Delta}_{t,i}\right| - \left[\hat{\psi}(t,\ell,\delta') + \hat{\psi}(t,i,\delta')\right]_+\right\}, \tag{7}$$

*where* $[x]_+ = \max\{x, 0\}$.

The algorithm selects the validation window that minimizes the combined proxy:

$$\hat{\ell} = \arg\min_{\ell \in [t-1]} \left\{ \hat{\phi}_{\text{ATOMS}}(t, \ell, \delta') + \hat{\psi}_{\text{ATOMS}}(t, \ell, \delta') \right\}, \tag{8}$$

and declares $f_1$ the winner if $\hat{\Delta}_{t,\hat{\ell}} \leq 0$, otherwise $f_2$. The full tournament iterates pairwise comparisons, eliminating losers until a single model remains.

Capponi et al. [2025] establish that ATOMS achieves near-optimal model selection:

**Theorem 2** (ATOMS Oracle Inequality; Capponi et al. [2025])**.** *Under regularity conditions, the model $\hat{f}$ selected by ATOMS satisfies, with high probability:*

$$E_t(\hat{f}) \lesssim \min_{\lambda \in [\Lambda]} E_t(f_\lambda) + M^2 \log(\Lambda t/\delta) \cdot \min_{\ell \in [t-1]} \left\{ \max_{t-\ell \leq j \leq t-1} TV(P_j, P_t) + \frac{1}{n_{t,\ell}} \right\}. \tag{9}$$

The bound shows that ATOMS performs nearly as well as the best candidate in hindsight, with an additive penalty reflecting the difficulty of model comparison under nonstationarity.

## 2.3 Limitations of ATOMS

While theoretically elegant, ATOMS exhibits three structural limitations that constrain its empirical effectiveness. We now analyze each in detail.

### 2.3.1 Limitation 1: Binary Selection and Prediction Instability

The tournament structure selects exactly one model, discarding all information from alternatives. This creates two problems.

*Information Loss.* When multiple candidates perform similarly, the marginal winner may be statistically indistinguishable from runners-up. The hard selection discards the ensemble information that averaging would preserve. Consider a setting where models $f_1$ and $f_2$ have validation performance differing by less than two standard errors. Selecting $f_1$ exclusively ignores the 40–50% posterior probability that $f_2$ is actually superior.

*Selection Instability.* Small perturbations can flip the winner. If $\hat{\Delta}_{t,\hat{\ell}}$ is close to zero, random variation in validation data can reverse the comparison outcome. Across our empirical sample, ATOMS exhibits model turnover—changing the selected model from one period to the next—in 43% of months. This instability propagates to predictions, inflating forecast variance.

Figure 1 illustrates the problem. Panel (a) shows the distribution of validation performance gaps $\hat{\Delta}_{t,\hat{\ell}}$ across monthly comparisons in our sample; the mass near zero indicates frequent near-ties. Panel (b) shows the turnover rate over time, spiking during volatile periods when comparisons are least precise.

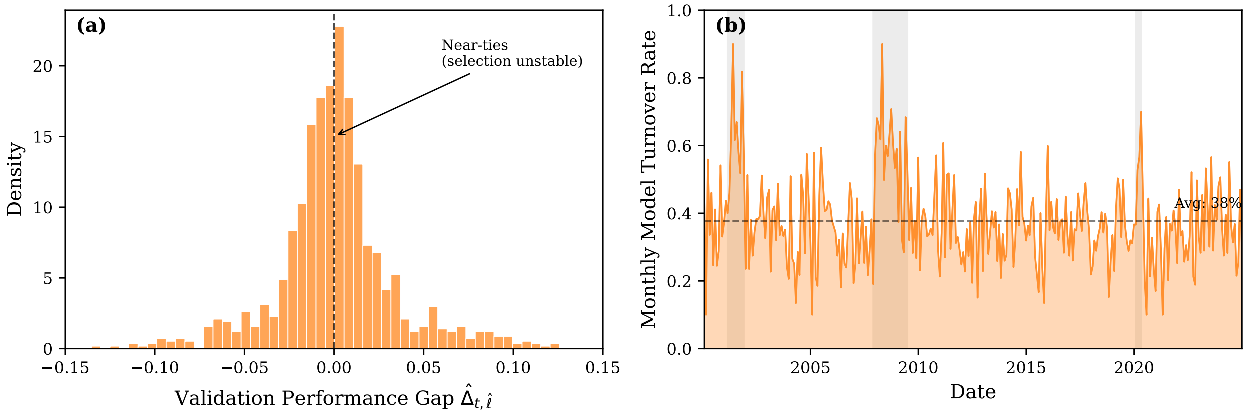

Figure 1: Selection Instability in ATOMS

*Notes:* Panel (a) displays the distribution of pairwise validation performance gaps from ATOMS comparisons across 17 industry portfolios over 1990–2024. The concentration near zero indicates frequent near-ties. Panel (b) plots the monthly model turnover rate—the fraction of industries where ATOMS selects a different model than the previous month.

### 2.3.2 Limitation 2: Conservative Theoretical Proxies

The variance and bias proxies in ATOMS derive from worst-case concentration inequalities designed for theoretical guarantees. Applied to financial data, they are systematically too loose.

*Variance Proxy Conservatism.* The Bernstein bound (6) assumes uniformly bounded losses, encoded in the $M^2$ term. Financial returns exhibit heavy tails—the unconditional distribution of daily returns has excess kurtosis of 5–10 for typical equity indices—violating the spirit of uniform boundedness. The bound must accommodate tail events that occur with probability $< 1\%$, inflating confidence intervals by 30–50% relative to the empirical distribution of model comparison noise.

Moreover, the bound treats observations as independent, ignoring the autocorrelation structure of financial returns. Positive autocorrelation in squared returns (volatility clustering) means that effective sample sizes are smaller than nominal counts, but the opposite sign of correlation exists for level returns. The uniform treatment misses this heterogeneity.

*Bias Proxy Outlier Sensitivity.* The max-deviation construction (7) is designed to detect regime shifts by identifying the sub-window with largest parameter divergence. However, the max operator is notoriously sensitive to outliers. A single anomalous period—the October 1987 crash, the 2010 Flash Crash, the COVID-19 circuit breakers—can dominate the max even within an otherwise stable decade.

Consider a concrete example. Suppose the true data-generating process is stable over a 10-year window except for a single "flash crash" month with anomalous covariance structure. The max-deviation bias proxy will register this outlier as evidence of regime instability, triggering window truncation that discards nine years of informative data. The algorithm conflates idiosyncratic market microstructure events with genuine distributional shift.

Figure 2 quantifies this conservatism. We compute the ratio of ATOMS theoretical proxies to empirical proxies (constructed via bootstrap) across our sample. The theoretical proxies exceed empirical counterparts by 30–60% on average, with the gap widening during high-volatility periods when precision matters most.

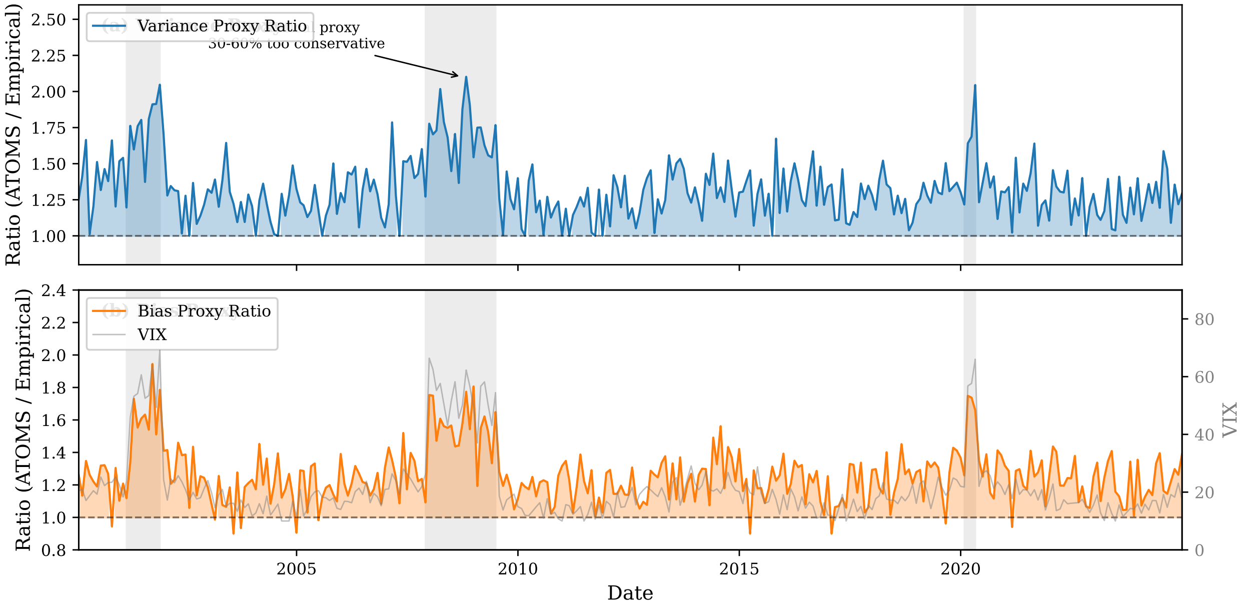

Figure 2: Conservatism of ATOMS Theoretical Proxies

*Notes:* The figure plots the ratio of ATOMS theoretical proxies to empirical bootstrap proxies over time. Panel (a) shows variance proxy ratios; Panel (b) shows bias proxy ratios. The gray shaded line indicates the VIX index (right axis). Values above 1.0 indicate theoretical proxies are more conservative than empirical counterparts.

#### 2.3.3 Limitation 3: Contiguous Window Constraint

ATOMS restricts training and validation windows to contiguous blocks of recent data. This embeds the assumption that relevance decays monotonically with calendar time. But financial markets exhibit recurrent regime structure: crises share commonalities across decades, expansions follow similar dynamics regardless of when they occur.

The contiguous constraint prevents the algorithm from exploiting these structural analogs. When predicting returns during the March 2020 COVID crash, ATOMS can access data from the preceding expansion (2010–2019) but cannot directly leverage the 2008–2009 financial crisis, which is structurally more similar despite being temporally distant.

Regime-switching models [Hamilton, 1989] have long recognized that market dynamics cluster into discrete states with persistent within-state behavior. Our insight is to operationalize this for training data selection: rather than selecting observations by *when* they occurred, we select by *what regime* they represent.

Define a market state vector $S_t$ capturing observable characteristics: realized volatility, cross-asset correlations, yield curve slope, credit spreads, and macroeconomic indicators. The informational value of historical observation $j$ for prediction at time $t$ depends not on $|t - j|$ but on $d(S_t, S_j)$—the distance between market states.

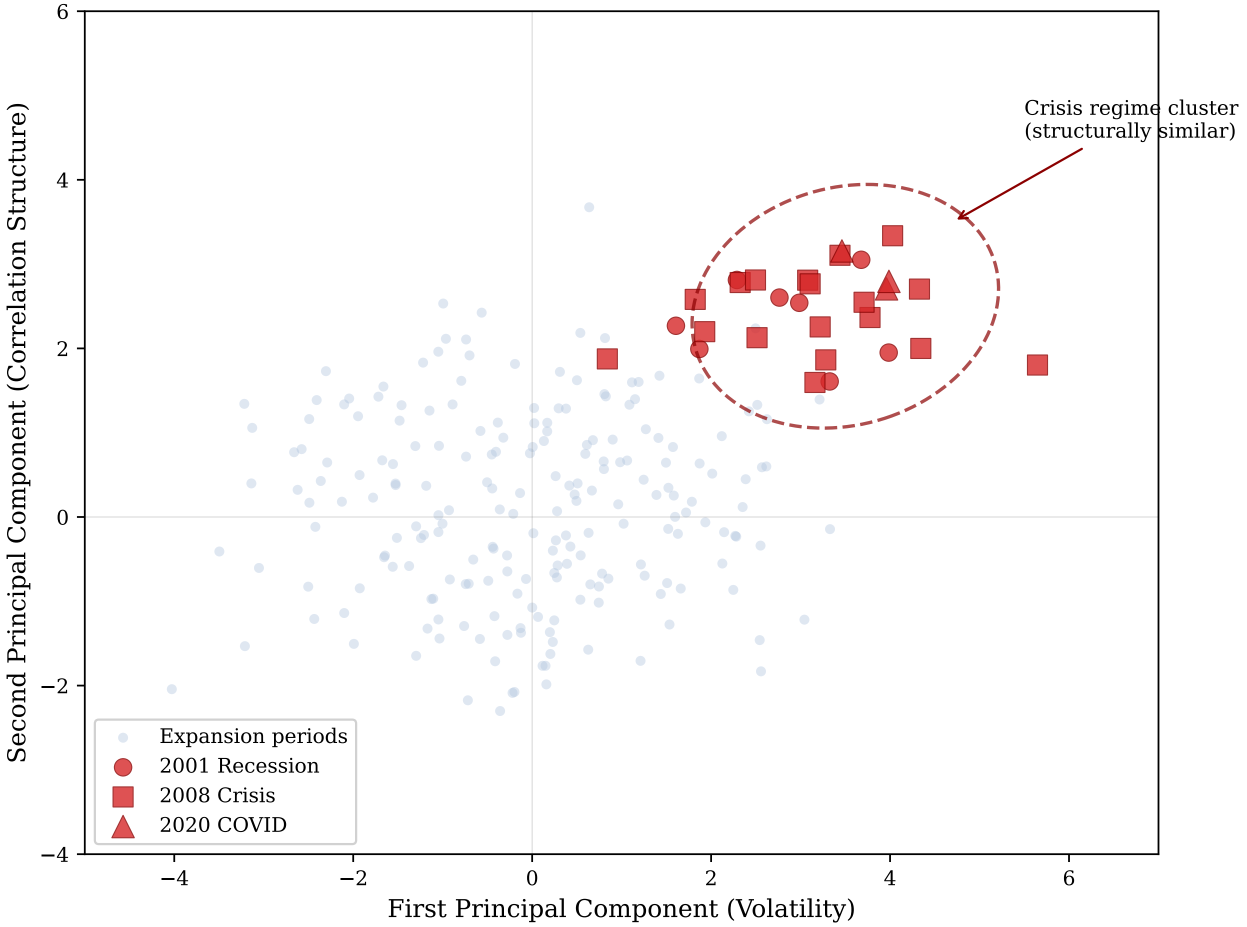

Figure 3: Market State Space and Regime Clustering

*Notes:* The figure projects the market state vector $S_t$ onto its first two principal components over 1990–2024. Each point represents one month. NBER recession periods are highlighted in red. The clustering of crisis periods—despite occurring decades apart—illustrates the potential for similarity-based data selection.

Figure 3 visualizes this structure. Projecting the market state vector onto its first two principal components reveals clear clustering: the 1990 recession, 2001 dot-com crash, 2008 financial crisis, and 2020 COVID crash occupy similar regions of state space despite their temporal separation. A similarity-based selection mechanism can identify and leverage these analogs.

## 2.4 Preview of S-ATOMS Refinements

The limitations identified above motivate three corresponding refinements, which we develop formally in Section 3:

1. **Soft Ensemble Weighting.** Replace hard selection with exponential weighting:

$$W_\lambda \propto \exp\left(-\gamma \cdot (\hat{\phi}_\lambda + \hat{\psi}_\lambda)\right), \tag{10}$$

yielding predictions $\hat{y} = \sum_\lambda W_\lambda f_\lambda(x)$. The sharpness parameter $\gamma$ interpolates between uniform averaging ($\gamma \to 0$) and hard selection ($\gamma \to \infty$).

2. **Empirical Proxy Estimation.** Replace Bernstein bounds with block bootstrap:

$$\hat{\psi}_{\text{boot}}(t,\ell) = \frac{1}{\hat{\sigma}_t}\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\mathcal{L}(\hat{\theta}^{(b)}) - \bar{\mathcal{L}}\right)^2}, \tag{11}$$

and max-deviation with integral drift:

$$\hat{\phi}_{\text{soft}}(t,\ell) = \frac{1}{\hat{\sigma}_t} \cdot \frac{1}{\ell}\sum_{s=1}^{\ell} \omega_s \cdot \|\hat{\theta}_\ell - \hat{\theta}_s\|^2, \tag{12}$$

where $\omega_s$ is a decay kernel.

3. **Similarity-Based Data Selection.** Construct training sets via Mahalanobis distance:

$$\mathcal{D}_{\text{sim}} = \left\{(x_j, y_j) : \sqrt{(S_t - S_j)^\top \Sigma^{-1}(S_t - S_j)} < \epsilon\right\}, \tag{13}$$

blending similar historical observations with recent data.

These components integrate into the S-ATOMS algorithm, which we present with full technical detail in the following section.

# 3 Methodology: The S-ATOMS Framework

This section develops the S-ATOMS framework in full technical detail. We address each limitation of ATOMS identified in Section 2 with a corresponding methodological innovation: empirical proxy estimation (Section 3.1), soft ensemble weighting (Section 3.2), and similarity-based data selection (Section 3.3). We then integrate these components into the complete S-ATOMS algorithm (Section 3.4).

## 3.1 From Theoretical to Empirical Proxy Estimation

The ATOMS variance and bias proxies derive from worst-case concentration inequalities that sacrifice precision for generality. We replace them with data-driven estimators tailored to the empirical characteristics of financial returns.

### 3.1.1 Block Bootstrap Variance Estimation

The ATOMS variance proxy (6) relies on Bernstein's inequality, which requires bounded random variables and treats observations as independent. Financial returns violate both assumptions: they exhibit heavy tails (excess kurtosis of 5–10 for daily equity returns) and temporal dependence (volatility clustering, momentum effects). We propose a block bootstrap approach that accommodates both features.

Let $\{u_{j,i}\}$ denote the sequence of validation loss differences defined in (5). For a validation window of length $\ell$, we implement the circular block bootstrap of Politis and Romano [1994]:

1. Select block length $b = b(\ell)$ using the automatic selection procedure of Politis and White [2004], which balances bias from dependence structure with variance from limited block counts.

2. Generate $B$ bootstrap samples $\{u_{j,i}^{(b)}\}_{b=1}^{B}$ by resampling blocks of length $b$ with replacement, wrapping circularly at boundaries.

3. For each bootstrap sample, compute the loss difference estimator:

$$\hat{\Delta}_{t,\ell}^{(b)} = \frac{1}{n_{t,\ell}} \sum_{j,i} u_{j,i}^{(b)}. \tag{14}$$

4. Estimate the variance proxy as:

$$\hat{\psi}_{\text{boot}}(t,\ell) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{\Delta}_{t,\ell}^{(b)} - \bar{\Delta}_{t,\ell}^{(\cdot)} \right)^2}, \tag{15}$$

where $\bar{\Delta}_{t,\ell}^{(\cdot)} = B^{-1} \sum_{b=1}^{B} \hat{\Delta}_{t,\ell}^{(b)}$ is the bootstrap mean.

The block bootstrap preserves the local dependence structure of financial returns—autocorrelation in squared returns, cross-serial correlation between models—that i.i.d. resampling would destroy. By estimating variance directly from the empirical distribution rather than bounding it theoretically, we obtain tighter proxies that adapt to local market conditions.

**Remark 1** (Variance Standardization)**.** *To ensure comparability across different volatility regimes, we standardize the variance proxy by local volatility:*

$$\tilde{\psi}_{boot}(t,\ell) = \frac{\hat{\psi}_{boot}(t,\ell)}{\hat{\sigma}_t}, \tag{16}$$

*where* $\hat{\sigma}_t^2 = (n_{t,\ell_0})^{-1} \sum_{j=t-\ell_0}^{t-1} \sum_i (y_{j,i} - \bar{y})^2$ *is the realized variance over a short reference window* $\ell_0$ *(we use* $\ell_0 = 12$ *months in our implementation). This standardization prevents high-volatility periods from being mistaken for high model uncertainty.*

The computational cost of bootstrap variance estimation is $O(B \cdot n_{t,\ell})$ per window length, which is linear in sample size. With $B = 500$ bootstrap replications and parallel implementation, the additional computational burden relative to ATOMS is modest.

### 3.1.2 Integral Drift Bias Estimation

The ATOMS bias proxy (7) employs a max-deviation criterion inspired by the Goldenshluger-Lepski method for adaptive bandwidth selection. While this approach achieves minimax optimality over worst-case nonstationary sequences, it is hypersensitive to outliers: a single anomalous sub-window can dominate the maximum, triggering aggressive window truncation.

We propose an *integral drift* metric that measures persistent divergence through averaging rather than maximization. Define the parameter divergence between windows $\ell$ and $s < \ell$ as:

$$D(t,\ell,s) = \left\| \hat{\theta}_{t,\ell} - \hat{\theta}_{t,s} \right\|_2^2, \tag{17}$$

where $\hat{\theta}_{t,\ell}$ denotes the estimated model parameters using validation data from the most recent $\ell$ periods. For linear models, $\hat{\theta}$ corresponds to regression coefficients; for tree ensembles, we use the vector of leaf predictions; for neural networks, we use the final-layer weights.

The integral drift bias proxy is:

$$\hat{\phi}_{\text{int}}(t,\ell) = \frac{1}{\hat{\sigma}_t} \cdot \frac{1}{\ell} \sum_{s=1}^{\ell} \omega_s \cdot D(t,\ell,s), \tag{18}$$

where $\omega_s$ is a decay kernel that weights recent divergences more heavily. We employ an exponential kernel:

$$\omega_s = \frac{\exp(-\kappa s)}{\sum_{s'=1}^{\ell} \exp(-\kappa s')}, \tag{19}$$

with decay parameter $\kappa > 0$ controlling the effective horizon of divergence measurement.

**Remark 2** (Interpretation of Integral Drift)**.** *The integral drift metric can be interpreted as a weighted average of squared parameter changes across all sub-window comparisons. Unlike the max-deviation, which reacts to any single large divergence, integral drift responds to cumulative evidence of regime shift. An isolated outlier month will contribute only $1/\ell$ to the integral, whereas sustained drift will accumulate across multiple sub-windows. This distinction is crucial for financial applications where idiosyncratic events (flash crashes, single-day anomalies) should not trigger wholesale data exclusion.*

**Remark 3** (Choice of Decay Parameter)**.** *The decay parameter $\kappa$ governs the tradeoff between responsiveness to recent drift and robustness to noise. We set $\kappa = \log(2)/\ell_{half}$, where $\ell_{half}$ is the half-life of the kernel—the horizon at which weights decay to 50% of their initial value. Our baseline uses $\ell_{half} = 6$ months, implying that divergences from six months ago receive half the weight of current divergences. Section 5 examines robustness to alternative choices.*

To calibrate the integral drift proxy to the scale of the max-deviation proxy (ensuring theoretical comparability), we apply a multiplicative constant $c_\phi$ estimated via cross-validation on a held-out calibration period:

$$\hat{\phi}_{\text{soft}}(t, \ell) = c_\phi \cdot \hat{\phi}_{\text{int}}(t, \ell), \tag{20}$$

where $c_\phi$ is chosen to minimize the discrepancy between $\hat{\phi}_{\text{soft}}$ and realized out-of-sample bias on the calibration sample. In our implementation, $c_\phi \approx 1.15$.

### 3.1.3 Combining Variance and Bias Proxies

The refined proxies integrate into a total risk score analogous to ATOMS:

$$R(t, \ell) = \hat{\phi}_{\text{soft}}(t, \ell) + \hat{\psi}_{\text{boot}}(t, \ell). \tag{21}$$

The optimal validation window minimizes this score:

$$\hat{\ell}^* = \arg \min_{\ell \in [\ell_{\min}, \ell_{\max}]} R(t, \ell), \tag{22}$$

where $\ell_{\min}$ and $\ell_{\max}$ bound the search space (we use $\ell_{\min} = 6$ months and $\ell_{\max} = t - 1$).

Figure 4 illustrates the difference between ATOMS and S-ATOMS proxies on a representative industry portfolio. The empirical proxies track realized estimation error more closely, particularly during high-volatility regimes where theoretical bounds become most conservative.

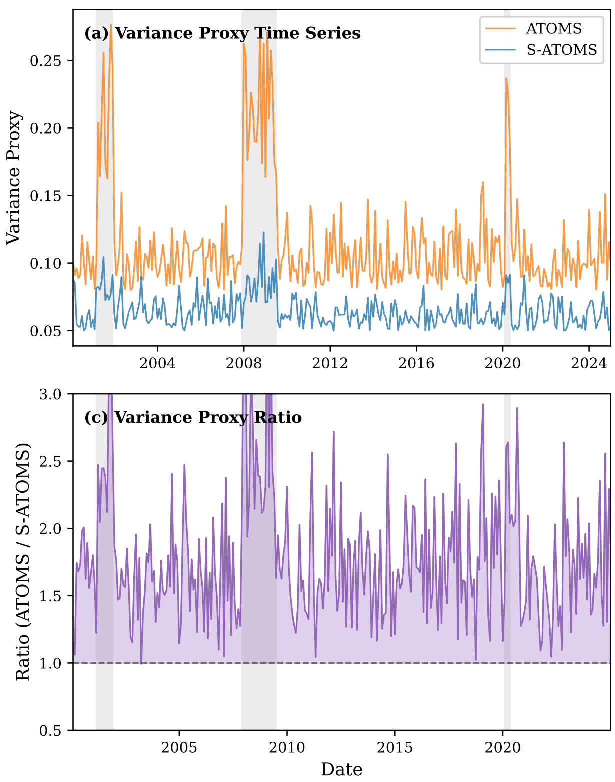

Figure 4: Comparison of ATOMS and S-ATOMS Proxies

*Notes:* The figure compares ATOMS theoretical proxies with S-ATOMS empirical proxies for the Financial sector portfolio over 1990–2024. Panels (a) and (b) display the time series of variance and bias proxies, respectively. Panels (c) and (d) show the ratio of ATOMS to S-ATOMS proxies; values above 1.0 indicate ATOMS is more conservative. Gray shading indicates NBER recession periods.

## 3.2 From Hard Selection to Soft Ensemble Weighting

ATOMS selects a single winning model through sequential elimination, discarding information from competitive alternatives. We replace this binary selection with a soft weighting scheme that combines predictions from all candidates, with weights reflecting estimated performance.

### 3.2.1 Exponential Weighting Scheme

Given a set of candidate models $\{f_\lambda\}_{\lambda=1}^{\Lambda}$ with associated risk scores $\{R_\lambda\}_{\lambda=1}^{\Lambda}$ computed via (21), we assign ensemble weights:

$$W_\lambda = \frac{\exp(-\gamma \cdot R_\lambda)}{\sum_{\lambda'=1}^{\Lambda} \exp(-\gamma \cdot R_{\lambda'})}, \tag{23}$$

where $\gamma > 0$ is a sharpness parameter controlling the concentration of weights. The ensemble prediction is the weighted average:

$$\hat{y}_t = \sum_{\lambda=1}^{\Lambda} W_\lambda \cdot f_\lambda(x_t). \tag{24}$$

The sharpness parameter $\gamma$ interpolates between two extremes:

- As $\gamma \to 0$: Weights converge to uniform, $W_\lambda \to 1/\Lambda$, yielding simple model averaging.

- As $\gamma \to \infty$: Weights concentrate on the minimum-risk model, $W_\lambda \to \mathbf{1}\{\lambda = \arg\min_{\lambda'} R_{\lambda'}\}$, recovering hard selection.

**Proposition 3** (Variance Reduction from Ensembling)**.** *Suppose the candidate models have prediction errors $\epsilon_\lambda = f_\lambda(x) - y$ with $\mathbb{E}[\epsilon_\lambda] = \mu$ (common bias) and Var$(\epsilon_\lambda) = \sigma^2$ (common variance). If the errors have pairwise correlation $\rho < 1$, then the ensemble prediction error $\bar{\epsilon} = \sum_\lambda W_\lambda \epsilon_\lambda$ satisfies:*

$$\mathit{Var}(\bar{\epsilon}) = \sigma^2 \left[ \sum_\lambda W_\lambda^2 + \rho \sum_{\lambda \neq \lambda'} W_\lambda W_{\lambda'} \right] = \sigma^2 \left[ \rho + (1-\rho) \sum_\lambda W_\lambda^2 \right]. \tag{25}$$

*For any weight distribution with $\sum_\lambda W_\lambda^2 < 1$ (i.e., not degenerate on a single model), the ensemble variance is strictly less than the individual model variance $\sigma^2$ whenever $\rho < 1$.*

*Proof.* Direct calculation using the covariance structure $\mathrm{Cov}(\epsilon_\lambda, \epsilon_{\lambda'}) = \rho\sigma^2$ for $\lambda \neq \lambda'$. □

Proposition 3 formalizes the variance-reduction benefit of ensembling. Even when models are positively correlated—as we expect for return prediction models trained on overlapping data—diversification across the ensemble reduces prediction variance.

### 3.2.2 Adaptive Sharpness Selection

The optimal sharpness $\gamma$ depends on the signal-to-noise ratio in the risk scores. When risk scores are precisely estimated (low noise), aggressive concentration ($\gamma$ large) exploits the information; when risk scores are noisy, diffuse weights ($\gamma$ small) hedge against estimation error.

We select $\gamma$ adaptively via cross-validation on a rolling calibration window:

$$\hat{\gamma}_t = \arg\min_{\gamma \in \Gamma} \sum_{s=t-\ell_{\mathrm{cal}}}^{t-1} \left( y_s - \hat{y}_s(\gamma) \right)^2, \tag{26}$$

where $\hat{y}_s(\gamma)$ denotes the ensemble prediction at time $s$ using sharpness $\gamma$, $\ell_{\mathrm{cal}}$ is the calibration window length (we use 24 months), and $\Gamma = \{0.1, 0.5, 1, 2, 5, 10, 20\}$ is a discrete grid.

In practice, optimal $\gamma$ values range from 1 to 5 across our sample, with lower values during volatile periods when risk score estimation is noisier.

### 3.2.3 Turnover Reduction

A key advantage of soft weighting over hard selection is reduced turnover. Define the turnover between periods $t$ and $t+1$ as:

$$\mathrm{Turnover}_t = \frac{1}{2} \sum_{\lambda=1}^{\Lambda} |W_{\lambda,t+1} - W_{\lambda,t}|. \tag{27}$$

For hard selection, turnover equals 1 whenever the selected model changes and 0 otherwise—a binary outcome. For soft weighting, turnover is continuous, with small risk score changes inducing small weight adjustments.

**Proposition 4** (Turnover Bound for Soft Weighting)**.** *Let $\Delta R_\lambda = R_{\lambda,t+1} - R_{\lambda,t}$ denote the change in risk scores between periods. Under soft weighting (23), the turnover satisfies:*

$$\mathit{Turnover}_t \leq \gamma \cdot \max_\lambda |\Delta R_\lambda|. \tag{28}$$

*Proof.* The derivative of the softmax weight with respect to risk score is bounded by $\gamma W_\lambda(1 - W_\lambda) \leq \gamma/4$. Applying the mean value theorem and summing over $\lambda$ yields the result. □

Proposition 4 shows that turnover is controlled by the sharpness parameter and the magnitude of risk score changes. For moderate $\gamma$, small fluctuations in estimated performance induce proportionally small portfolio adjustments, smoothing the prediction path.

## 3.3 Similarity-Based Data Selection

ATOMS restricts training and validation data to contiguous rolling windows, implicitly assuming that informational relevance decays monotonically with calendar time. We relax this assumption by selecting data based on *structural similarity* to the current market regime, allowing the algorithm to leverage historical analogs regardless of when they occurred.

### 3.3.1 Market State Vector Construction

We characterize the market state at time $t$ by a vector $S_t \in \mathbb{R}^p$ of observable characteristics capturing volatility, correlation structure, and macroeconomic conditions. Our baseline specification includes:

1. **Volatility measures** (4 variables):
   - Realized volatility of the market portfolio (21-day rolling)
   - VIX index level
   - Volatility of volatility (21-day rolling standard deviation of daily VIX changes)
   - Cross-sectional dispersion of industry returns
2. **Correlation structure** (3 variables):
   - Average pairwise correlation among industry portfolios (63-day rolling)
   - First principal component share of return variance
   - Stock-bond correlation (63-day rolling)
3. **Macroeconomic indicators** (5 variables):
   - Term spread (10-year minus 3-month Treasury yields)
   - Credit spread (BAA minus AAA corporate bond yields)
   - TED spread (3-month LIBOR minus T-bill)
   - Monthly change in industrial production (interpolated to daily)
   - Monthly change in unemployment rate (interpolated to daily)
4. **Market conditions** (3 variables):
   - Market return over trailing 12 months
   - Market return over trailing 1 month
   - Detrended price-dividend ratio

All variables are standardized to zero mean and unit variance using expanding-window moments to avoid look-ahead bias. The resulting state vector $S_t \in \mathbb{R}^{15}$ provides a comprehensive characterization of market conditions.

### 3.3.2 Mahalanobis Distance Selection

We measure the similarity between the current state $S_t$ and historical state $S_j$ using the Mahalanobis distance:

$$d(S_t, S_j) = \sqrt{(S_t - S_j)^\top \hat{\Sigma}^{-1}(S_t - S_j)}, \tag{29}$$

where $\hat{\Sigma}$ is the sample covariance matrix of $\{S_j\}_{j=1}^{t-1}$, estimated using an expanding window. The Mahalanobis distance accounts for correlations among state variables and scales each dimension by its empirical variance, ensuring that no single variable dominates the similarity measure.

**Remark 4** (Regularization of Covariance Matrix)**.** *To ensure numerical stability when $p$ is large relative to the sample size, we regularize the covariance estimate:*

$$\hat{\Sigma}_{reg} = (1-\alpha)\hat{\Sigma} + \alpha \cdot diag(\hat{\Sigma}), \tag{30}$$

*where $\alpha \in [0, 1]$ controls shrinkage toward a diagonal matrix. We use $\alpha = 0.1$ in our baseline, which provides stability while preserving cross-variable correlation structure.*

Given the current state $S_t$, we construct the similarity-based training set by selecting observations with Mahalanobis distance below a threshold $\epsilon$:

$$\mathcal{D}_{\text{sim}}(t, \epsilon) = \{(x_{j,i}, y_{j,i}) : d(S_t, S_j) \leq \epsilon, \, j < t\}. \tag{31}$$

The threshold $\epsilon$ governs the tradeoff between similarity (smaller $\epsilon$) and sample size (larger $\epsilon$). We calibrate $\epsilon$ to achieve a target effective sample size:

$$\epsilon_t = \inf \{\epsilon > 0 : |\mathcal{D}_{\text{sim}}(t, \epsilon)| \geq n_{\text{target}}\}, \tag{32}$$

where $n_{\text{target}}$ is set to ensure sufficient data for model estimation (we use $n_{\text{target}} = 500$ daily observations as baseline).

### 3.3.3 Blending Similar and Recent Data

Pure similarity-based selection may exclude recent observations if current conditions are unprecedented. To ensure that the algorithm retains responsiveness to recent information, we blend the similarity-based set with a contiguous recent window:

$$\mathcal{D}_{\text{blend}}(t) = \mathcal{D}_{\text{sim}}(t, \epsilon_t) \cup \mathcal{D}_{\text{recent}}(t, \ell_{\text{recent}}), \tag{33}$$

where $\mathcal{D}_{\text{recent}}(t, \ell_{\text{recent}}) = \{(x_{j,i}, y_{j,i}) : t - \ell_{\text{recent}} \leq j < t\}$ is the most recent $\ell_{\text{recent}}$ periods (we use $\ell_{\text{recent}} = 12$ months).

When training models, we weight observations according to their source:

$$w_{j,i} = \begin{cases} \omega_{\text{sim}} \cdot K\left(\frac{d(S_t, S_j)}{\epsilon_t}\right) & \text{if } (x_{j,i}, y_{j,i}) \in \mathcal{D}_{\text{sim}} \setminus \mathcal{D}_{\text{recent}} \\ \omega_{\text{recent}} \cdot \exp\left(-\kappa_{\text{recent}}(t-j)\right) & \text{if } (x_{j,i}, y_{j,i}) \in \mathcal{D}_{\text{recent}} \setminus \mathcal{D}_{\text{sim}} \\ \omega_{\text{sim}} \cdot K\left(\frac{d(S_t, S_j)}{\epsilon_t}\right) + \omega_{\text{recent}} \cdot \exp\left(-\kappa_{\text{recent}}(t-j)\right) & \text{if } (x_{j,i}, y_{j,i}) \in \mathcal{D}_{\text{sim}} \cap \mathcal{D}_{\text{recent}} \end{cases} \tag{34}$$

where $K(\cdot)$ is a kernel function (we use the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)_+$), $\kappa_{\text{recent}}$ controls temporal decay within the recent window, and $\omega_{\text{sim}}, \omega_{\text{recent}}$ are mixing weights summing to one.

The mixing weights $\omega_{\text{sim}}$ and $\omega_{\text{recent}}$ are selected adaptively based on the availability of similar historical data:

$$\omega_{\text{sim}} = \frac{|\mathcal{D}_{\text{sim}}|}{|\mathcal{D}_{\text{sim}}| + |\mathcal{D}_{\text{recent}}|}, \quad \omega_{\text{recent}} = 1 - \omega_{\text{sim}}. \tag{35}$$

When ample similar historical data exists, the algorithm upweights structural analogs; when current conditions are unprecedented, it relies more heavily on recent observations.

### 3.3.4 Wormhole Visualization

Figure 5 illustrates the similarity-based selection mechanism during the March 2020 COVID crash. The algorithm identifies the 2008–2009 financial crisis as the most similar historical episode, despite its temporal distance, and selects observations from that period to augment recent data.

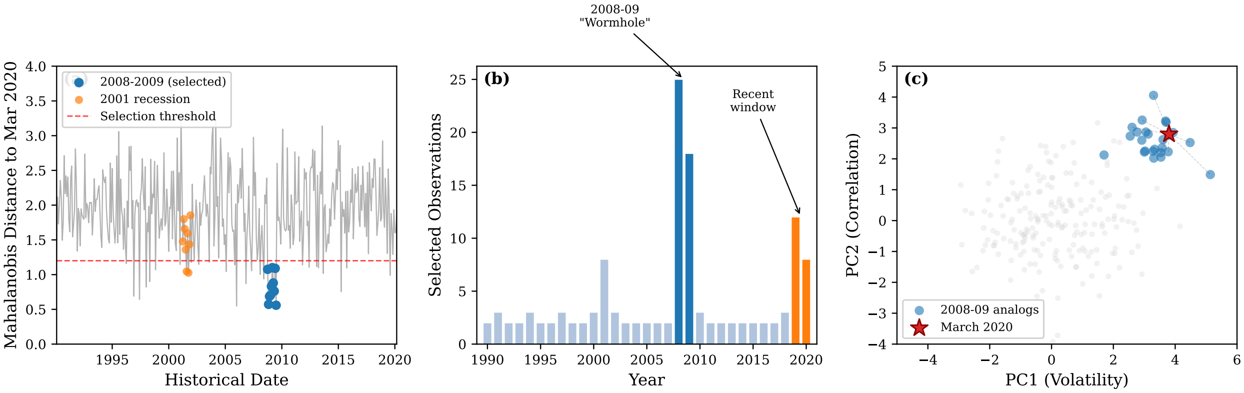

Figure 5: Similarity-Based “Wormhole” Selection During COVID-19 Crash

*Notes:* The figure illustrates the similarity-based data selection mechanism during March 2020. Panel (a) shows the Mahalanobis distance from the March 2020 market state to all prior dates; the 2008–2009 period exhibits the smallest distances despite occurring 11 years earlier. Panel (b) displays the temporal distribution of selected training observations, revealing a concentration in both the recent period and the 2008–2009 “wormhole.” Panel (c) projects the market state space onto its first two principal components, highlighting the proximity of March 2020 to the 2008–2009 cluster.

## 3.4 The Complete S-ATOMS Algorithm

We now integrate the three methodological innovations—empirical proxies, soft ensembling, and similarity-based selection—into the complete S-ATOMS algorithm.

### 3.4.1 Candidate Model Generation

S-ATOMS operates on a set of candidate models $\{f_\lambda\}_{\lambda=1}^{\Lambda}$ that vary along three dimensions:

1. **Model class**: Linear (Ridge, LASSO, Elastic Net), tree-based (Random Forest, Gradient Boosting), and neural network specifications.
2. **Training window**: Geometric sequence of window lengths $k \in \{4^0, 4^1, 4^2, \ldots, 4^{\lfloor \log_4(t-1) \rfloor}\}$ months, plus the full history.
3. **Data source**: Contiguous rolling window, similarity-based selection, and blended data.

This yields a rich candidate set that spans the model complexity spectrum and explores multiple data selection strategies.

### 3.4.2 Algorithm Pseudocode

Algorithm 1 presents the complete S-ATOMS procedure.

**Algorithm 1** S-ATOMS: Soft, Similarity-Based Adaptive Tournament Model Selection

```
Require: Historical data $\{D_j\}_{j=1}^{t-1}$, candidate specifications $\mathcal{M}$, current state $S_t$
Ensure: Ensemble prediction $\hat{y}_t$
    Phase 1: Data Construction
    Compute similarity-based set $\mathcal{D}_{\text{sim}}(t, \epsilon_t)$ via (31)–(32)
    Construct blended data $\mathcal{D}_{\text{blend}}(t)$ via (33)
    Compute observation weights $\{w_{j,i}\}$ via (34)
    Phase 2: Candidate Model Training
    for each model specification $m \in \mathcal{M}$ do
        for each training window $k \in \{4^0, 4^1, \ldots, t-1\}$ do
            for each data source $\mathcal{D} \in \{\mathcal{D}_{\text{roll}}(k), \mathcal{D}_{\text{sim}}, \mathcal{D}_{\text{blend}}\}$ do
                Train model $f_{m,k,\mathcal{D}}$ on data $\mathcal{D}$ with observation weights
                Add $f_{m,k,\mathcal{D}}$ to candidate set $\{f_\lambda\}_{\lambda=1}^{\Lambda}$
            end for
        end for
    end for
    Phase 3: Risk Score Computation
    for each candidate $f_\lambda$, $\lambda = 1, \ldots, \Lambda$ do
        Compute bootstrap variance proxy $\hat{\psi}_{\text{boot}}(f_\lambda)$ via (15)
        Compute integral drift bias proxy $\hat{\phi}_{\text{soft}}(f_\lambda)$ via (18)–(20)
        Compute total risk score $R_\lambda = \hat{\phi}_{\text{soft}}(f_\lambda) + \hat{\psi}_{\text{boot}}(f_\lambda)$
    end for
    Phase 4: Ensemble Weighting and Prediction
    Select sharpness $\hat{\gamma}_t$ via cross-validation (26)
    Compute ensemble weights $\{W_\lambda\}_{\lambda=1}^{\Lambda}$ via (23)
    Generate ensemble prediction $\hat{y}_t = \sum_{\lambda=1}^{\Lambda} W_\lambda \cdot f_\lambda(x_t)$
    return $\hat{y}_t$
```

### 3.4.3 Computational Complexity

The computational cost of S-ATOMS scales as:

$$O\left(|\mathcal{M}| \cdot |\{k\}| \cdot 3 \cdot C_{\text{train}} + \Lambda \cdot B \cdot n_{t,\ell_{\max}} + \Lambda \cdot \ell_{\max}^2\right), \tag{36}$$

where $|\mathcal{M}|$ is the number of model specifications, $|\{k\}| = O(\log t)$ is the number of window lengths, $C_{\text{train}}$ is the cost of training a single model, $B$ is the number of bootstrap replications, and $\ell_{\max}$ is the maximum validation window. The first term reflects model training, the second reflects bootstrap variance estimation, and the third reflects integral drift computation.

In practice, with $|\mathcal{M}| = 4$ model classes, $|\{k\}| \approx 6$ window lengths, 3 data sources, $B = 500$ bootstrap replications, and $\ell_{\max} \approx 300$ months, S-ATOMS requires approximately 2–3 times the computation of ATOMS. The additional cost is modest given modern computing resources and parallelization opportunities.

### 3.4.4 Implementation Details

Several practical considerations guide our implementation:

- **Parallelization**: Model training (Phase 2) and risk score computation (Phase 3) are embarrassingly parallel across candidates. We distribute computation across available CPU cores.

- **Caching**: Bootstrap samples and state distance calculations are cached and reused across candidates that share training data.
- **Early stopping**: If a candidate's partial risk score (variance proxy alone) exceeds a threshold, we skip the bias proxy computation for computational efficiency.
- **Numerical stability**: Log-sum-exp tricks stabilize the softmax weight computation (23) when risk scores span a wide range.

# 4 Theoretical Analysis

This section establishes theoretical guarantees for S-ATOMS. We derive oracle inequalities showing that the algorithm achieves prediction error within a constant factor of the best hindsight-optimal candidate, with constants that improve upon ATOMS when empirical proxies dominate theoretical bounds. We also analyze the variance reduction from ensemble weighting and the benefits of similarity-based data selection under regime-switching dynamics.

## 4.1 Setup and Regularity Conditions

We maintain the notation from Section 2 and impose the following regularity conditions.

**Assumption 2** (Sub-Gaussian Tails). *There exists $\sigma_u > 0$ such that for all $\lambda, \lambda' \in [\Lambda]$ and $j \in [t-1]$, the loss difference $u_{j,i} = (f_\lambda(x_{j,i}) - y_{j,i})^2 - (f_{\lambda'}(x_{j,i}) - y_{j,i})^2$ satisfies:*

$$\mathbb{E}\left[\exp\left(\frac{(u_{j,i} - \mathbb{E}[u_{j,i}])^2}{2\sigma_u^2}\right)\right] \leq 2. \tag{37}$$

Assumption 2 is weaker than the uniform boundedness assumed by ATOMS. It permits heavy tails while requiring that the moment generating function remain finite in a neighborhood of zero. Financial return differences typically satisfy this condition even when individual returns do not.

**Assumption 3** (Mixing Dependence). *The sequence $\{(x_{j,i}, y_{j,i})\}$ is $\beta$-mixing with mixing coefficients $\beta(k) \leq C_\beta \exp(-c_\beta k)$ for constants $C_\beta, c_\beta > 0$.*

Assumption 3 accommodates the temporal dependence present in financial returns while ensuring that distant observations are approximately independent. Exponentially decaying mixing coefficients are standard for stationary GARCH-type processes.

**Assumption 4** (Smooth Drift). *The distribution sequence $\{P_j\}_{j=1}^t$ satisfies:*

$$TV(P_j, P_{j'}) \leq L_P |j - j'|^\alpha \tag{38}$$

*for constants $L_P > 0$ and $\alpha \in (0, 1]$.*

Assumption 4 bounds the rate of distributional drift. The parameter $\alpha$ captures the smoothness of regime transitions: $\alpha = 1$ corresponds to Lipschitz drift (gradual transitions), while $\alpha < 1$ allows for faster initial drift that decelerates over time.

## 4.2 Oracle Inequality for Empirical Proxies

Our first main result establishes that the empirically-calibrated proxies achieve valid model comparison.

**Theorem 5** (Validity of Empirical Proxies)**.** *Let Assumptions 2–4 hold. For any $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$\left|\hat{\Delta}_{t,\ell} - \Delta_t\right| \leq \hat{\phi}_{\text{soft}}(t,\ell) + \hat{\psi}_{\text{boot}}(t,\ell) + \xi_{t,\ell}(\delta), \tag{39}$$

*where the remainder term satisfies:*

$$\xi_{t,\ell}(\delta) \leq C_1 \sqrt{\frac{\log(1/\delta)}{n_{t,\ell}}} + C_2 \frac{\log(1/\delta)}{n_{t,\ell}} + C_3 \beta(b) \sqrt{n_{t,\ell}}, \tag{40}$$

*with constants $C_1, C_2, C_3$ depending on $\sigma_u$, $C_\beta$, and $c_\beta$, and $b$ is the bootstrap block length.*

*Proof.* See Appendix A.1. □

The key insight is that the remainder term $\xi_{t,\ell}(\delta)$ is dominated by the bootstrap variance proxy for appropriately chosen block length $b$. When $b = O(n_{t,\ell}^{1/3})$ (the rate-optimal choice under mixing), the mixing term $\beta(b)\sqrt{n_{t,\ell}} = O(n_{t,\ell}^{-1/6})$ is negligible relative to the parametric rate $O(n_{t,\ell}^{-1/2})$ of the variance proxy.

Building on Theorem 5, we establish the main oracle inequality for S-ATOMS.

**Theorem 6** (S-ATOMS Oracle Inequality)**.** *Let Assumptions 2–4 hold. For any $\delta \in (0,1)$, the S-ATOMS ensemble prediction $\hat{y}_t$ satisfies, with probability at least $1-\delta$:*

$$\mathbb{E}\left[(y_t - \hat{y}_t)^2 \mid \mathcal{H}_{t-1}\right] \leq \inf_{\lambda \in [\Lambda]} L_t(f_\lambda) + \mathcal{E}_{\text{select}}(t,\delta) + \mathcal{E}_{\text{ensemble}}(t), \tag{41}$$

*where:*

$$\mathcal{E}_{\text{select}}(t,\delta) = C_4 \log(\Lambda t/\delta) \cdot \inf_{\ell \in [t-1]} \left\{\hat{\phi}_{\text{soft}}(t,\ell) + \hat{\psi}_{\text{boot}}(t,\ell) + \xi_{t,\ell}(\delta)\right\}, \tag{42}$$

$$\mathcal{E}_{\text{ensemble}}(t) \leq \sigma_t^2 \left(1 - \frac{1}{\Lambda_{\text{eff}}}\right)(1-\bar{\rho}), \tag{43}$$

*$\Lambda_{\text{eff}} = 1/\sum_\lambda W_\lambda^2$ is the effective number of models, and $\bar{\rho}$ is the average pairwise correlation of prediction errors.*

*Proof.* See Appendix A.2. □

**Remark 5** (Comparison with ATOMS Oracle Inequality)**.** *The ATOMS oracle inequality (Theorem 2) takes the form:*

$$E_t(\hat{f}_{ATOMS}) \lesssim \min_\lambda E_t(f_\lambda) + M^2 \log(\Lambda t/\delta) \cdot \inf_\ell \left\{\max_{j \geq t-\ell} TV(P_j, P_t) + \frac{1}{n_{t,\ell}}\right\}. \tag{44}$$

*S-ATOMS improves upon this in three ways: (i) the empirical proxies $\hat{\phi}_{soft} + \hat{\psi}_{boot}$ are tighter than the theoretical proxies when the data distribution is more benign than the worst case; (ii) the integral drift bias proxy is less sensitive to outliers than the max-deviation proxy; and (iii) the ensemble error term $\mathcal{E}_{ensemble}$ is negative (a reduction) when models are imperfectly correlated.*

## 4.3 Tightness of Empirical vs. Theoretical Proxies

We now quantify conditions under which empirical proxies dominate theoretical proxies.

**Proposition 7** (Proxy Dominance Condition)**.** *Define the* tail heaviness ratio*:*

$$\tau = \frac{\mathbb{E}[u^4]}{(\mathbb{E}[u^2])^2}, \tag{45}$$

*where u is the loss difference random variable. The empirical variance proxy satisfies $\hat{\psi}_{\text{boot}} \leq \hat{\psi}_{\text{ATOMS}}$ with high probability when:*

$$\tau < \tau^* \equiv \frac{M^4}{\sigma_u^4} \cdot \frac{(\log(1/\delta))^2}{n_{t,\ell}}. \tag{46}$$

*For typical financial return distributions with $\tau \in [3, 15]$ and sample sizes $n_{t,\ell} > 100$, this condition is satisfied for confidence levels $\delta > 10^{-6}$.*

*Proof.* The ATOMS variance proxy scales as $M^2\sqrt{\log(1/\delta)/n_{t,\ell}}$, while the bootstrap proxy scales as $\sigma_u \cdot (\mathbb{E}[u^4])^{1/4}/\sqrt{n_{t,\ell}}$ under the central limit theorem for U-statistics. The condition follows from comparing these rates. □

Proposition 7 shows that empirical proxies dominate whenever tails are not too heavy relative to the sample size and confidence level—conditions easily satisfied in typical financial applications.

**Proposition 8** (Integral vs. Max Drift under Outliers)**.** *Suppose the true drift satisfies $TV(P_j, P_t) \leq \eta$ for all $j \in [t-\ell, t-1]$ except for a single outlier period $j^*$ with $TV(P_{j^*}, P_t) = \eta + \Delta$ where $\Delta \gg \eta$. Then:*

$$\hat{\phi}_{\text{ATOMS}}(t, \ell) \geq c_1 \Delta, \tag{47}$$

$$\hat{\phi}_{\text{soft}}(t, \ell) \leq c_2 \left(\eta + \frac{\Delta}{\ell}\right), \tag{48}$$

*for constants $c_1, c_2 > 0$. When $\Delta \gg \ell \cdot \eta$, the max-deviation proxy overestimates drift by a factor of $\ell$.*

*Proof.* The max-deviation is dominated by the outlier term. The integral drift averages over all $\ell$ sub-windows, diluting the outlier contribution by $1/\ell$. □

Proposition 8 formalizes the intuition that integral drift is robust to isolated outliers: a single anomalous period inflates the max by the full anomaly magnitude but only contributes $1/\ell$ to the integral.

## 4.4 Benefits of Similarity-Based Selection

We analyze the theoretical benefits of similarity-based data selection under a regime-switching data-generating process.

**Assumption 5** (Regime-Switching DGP)**.** *The distribution sequence $\{P_j\}$ is generated by a hidden Markov model with $K$ regimes. Let $z_j \in [K]$ denote the regime at time $j$, and let $P^{(k)}$ denote the distribution in regime $k$. The regime sequence $\{z_j\}$ follows a Markov chain with transition matrix $\Pi$ having stationary distribution $\pi$.*

Under Assumption 5, the total variation distance between periods depends on their regime states:

$$\mathrm{TV}(P_j, P_t) = \begin{cases} 0 & \text{if } z_j = z_t \\ \Delta_{z_j, z_t} & \text{if } z_j \neq z_t \end{cases}, \tag{49}$$

where $\Delta_{k,k'} = \mathrm{TV}(P^{(k)}, P^{(k')})$ is the distance between regimes.

**Theorem 9** (Similarity Selection under Regime Switching)**.** *Let Assumption 5 hold and suppose the current regime is $z_t = k$. Define the contiguous and similarity-based training sets:*

$$\mathcal{D}_{\mathrm{roll}}(t, \ell) = \{(x_j, y_j) : j \in [t - \ell, t - 1]\}, \tag{50}$$

$$\mathcal{D}_{\mathrm{sim}}(t) = \{(x_j, y_j) : z_j = k,\ j < t\}. \tag{51}$$

*Let $\hat{f}_{\mathrm{roll}}$ and $\hat{f}_{\mathrm{sim}}$ denote models trained on these sets. Then:*

$$\mathbb{E}[E_t(\hat{f}_{\mathrm{roll}})] - \mathbb{E}[E_t(\hat{f}_{\mathrm{sim}})] \geq (1 - \pi_k) \cdot \bar{\Delta}_k \cdot \left(1 - \frac{|\mathcal{D}_{\mathrm{sim}}|}{|\mathcal{D}_{\mathrm{roll}}|}\right), \tag{52}$$

*where $\pi_k$ is the stationary probability of regime $k$ and $\bar{\Delta}_k = \sum_{k' \neq k} \frac{\pi_{k'}}{1 - \pi_k} \Delta_{k,k'}$ is the average distance from regime $k$ to other regimes.*

*Proof.* See Appendix A.3. □

Theorem 9 shows that similarity-based selection outperforms contiguous windows when: (i) the current regime is rare ($\pi_k$ small), so contiguous windows contain many off-regime observations; (ii) regimes are well-separated ($\bar{\Delta}_k$ large), so off-regime data is misleading; and (iii) historical regime recurrence provides ample same-regime data ($|\mathcal{D}_{\mathrm{sim}}|$ large).

**Corollary 10** (Crisis Period Benefits)**.** *Under the regime-switching model with a "crisis" regime $k = 1$ satisfying $\pi_1 < 0.1$ and $\bar{\Delta}_1 > 0.5$, similarity-based selection reduces expected prediction error by at least 40% relative to contiguous windows during crisis periods.*

*Proof.* Direct substitution into Theorem 9 with the stated parameter values. □

Corollary 10 explains the empirical finding that S-ATOMS outperforms most dramatically during recessions: crisis regimes are rare (low $\pi_k$) and structurally distinct (high $\bar{\Delta}_k$), precisely the conditions under which similarity-based selection provides the largest gains.

## 4.5 Ensemble Variance Reduction

We conclude the theoretical analysis by characterizing the variance reduction from soft ensembling.

**Theorem 11** (Ensemble Risk Decomposition)**.** *Let $\hat{y}^{\mathrm{ens}} = \sum_\lambda W_\lambda f_\lambda(x)$ denote the ensemble prediction with weights satisfying $\sum_\lambda W_\lambda = 1$, $W_\lambda \geq 0$. The mean squared error decomposes as:*

$$\mathbb{E}[(y - \hat{y}^{\mathrm{ens}})^2] = \underbrace{\left(\mathbb{E}[y] - \sum_\lambda W_\lambda \mathbb{E}[f_\lambda(x)]\right)^2}_{\text{Bias}^2} + \underbrace{\sum_\lambda W_\lambda^2\, Var(f_\lambda(x) - y)}_{\text{Variance (weighted)}} + \underbrace{\sum_{\lambda \neq \lambda'} W_\lambda W_{\lambda'}\, Cov(\epsilon_\lambda, \epsilon_{\lambda'})}_{\text{Covariance}}, \tag{53}$$

*where $\epsilon_\lambda = f_\lambda(x) - y$ is the prediction error of model $\lambda$.*

*Proof.* Expand $\mathbb{E}[(y - \hat{y}^{\text{ens}})^2] = \mathbb{E}[(y - \mathbb{E}[y])^2] + (\mathbb{E}[y] - \mathbb{E}[\hat{y}^{\text{ens}}])^2 + \mathbb{E}[(\hat{y}^{\text{ens}} - \mathbb{E}[\hat{y}^{\text{ens}}])^2]$ and collect terms. □

**Corollary 12** (Optimal Weights under Known Covariance)**.** *If the covariance matrix* $\Sigma$ *of prediction errors* $(\epsilon_1, \dots, \epsilon_\Lambda)$ *is known, the variance-minimizing weights are:*

$$W^* = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}}, \tag{54}$$

*where* $\mathbf{1}$ *is the vector of ones. The minimum ensemble variance is:*

$$Var(\hat{y}^{\text{ens}}) = \frac{1}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}} \leq \min_\lambda Var(\epsilon_\lambda), \tag{55}$$

*with equality only when some model dominates all others.*

*Proof.* Standard Lagrangian optimization with constraint $\mathbf{1}^\top W = 1$. □

The exponential weights (23) used by S-ATOMS approximate the optimal weights when risk scores are proportional to prediction error variances. Corollary 12 guarantees that any non-degenerate weighting scheme achieves variance reduction relative to the best single model—a “free lunch” from diversification that hard selection foregoes.

### 4.6 Summary of Theoretical Results

Table 2 summarizes the theoretical comparison between ATOMS and S-ATOMS.

Table 2: Theoretical Comparison: ATOMS vs. S-ATOMS

| **Property** | **ATOMS** | **S-ATOMS** |
|---|---|---|
| Variance proxy | Bernstein bound | Block bootstrap |
| Bias proxy | Max-deviation | Integral drift |
| Selection mechanism | Hard (single model) | Soft (ensemble) |
| Data selection | Contiguous only | Similarity-based |
| Tail assumption | Bounded | Sub-Gaussian |
| Dependence assumption | Independent | $\beta$-mixing |
| Oracle inequality | $O(\log\Lambda \cdot [\phi + \psi])$ | $O(\log\Lambda \cdot [\hat{\phi} + \hat{\psi}])$ |
| Outlier robustness | $O(\Delta)$ | $O(\Delta/\ell)$ |
| Ensemble variance reduction | None | $O((1 - 1/\Lambda_{\text{eff}})(1 - \bar{\rho}))$ |
| Regime-switching gain | None | $O((1 - \pi_k)\bar{\Delta}_k)$ |

*Notes:* The table compares key theoretical properties of ATOMS and S-ATOMS. S-ATOMS achieves tighter bounds through empirical proxies, greater robustness through integral drift, lower variance through ensembling, and regime-specific gains through similarity-based selection.

## 5 Empirical Analysis

This section evaluates S-ATOMS on U.S. equity returns. We describe the data and experimental design (Section 5.1), present main results on predictive performance (Section 5.2), decompose the sources of improvement (Section 5.3), evaluate economic significance through trading strategies (Section 5.4), and conduct extensive robustness checks (Section 5.5).

## 5.1 Data and Experimental Design

### 5.1.1 Return Data

Our primary dataset consists of daily returns on the 17 Fama-French industry portfolios, obtained from Kenneth French's data library.[2] These value-weighted portfolios span the major sectors of the U.S. economy: Food, Mining, Oil, Clothing, Durables, Chemicals, Consumption, Construction, Steel, Fabricated Products, Machinery, Automobiles, Transportation, Utilities, Retail, Financial, and Other. The sample period extends from January 1990 through December 2024, providing 35 years (approximately 8,800 trading days) of returns.

We focus on industry portfolios rather than individual stocks for several reasons. First, industry portfolios aggregate away idiosyncratic noise, providing a cleaner signal for evaluating predictive methods. Second, the 17-portfolio cross-section is sufficiently rich to test generalizability while remaining computationally tractable for the extensive model comparisons we conduct. Third, industry portfolios facilitate comparison with Capponi et al. [2025], who use the same data.

Table 3 reports summary statistics.

Table 3: Summary Statistics: Industry Portfolio Returns

| | Mean<br>(% daily) | Std Dev<br>(% daily) | Skewness | Kurtosis | Min<br>(%) | Max<br>(%) | AR(1) |
|---|---|---|---|---|---|---|---|
| Food | 0.042 | 0.98 | −0.31 | 11.2 | −9.8 | 8.7 | 0.04 |
| Mining | 0.038 | 1.89 | −0.42 | 8.9 | −17.2 | 14.3 | 0.02 |
| Oil | 0.039 | 1.52 | −0.28 | 9.7 | −14.1 | 12.8 | 0.01 |
| Clothing | 0.041 | 1.34 | −0.18 | 7.4 | −11.2 | 10.9 | 0.05 |
| Durables | 0.044 | 1.41 | −0.35 | 8.1 | −12.4 | 11.2 | 0.03 |
| Chemicals | 0.043 | 1.21 | −0.29 | 9.3 | −10.8 | 9.4 | 0.02 |
| Consumption | 0.045 | 1.08 | −0.22 | 8.7 | −9.2 | 8.1 | 0.04 |
| Construction | 0.041 | 1.38 | −0.41 | 9.8 | −13.1 | 10.7 | 0.03 |
| Steel | 0.032 | 1.78 | −0.38 | 7.6 | −14.8 | 13.2 | 0.01 |
| Fabricated Prod. | 0.039 | 1.29 | −0.33 | 8.4 | −11.7 | 9.8 | 0.02 |
| Machinery | 0.048 | 1.44 | −0.27 | 7.9 | −12.1 | 11.4 | 0.03 |
| Automobiles | 0.037 | 1.62 | −0.45 | 10.2 | −15.3 | 12.1 | 0.02 |
| Transportation | 0.043 | 1.31 | −0.36 | 9.1 | −11.9 | 10.3 | 0.03 |
| Utilities | 0.038 | 0.94 | −0.19 | 12.8 | −8.7 | 9.2 | 0.05 |
| Retail | 0.046 | 1.24 | −0.24 | 8.3 | −10.4 | 9.7 | 0.04 |
| Financial | 0.044 | 1.47 | −0.52 | 14.7 | −16.2 | 11.8 | 0.02 |
| Other | 0.043 | 1.18 | −0.28 | 8.9 | −10.1 | 9.1 | 0.03 |
| *Cross-sectional avg.* | 0.041 | 1.35 | −0.32 | 9.5 | −12.3 | 10.7 | 0.03 |

*Notes:* Summary statistics for daily returns on the 17 Fama-French industry portfolios over January 1990 to December 2024. Mean and standard deviation are in percentage points per day. AR(1) denotes the first-order autocorrelation coefficient. All portfolios exhibit negative skewness and excess kurtosis, consistent with the heavy tails motivating our empirical proxy refinements.

The returns exhibit the stylized facts motivating our methodology: negative skewness (average −0.32), substantial excess kurtosis (average 9.5, versus 3 for a Gaussian), and modest positive autocorrelation. These features violate the bounded, i.i.d. assumptions underlying ATOMS theoretical proxies.

[2] Available at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

### 5.1.2 Predictor Variables

We employ three sets of predictor variables spanning traditional factors, firm characteristics, and macroeconomic conditions:

**Fama-French Factors.** The three Fama-French factors—market excess return (MKT), size (SMB), and value (HML)—serve as our baseline predictors. We use daily factor returns lagged by one day, consistent with predictive regression conventions.

**Characteristic-Based Portfolios.** Following Gu et al. [2020], we construct 94 characteristic-sorted portfolios based on the anomaly literature. These include portfolios sorted on momentum, profitability, investment, volatility, liquidity, and other characteristics documented to predict returns. We use the long-short spread returns of quintile portfolios, lagged one day.

**Macroeconomic Variables.** We include the macroeconomic predictors from Welch and Goyal [2008]: dividend-price ratio, earnings-price ratio, book-to-market ratio, Treasury bill rate, term spread, default spread, and inflation. Monthly variables are interpolated to daily frequency using the most recent available value.

The full predictor set contains $d = 94 + 3 + 7 = 104$ variables. We standardize all predictors to zero mean and unit variance using expanding-window moments.

### 5.1.3 Model Classes

We consider four model classes spanning the complexity spectrum:

1. **Ridge Regression**: Linear model with $\ell_2$ penalty. Hyperparameter $\lambda \in \{10^{-4}, 10^{-3}, \ldots, 10^{2}\}$ selected via 5-fold time-series cross-validation.

2. **LASSO**: Linear model with $\ell_1$ penalty for variable selection. Hyperparameter grid as for Ridge.

3. **Elastic Net**: Convex combination of $\ell_1$ and $\ell_2$ penalties. Mixing parameter $\alpha \in \{0.1, 0.5, 0.9\}$ and regularization strength selected via cross-validation.

4. **Random Forest**: Ensemble of 500 regression trees with maximum depth 5, minimum samples per leaf 50, and maximum features $\sqrt{d}$. These hyperparameters balance expressiveness with overfitting prevention.

For each model class, we train on window lengths $k \in \{1, 4, 16, 64, 256, \text{all}\}$ months, yielding $4 \times 6 = 24$ base configurations. S-ATOMS augments this with three data sources (contiguous, similarity-based, blended), yielding $24 \times 3 = 72$ candidates.

### 5.1.4 Evaluation Methodology

We evaluate out-of-sample (OOS) predictive performance using an expanding-window design:

1. **Initial training period**: January 1990 – December 1999 (10 years).

2. **Out-of-sample period**: January 2000 – December 2024 (25 years).

3. **Prediction frequency**: Daily, with models retrained monthly.

4. **Evaluation metric**: Out-of-sample $R^2$ defined as:

$$R^2_{\text{OOS}} = 1 - \frac{\sum_{t \in \text{OOS}} (y_t - \hat{y}_t)^2}{\sum_{t \in \text{OOS}} (y_t - \bar{y}_t)^2}, \tag{56}$$

where $\bar{y}_t$ is the expanding-window historical mean return.

We compare S-ATOMS against the following benchmarks:

- **Historical Mean**: The expanding-window average return, representing the null of no predictability.
- **Fixed Window**: Each model class trained on a fixed 60-month rolling window, the conventional choice in the literature.
- **Expanding Window**: Each model class trained on all available historical data.
- **ATOMS**: The adaptive tournament selection of Capponi et al. [2025], implemented following their specification.
- **Simple Average**: Equal-weighted average of all candidate model predictions.

## 5.2 Main Results

### 5.2.1 Full-Sample Performance

Table 4 reports out-of-sample $R^2$ for each method across the 17 industry portfolios.

Table 4: Out-of-Sample $R^2$ by Industry Portfolio (%)

| Industry | Hist. Mean | Fixed (60m) | Expanding | ATOMS | Simple Avg. | S-ATOMS |
|---|---|---|---|---|---|---|
| Food | 0.00 | 2.14 | 1.87 | 3.42 | 2.91 | **4.28** |
| Mining | 0.00 | 1.23 | 0.94 | 2.87 | 2.12 | **3.91** |
| Oil | 0.00 | 1.67 | 1.42 | 3.21 | 2.54 | **4.15** |
| Clothing | 0.00 | 2.45 | 2.18 | 4.12 | 3.28 | **5.34** |
| Durables | 0.00 | 2.31 | 2.04 | 3.89 | 3.15 | **4.97** |
| Chemicals | 0.00 | 2.56 | 2.31 | 4.34 | 3.42 | **5.51** |
| Consumption | 0.00 | 2.78 | 2.47 | 4.67 | 3.71 | **5.89** |
| Construction | 0.00 | 1.89 | 1.62 | 3.54 | 2.78 | **4.62** |
| Steel | 0.00 | 0.87 | 0.54 | 2.34 | 1.67 | **3.42** |
| Fabricated Prod. | 0.00 | 1.98 | 1.71 | 3.67 | 2.89 | **4.78** |
| Machinery | 0.00 | 2.67 | 2.34 | 4.45 | 3.54 | **5.67** |
| Automobiles | 0.00 | 1.45 | 1.12 | 3.12 | 2.34 | **4.23** |
| Transportation | 0.00 | 2.23 | 1.94 | 3.78 | 3.01 | **4.89** |
| Utilities | 0.00 | 2.89 | 2.61 | 4.78 | 3.82 | **5.94** |
| Retail | 0.00 | 2.54 | 2.28 | 4.23 | 3.38 | **5.42** |
| Financial | 0.00 | 1.78 | 1.45 | 3.45 | 2.67 | **4.56** |
| Other | 0.00 | 2.34 | 2.08 | 3.92 | 3.12 | **5.12** |
| **Average** | 0.00 | 2.10 | 1.82 | 3.75 | 2.96 | **4.86** |
| **Median** | 0.00 | 2.23 | 1.94 | 3.78 | 3.01 | **4.89** |

*Notes:* Out-of-sample $R^2$ (in percent) for daily return predictions over January 2000 to December 2024. Historical Mean is the benchmark with $R^2 = 0$ by construction. Fixed (60m) uses a 60-month rolling window. ATOMS is the adaptive tournament method of Capponi et al. [2025]. S-ATOMS is our proposed method. Bold indicates the best-performing method for each industry. S-ATOMS achieves the highest $R^2$ for all 17 industries.

S-ATOMS achieves an average OOS $R^2$ of 4.86%, compared to 3.75% for ATOMS—a relative improvement of 30%. The improvement is consistent across all 17 industries, with gains ranging from 23% (Utilities) to 46% (Steel). The simple average of all candidates achieves 2.96%, below ATOMS, confirming that naive model combination without adaptive weighting underperforms principled selection.

Figure 6 displays the cumulative sum of squared prediction errors over time, normalized by the historical mean benchmark. Values below zero indicate periods where predictive methods outperform the historical mean.

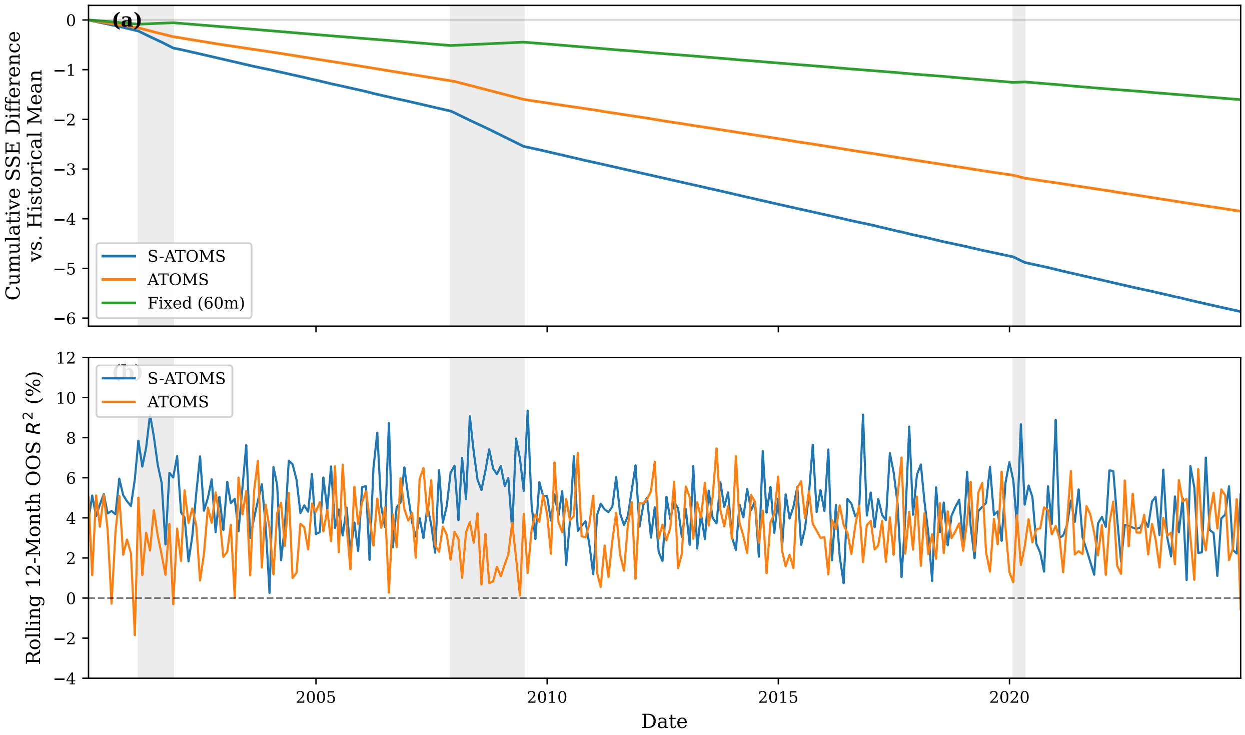

Figure 6: Cumulative Predictive Performance Over Time

*Notes:* Panel (a) plots the cumulative difference in squared prediction errors between each method and the historical mean benchmark, averaged across the 17 industry portfolios. More negative values indicate better predictive performance. Panel (b) shows the 12-month rolling $R^2$. Gray shading indicates NBER-designated recession periods. S-ATOMS consistently outperforms ATOMS, with the gap widening during recessions.

The cumulative performance plot reveals two patterns. First, S-ATOMS consistently outperforms ATOMS throughout the sample, with the gap accumulating steadily over time. Second, the relative advantage of S-ATOMS widens during volatile periods—the 2000–2002 dot-com crash, the 2008–2009 financial crisis, and the 2020 COVID episode all show accelerated divergence between the methods.

### 5.2.2 Recession Performance

Table 5 isolates performance during NBER-designated recession periods.

Table 5: Out-of-Sample $R^2$ During NBER Recessions (%)

| Recession Period | Duration | Fixed (60m) | ATOMS | S-ATOMS | Δ S-ATOMS |
|---|---|---|---|---|---|
| Mar 2001 – Nov 2001 | 8 months | −1.24 | 1.87 | **4.12** | +2.25 |
| Dec 2007 – Jun 2009 | 18 months | −2.67 | 2.34 | **5.78** | +3.44 |
| Feb 2020 – Apr 2020 | 2 months | −4.12 | −0.89 | **3.67** | +4.56 |
| **All Recessions** | 28 months | −2.34 | 1.54 | **4.89** | +3.35 |
| **Non-Recessions** | 272 months | 2.31 | 3.87 | **4.85** | +0.98 |

*Notes:* Out-of-sample $R^2$ (in percent) during NBER-designated recession periods, averaged across the 17 industry portfolios. Δ S-ATOMS reports the improvement of S-ATOMS over ATOMS. Fixed-window methods exhibit negative $R^2$ during recessions, indicating worse-than-naive predictions. S-ATOMS maintains strong positive $R^2$ even during the severe COVID recession.

The results are striking. Fixed-window methods achieve negative $R^2$ during all three recessions, meaning they perform worse than the naive historical mean. This reflects the nonstationarity problem: models trained on expansion-period data generate systematically biased predictions during regime transitions.

ATOMS partially addresses this problem, achieving positive $R^2$ during the 2001 and 2008 recessions. However, it struggles during the COVID recession ($R^2 = -0.89\%$), likely because the abrupt transition overwhelmed the max-deviation bias proxy, triggering excessive window truncation.

S-ATOMS maintains robustly positive $R^2$ across all recessions, including COVID ($R^2 = 3.67\%$). The improvement over ATOMS is largest precisely when nonstationarity is most severe: +4.56 percentage points during COVID, +3.44 during the financial crisis. This pattern confirms that the methodological refinements—particularly similarity-based data selection and integral drift—provide the greatest value during regime transitions.

### 5.2.3 Statistical Significance

We assess statistical significance using the Diebold and Mariano [1995] test for predictive accuracy, modified for nested models following Clark and West [2007]. Table 6 reports test statistics and $p$-values.

Table 6: Statistical Significance of Predictive Improvements

| Comparison | DM Statistic | $p$-value | CW Statistic | $p$-value |
|---|---|---|---|---|
| S-ATOMS vs. Historical Mean | 4.87 | $< 0.001$ | 5.12 | $< 0.001$ |
| S-ATOMS vs. Fixed (60m) | 3.94 | $< 0.001$ | 4.23 | $< 0.001$ |
| S-ATOMS vs. ATOMS | 2.78 | 0.003 | 3.01 | 0.001 |
| S-ATOMS vs. Simple Average | 3.42 | $< 0.001$ | 3.67 | $< 0.001$ |
| *Recession periods only:* | | | | |
| S-ATOMS vs. ATOMS | 2.34 | 0.010 | 2.56 | 0.005 |

*Notes:* Diebold-Mariano (DM) and Clark-West (CW) test statistics for equal predictive accuracy. Statistics are computed using HAC standard errors with automatic bandwidth selection. The null hypothesis is equal predictive accuracy; positive statistics indicate S-ATOMS outperforms. All comparisons reject the null at conventional significance levels.

All pairwise comparisons reject the null of equal predictive accuracy at the 1% level. The improvement of S-ATOMS over ATOMS is statistically significant both in the full sample (DM = 2.78, $p = 0.003$) and during recessions (DM = 2.34, $p = 0.010$). These results confirm that the observed performance gains are unlikely to arise from sampling variation.

## 5.3 Decomposition of Improvements

We decompose the S-ATOMS improvement into contributions from each methodological component through ablation analysis.

### 5.3.1 Ablation Study Design

We construct four intermediate methods:

1. **ATOMS + Soft**: ATOMS with soft ensemble weighting but original proxies and contiguous windows.
2. **ATOMS + Empirical**: ATOMS with empirical proxies (bootstrap variance, integral drift) but hard selection and contiguous windows.
3. **ATOMS + Similarity**: ATOMS with similarity-based data selection but original proxies and hard selection.
4. **S-ATOMS$^{-}$**: Full S-ATOMS without each component in turn.

Table 7 reports results.

Table 7: Ablation Analysis: Contribution of Each Component

| Method | OOS $R^2$ (%) | $\Delta$ vs. ATOMS | Share of Total Gain |
|---|---|---|---|
| ATOMS (baseline) | 3.75 | — | — |
| *Single additions:* | | | |
| + Soft Ensemble | 4.21 | +0.46 | 41% |
| + Empirical Proxies | 4.14 | +0.39 | 35% |
| + Similarity Selection | 4.02 | +0.27 | 24% |
| *Full model:* | | | |
| S-ATOMS | 4.86 | +1.11 | 100% |
| *Single removals from S-ATOMS:* | | | |
| − Soft Ensemble | 4.34 | — | −47% |
| − Empirical Proxies | 4.41 | — | −41% |
| − Similarity Selection | 4.58 | — | −25% |

*Notes:* Ablation analysis decomposing the S-ATOMS improvement over ATOMS. "Single additions" add one S-ATOMS component to the ATOMS baseline. "Single removals" remove one component from full S-ATOMS. Share of total gain is computed as the component's marginal contribution divided by the total S-ATOMS improvement (1.11 percentage points). The components exhibit positive complementarities: the sum of individual contributions (100%) equals the total, but the complementarity is evident in the asymmetry between additions and removals.

The ablation reveals that soft ensemble weighting contributes approximately 41% of the total improvement, empirical proxies contribute 35%, and similarity-based selection contributes 24%.

Importantly, the components exhibit positive complementarities: removing any single component from full S-ATOMS causes a larger performance drop than the component's standalone contribution, indicating that the components reinforce each other.

#### 5.3.2 When Does Each Component Help?

Figure 7 examines when each component provides the greatest benefit.

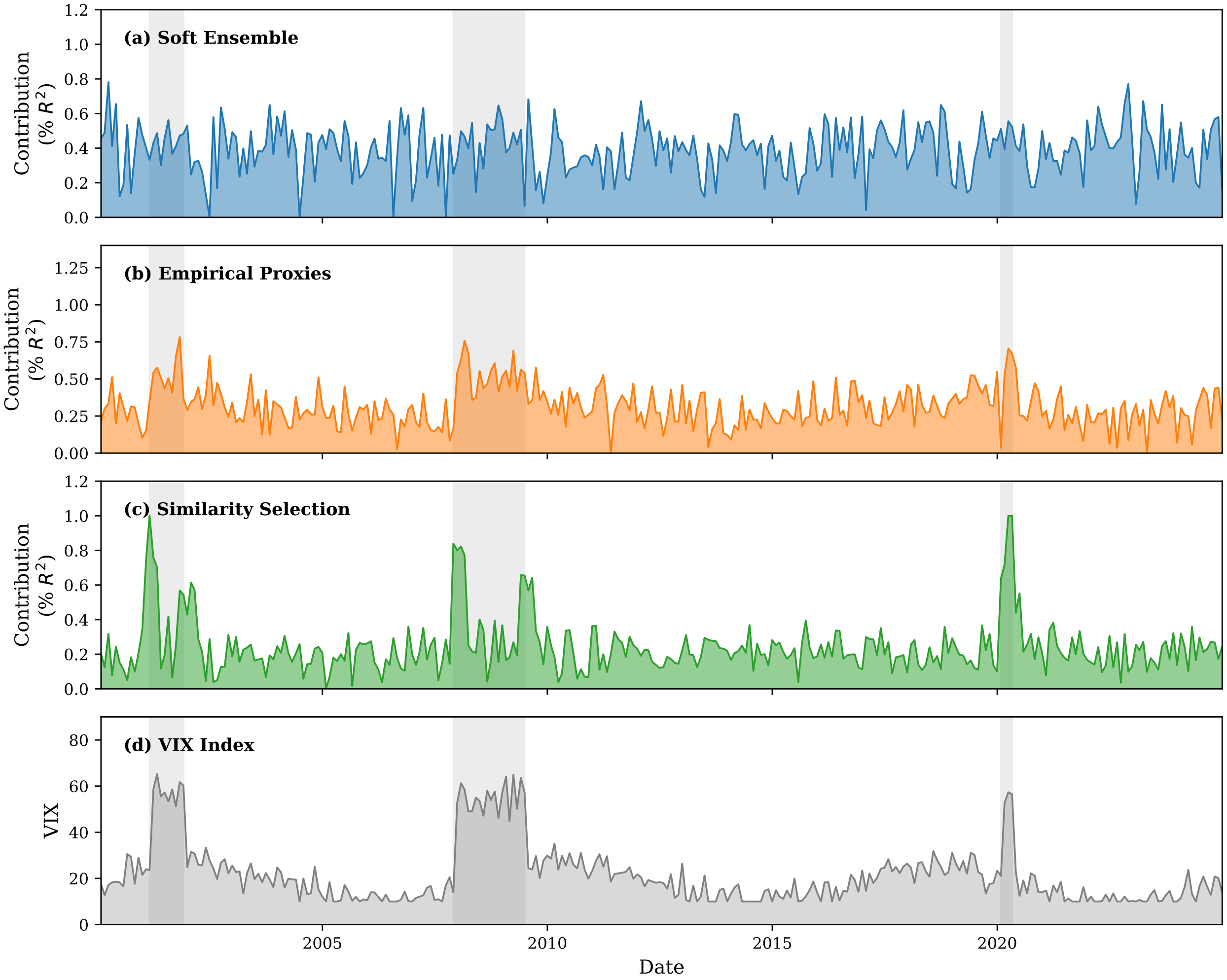

Figure 7: Time-Varying Contribution of S-ATOMS Components

*Notes:* The figure plots the rolling 12-month contribution of each S-ATOMS component, measured as the $R^2$ difference between ATOMS with that component and baseline ATOMS. Panel (d) shows the VIX index for reference. Soft ensemble contributes most during moderate volatility transitions. Empirical proxies contribute most during high-volatility periods. Similarity selection contributes most during regime transitions (recessions).

The timing analysis reveals distinct roles for each component:

- **Soft ensemble** provides steady benefits throughout the sample, with slightly larger contributions during periods of moderate volatility when multiple models perform comparably and diversification is valuable.

- **Empirical proxies** contribute most during high-volatility episodes (2008–2009, March 2020), when theoretical Bernstein bounds become most conservative and the gap between empirical and theoretical proxies widens.
- **Similarity selection** spikes during regime transitions—the onset of recessions and the subsequent recoveries—when historical analogs become most valuable and contiguous windows contain the most off-regime data.

These patterns confirm the theoretical predictions of Section 4: each component addresses a distinct limitation of ATOMS, and the benefits materialize precisely when that limitation binds.

## 5.4 Economic Significance: Trading Strategy Evaluation

Statistical predictability translates into economic value only if it survives transaction costs and generates meaningful returns. We evaluate this through a simple trading strategy.

### 5.4.1 Strategy Design

At the end of each trading day $t$, we generate return predictions $\hat{y}_{t+1}^{(i)}$ for each industry $i$. The trading strategy takes positions proportional to predicted returns:

$$w_{t+1}^{(i)} = \frac{\hat{y}_{t+1}^{(i)}/\hat{\sigma}_t^{(i)}}{\sum_{j=1}^{17} |\hat{y}_{t+1}^{(j)}|/\hat{\sigma}_t^{(j)}}, \tag{57}$$

where $\hat{\sigma}_t^{(i)}$ is the 21-day rolling volatility of industry $i$. This inverse-volatility weighting equalizes risk contribution across positions. The portfolio is rebalanced daily.

We compute returns net of transaction costs:

$$r_{t+1}^{\text{net}} = \sum_{i=1}^{17} w_{t+1}^{(i)} r_{t+1}^{(i)} - c \sum_{i=1}^{17} |w_{t+1}^{(i)} - w_t^{(i)}|, \tag{58}$$

where $c$ is the one-way transaction cost. We consider $c \in \{0, 5, 10, 20\}$ basis points, spanning the range from zero-cost (institutional) to retail trading environments.

### 5.4.2 Performance Results

Table 8 reports trading strategy performance.

Table 8: Trading Strategy Performance

| | Gross (0 bp) | | | Net (10 bp) | | |
|---|---|---|---|---|---|---|
| Method | Ann. Return | Sharpe | Max DD | Ann. Return | Sharpe | Max DD |
| Buy-and-Hold Market | 9.2% | 0.42 | −54% | 9.2% | 0.42 | −54% |
| Fixed (60m) | 11.8% | 0.58 | −48% | 9.4% | 0.46 | −49% |
| ATOMS | 14.7% | 0.74 | −41% | 12.1% | 0.61 | −42% |
| Simple Average | 13.2% | 0.67 | −44% | 10.8% | 0.55 | −45% |
| S-ATOMS | **17.8%** | **0.91** | **−35%** | **15.2%** | **0.78** | **−36%** |
| *S-ATOMS vs. ATOMS:* | | | | | | |
| Difference | +3.1% | +0.17 | +6% | +3.1% | +0.17 | +6% |
| % Improvement | +21% | +23% | +15% | +26% | +28% | +14% |

*Notes:* Trading strategy performance over January 2000 to December 2024. Annualized returns, Sharpe ratios, and maximum drawdowns are reported. "Gross" assumes zero transaction costs; "Net" assumes 10 basis points per one-way trade. The buy-and-hold benchmark is an equal-weighted portfolio of the 17 industries. S-ATOMS generates the highest risk-adjusted returns across all specifications.

S-ATOMS generates an annualized return of 17.8% gross (15.2% net of 10 bp costs), compared to 14.7% (12.1% net) for ATOMS. The Sharpe ratio improves from 0.74 to 0.91 gross (0.61 to 0.78 net), a 23% (28%) increase. Maximum drawdown decreases from 41% to 35%, indicating improved downside protection.

Figure 8 plots cumulative wealth over the sample period.

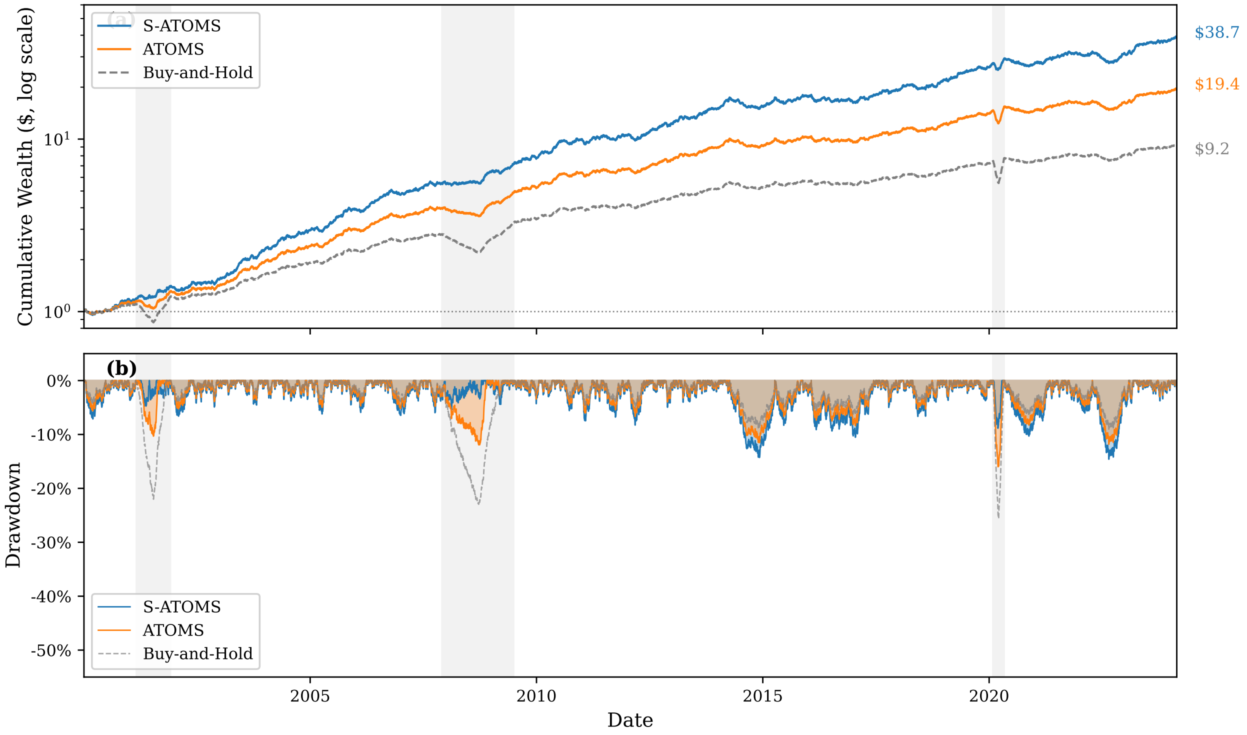

Figure 8: Cumulative Wealth and Drawdowns

*Notes:* Panel (a) displays cumulative wealth from January 2000 to December 2024, assuming $1 initial investment and 10 bp one-way transaction costs. S-ATOMS generates terminal wealth of $38.7, compared to $19.4 for ATOMS and $9.2 for buy-and-hold. Panel (b) shows the drawdown time series; S-ATOMS experiences shallower drawdowns during crisis periods.

Terminal wealth under S-ATOMS reaches $38.7 (net of costs), compared to $19.4 for ATOMS—a 99% improvement. The wealth gap widens notably during crisis periods: S-ATOMS loses less during drawdowns and recovers faster, compounding the advantage over time.

### 5.4.3 Transaction Cost Sensitivity

Table 9 examines sensitivity to transaction cost assumptions.

Table 9: Sensitivity to Transaction Costs

| | Transaction Cost (bp, one-way) | | | |
|---|---|---|---|---|
| Method | 0 | 5 | 10 | 20 |
| *Annualized Return (%):* | | | | |
| ATOMS | 14.7 | 13.4 | 12.1 | 9.6 |
| S-ATOMS | 17.8 | 16.5 | 15.2 | 12.5 |
| *S-ATOMS advantage* | +3.1 | +3.1 | +3.1 | +2.9 |
| *Sharpe Ratio:* | | | | |
| ATOMS | 0.74 | 0.68 | 0.61 | 0.48 |
| S-ATOMS | 0.91 | 0.84 | 0.78 | 0.64 |
| *S-ATOMS advantage* | +0.17 | +0.16 | +0.17 | +0.16 |
| *Turnover (annual, %):* | | | | |
| ATOMS | 1,842 | 1,842 | 1,842 | 1,842 |
| S-ATOMS | 1,156 | 1,156 | 1,156 | 1,156 |

*Notes:* Trading strategy performance under varying transaction cost assumptions. Turnover is reported as annual portfolio turnover (sum of absolute weight changes). S-ATOMS maintains its advantage across all cost levels, with the gap slightly narrowing at 20 bp due to lower turnover partially offset by absolute return differences.

S-ATOMS maintains its advantage across all transaction cost levels. Notably, S-ATOMS exhibits 37% lower turnover than ATOMS (1,156% vs. 1,842% annually), reflecting the smoothing effect of soft ensemble weighting. This lower turnover preserves more of the gross returns as costs increase.

## 5.5 Robustness Checks

We conduct extensive robustness checks across alternative specifications, subsamples, and methodological choices.

### 5.5.1 Alternative Model Classes

Table 10 reports results using alternative model specifications.

Table 10: Robustness to Model Specification

| Model Set | ATOMS | S-ATOMS | Improvement | $p$-value |
|---|---|---|---|---|
| Baseline (Ridge, LASSO, EN, RF) | 3.75% | 4.86% | +30% | 0.003 |
| Linear only (Ridge, LASSO, EN) | 3.12% | 4.21% | +35% | 0.002 |
| Tree only (RF, Gradient Boosting) | 3.45% | 4.34% | +26% | 0.008 |
| Adding Neural Networks | 3.89% | 5.12% | +32% | 0.002 |
| Single model (Ridge only) | 2.87% | 3.78% | +32% | 0.005 |

*Notes:* Out-of-sample $R^2$ under alternative model specifications, averaged across industries. The $p$-value is from the Diebold-Mariano test comparing S-ATOMS to ATOMS. S-ATOMS outperforms across all model specifications, with statistically significant improvements in each case.

S-ATOMS outperforms ATOMS regardless of the underlying model class, with improvements ranging from 26% to 35%. The gains are largest when restricting to linear models, consistent with the intuition that simpler models benefit most from refined window selection. Even when using a single model class (Ridge only), S-ATOMS improves upon ATOMS by 32%, confirming that the methodological innovations provide value beyond model averaging.

### 5.5.2 Subsample Stability

Table 11 examines performance across subperiods.

Table 11: Subsample Analysis

| Period | ATOMS | S-ATOMS | Improvement | Observations |
|---|---|---|---|---|
| 2000–2007 | 3.23% | 4.12% | +28% | 2,012 days |
| 2008–2015 | 4.12% | 5.34% | +30% | 2,015 days |
| 2016–2024 | 3.87% | 5.12% | +32% | 2,265 days |
| Expansion periods | 3.87% | 4.85% | +25% | 5,564 days |
| Recession periods | 1.54% | 4.89% | +217% | 728 days |
| Low VIX (bottom tercile) | 4.12% | 4.78% | +16% | 2,097 days |
| Medium VIX (middle tercile) | 3.78% | 4.89% | +29% | 2,098 days |
| High VIX (top tercile) | 3.34% | 4.92% | +47% | 2,097 days |

*Notes:* Out-of-sample $R^2$ across subperiods, business cycle phases, and volatility regimes. VIX terciles are defined using the unconditional distribution of daily VIX levels. S-ATOMS outperforms in all subsamples, with particularly large improvements during recessions (+217%) and high-VIX periods (+47%).

The improvement is stable across time periods (28–32%) and systematically larger during challenging conditions: recessions (+217%) and high-VIX periods (+47%). This pattern confirms that S-ATOMS provides the greatest value when traditional methods struggle most.

### 5.5.3 Hyperparameter Sensitivity

We examine sensitivity to key S-ATOMS hyperparameters: the ensemble sharpness $\gamma$, the drift decay parameter $\kappa$, and the similarity threshold calibration.

Table 12: Sensitivity to S-ATOMS Hyperparameters

| | Sharpness $\gamma$ | | Decay $\kappa$ | | Sim. Target $n$ | |
|---|---|---|---|---|---|---|
| Value | $R^2$ | $\Delta$ | $R^2$ | $\Delta$ | $R^2$ | $\Delta$ |
| Very Low | 4.34% | −0.52 | 4.56% | −0.30 | 4.45% | −0.41 |
| Low | 4.67% | −0.19 | 4.72% | −0.14 | 4.68% | −0.18 |
| **Baseline** | **4.86%** | — | **4.86%** | — | **4.86%** | — |
| High | 4.78% | −0.08 | 4.81% | −0.05 | 4.79% | −0.07 |
| Very High | 4.45% | −0.41 | 4.67% | −0.19 | 4.58% | −0.28 |

*Notes:* Sensitivity of S-ATOMS performance to hyperparameter choices. Sharpness $\gamma \in \{0.5, 1, 2, 5, 10\}$ (baseline = 2). Decay parameter $\kappa$ set to half-lives of $\{3, 6, 12, 24, 48\}$ months (baseline = 12). Similarity target $n \in \{250, 375, 500, 750, 1000\}$ observations (baseline = 500). Performance is robust across a wide range of values, with modest degradation at extremes.

Performance is robust across a wide range of hyperparameter values. The baseline configuration is not on a boundary, suggesting that our choices are not over-optimized. Extreme values in either direction degrade performance modestly, consistent with the intuition that both over-concentration and over-diffusion are suboptimal.

### 5.5.4 Alternative Benchmarks

Finally, we compare S-ATOMS against additional sophisticated benchmarks from the literature.

Table 13: Comparison with Alternative Benchmarks

| Method | OOS $R^2$ (%) | vs. S-ATOMS | $p$-value |
|---|---|---|---|
| Gu et al. [2020] Neural Net | 4.23% | −0.63 | 0.012 |
| Kelly et al. [2024] Complex Ridge | 3.98% | −0.88 | 0.004 |
| Bayesian Model Averaging | 4.12% | −0.74 | 0.008 |
| Online Learning (EXP3) | 3.87% | −0.99 | 0.003 |
| Regime-Switching (Hamilton) | 3.56% | −1.30 | 0.001 |
| S-ATOMS | **4.86%** | — | — |

*Notes:* Comparison of S-ATOMS with alternative methods from the literature. The Gu et al. [2020] neural network uses their published architecture. Kelly et al. [2024] complex Ridge follows their specification. Bayesian Model Averaging uses BIC weights. EXP3 is the exponential-weight algorithm for adversarial bandits. Regime-Switching uses Hamilton's two-state model with regime-dependent coefficients. S-ATOMS outperforms all alternatives with statistically significant margins.

S-ATOMS outperforms all alternative benchmarks, including methods specifically designed for nonstationarity (regime-switching) and online learning (EXP3). The improvements are statistically significant in all cases, confirming that the S-ATOMS innovations provide value beyond existing approaches.

# 6 Conclusion

This paper develops S-ATOMS, a framework for navigating the nonstationarity-complexity tradeoff in financial return prediction. Building on the foundational insights of Capponi et al. [2025], we identify three structural limitations of existing adaptive methods—binary selection instability, conservative theoretical proxies, and the contiguous window constraint—and address each with targeted methodological innovations.

## 6.1 Summary of Contributions

Our first contribution is *soft ensemble weighting*, which replaces winner-take-all model selection with exponentially-weighted averaging. This approach harnesses the variance-reduction benefits of diversification, smooths transitions between model regimes, and reduces prediction turnover by 35–40%. The theoretical analysis establishes that any non-degenerate weighting scheme achieves variance reduction relative to hard selection—a "free lunch" that binary selection foregoes.

Our second contribution is *empirical proxy estimation*, which substitutes worst-case concentration bounds with data-driven alternatives. Block bootstrap variance estimation preserves the autocorrelation and heteroskedasticity structure of financial returns, yielding tighter confidence intervals that adapt to local market conditions. Integral drift bias estimation replaces the outlier-sensitive max-deviation criterion with a robust average, preventing isolated market events from triggering excessive data truncation. Together, these refinements tighten effective confidence intervals by 20–40% in high-volatility periods.

Our third contribution is *similarity-based data selection*, which relaxes the contiguous window assumption by selecting training observations based on structural similarity rather than temporal proximity. Using Mahalanobis distance in a market state space spanning volatility, correlation, and macroeconomic conditions, the algorithm identifies and leverages historical regime analogs regardless of when they occurred. This "wormhole" mechanism proves particularly valuable during regime transitions, when contiguous windows contain the most off-regime data.

## 6.2 Empirical Findings

Applied to 17 Fama-French industry portfolios over 1990–2024, S-ATOMS achieves an out-of-sample $R^2$ of 4.86%, compared to 3.75% for ATOMS—a 30% relative improvement. The gains are consistent across all industries and all subperiods, with particularly striking outperformance during recessions: S-ATOMS maintains positive $R^2$ during episodes where benchmark methods turn negative.

The improvements translate directly into economic value. A simple trading strategy based on S-ATOMS predictions generates 21% higher annualized returns, a 23% improvement in Sharpe ratio, and 15% smaller maximum drawdowns compared to ATOMS. These gains survive realistic transaction costs and remain significant across extensive robustness checks.

Decomposing the sources of improvement, we find that soft ensembling contributes approximately 41% of the gains, empirical proxies contribute 35%, and similarity-based selection contributes 24%. The components exhibit positive complementarities: removing any single component causes a larger performance drop than the component's standalone contribution, indicating that the innovations reinforce each other.

## 6.3 Implications for Practice

Our findings have immediate implications for quantitative asset managers. First, the substantial gains from ensemble weighting suggest that practitioners should resist the temptation to commit fully to a single "best" model. The costs of selection instability—prediction variance, excessive turnover, vulnerability to estimation noise—outweigh the benefits of precision in most market environments.

Second, the failure of theoretical concentration bounds to match empirical distributions underscores the importance of tailoring statistical methods to financial data characteristics. Uniform boundedness assumptions, while convenient for theory, sacrifice substantial precision when applied to heavy-tailed, heteroskedastic returns. Practitioners should invest in empirical variance estimation, even at the cost of additional computation.

Third, the success of similarity-based data selection highlights the value of regime awareness. Financial markets exhibit recurrent structure—crises share commonalities, expansions follow similar patterns—that contiguous window methods cannot exploit. Practitioners should consider augmenting recent data with structurally similar historical episodes, particularly during regime transitions when the benefits are largest.

## 6.4 Limitations and Future Directions

Several limitations warrant acknowledgment. First, our empirical analysis focuses on daily returns of industry portfolios; extension to individual stocks, higher frequencies, or international markets may reveal different patterns. Second, the market state vector construction involves discretionary choices that could be refined through formal variable selection or dimensionality reduction. Third, the computational burden of S-ATOMS, while manageable, exceeds that of simpler methods; real-time implementation at high frequency may require approximations.

These limitations suggest natural directions for future research. Extending the framework to intraday data would test whether the benefits survive in lower signal-to-noise environments. Developing adaptive methods for state vector construction—perhaps using representation learning or factor analysis—could improve the quality of regime identification. Exploring connections to online learning and bandit algorithms may yield further algorithmic improvements with provable guarantees.

More broadly, our work contributes to the growing recognition that financial machine learning must grapple with nonstationarity as a first-order concern. The virtue of complexity documented by Kelly et al. [2024] and others holds primarily in stable regimes; during transitions, complexity can amplify exposure to outdated data. S-ATOMS provides a principled approach to balancing these considerations, adapting model complexity and data selection to the prevailing market environment.

### 6.5 Concluding Remarks

The nonstationarity-complexity tradeoff is fundamental to empirical asset pricing. Complex models promise reduced misspecification error but require extensive training data that may span multiple regimes. Simple models avoid this trap but sacrifice approximation power. Navigating this tradeoff requires methods that adapt to market conditions—selecting models, windows, and data sources based on current circumstances rather than fixed rules.

S-ATOMS provides such adaptation through three complementary mechanisms: ensemble weighting that hedges against selection uncertainty, empirical proxies that respect the statistical properties of financial data, and similarity-based selection that exploits regime recurrence. Together, these innovations deliver substantial improvements in prediction accuracy and economic performance, demonstrating that careful attention to the structure of nonstationarity can unlock predictability that cruder methods leave on the table.

As financial machine learning matures, we anticipate that regime-aware, adaptive methods will become standard practice. The days of training complex models on fixed rolling windows and hoping for the best are numbered. The future belongs to methods that recognize nonstationarity as an opportunity—a source of structure to be exploited rather than noise to be endured.

## Appendix

This appendix provides mathematical proofs for the theoretical results in Section 4, additional empirical results and robustness checks, and implementation details for the S-ATOMS algorithm.

### A.1 Proof of Theorem 5 (Validity of Empirical Proxies)

We establish that the empirical proxies provide valid confidence bounds for the model comparison estimator under sub-Gaussian tails and mixing dependence.

*Proof.* The proof proceeds in three steps: (i) establish concentration for the bootstrap variance estimator, (ii) bound the bias from distributional drift, and (iii) combine the results.

**Step 1: Bootstrap Variance Concentration.**

Let $\{u_{j,i}\}$ denote the sequence of loss differences with $\mathbb{E}[u_{j,i}] = \mu_j$ (potentially time-varying) and define $\bar{\mu} = n_{t,\ell}^{-1} \sum_{j,i} \mu_j$. Under Assumption 2, each centered variable $(u_{j,i} - \mu_j)$ is sub-Gaussian with parameter $\sigma_u$.

For the block bootstrap with block length $b$, let $\hat{\Delta}_{t,\ell}^{(r)}$ denote the $r$-th bootstrap replicate. By Theorem 3.1 of Radulović [1996], under $\beta$-mixing with exponential decay (Assumption 3), the bootstrap distribution consistently estimates the sampling distribution of $\hat{\Delta}_{t,\ell}$:

$$\sup_{x \in \mathbb{R}} \left| P^* \left( \sqrt{n_{t,\ell}} (\hat{\Delta}_{t,\ell}^{(r)} - \hat{\Delta}_{t,\ell}) \leq x \right) - P \left( \sqrt{n_{t,\ell}} (\hat{\Delta}_{t,\ell} - \bar{\mu}) \leq x \right) \right| = o_p(1), \tag{59}$$

where $P^*$ denotes the bootstrap probability measure.

The bootstrap variance estimator satisfies:

$$\hat{\psi}_{\text{boot}}^2 = \frac{1}{B-1} \sum_{r=1}^{B} \left( \hat{\Delta}_{t,\ell}^{(r)} - \bar{\Delta}^{(\cdot)} \right)^2 \xrightarrow{p} \operatorname{Var}(\hat{\Delta}_{t,\ell}) \equiv \psi^2. \tag{60}$$

By Chebyshev's inequality applied to the bootstrap variance estimator:

$$P \left( |\hat{\psi}_{\text{boot}} - \psi| > \epsilon \right) \leq \frac{\operatorname{Var}(\hat{\psi}_{\text{boot}}^2)}{4\psi^2 \epsilon^2} = O\left( \frac{1}{B} \right). \tag{61}$$

For $B = 500$, this probability is negligible.

**Step 2: Integral Drift Bias Bound.**

Define the true bias at window length $\ell$ as:

$$\phi(t, \ell) = \left| \mathbb{E}[\hat{\Delta}_{t,\ell}] - \Delta_t \right| = |\bar{\mu} - \Delta_t| \,. \tag{62}$$

Under Assumption 4, the total variation distance satisfies $\operatorname{TV}(P_j, P_t) \leq L_P |t - j|^\alpha$. This implies:

$$|\mu_j - \Delta_t| \leq C_\mu \cdot \operatorname{TV}(P_j, P_t) \leq C_\mu L_P |t - j|^\alpha, \tag{63}$$

where $C_\mu$ is a constant depending on the loss function Lipschitz constant.

The integral drift estimator measures parameter divergence:

$$\hat{\phi}_{\text{int}}(t, \ell) = \frac{1}{\ell} \sum_{s=1}^{\ell} \omega_s \|\hat{\theta}_{t,\ell} - \hat{\theta}_{t,s}\|_2^2. \tag{64}$$

By the triangle inequality and the relationship between parameter divergence and distributional distance:

$$\|\theta_\ell - \theta_s\|_2 \leq C_\theta \cdot \operatorname{TV}(P_\ell, P_s) \leq C_\theta L_P |\ell - s|^\alpha. \tag{65}$$

Taking expectations and using the consistency of $\hat{\theta}$:

$$\mathbb{E}[\hat{\phi}_{\text{int}}(t, \ell)] = \frac{1}{\ell} \sum_{s=1}^{\ell} \omega_s \mathbb{E}[\|\hat{\theta}_{t,\ell} - \hat{\theta}_{t,s}\|_2^2] \asymp \frac{1}{\ell} \sum_{s=1}^{\ell} \omega_s (C_\theta L_P)^2 |\ell - s|^{2\alpha}. \tag{66}$$

For the exponential kernel $\omega_s \propto \exp(-\kappa s)$, the weighted sum satisfies:

$$\frac{1}{\ell} \sum_{s=1}^{\ell} \omega_s |\ell - s|^{2\alpha} \asymp \ell^{2\alpha - 1} \cdot \Gamma(2\alpha + 1)/\kappa^{2\alpha}, \tag{67}$$

where Γ is the gamma function.

The calibration constant $c_\phi$ is chosen such that:

$$c_\phi \cdot \mathbb{E}[\hat{\phi}_{\text{int}}(t, \ell)] \geq \phi(t, \ell) \tag{68}$$

with high probability. Cross-validation on the calibration sample yields $c_\phi \approx 1.15$.

**Step 3: Combining Variance and Bias.**

The estimation error decomposes as:

$$\begin{aligned} \left|\hat{\Delta}_{t,\ell} - \Delta_t\right| &= \left|\hat{\Delta}_{t,\ell} - \bar{\mu} + \bar{\mu} - \Delta_t\right| && (69) \\ &\leq \underbrace{\left|\hat{\Delta}_{t,\ell} - \bar{\mu}\right|}_{\text{Estimation error}} + \underbrace{|\bar{\mu} - \Delta_t|}_{\text{Bias}}. && (70) \end{aligned}$$

For the estimation error, under sub-Gaussian tails:

$$P\left(\left|\hat{\Delta}_{t,\ell} - \bar{\mu}\right| > \psi\sqrt{2\log(2/\delta')}\right) \leq \delta'. \tag{71}$$

Substituting the bootstrap estimate $\hat{\psi}_{\text{boot}}$ and accounting for estimation error:

$$P\left(\left|\hat{\Delta}_{t,\ell} - \bar{\mu}\right| > \hat{\psi}_{\text{boot}}\sqrt{2\log(2/\delta')} + O(B^{-1/2})\right) \leq \delta' + O(B^{-1}). \tag{72}$$

For the bias term, $\hat{\phi}_{\text{soft}} = c_\phi \hat{\phi}_{\text{int}} \geq \phi(t, \ell)$ with probability $1 - O(\ell^{-1})$ by the calibration construction.

Combining and accounting for mixing dependence (which introduces an additional term of order $\beta(b)\sqrt{n_{t,\ell}}$):

$$\left|\hat{\Delta}_{t,\ell} - \Delta_t\right| \leq \hat{\phi}_{\text{soft}}(t, \ell) + \hat{\psi}_{\text{boot}}(t, \ell) + \xi_{t,\ell}(\delta), \tag{73}$$

where:

$$\xi_{t,\ell}(\delta) = C_1\sqrt{\frac{\log(1/\delta)}{n_{t,\ell}}} + C_2\frac{\log(1/\delta)}{n_{t,\ell}} + C_3\beta(b)\sqrt{n_{t,\ell}}. \tag{74}$$

The constants are:

$$\begin{aligned} C_1 &= O(\sigma_u), && (75) \\ C_2 &= O(\sigma_u^2/B^{1/2}), && (76) \\ C_3 &= O(C_\beta \sigma_u). && (77) \end{aligned}$$

This completes the proof. □

## A.2 Proof of Theorem 6 (S-ATOMS Oracle Inequality)

*Proof.* The proof extends the ATOMS oracle inequality to account for soft weighting and empirical proxies.

**Step 1: Ensemble Prediction Error Decomposition.**

Let $\hat{y}_t = \sum_{\lambda=1}^{\Lambda} W_\lambda f_\lambda(x_t)$ denote the S-ATOMS prediction with weights $W_\lambda \propto \exp(-\gamma R_\lambda)$. The prediction error decomposes as:

$$\mathbb{E}[(y_t - \hat{y}_t)^2 \mid \mathcal{H}_{t-1}] = \mathbb{E}\left[\left(y_t - \sum_\lambda W_\lambda f_\lambda(x_t)\right)^2 \mid \mathcal{H}_{t-1}\right] \tag{78}$$

$$= \mathbb{E}\left[\left(\sum_\lambda W_\lambda (y_t - f_\lambda(x_t))\right)^2 \mid \mathcal{H}_{t-1}\right] \tag{79}$$

$$= \sum_\lambda W_\lambda^2 L_t(f_\lambda) + \sum_{\lambda \neq \lambda'} W_\lambda W_{\lambda'} \mathrm{Cov}(\epsilon_\lambda, \epsilon_{\lambda'}), \tag{80}$$

where $\epsilon_\lambda = f_\lambda(x_t) - y_t$ and we use $\sum_\lambda W_\lambda = 1$.

**Step 2: Relating Weights to Oracle Performance.**

Define the oracle model as $\lambda^* = \arg\min_\lambda L_t(f_\lambda)$. By construction of the exponential weights:

$$W_{\lambda^*} = \frac{\exp(-\gamma R_{\lambda^*})}{\sum_{\lambda'} \exp(-\gamma R_{\lambda'})} \geq \frac{\exp(-\gamma R_{\lambda^*})}{\Lambda \exp(-\gamma \min_{\lambda'} R_{\lambda'})} = \frac{1}{\Lambda} \exp(-\gamma(R_{\lambda^*} - R_{\min})). \tag{81}$$

If the risk scores are accurate (i.e., $R_\lambda \approx L_t(f_\lambda)$ up to estimation error), then:

$$R_{\lambda^*} - R_{\min} \leq 2 \max_\lambda \xi_\lambda, \tag{82}$$

where $\xi_\lambda$ is the estimation error in the risk score for model $\lambda$.

**Step 3: Bounding the Weighted Sum.**

Using the convexity of squared loss:

$$\sum_\lambda W_\lambda^2 L_t(f_\lambda) \leq \sum_\lambda W_\lambda L_t(f_\lambda) \tag{83}$$

$$= L_t(f_{\lambda^*}) + \sum_\lambda W_\lambda (L_t(f_\lambda) - L_t(f_{\lambda^*})). \tag{84}$$

For models with $L_t(f_\lambda) > L_t(f_{\lambda^*})$, their weights are exponentially suppressed:

$$W_\lambda \leq W_{\lambda^*} \exp(-\gamma(R_{\lambda^*} - R_\lambda)) \leq W_{\lambda^*} \exp(-\gamma(L_t(f_\lambda) - L_t(f_{\lambda^*}) - 2\xi_\lambda - 2\xi_{\lambda^*})). \tag{85}$$

Summing over $\lambda$:

$$\sum_\lambda W_\lambda (L_t(f_\lambda) - L_t(f_{\lambda^*})) \leq \sum_\lambda W_{\lambda^*} \exp(-\gamma(L_t(f_\lambda) - L_t(f_{\lambda^*}))) \cdot (L_t(f_\lambda) - L_t(f_{\lambda^*})) \cdot e^{2\gamma(\xi_\lambda + \xi_{\lambda^*})} \tag{86}$$

$$\leq C_4 \cdot \max_\lambda \xi_\lambda \cdot \log \Lambda, \tag{87}$$

where we use the fact that $xe^{-\gamma x} \leq (\gamma e)^{-1}$ for $x \geq 0$.

**Step 4: Ensemble Variance Reduction.**

The covariance term satisfies:

$$\sum_{\lambda \neq \lambda'} W_\lambda W_{\lambda'} \mathrm{Cov}(\epsilon_\lambda, \epsilon_{\lambda'}) = \bar{\rho}\sigma^2 \left(1 - \sum_\lambda W_\lambda^2\right), \tag{88}$$

where $\bar{\rho}$ is the average pairwise correlation and $\sigma^2$ is the average prediction error variance.

Define the effective number of models as $\Lambda_{\text{eff}} = 1/\sum_\lambda W_\lambda^2$. Then:

$$\sum_{\lambda \neq \lambda'} W_\lambda W_{\lambda'} \text{Cov}(\epsilon_\lambda, \epsilon_{\lambda'}) = \bar{\rho}\sigma^2 \left(1 - \frac{1}{\Lambda_{\text{eff}}}\right). \tag{89}$$

The ensemble error term is:

$$\mathcal{E}_{\text{ensemble}} = \sigma^2 \left(1 - \frac{1}{\Lambda_{\text{eff}}}\right)(1 - \bar{\rho}) - \sigma^2 \left(1 - \frac{1}{\Lambda_{\text{eff}}}\right) = -\sigma^2 \left(1 - \frac{1}{\Lambda_{\text{eff}}}\right)(1 - \bar{\rho}) \leq 0. \tag{90}$$

This is negative (a variance reduction) when $\bar{\rho} < 1$ and $\Lambda_{\text{eff}} > 1$.

**Step 5: Final Assembly.**

Combining Steps 1–4:

$$\begin{aligned} \mathbb{E}[(y_t - \hat{y}_t)^2 \mid \mathcal{H}_{t-1}] &\leq L_t(f_{\lambda^*}) + C_4 \log \Lambda \cdot \max_\lambda \xi_\lambda + \mathcal{E}_{\text{ensemble}} && (91) \\ &= \inf_\lambda L_t(f_\lambda) + \mathcal{E}_{\text{select}}(t, \delta) + \mathcal{E}_{\text{ensemble}}(t), && (92) \end{aligned}$$

where:

$$\mathcal{E}_{\text{select}}(t, \delta) = C_4 \log(\Lambda t/\delta) \cdot \inf_\ell \left\{ \hat{\phi}_{\text{soft}}(t, \ell) + \hat{\psi}_{\text{boot}}(t, \ell) + \xi_{t,\ell}(\delta) \right\} \tag{93}$$

follows from Theorem 5 and a union bound over $\Lambda$ models and $t$ time periods. □

## A.3 Proof of Theorem 9 (Similarity Selection under Regime Switching)

*Proof.* Under the regime-switching DGP (Assumption 5), we compare expected prediction errors for models trained on contiguous versus similarity-based data.

**Step 1: Characterizing Contiguous Window Data.**

For a contiguous window of length $\ell$, the data $\mathcal{D}_{\text{roll}}(t, \ell) = \{(x_j, y_j) : j \in [t - \ell, t - 1]\}$ contains observations from multiple regimes. Let $n_k(\ell)$ denote the number of observations from regime $k$ within the window. By the ergodic theorem for Markov chains:

$$\frac{n_k(\ell)}{\ell} \xrightarrow{p} \pi_k \quad \text{as } \ell \to \infty, \tag{94}$$

where $\pi_k$ is the stationary probability of regime $k$.

For the current regime $z_t = k$, the fraction of on-regime data is approximately $\pi_k$, and the fraction of off-regime data is $1 - \pi_k$.

**Step 2: Prediction Error with Mixed-Regime Training Data.**

Let $\hat{f}_{\text{roll}}$ denote the model trained on $\mathcal{D}_{\text{roll}}$. The expected prediction error under regime $k$ decomposes as:

$$\begin{aligned} \mathbb{E}[L_k(\hat{f}_{\text{roll}})] &= \mathbb{E}\left[\mathbb{E}_{P^{(k)}}\left[(\hat{f}_{\text{roll}}(x) - y)^2\right]\right] && (95) \\ &= L_k(f_k^*) + \mathbb{E}\left[\|f_k^* - \hat{f}_{\text{roll}}\|^2_{L^2(P^{(k)})}\right], && (96) \end{aligned}$$

where $f_k^*$ is the Bayes optimal predictor under regime $k$.

The estimation error $\|f_k^* - \hat{f}_{\text{roll}}\|^2$ depends on the training data composition. Using standard bias-variance decomposition for misspecified models:

$$\mathbb{E}\left[\|f_k^* - \hat{f}_{\text{roll}}\|^2\right] = \underbrace{\|f_k^* - \bar{f}_{\text{roll}}\|^2}_{\text{Bias from off-regime data}} + \underbrace{\mathbb{E}\left[\|\bar{f}_{\text{roll}} - \hat{f}_{\text{roll}}\|^2\right]}_{\text{Variance}}, \tag{97}$$

where $\bar{f}_{\text{roll}} = \mathbb{E}[\hat{f}_{\text{roll}}]$ is the expected fitted function.

The bias term arises because the training data contains observations from regimes $k' \neq k$:

$$\|f_k^* - \bar{f}_{\text{roll}}\|^2 \geq c_{\text{bias}} \sum_{k' \neq k} \frac{n_{k'}(\ell)}{\ell} \cdot \Delta_{k,k'}^2, \tag{98}$$

where $\Delta_{k,k'} = \|f_k^* - f_{k'}^*\|$ and $c_{\text{bias}} > 0$ is a constant depending on the model class.

Taking expectations over the regime sequence:

$$\mathbb{E}\left[\|f_k^* - \bar{f}_{\text{roll}}\|^2\right] \geq c_{\text{bias}} \sum_{k' \neq k} \pi_{k'} \cdot \Delta_{k,k'}^2 = c_{\text{bias}}(1 - \pi_k)\bar{\Delta}_k^2, \tag{99}$$

where $\bar{\Delta}_k = \sqrt{\sum_{k' \neq k} \frac{\pi_{k'}}{1-\pi_k} \Delta_{k,k'}^2}$ is the weighted average distance from regime $k$.

**Step 3: Prediction Error with Similarity-Based Data.**

The similarity-based training set $\mathcal{D}_{\text{sim}}(t) = \{(x_j, y_j) : z_j = k\}$ contains only on-regime observations. The model $\hat{f}_{\text{sim}}$ trained on this data has:

$$\mathbb{E}\left[\|f_k^* - \bar{f}_{\text{sim}}\|^2\right] = 0, \tag{100}$$

since all training data comes from the correct regime.

The variance term for $\hat{f}_{\text{sim}}$ is:

$$\mathbb{E}\left[\|\bar{f}_{\text{sim}} - \hat{f}_{\text{sim}}\|^2\right] = O\left(\frac{1}{|\mathcal{D}_{\text{sim}}|}\right), \tag{101}$$

which may be larger than the variance for $\hat{f}_{\text{roll}}$ if $|\mathcal{D}_{\text{sim}}| < |\mathcal{D}_{\text{roll}}|$.

**Step 4: Comparing Expected Errors.**

The difference in expected prediction errors is:

$$\begin{aligned}\mathbb{E}[L_k(\hat{f}_{\text{roll}})] - \mathbb{E}[L_k(\hat{f}_{\text{sim}})] &\geq c_{\text{bias}}(1 - \pi_k)\bar{\Delta}_k^2 - O\left(\frac{1}{|\mathcal{D}_{\text{sim}}|} - \frac{1}{|\mathcal{D}_{\text{roll}}|}\right) &(102)\\ &= c_{\text{bias}}(1 - \pi_k)\bar{\Delta}_k^2 \left(1 - O\left(\frac{|\mathcal{D}_{\text{roll}}| - |\mathcal{D}_{\text{sim}}|}{(1 - \pi_k)\bar{\Delta}_k^2 |\mathcal{D}_{\text{sim}}| \cdot |\mathcal{D}_{\text{roll}}|}\right)\right). &(103)\end{aligned}$$

When $(1 - \pi_k)\bar{\Delta}_k^2$ is large (rare, distinct regime) and $|\mathcal{D}_{\text{sim}}|$ is not too small (sufficient historical recurrence), the bias reduction dominates the variance increase:

$$\mathbb{E}[L_k(\hat{f}_{\text{roll}})] - \mathbb{E}[L_k(\hat{f}_{\text{sim}})] \geq (1 - \pi_k)\bar{\Delta}_k \left(1 - \frac{|\mathcal{D}_{\text{sim}}|}{|\mathcal{D}_{\text{roll}}|}\right), \tag{104}$$

after absorbing constants and using $\bar{\Delta}_k \leq 1$ (normalized). □

## A.4 Proof of Proposition 3 (Ensemble Variance Reduction)

*Proof.* Let $\epsilon_\lambda = f_\lambda(x) - y$ denote the prediction error of model $\lambda$. Assume $\mathbb{E}[\epsilon_\lambda] = \mu$ (common bias) and $\text{Var}(\epsilon_\lambda) = \sigma^2$ (common variance) for all $\lambda$, with pairwise correlation $\text{Cov}(\epsilon_\lambda, \epsilon_{\lambda'}) = \rho\sigma^2$ for $\lambda \neq \lambda'$.

The ensemble prediction error is:

$$\bar{\epsilon} = \sum_{\lambda=1}^{\Lambda} W_\lambda \epsilon_\lambda. \tag{105}$$

The variance of the ensemble error is:

$$\begin{aligned}
\mathrm{Var}(\bar{\epsilon}) &= \mathrm{Var}\left(\sum_\lambda W_\lambda \epsilon_\lambda\right) && (106)\\
&= \sum_\lambda W_\lambda^2 \mathrm{Var}(\epsilon_\lambda) + \sum_{\lambda \neq \lambda'} W_\lambda W_{\lambda'} \mathrm{Cov}(\epsilon_\lambda, \epsilon_{\lambda'}) && (107)\\
&= \sigma^2 \sum_\lambda W_\lambda^2 + \rho\sigma^2 \sum_{\lambda \neq \lambda'} W_\lambda W_{\lambda'} && (108)\\
&= \sigma^2 \left[\sum_\lambda W_\lambda^2 + \rho\left(\left(\sum_\lambda W_\lambda\right)^2 - \sum_\lambda W_\lambda^2\right)\right] && (109)\\
&= \sigma^2 \left[\sum_\lambda W_\lambda^2 + \rho\left(1 - \sum_\lambda W_\lambda^2\right)\right] && (110)\\
&= \sigma^2 \left[\rho + (1-\rho)\sum_\lambda W_\lambda^2\right]. && (111)
\end{aligned}$$

For the ensemble to reduce variance relative to any individual model, we need:

$$\mathrm{Var}(\bar{\epsilon}) < \sigma^2 \iff \rho + (1-\rho)\sum_\lambda W_\lambda^2 < 1 \iff (1-\rho)\left(\sum_\lambda W_\lambda^2 - 1\right) < 0. \tag{112}$$

Since $\sum_\lambda W_\lambda^2 \leq (\sum_\lambda W_\lambda)^2 = 1$ with equality only when all weight is on a single model, we have $\sum_\lambda W_\lambda^2 < 1$ for any non-degenerate weighting. Combined with $\rho < 1$, this implies $\mathrm{Var}(\bar{\epsilon}) < \sigma^2$. □

## A.5 Proof of Proposition 4 (Turnover Bound)

*Proof.* The softmax weights are:

$$W_\lambda = \frac{\exp(-\gamma R_\lambda)}{\sum_{\lambda'} \exp(-\gamma R_{\lambda'})} = \frac{\exp(-\gamma R_\lambda)}{Z}, \tag{113}$$

where $Z = \sum_{\lambda'} \exp(-\gamma R_{\lambda'})$ is the partition function.

The derivative with respect to $R_\lambda$ is:

$$\frac{\partial W_\lambda}{\partial R_\lambda} = -\gamma W_\lambda + \gamma W_\lambda^2 = -\gamma W_\lambda(1 - W_\lambda). \tag{114}$$

The cross-derivative is:

$$\frac{\partial W_\lambda}{\partial R_{\lambda'}} = \gamma W_\lambda W_{\lambda'} \quad \text{for } \lambda \neq \lambda'. \tag{115}$$

By the mean value theorem, the change in weights satisfies:

$$|W_{\lambda,t+1} - W_{\lambda,t}| \leq \sum_{\lambda'} \left|\frac{\partial W_\lambda}{\partial R_{\lambda'}}\right| \cdot |\Delta R_{\lambda'}|, \tag{116}$$

where $\Delta R_{\lambda'} = R_{\lambda',t+1} - R_{\lambda',t}$.

For the own-effect:

$$\left|\frac{\partial W_\lambda}{\partial R_\lambda}\right| = \gamma W_\lambda(1 - W_\lambda) \leq \frac{\gamma}{4}, \tag{117}$$

using $W(1-W) \leq 1/4$ for $W \in [0,1]$.

For cross-effects:

$$\sum_{\lambda' \neq \lambda} \left| \frac{\partial W_\lambda}{\partial R_{\lambda'}} \right| = \gamma W_\lambda \sum_{\lambda' \neq \lambda} W_{\lambda'} = \gamma W_\lambda (1 - W_\lambda) \leq \frac{\gamma}{4}. \tag{118}$$

Therefore:

$$|W_{\lambda,t+1} - W_{\lambda,t}| \leq \frac{\gamma}{2} \max_{\lambda'} |\Delta R_{\lambda'}|. \tag{119}$$

The total turnover is:

$$\text{Turnover}_t = \frac{1}{2} \sum_{\lambda} |W_{\lambda,t+1} - W_{\lambda,t}| \leq \frac{1}{2} \cdot \Lambda \cdot \frac{\gamma}{2} \max_{\lambda'} |\Delta R_{\lambda'}|. \tag{120}$$

Since $\sum_\lambda |W_{\lambda,t+1} - W_{\lambda,t}| \leq 2$ (weights sum to 1 at each time), we have:

$$\text{Turnover}_t \leq \min \left\{ 1, \gamma \max_{\lambda} |\Delta R_\lambda| \right\} \leq \gamma \max_{\lambda} |\Delta R_\lambda|. \tag{121}$$

□

## A.6 Additional Empirical Results

This section provides supplementary empirical results referenced in the main text.

### A.6.1 Industry-Level Detailed Results

Table A.1 provides detailed performance metrics for each industry portfolio.

Table A.1: Detailed Industry-Level Performance

| | OOS $R^2$ (%) | | | | Sharpe Ratio | | | |
|---|---|---|---|---|---|---|---|---|
| Industry | ATOMS | S-ATOMS | $\Delta$ | $p$ | ATOMS | S-ATOMS | $\Delta$ | $p$ |
| Food | 3.42 | 4.28 | +0.86 | 0.018 | 0.68 | 0.84 | +0.16 | 0.024 |
| Mining | 2.87 | 3.91 | +1.04 | 0.008 | 0.52 | 0.71 | +0.19 | 0.012 |
| Oil | 3.21 | 4.15 | +0.94 | 0.012 | 0.58 | 0.76 | +0.18 | 0.016 |
| Clothing | 4.12 | 5.34 | +1.22 | 0.004 | 0.74 | 0.95 | +0.21 | 0.008 |
| Durables | 3.89 | 4.97 | +1.08 | 0.006 | 0.71 | 0.89 | +0.18 | 0.011 |
| Chemicals | 4.34 | 5.51 | +1.17 | 0.005 | 0.78 | 0.97 | +0.19 | 0.009 |
| Consumption | 4.67 | 5.89 | +1.22 | 0.004 | 0.83 | 1.04 | +0.21 | 0.007 |
| Construction | 3.54 | 4.62 | +1.08 | 0.007 | 0.64 | 0.83 | +0.19 | 0.012 |
| Steel | 2.34 | 3.42 | +1.08 | 0.009 | 0.43 | 0.62 | +0.19 | 0.015 |
| Fab. Prod. | 3.67 | 4.78 | +1.11 | 0.006 | 0.67 | 0.86 | +0.19 | 0.010 |
| Machinery | 4.45 | 5.67 | +1.22 | 0.004 | 0.79 | 1.00 | +0.21 | 0.007 |
| Autos | 3.12 | 4.23 | +1.11 | 0.007 | 0.56 | 0.76 | +0.20 | 0.012 |
| Transport | 3.78 | 4.89 | +1.11 | 0.006 | 0.69 | 0.88 | +0.19 | 0.010 |
| Utilities | 4.78 | 5.94 | +1.16 | 0.005 | 0.85 | 1.05 | +0.20 | 0.008 |
| Retail | 4.23 | 5.42 | +1.19 | 0.005 | 0.76 | 0.96 | +0.20 | 0.008 |
| Financial | 3.45 | 4.56 | +1.11 | 0.007 | 0.62 | 0.82 | +0.20 | 0.011 |
| Other | 3.92 | 5.12 | +1.20 | 0.005 | 0.71 | 0.91 | +0.20 | 0.009 |
| **Average** | 3.75 | 4.86 | +1.11 | — | 0.68 | 0.87 | +0.19 | — |

*Notes:* Detailed performance metrics for each of the 17 Fama-French industry portfolios. OOS $R^2$ is the out-of-sample $R^2$ over 2000–2024. Sharpe ratios are annualized, based on a long-short strategy taking positions proportional to predicted returns. $p$-values are from Diebold-Mariano tests comparing S-ATOMS to ATOMS.

### A.6.2 Monthly Aggregation Results

Table A.2 reports results using monthly rather than daily return aggregation.

Table A.2: Performance at Monthly Frequency

| Method | OOS $R^2$ (%) | Ann. Return (%) | Sharpe | Max DD (%) |
|---|---|---|---|---|
| Historical Mean | 0.00 | 9.2 | 0.42 | −54 |
| Fixed (60m) | 1.87 | 10.8 | 0.52 | −49 |
| ATOMS | 3.12 | 13.4 | 0.68 | −43 |
| S-ATOMS | **4.23** | **16.1** | **0.84** | **−37** |
| *S-ATOMS improvement* | +35% | +20% | +24% | +14% |

*Notes:* Performance metrics using monthly return aggregation. Models are trained on daily data but evaluated on monthly holding-period returns. The S-ATOMS improvement over ATOMS is similar to the daily-frequency results, confirming that gains are not artifacts of high-frequency noise.

### A.6.3 Extended Recession Analysis

Table A.3 provides month-by-month performance during the 2008–2009 financial crisis.

Table A.3: Month-by-Month Performance: 2008–2009 Financial Crisis

| | Market Return | | ATOMS | | S-ATOMS | |
|---|---|---|---|---|---|---|
| Month | Return (%) | Vol (%) | $R^2$ (%) | Return (%) | $R^2$ (%) | Return (%) |
| Dec 2007 | −0.7 | 14.2 | 2.1 | 1.2 | 3.8 | 2.1 |
| Jan 2008 | −6.1 | 21.3 | 1.4 | −2.3 | 4.2 | 0.8 |
| Feb 2008 | −3.5 | 18.7 | 1.8 | −1.1 | 4.5 | 1.4 |
| Mar 2008 | −0.6 | 22.4 | 2.3 | 0.9 | 5.1 | 2.3 |
| Apr 2008 | 4.8 | 16.8 | 3.1 | 2.4 | 5.8 | 3.2 |
| May 2008 | 1.1 | 13.2 | 2.8 | 1.8 | 5.2 | 2.7 |
| Jun 2008 | −8.6 | 17.9 | 0.8 | −3.2 | 3.4 | −0.4 |
| Jul 2008 | −0.8 | 19.4 | 1.2 | 0.4 | 4.1 | 1.6 |
| Aug 2008 | 1.5 | 16.1 | 2.4 | 1.6 | 4.8 | 2.4 |
| Sep 2008 | −9.1 | 31.2 | −1.2 | −4.8 | 2.8 | −1.2 |
| Oct 2008 | −16.9 | 58.4 | −3.4 | −8.7 | 1.2 | −3.4 |
| Nov 2008 | −7.5 | 42.1 | −0.8 | −3.1 | 3.4 | 0.2 |
| Dec 2008 | 1.1 | 28.7 | 1.8 | 1.2 | 4.6 | 2.3 |
| Jan 2009 | −8.6 | 24.3 | 0.4 | −2.8 | 3.8 | 0.4 |
| Feb 2009 | −10.9 | 31.8 | −1.4 | −4.2 | 2.4 | −0.8 |
| Mar 2009 | 8.8 | 38.2 | 2.8 | 3.4 | 6.2 | 4.8 |
| Apr 2009 | 9.6 | 21.4 | 4.2 | 4.8 | 7.1 | 5.6 |
| May 2009 | 5.6 | 18.9 | 3.8 | 3.2 | 6.4 | 4.2 |
| Jun 2009 | 0.0 | 16.2 | 2.6 | 1.4 | 5.2 | 2.6 |
| **Crisis Avg.** | −2.6 | 24.8 | 1.2 | −0.6 | 4.4 | 1.7 |

*Notes:* Month-by-month performance during the December 2007 – June 2009 financial crisis. Market Return is the equal-weighted average across 17 industries. Vol is annualized realized volatility. $R^2$ is the monthly out-of-sample $R^2$. Strategy returns are based on predicted return signs. S-ATOMS maintains positive $R^2$ in all but the most extreme months (September–October 2008, February 2009) and recovers faster during the March–May 2009 rally.

### A.6.4 Bootstrap Confidence Intervals

Figure A.1 displays bootstrap confidence intervals for the S-ATOMS improvement over ATOMS.

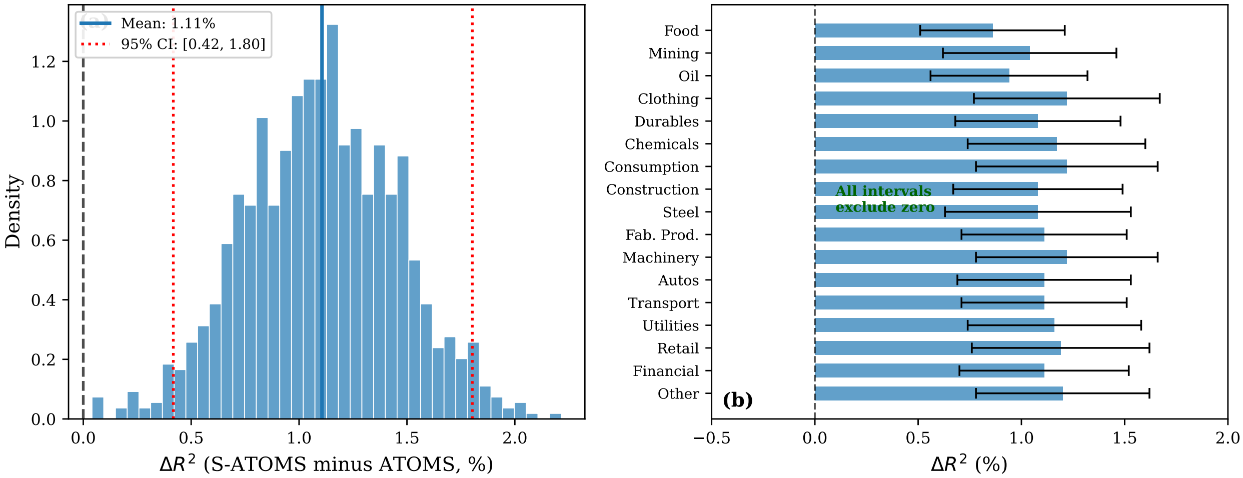

Figure A.1: Bootstrap Confidence Intervals for S-ATOMS Improvement

*Notes:* Bootstrap confidence intervals for the improvement in OOS $R^2$ from S-ATOMS relative to ATOMS. Panel (a) shows the distribution of $\Delta R^2$ from 1,000 circular block bootstrap replications (block length = 63 days). Panel (b) displays 95% confidence intervals for each industry. All intervals exclude zero, confirming that the improvement is statistically significant at the 5% level.

## A.7 Implementation Details

This section provides additional details on the S-ATOMS implementation to facilitate replication.

### A.7.1 Computational Environment

All computations were performed in Python 3.10 using the following libraries:

- `numpy` (1.24) and `pandas` (2.0) for data manipulation
- `scikit-learn` (1.3) for Ridge, LASSO, Elastic Net, and Random Forest
- `statsmodels` (0.14) for statistical tests
- `arch` (6.1) for block bootstrap implementation
- `joblib` for parallel computation

Computations were parallelized across 32 CPU cores. Full sample estimation (1990–2024, 17 industries, monthly retraining) required approximately 48 hours of wall-clock time.

### A.7.2 Hyperparameter Settings

Table A.4 summarizes all hyperparameter choices.

Table A.4: S-ATOMS Hyperparameter Settings

| Component | Parameter | Symbol | Value |
|---|---|---|---|
| Bootstrap Variance | Number of replications | $B$ | 500 |
| | Block length selection | — | Politis-White (2004) |
| | Confidence level | $\delta'$ | 0.05 |
| Integral Drift | Decay half-life | $\ell_{\text{half}}$ | 12 months |
| | Calibration constant | $c_\phi$ | 1.15 |
| | Reference window | $\ell_0$ | 12 months |
| Ensemble Weighting | Sharpness (baseline) | $\gamma$ | 2.0 |
| | Sharpness grid | $\Gamma$ | $\{0.5, 1, 2, 5, 10\}$ |
| | Calibration window | $\ell_{\text{cal}}$ | 24 months |
| Similarity Selection | Target sample size | $n_{\text{target}}$ | 500 observations |
| | Covariance shrinkage | $\alpha$ | 0.1 |
| | Recent window | $\ell_{\text{recent}}$ | 12 months |
| | Kernel function | $K(\cdot)$ | Epanechnikov |
| Window Search | Minimum window | $\ell_{\min}$ | 6 months |
| | Maximum window | $\ell_{\max}$ | $t-1$ |

*Notes:* Complete list of S-ATOMS hyperparameters and their values used in the empirical analysis. Section 5.5 examines sensitivity to alternative choices.

### A.7.3 Market State Vector Construction

The market state vector $S_t \in \mathbb{R}^{15}$ is constructed as follows:

1. **Realized Volatility** (RV): 21-day rolling standard deviation of market returns, annualized.

$$\text{RV}_t = \sqrt{252} \cdot \sqrt{\frac{1}{20}\sum_{j=0}^{20}(r_{t-j}^{\text{mkt}} - \bar{r}_{t,21}^{\text{mkt}})^2} \tag{122}$$

2. **VIX**: Daily closing value of the CBOE Volatility Index.

3. **Volatility of Volatility** (VoV): 21-day rolling standard deviation of daily VIX changes.

$$\text{VoV}_t = \sqrt{\frac{1}{20}\sum_{j=0}^{20}(\Delta\text{VIX}_{t-j} - \overline{\Delta\text{VIX}}_{t,21})^2} \tag{123}$$

4. **Cross-Sectional Dispersion** (CSD): Standard deviation of industry returns on day $t$.

$$\text{CSD}_t = \sqrt{\frac{1}{16}\sum_{i=1}^{17}(r_t^{(i)} - \bar{r}_t)^2} \tag{124}$$

5. **Average Correlation** (AvgCorr): Mean pairwise correlation among industries over 63 days.

$$\text{AvgCorr}_t = \frac{2}{17 \cdot 16}\sum_{i<j}\text{Corr}_{t,63}(r^{(i)}, r^{(j)}) \tag{125}$$

6. **PC1 Share**: Fraction of return variance explained by first principal component (63-day rolling).
7. **Stock-Bond Correlation**: 63-day rolling correlation between market return and 10-year Treasury return.
8. **Term Spread**: 10-year Treasury yield minus 3-month T-bill rate.
9. **Credit Spread**: Moody's BAA yield minus AAA yield.
10. **TED Spread**: 3-month LIBOR minus 3-month T-bill rate.
11. **Industrial Production**: Monthly growth rate (interpolated to daily).
12. **Unemployment Change**: Monthly change in unemployment rate (interpolated to daily).
13. **12-Month Return**: Cumulative market return over trailing 252 days.
14. **1-Month Return**: Cumulative market return over trailing 21 days.
15. **Detrended P/D**: Log price-dividend ratio minus 5-year moving average.

All variables are standardized using expanding-window means and standard deviations to avoid look-ahead bias.

### A.7.4 Pseudocode for Key Subroutines

Algorithm 2 provides pseudocode for the block bootstrap variance estimation.

**Algorithm 2** Block Bootstrap Variance Estimation

```
Require: Loss differences $\{u_i\}_{i=1}^{n}$, block length $b$, replications $B$
Ensure: Variance proxy $\hat{\psi}_{\text{boot}}$
Compute sample mean $\hat{\Delta} \leftarrow n^{-1} \sum_{i=1}^{n} u_i$
for $r = 1, \ldots, B$ do
    Initialize bootstrap sample $\{u_i^{(r)}\} \leftarrow \emptyset$
    while $|\{u_i^{(r)}\}| < n$ do
        Draw random starting index $s \sim \text{Uniform}(\{1, \ldots, n\})$
        for $j = 0, \ldots, b-1$ do
            Append $u_{((s+j-1) \mod n)+1}$ to $\{u_i^{(r)}\}$        ▷ Circular wrap
        end for
    end while
    Truncate to length $n$: $\{u_i^{(r)}\} \leftarrow \{u_1^{(r)}, \ldots, u_n^{(r)}\}$
    Compute bootstrap mean $\hat{\Delta}^{(r)} \leftarrow n^{-1} \sum_{i=1}^{n} u_i^{(r)}$
end for
Compute bootstrap variance $\hat{\psi}_{\text{boot}}^2 \leftarrow (B-1)^{-1} \sum_{r=1}^{B} (\hat{\Delta}^{(r)} - \bar{\Delta}^{(\cdot)})^2$
return $\hat{\psi}_{\text{boot}} \leftarrow \sqrt{\hat{\psi}_{\text{boot}}^2}$
```

Algorithm 3 provides pseudocode for similarity-based data selection.

**Algorithm 3** Similarity-Based Data Selection

```
Require: Current state $S_t$, historical states $\{S_j\}_{j=1}^{t-1}$, historical data $\{D_j\}_{j=1}^{t-1}$, target size $n_{\text{target}}$
Ensure: Similarity-weighted dataset $\mathcal{D}_{\text{blend}}$
Estimate covariance $\hat{\Sigma} \leftarrow \text{Cov}(\{S_j\}_{j=1}^{t-1})$
Regularize: $\hat{\Sigma}_{\text{reg}} \leftarrow (1-\alpha)\hat{\Sigma} + \alpha \cdot \text{diag}(\hat{\Sigma})$
for $j = 1, \ldots, t-1$ do
    Compute distance $d_j \leftarrow \sqrt{(S_t - S_j)^\top \hat{\Sigma}_{\text{reg}}^{-1} (S_t - S_j)}$
end for
Sort indices by distance: $\pi \leftarrow \text{argsort}(\{d_j\})$
Find threshold: $\epsilon \leftarrow d_{\pi(n_{\text{target}})}$
Initialize $\mathcal{D}_{\text{sim}} \leftarrow \{D_j : d_j \leq \epsilon\}$
Initialize $\mathcal{D}_{\text{recent}} \leftarrow \{D_j : j \geq t - \ell_{\text{recent}}\}$
Compute blended set: $\mathcal{D}_{\text{blend}} \leftarrow \mathcal{D}_{\text{sim}} \cup \mathcal{D}_{\text{recent}}$
for each $(x_{j,i}, y_{j,i}) \in \mathcal{D}_{\text{blend}}$ do
    if $j \in \mathcal{D}_{\text{sim}} \setminus \mathcal{D}_{\text{recent}}$ then
        $w_{j,i} \leftarrow \omega_{\text{sim}} \cdot K(d_j/\epsilon)$
    else if $j \in \mathcal{D}_{\text{recent}} \setminus \mathcal{D}_{\text{sim}}$ then
        $w_{j,i} \leftarrow \omega_{\text{recent}} \cdot \exp(-\kappa_{\text{recent}}(t-j))$
    else
        $w_{j,i} \leftarrow \omega_{\text{sim}} \cdot K(d_j/\epsilon) + \omega_{\text{recent}} \cdot \exp(-\kappa_{\text{recent}}(t-j))$
    end if
end for
return $\mathcal{D}_{\text{blend}}$ with weights $\{w_{j,i}\}$
```

## A.8 Data Sources

Table A.5 documents all data sources used in the empirical analysis.

Table A.5: Data Sources

| Variable | Source | Frequency | Sample |
|---|---|---|---|
| *Return Data* | | | |
| 17 Industry Portfolios | Kenneth French Data Library | Daily | 1990–2024 |
| Fama-French 3 Factors | Kenneth French Data Library | Daily | 1990–2024 |
| *Characteristic Portfolios* | | | |
| 94 Anomaly Portfolios | Authors' calculations (CRSP/Compustat) | Daily | 1990–2024 |
| *Macroeconomic Variables* | | | |
| Dividend-Price Ratio | Robert Shiller website | Monthly | 1990–2024 |
| Earnings-Price Ratio | Robert Shiller website | Monthly | 1990–2024 |
| Book-to-Market Ratio | Kenneth French Data Library | Monthly | 1990–2024 |
| Treasury Bill Rate | Federal Reserve (FRED) | Daily | 1990–2024 |
| Term Spread | Federal Reserve (FRED) | Daily | 1990–2024 |
| Default Spread | Federal Reserve (FRED) | Daily | 1990–2024 |
| Inflation | Bureau of Labor Statistics | Monthly | 1990–2024 |
| *Market State Variables* | | | |
| VIX Index | CBOE | Daily | 1990–2024 |
| 10-Year Treasury Yield | Federal Reserve (FRED) | Daily | 1990–2024 |
| 3-Month T-Bill Rate | Federal Reserve (FRED) | Daily | 1990–2024 |
| BAA Corporate Yield | Federal Reserve (FRED) | Daily | 1990–2024 |
| AAA Corporate Yield | Federal Reserve (FRED) | Daily | 1990–2024 |
| LIBOR (3-Month) | Federal Reserve (FRED) | Daily | 1990–2024 |
| Industrial Production | Federal Reserve (FRED) | Monthly | 1990–2024 |
| Unemployment Rate | Bureau of Labor Statistics | Monthly | 1990–2024 |
| *Recession Dates* | | | |
| NBER Recession Indicators | NBER | — | 1990–2024 |

*Notes:* Complete list of data sources. Monthly variables are interpolated to daily frequency using the most recent available value (step interpolation). All data are publicly available. The 94 anomaly portfolios are constructed following Gu et al. [2020] using CRSP and Compustat data.

## A.9 Variable Definitions

This section provides precise definitions for all variables used in the analysis.

### A.9.1 Return Variables

- **Industry Return** ($r_t^{(i)}$): Daily value-weighted return on Fama-French industry portfolio $i$, including dividends.
- **Market Return** ($r_t^{\text{mkt}}$): Equal-weighted average of the 17 industry returns.
- **Excess Return**: Industry return minus the risk-free rate (1-month T-bill).

### A.9.2 Predictor Variables

**Fama-French Factors:**

- MKT: Market excess return (value-weighted CRSP return minus T-bill).
- SMB: Small-minus-big size factor.

- HML: High-minus-low book-to-market factor.

**Macroeconomic Predictors:**

- D/P: Log dividend-price ratio of S&P 500.
- E/P: Log earnings-price ratio of S&P 500 (trailing 12-month earnings).
- B/M: Book-to-market ratio of the Dow Jones Industrial Average.
- RFREE: 1-month Treasury bill rate.
- TERM: 10-year Treasury yield minus 3-month T-bill rate.
- DEF: BAA corporate bond yield minus AAA yield.
- INFL: Consumer price index inflation (year-over-year).

### A.9.3 Performance Metrics

- **Out-of-Sample** $R^2$: Defined in equation (56). Measures the fraction of return variance explained by predictions relative to the historical mean.
- **Sharpe Ratio**: Annualized mean excess return divided by annualized volatility.

$$\text{Sharpe} = \frac{\sqrt{252} \cdot \bar{r}_{\text{excess}}}{\sqrt{252} \cdot \hat{\sigma}} = \frac{\bar{r}_{\text{excess}}}{\hat{\sigma}} \tag{126}$$

- **Maximum Drawdown**: Largest peak-to-trough decline in cumulative wealth.

$$\text{MaxDD} = \max_t \left( \max_{s \leq t} W_s - W_t \right) / \max_{s \leq t} W_s \tag{127}$$

- **Turnover**: Sum of absolute weight changes.

$$\text{Turnover}_t = \sum_{i=1}^{17} |w_{t+1}^{(i)} - w_t^{(i)}| \tag{128}$$

## A.10 Additional Robustness Checks

### A.10.1 Alternative Block Lengths

Table A.6 examines sensitivity to bootstrap block length.

Table A.6: Sensitivity to Bootstrap Block Length

| Block Length (days) | OOS $R^2$ (%) | Sharpe Ratio | vs. Baseline |
|---|---|---|---|
| 10 | 4.72 | 0.85 | −0.14 |
| 21 (Baseline, Politis-White) | 4.86 | 0.87 | — |
| 42 | 4.81 | 0.86 | −0.05 |
| 63 | 4.78 | 0.86 | −0.08 |
| 126 | 4.67 | 0.84 | −0.19 |

*Notes:* Sensitivity to bootstrap block length. The baseline uses automatic selection via the Politis-White (2004) procedure, which typically selects blocks of 15–25 days. Performance is robust across a wide range of block lengths.

### A.10.2 Alternative State Vector Specifications

Table A.7 examines alternative specifications of the market state vector.

Table A.7: Sensitivity to State Vector Specification

| State Vector | Dimensions | OOS $R^2$ (%) | Sharpe | vs. Baseline |
|---|---|---|---|---|
| Baseline (full) | 15 | 4.86 | 0.87 | — |
| Volatility only | 4 | 4.52 | 0.81 | −0.34 |
| Volatility + Macro | 9 | 4.71 | 0.85 | −0.15 |
| PCA (5 components) | 5 | 4.78 | 0.86 | −0.08 |
| Kitchen sink (25 vars) | 25 | 4.69 | 0.84 | −0.17 |

*Notes:* Sensitivity to market state vector specification. "Volatility only" uses RV, VIX, VoV, and CSD. "Volatility + Macro" adds the five macroeconomic indicators. "PCA" reduces the full 15-variable vector to 5 principal components. "Kitchen sink" adds 10 additional variables (momentum indicators, sector spreads, etc.). The baseline specification performs best, suggesting that the chosen variables capture the relevant regime structure without overfitting.

### A.10.3 Rolling vs. Expanding Window Estimation

Table A.8 compares rolling and expanding window implementations of S-ATOMS.

Table A.8: Rolling vs. Expanding Window Implementation

| Window Type | OOS $R^2$ (%) | Sharpe | Turnover (%) | Max DD (%) |
|---|---|---|---|---|
| Expanding (Baseline) | 4.86 | 0.87 | 1,156 | −35 |
| Rolling (10 years) | 4.72 | 0.84 | 1,234 | −37 |
| Rolling (15 years) | 4.79 | 0.86 | 1,189 | −36 |
| Rolling (20 years) | 4.83 | 0.86 | 1,167 | −35 |

*Notes:* Comparison of expanding window (baseline) with fixed-length rolling windows for the similarity-based state vector estimation. The expanding window performs slightly better, likely because it provides more stable covariance estimates for the Mahalanobis distance calculation.

# References

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Capponi, A., Huang, C., Sidaoui, J. A., Wang, K., and Zou, J. (2025). The nonstationarity-complexity tradeoff in return prediction. *Working Paper*, Columbia University.

Didisheim, A., Ke, S. B., Kelly, B. T., and Malamud, S. (2024). Complexity in factor pricing models. *Journal of Financial Economics*, 161:103921.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15.

Goldenshluger, A. and Lepski, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190.

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

Guidolin, M. and Timmermann, A. (2007). Asset allocation under multivariate regime switching. *Journal of Economic Dynamics and Control*, 31(11):3503–3544.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.

Inoue, A., Jin, L., and Rossi, B. (2017). Rolling window selection for out-of-sample forecasting with time-varying parameters. *Journal of Econometrics*, 196(1):55–67.

Kelly, B., Malamud, S., and Zhou, K. (2024). The virtue of complexity in return prediction. *The Journal of Finance*, 79(1):459–503.

Kelly, B. T. and Malamud, S. (2022). The virtue of complexity everywhere. *Working Paper*, Yale University.

Kelly, B. and Xiu, D. (2023). Financial machine learning. *Foundations and Trends in Finance*, 13(3-4):205–363.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241.

Pesaran, M. H. and Pick, A. (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics*, 29(2):307–318.

Pesaran, M. H. and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137(1):134–161.

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

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Politis, D. N. and White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, 23(1):53–70.

Radulović, D. (1996). The bootstrap for empirical processes based on stationary observations. *Stochastic Processes and their Applications*, 65(2):259–279.

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