Financial Innovation

RESEARCH

Open Access

Exploiting market state data for forecasting stock prices: an analysis using predictive algorithms for the Dow Jones and Nasdaq indexes

Josué Bustarviejo[1*], Carlos Bousoño-Calzón[1], Pedro Aceituno-Aceituno[2] and Jose A. Zuñiga-Vicente[3]

*Correspondence:
jbmunoz@tsc.uc3m.es

[1] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, 28911 Leganés, Madrid, Spain
[2] Department of Business Administration and Management and Economics, Universidad a Distancia de Madrid, 28400 Collado Villalba, Madrid, Spain
[3] Business Administration (ADO), Applied Economics II and Fundaments of Economic Analysis, Universidad Rey Juan Carlos, 28933 Madrid, Spain

Abstract

Predicting asset price movements is relevant for investment analysis, risk monitoring, and economic forecasting. This study evaluates whether a cross-sectional representation of the market state can improve short-horizon index forecasting relative to a univariate historical-price benchmark. The paper is framed as a forecasting-comparison study, while market-efficiency testing, asset-pricing restrictions, portfolio efficiency, and trading profitability are treated as complementary evaluation layers. We compare market state data and historical price data for the Dow Jones Industrial Average and the Nasdaq-100. The dataset spans 2005 to 2024, and the empirical evaluation focuses on two single-year experiments: 2019, as the base pre-COVID implementation period, and 2024, as a post-COVID robustness check in a bullish market environment. We introduce a Kalman-inspired linear forecasting framework that uses deterministic matrix estimation and dimensionality reduction. Three market state estimation variants are compared with a univariate time series Kalman benchmark. The findings show that the relative usefulness of historical price data and market state data depends on the forecasting target and the market context, consistent with the no free lunch intuition that performance depends on the match between method, target, and data context.

Keywords: Index Price forecasting, Market state data, Dimensionality reduction, Kalman filter, Decision systems in financial forecasting, No free lunch theorem, Market-state-based forecasting methods

Introduction

In the ever-changing landscape of today's financial markets, the accurate prediction of stock market movements and/or stock returns across multiple assets is not only a matter of monetary gain but also a reflection of our understanding of the complex interdependencies between market dynamics and the performance of real-world firms and economies. More specifically, stock-market forecasting plays a central role in investment decision-making, risk assessment, and macroeconomic analysis, although improvements

in statistical accuracy do not necessarily translate into superior trading performance (Fischer and Krauss 2018; Patel et al. 2015; Shao et al. 2025).

The efficient market hypothesis (EMH) posits that stock prices fully reflect the available information, making it difficult to obtain persistent risk-adjusted excess returns (Fama 1970, 1991). This view has been widely debated, with prior work documenting return predictability, regime dependence, and behavioral factors that motivate alternative forecasting representations (Campbell et al. 1997; Shiller 2003; Pesaran and Timmermann 1995; Malkiel 2003). The adaptive market hypothesis (AMH) further suggests that market efficiency may evolve with changing market conditions (Lo 2004, 2005; Gao et al. 2024). In this study, the EMH, the AMH, and the No Free Lunch (NFL) theorem are used as conceptual lenses for understanding why forecasting performance may depend on the information set, the asset universe, and the market context. This study focuses on the forecasting layer, while formal tests of market efficiency, adaptive efficiency, Capital Asset Pricing Model (CAPM) validity, beta instability, portfolio efficiency, and trading profitability are positioned as complementary extensions. Our empirical aim is to compare the predictive usefulness and computational cost of historical index data and cross-sectional market-state representations within a daily one-step-ahead forecasting framework.

Related strands of the finance literature address questions that are adjacent to, yet distinct from, the main objective of this study. Portfolio-efficiency and asset-pricing tests examine whether portfolios or pricing models satisfy mean–variance efficiency, pricing error, or stochastic discount factor restrictions (Jobson and Korkie 1982; Gibbons et al. 1989; Hansen and Jagannathan 1997), while conditional CAPM frameworks provide evidence that beta exposures may vary across macroeconomic regimes such as inflationary periods (Valadkhani 2025). Complementary empirical approaches show how dummy variables and interaction terms can be used to assess whether predictive relationships differ across signals or conditions (Waggle et al. 2001). Event-study approaches, in turn, evaluate abnormal returns around specific information releases, including long-term cumulative abnormal return (CAR) responses to influential market agents (Agrrawal and Agarwal 2025). Although these perspectives help in the interpretation of the economic significance of predictive models and offer natural extensions, they operate on a different analytical layer. This study, by contrast, focuses on daily index forecasting, treating alphas, risk premia, beta instability, and event-driven abnormal returns as subsequent asset pricing or event study considerations. Specifically, we examine whether the cross-sectional configuration of constituent prices contains exploitable information for one-step-ahead index prediction.

In this context, the relationship between price time series data and the market state merits closer examination. In this study, we define the market state as the instantaneous cross-sectional price vector of an index's constituents on day k (after adjustments such as splits and normalization), optionally reduced via a linear projection; this structural snapshot represents the system's configuration at that moment and differs from correlation-based regime definitions, such as states inferred from correlation matrices, which are complementary and left for future integration. The definition is also distinct from market-state concepts used in momentum studies in which states are often defined by prior market performance rather than by the contemporaneous constituent-price vector

(Cooper et al. 2004).[1] Thus, our objective is not to claim that level-based market states are universally superior to return-based representations but rather to test whether the cross-sectional configuration of constituent prices contains predictive information that is not fully captured by a purely univariate historical treatment of the index.

It should be noted that recent AI-based approaches to stock-market forecasting span neural networks, ensemble methods, alternative data and sentiment systems, graph-based models, quantum-inspired representations, attention mechanisms, and hybrid architectures (Bustos and Pomares-Quimbaya 2020; Kumar et al. 2021; Kumbure et al. 2022; Sun et al. 2024, 2025; Lin and Marques 2024; Rundo et al. 2019; Cheng et al. 2022; Sezer et al. 2020; Xu et al. 2024; Sai et al. 2025). Despite these advances, prior financial innovation research emphasizes that forecasting remains inherently challenging because of ambiguity, heterogeneous information, and evolving institutional contexts (Li et al. 2025). Against this backdrop, this study adopts a complementary perspective by evaluating a transparent cross-sectional market-state representation within a linear and computationally efficient framework. The baseline market-state specification relies on normalized price levels to capture the contemporaneous cross-sectional configuration of the index constituents. Return-based representations are considered in the sensitivity analysis, while extensions involving excess-return pricing tests, CAPM beta estimation, and risk-free rate dynamics are reserved for future asset-pricing analysis.

Our analysis focuses on index-level targets, while tradable implementations would require proxies such as exchange traded funds (ETFs), futures, or derivatives (e.g., DIA and QQQ[2]) (State Street Global Advisors 2026; Invesco 2026). Such extensions must account explicitly for liquidity, bid–ask spreads, transaction costs, tracking error, and market impact since liquidity affects asset prices and ETF execution costs (Agrrawal and Clark 2009; Amihud 2002; Amihud et al. 2005; Marshall et al. 2018). In contrast to traditional models based on the temporal autocorrelation of a single variable (e.g., the index value), our approach defines the predictive state through the contemporaneous cross-sectional market structure. This enables the incorporation of systemic dynamics and collective market behavior into the forecasting process, offering a fundamentally distinct perspective on index evolution. In particular, the market state may implicitly encode episodes of herding behavior or coordinated investor responses that influence price formation outside individual trends (Chiang and Zheng 2010).

This study makes several contributions. First, it proposes the use of market state estimation (MSE), a deterministic linear framework for daily index prediction inspired by Kalman filtering. MSE bases its predictions precisely on cross-sectional market states. Second, it introduces three MSE variants, which differ by whether they use contemporaneous or forecasted market-state information. Third, it explores random projections as a computationally efficient means of reducing dimensionality. Finally, it benchmarks MSE forecasts against a univariate Kalman model for the Dow Jones and Nasdaq-100 indexes in 2019 and 2024. This shows how the value of historical data and market state

[1] The concept of market state is commonly used in finance to describe the overall condition of a financial market (like the stock market) at a given moment. This condition is typically influenced by various factors and the actions of different participants (see, for example, (Cooper et al. 2004; Münnix et al. 2012; Song et al. 2025).

[2] ETFs like DIA (Dow Jones) and QQQ (Nasdaq-100) are standard empirical proxies for trading.

data varies with index characteristics, consistent with NFL-type reasoning about problem-dependent performance.

Accordingly, this study contributes to the financial forecasting and financial innovation literature by examining a transparent market-state representation within a computationally efficient framework. By incorporating dimensionality reduction techniques, the approach enhances the scalability of large-scale daily forecasting (Jia et al. 2022; Reddy et al. 2020). The results indicate that models based on the market state can serve as useful components in analytical and decision-support workflows for short-horizon index forecasting. In particular, they can help to streamline data processing and reduce computational burden under time or resource constraints (Krauss et al. 2017; Wang and Wang 2016). Translating this framework into tradable systems requires an additional implementation layer addressing ETF tracking, transaction costs, bid–ask spreads, market impact, liquidity constraints, portfolio turnover, and execution latency. Thus, the findings should be interpreted as providing forecasting evidence that can inform, but not substitute, trading-performance and production-deployment evaluations. More broadly, the results support the development of data-driven decision-support tools suited to evolving market environments (Kou et al. 2014; Gottschlich and Hinz 2014). In short, the contribution is primarily methodological and comparative, with portfolio optimality, tradable strategy performance, and formal asset-pricing tests deferred to further extensions.

A Kalman filter and dimensionality reduction approach to market-state forecasting

The NFL theorem suggests that no single predictive method should be expected to dominate across all data-generating processes (Sharma et al. 2022; Schurz and Thorn 2024; Wolpert and Macready 1997). We adopt this insight as a conceptual guide rather than a testable proposition. Stock price forecasting has traditionally relied on historical time series data, technical indicators, nonlinear machine learning models, or hybrid systems (Kumbure et al. 2022; Rundo et al. 2019; Balasubramanian et al. 2024). In contrast, this study focuses on the predictive value of the contemporaneous cross-sectional configuration of index constituents. We treat this representation as one informative encoding within a broader forecasting toolkit and assess whether it provides incremental structure in daily forecasts relative to a univariate historical-price benchmark.

While studying the time series of a specific stock or price index requires a smaller dataset than analyzing the complete set of market prices, the computational and energy demands associated with the latter approach pose greater challenges (Kumar et al. 2021; Balasubramanian et al. 2024). Many forecasting methods used in this area, such as support vector machines (SVMs), long short-term memory (LSTM) networks, artificial neural networks (ANNs), and Kalman filtering (Chhajer et al. 2022; Gao et al. 2019; Kurani et al. 2023; Zhang et al. 2020), can be computationally demanding and time-consuming, which may limit their use in resource-constrained applications (Karmiani et al. 2019). To overcome these challenges, recent advances using Kalman filters to process large volumes of data have revealed that random projections for data dimensionality reduction may enhance filter performance (Berberidis and Giannakis 2017). The application of dimensionality reduction techniques can significantly reduce the size of the input

representation while preserving selected geometric properties of the data (Donoho and Tanner 2009a).

Accordingly, this study explores the feasibility of exploiting the market state data obtained from online sources through dimensionality reduction. Beginning with the linear relationship between market information and prices postulated in traditional financial theory (Campbell et al. 1997; LeRoy and Werner 2014), we adopt a linear Kalman-like model that incorporates the explicit consideration of a state and the potential for dimensionality reduction (Berberidis and Giannakis 2017). While nonlinear treatments may give more accurate predictions in some settings, our study focuses on price indexes and daily forecasts, for which linear approximations offer transparency, computational efficiency, and a controlled way to isolate the contribution of the market-state representation. Similar trade-offs have been observed in other areas of financial modeling, such as credit risk, where linear models prevail because of their transparency and ease of deployment, despite the availability of complex machine learning approaches (Khandani et al. 2010).

The Kalman filter consists of two linear equations, namely, the "system equation" and the "observation equation" (Bai et al. 2023; Khodarahmi and Maihami 2023; Meinhold and Singpurwalla 1983). The former describes the evolution of a state vector over time using a known matrix, while the latter computes an observable using the system state through another known matrix. Both equations incorporate noise components with known statistics. The Kalman filter algorithm recursively calculates the system state based on observations, minimizing errors in the presence of noise (Bai et al. 2023; Mendel 1971). Specific versions of the Kalman filter are available for time series prediction, assuming structured matrices (Khodarahmi and Maihami 2023; Durbin and Koopman 2012). Stock market prices usually fluctuate in a nonlinear way, but the Kalman filter provides reliable results when fitting into the series (Alp et al. 2023). We employ one of these algorithms for comparative purposes and as an auxiliary procedure for developing hybrid algorithms.

Our framework also includes a predictive algorithm called market state estimation (MSE). The structural equations of the Kalman filter form the basis of the MSE, with the market state acting as the system state, focused primarily on the index in question. We define the state as the set of prices constituting the stock market index or its dimensionally reduced representation. In this interpretation, both the matrix linking the state from one instant to the next in the system equation and the matrix connecting the market state to the prediction in the observation equation are unknown. Accordingly, we must estimate these matrices (Shumway and Stoffer 2017) and select suitable dimensionality reduction methods. We use random projections because of their universality, efficiency, and conceptual simplicity (Donoho and Tanner 2009a). More specifically, we propose three versions of our algorithm (i.e., MSE-1, MSE-2, and MSE-3), each of which explores a different approach to estimating the equation matrices and incorporating dimensionality reduction. Furthermore, we compare these three market-state-based forecasting methods with each other and with direct predictions from the time series Kalman filter for the index. We evaluate their performance based on the mean squared error of the prediction and the processing times. The analysis of these indexes sheds light on the relevance of different data sources according to the object of prediction, in line with the

NFL theorem. In fact, the DJ and the Nasdaq-100 are two flagship indexes of the US economy. In 2022, the market value of US domestic listed companies amounted to $40.3 trillion, which is more than 1.6 times the national GDP (Worldbank 2024).

Methods

This section outlines the methodological framework used to assess the predictive capabilities of market state data. Figure 1 illustrates the sequential steps from initial data acquisition to the final dataset for the empirical analysis. Our approach began with the gathering of stock market data for the DJ and Nasdaq-100 indexes from publicly available financial data sources and aggregators, such as Yahoo Finance, Google Finance, Bloomberg, online media such as the Financial Times, and other major financial aggregators, with subsequent manual verification where necessary. We applied specific inclusion criteria—selecting companies that had been in existence for at least five years and were still operating as of the analysis date—to ensure the reliability and relevance of our dataset. Next, we screened and normalized the datasets prior to their analysis.

This normalization serves two purposes. First, it places constituents with very different nominal price scales on a comparable footing. Second, it preserves the cross-sectional geometry of the market-state vector at each time point, which is central to our level-based formulation. Although return transformations are standard in many forecasting and asset-pricing applications, they were not chosen as the primary representation here because they emphasize incremental variation and may discard part of the structural information contained in the relative configuration of the constituent price levels.

Adjustments were made for corporate actions such as stock splits and reverse splits to maintain consistency in the price series over time. Nontrading days, including weekends and holidays, were removed to prevent distortions in the temporal analysis. We then normalized the price data by scaling each company's daily prices relative to their historical average, facilitating comparability across different stocks regardless of their absolute price levels. The processed data were stored in a structured format suitable for simulation, with a portion used for training our predictive models and the remainder reserved for testing.

Dataset and initial data processing

As noted above, the market state refers to the collective set of prices that constitutes a specific stock market index at a specific time, reflecting the values of individual stocks or assets and providing a snapshot of the market trend at any given moment. This information provides crucial economic data on the current state of the market, rendering it valuable for predicting stock market indexes (Münnix et al. 2012; Aloud and Alkhamees 2021; Nobi and Lee 2016). Financial forecasting models traditionally rely heavily

Fig. 1 Schematic diagram highlighting the main steps in our empirical research

on historical time series data, analyzing past trends to predict future market movements (Cheng et al. 2022; Sezer et al. 2020). While these historical data are important, incorporating the contemporaneous market state adds a cross-sectional dimension that may enhance predictive accuracy. This provides a holistic view of market dynamics, complementing time series analysis. In fact, evidence from deep learning research suggests that structured patterns of price formation emerge from the collective behavior of market participants—patterns that may be captured through representations such as the market state (Sirignano and Cont 2019). The inclusion of market state data should not be interpreted here as a formal challenge to the EMH. Rather, it is examined as an alternative predictive representation of publicly observed information within a daily forecasting exercise. The present paper does not test market efficiency directly; it evaluates whether different information encodings improve one-step-ahead predictive accuracy. Our approach integrates market state data as a complementary tool, potentially enriching forecasting without violating EMH principles.

We recreated the market state for this study by gathering data from publicly available online sources, including stock exchanges and financial aggregators such as Yahoo Finance, as suggested in prior research (Blazquez and Domenech 2018). We focused on the US dollar market because it includes the most companies and ensures consistent currency across datasets. We maintained data consistency and relevance by applying the two filtering criteria mentioned above: (1) only firms that have existed for at least five years were included; and (2) the dataset contains solely firms that were still trading on the analysis date.

These criteria helped to ensure the availability of reliable historical data and a stable market presence. After applying them to the broad US dollar universe, the resulting filtered universe contained just over 4,300 companies. We then intersected this filtered universe with the official DJ and Nasdaq-100 constituent lists used for the study. Because not every official constituent satisfied the inclusion and data-availability criteria, the final index-level datasets contain partial constituent sets (25/30 for the DJ and 60/100 for the Nasdaq-100). The objective is not to replicate the official indexes mechanically but to maintain a consistent data framework across methods to test the incremental value of the market state. The following three datasets were therefore created:

Dow Jones Industrial Average Index (DJ): 25 of the original 30 companies met the inclusion criteria.
Nasdaq-100 Index: 60 out of 100 companies remained after applying the filters.
Aggregate USD Market: A comprehensive set of over 4,300 companies that met the criteria.

The DJ and Nasdaq-100 provide contrasting test cases: the former covers large multisector firms, while the latter is more technology-oriented and innovation-driven. The use of index levels is appropriate for the forecasting-comparison objective of this paper. A tradable implementation would require a separate ETF-, futures-, or derivatives-based evaluation incorporating liquidity, tracking error, execution constraints, and transaction costs (State Street Global Advisors 2026; Invesco 2026; Agrrawal and Clark 2009; Marshall et al. 2018; Fama and French 1993).

The dataset was assembled from various online sources, primarily through CSV exports, web scraping techniques, and manual verification. All the data were standardized and integrated into a structured MySQL database following a common format. Although the database spans 2005–2024, the empirical evaluation uses two single-year experiments: 2019 as the base pre-COVID period and 2024 as a post-COVID robustness check. Index membership is held fixed within each evaluation year. For the same reason, the empirical design does not include risk-free rates, excess returns, market-factor regressions, or CAPM beta estimation. These variables would be essential if the research objective was to test asset-pricing restrictions or risk-adjusted portfolio performance. However, the objective here is narrower: to compare daily index forecasts generated from historical index prices and cross-sectional constituent-price states under a common preprocessing pipeline. Introducing excess-return or beta-based regressions would change the object of analysis from statistical index forecasting to conditional asset-pricing evaluation.

Data screening, consistency adjustments, and normalization

Given the dynamic nature of stock markets, certain factors may disrupt the consistency of data over time. For example, stock splits and reverse splits may trigger sudden price changes. We corrected this by adjusting historical prices to account for these events, and we reflect the split-adjusted values consistently across the entire time series (Alexander 2008; Tsay 2010). For example, Netflix (NFLX; Nasdaq-100) performed a 7-for-1 stock split on 14 July 2015, and the price series was adjusted accordingly. We used one year of data for the simulations to balance the need for historical depth with computational efficiency. Missing values were addressed through forward filling, whereby the last available price for each stock was propagated to the next trading day. This was adopted as a pragmatic preprocessing choice to maintain a complete daily market-state vector and ensure that all methods were evaluated using the same data pipeline (Little and Rubin 2019; Schafer and Graham 2002; Collier et al. 2024).

Forward filling preserves continuity but has distributional implications: it can smooth short-term variability, induce artificial dependence, reduce apparent volatility, and distort cross-sectional relationships, particularly for less liquid assets or longer gaps. To address this, Sect. "Additional diagnostic robustness checks" reports a concise sensitivity check comparing forward filling with linear interpolation, sample dropping, and mean and median imputation. The analysis evaluates whether the main qualitative conclusions are robust to alternative preprocessing choices, while a full optimization of missing data methods is left for future work. More advanced approaches, including Kalman smoothing, model-based and multiple imputation, and missingness-ratio sensitivity analyses, remain promising extensions.

The baseline specification focuses on market-state information and does not include explicit seasonal/calendar terms such as month-end or quarter-end dummies. The rolling-window design exposes the models to observations from different calendar periods, while explicit calendar-effect modeling remains a useful extension. Month-ends, quarter-ends, holidays, and related timing effects are therefore not identified separately in the present study. Likewise, heavy tails, skewness, and higher-moment effects define a promising route for extending the present linear forecasting comparison toward

distributionally robust specifications. Distributionally robust extensions would be useful avenues for future work. The 2024 replication serves as an out-of-sample regime check relative to 2019. While we again hold index membership fixed within the calendar year to ensure comparability of the methods, the underlying market environment differs materially (for example, in terms of post-COVID dynamics, tighter monetary policy or sectoral concentration). Keeping preprocessing identical across years isolates the methodological comparison from data engineering artifacts; any performance drift therefore reflects regime sensitivity rather than pipeline changes.

One of the challenges posed by using stock price data involves the large variation in price scales between companies. We addressed this by normalizing each company's time series to a comparable scale by dividing the daily price by its historical average:

$$p'_k = \frac{p_k}{\bar{p}} \tag{1}$$

This normalization process ensures that all the price series are on a similar scale, allowing a more effective analysis regardless of differences in the absolute values of the stock prices. The same normalization method was applied to the stock indexes (Tsay 2010). In Eq. (1), the denominator denotes the historical average price of the corresponding company.

Data storage and simulation setup

All the data gathered were initially stored in an SQL database for structured querying and management. For compatibility with our custom R software, the data were subsequently exported in CSV format. Seventy percent of the data were used to train the model in the simulation, while the remaining 30% were used for testing purposes.[3]

Distortion introduced by the dimensionality reduction projection

Stock market data often involve high-dimensional time series because of the high number of companies in an index. Analyzing such data poses computational challenges and may lead to overfitting in predictive models. Moreover, companies' historical price series may behave in a markedly nonlinear manner, as reported in previous studies on different markets (Behrendt and Schmidt 2021; Bhandari et al. 2022; Thakkar and Chaudhari 2020; Yue et al. 2017). This complexity complicates traditional forecasting methods because of the high dimensionality and nonlinear nature of the data used. We therefore applied compressed sensing techniques based on the restricted isometry property (RIP) (Candes and Tao 2005; Candes and Wakin 2008). These techniques involve compressing information from an original high-dimensional space into a smaller space, provided the data are suitably incoherent; the lower the coherence is, the fewer the samples required to preserve relevant information in the resulting space.

In high-dimensional settings, random projections and compressed representations can preserve useful geometric information even when the number of retained dimensions

[3] To provide a consistent basis for comparison, all simulations used a system equipped with a 2.3 GHz quad-core Intel Core i5 processor and 8 GB of 2133 MHz LPDDR3 RAM. This specific setup serves as a uniform benchmark, although performance times may vary on other platforms.

is substantially smaller than the original dimension (Yaroslavsky 2015; Donoho and Tanner 2009b). Donoho and Tanner (2009a, b) examined phase-transition behavior in high-dimensional geometry and randomly projected polytopes. Their results show that high-dimensional structures can often be represented in spaces with substantially lower dimensions, while relevant geometric information was preserved up to critical threshold regimes. This capability is particularly useful in areas such as texture classification in images (Liu et al. 2019) or the recovery of missing information in signals or images (Eldar and Kutyniok 2012) and can be applied to stock market data (Sirignano and Cont 2019; Wang et al. 2023). We use the Johnson–Lindenstrauss lemma—the backbone of random projection techniques—as it offers mathematical assurance that a set of high-dimensional points can be embedded into a lower-dimensional space using random projections with minimal distortion of pairwise distances. Proof of this lemma can be found in Fama and French (1993); Achlioptas 2003; Chiong and Shum 2019; Dasgupta and Gupta 2003), among others.

Thus, for any $0<\varepsilon<1$ and any intermediate number n, k is a positive intermediate number whereby

$$k \geq \frac{24}{3\epsilon^2 - 2\epsilon^3} log(n) \tag{2}$$

Then, for any set A of n points $\in R^d$, there is a map f: $R^d \rightarrow R^k$ such that for x_i, $x_j \in$ A

$$(1-\epsilon)\left\|xi - xj\right\|^2 \leq \left\|f(xi) - f\left(xj\right)\right\|^2 \leq (1+\epsilon)\left\|xi - xj\right\|^2 \tag{3}$$

This result supports our use of random projections to reduce dimensionality while maintaining the essential geometric relationships in the data (Vempala 2005).

Application of dimensionality reduction

A random projection matrix A is used to map the original high-dimensional data $p_k \epsilon R^m$ into a lower-dimensional space $x_k \epsilon R^k$, where $k<m$. The projection is applied independently for each time point k, thus preserving the temporal structure of the data. The reduced market state x_k is obtained as:

$$x_k = \frac{Ap_k}{||A||_1} \tag{4}$$

where $||A||_1$ is a normalization factor for mitigating any scaling distortions introduced by the projection and is defined as:

$$||A||_1 = max_{1\leq j\leq n} \sum_{i=1}^{m} \left|a_{ij}\right| \tag{5}$$

where a_{ij} is the element in row i and column j of matrix A.

We experimented with various probability distributions for generating A, aiming to minimize distortion and ensure computational efficiency. It is crucial for the distribution to be symmetric and centered at zero to preserve the structure of the data. Noncentered or asymmetric distributions significantly increase distortion and prediction errors. Consequently, we select a uniform distribution between -1 and 1 for the entries of A, which

is both centered and provides computational advantages, being approximately 45% faster in R than using a normal distribution.

We initially sought to generate a matrix A that would generate positive final values, which would be aligned with the intuitive notion of stock prices. However, this approach introduced significant distortion and increased the errors compared to using zero-mean distributions. Accordingly, we adopt zero-mean random matrices, which effectively combine and rotate the data in space. While this may generate negative values in the projected states, these values represent abstract states rather than direct stock prices; this is similar to the approaches used in other studies (Eldar and Kutyniok 2012; Campisi et al. 2024; Li et al. 2023).

Measuring and analyzing distortion

The degree to which the data structure is preserved after projection is quantified by introducing measures of distortion and prediction error. On the one hand, the distortion (D) is measured by the maximum relative difference in the Euclidean norms of the original and projected data:

$$Distortion\ (D) = \max\left\{\frac{|\|p_k\| - \|Ap_k\||}{p_k}\right\} \tag{6}$$

where a lower distortion value indicates that the projection has effectively preserved the magnitude and structure of the original data. On the other hand, the impact of dimensionality reduction on forecasting accuracy is assessed by calculating the percentage prediction error (% error) for the index value y_{k+1} as:

$$\%_{error} = \frac{\left|y_{k+1} - y\wedge_{k+1}\right|}{y_{k+1}} \times 100 \tag{7}$$

where y_{k+1} is the actual index value at time $k+1$, and $y^{\wedge}_{k+1}$ is the predicted value.

Additionally, we compute the mean squared error (MSE) of the prediction:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(z_i - z_i\wedge)^2 \tag{8}$$

where z_i represents the actual values, $z_i^{\wedge}$ are the predicted values, and n is the number of observations. The MSE provides a quantitative measure of the average squared difference between the predicted and actual values, and lower values of MSE indicate better predictive accuracy. We report its logarithm, log(MSE), to facilitate comparisons across methods and reduction levels spanning different error scales. In parallel, percentage error provides a scale-relative measure of forecast deviation. Together, these two metrics are used here as forecasting-comparison criteria rather than as direct proxies for trading profitability or portfolio risk.

Empirical evaluation of distortion

We use the selected uniform distribution for A to conduct simulations to assess the distortion and prediction error at various levels of dimensionality reduction. We generate

at least 1,000 random projection matrices for each reduction ratio to average out the effects of randomness and obtain reliable estimates. Figure 2 illustrates the relationship between dimensionality reduction, distortion, and prediction error. Figure 2a shows the maximum distortion induced by the projection matrix for various levels of dimensionality reduction. Each quartile represents the distortion for that percentage of data. Figure 2b reports the percentage of prediction error as a function of dimensionality reduction. A reduction ratio of 1 indicates no reduction. Figure 2a reveals that the distortion remains relatively low until the dimensionality is reduced by approximately 60–70% (reduction ratios up to approximately 1:3). Beyond this point, the distortion increases sharply, indicating a significant loss of information. Figure 2b shows that the prediction error remains below 20% until a reduction ratio of approximately 1:4 (reducing the volume of data to 25% of its original size). These findings suggest that dimensionality can be substantially reduced without severely compromising the quality of the data or the predictive performance of our models.

The ability to reduce the dimensionality of the market state without incurring significant distortion has major implications for our forecasting framework. Compressing the data therefore results in the following:

Computational Efficiency: Reduced data dimensionality lowers the computational load, enabling faster processing within the daily forecasting framework considered here.

Potential Reduction of Overfitting Risk: Simplifying the data may reduce the risk of overfitting by limiting dimensionality, although the present study does not benchmark this effect against alternative reduction schemes such as Principal Component Analysis (PCA) or factor-based representations.

Preservation of Selected Geometric Information: The distortion analysis indicates that substantial dimensionality reduction can preserve norm-based geometric properties up to a threshold, while return correlations, factor exposures, and other joint-distributional features require additional finance-specific diagnostics.

However, it is crucial to balance the trade-off between reduced data size and the potential loss of information. Exceeding the optimal reduction threshold increases distortion

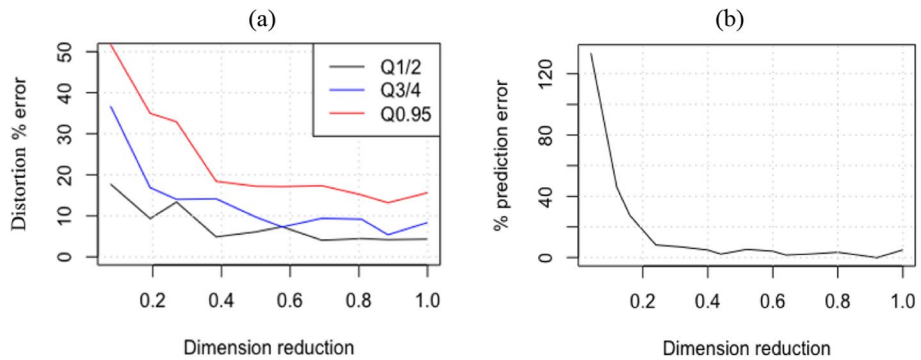


Fig. 2 The relationship between dimensionality reduction, distortion, and prediction error

and prediction errors, undermining the effectiveness of the model. In turn, implementing dimensionality reduction results in the following:

> Normalization: Dividing by A_1 ensures that the scale of the projected data is consistent with the original data, facilitating meaningful comparisons and analysis.
> Negative Values in Projected Data: The zero-mean random projection may result in negative values in x_k. These values represent abstract states rather than direct stock prices. As in image processing (Aloud and Alkhamees 2021), we interpret the projected states as features capturing essential information for forecasting.

In sum, carefully selecting the random projection matrix and measuring the induced distortion effectively reduces the dimensionality of high-dimensional stock market data. This reduction facilitates more efficient computation without significantly degrading the quality of the data or the performance of the predictive models. These insights guide our approach to the integration of dimensionality reduction into our forecasting framework, as detailed below.

The distortion analysis focuses on norm preservation and forecast error. Finance-specific structures—such as correlations, covariance, sector exposures, factor loadings, tail dependence, and higher moments—are left for future investigation. Sect. "Additional diagnostic robustness checks" provides a concise PCA-based diagnostic, while broader comparisons with eigenvector-, factor-, and sector-based reductions represent natural extensions.

The market state estimation forecasting framework

This section introduces the MSE forecasting framework, which uses Kalman-style state-space notation to incorporate market state data into daily index forecasting. The framework is designed to predict stock market indexes using the current market state, potentially reduced in dimensionality to enhance computational efficiency. Kalman-type algorithms have been widely used in financial forecasting and market-monitoring applications, including studies of crisis and stress periods such as the 2008 real estate bubble and the 2020 COVID-19 outbreak (Chu et al. 2019; Elsegai 2022).

Adapting the kalman filtering framework

The traditional Kalman filter is a recursive algorithm that estimates the state of a dynamic system from incomplete and noisy measurements (Meinhold and Singpurwalla 1983). Kalman-style equations are used here as a modeling template to describe how the observed market-state vector is propagated and mapped into an index forecast. In line with the linear discrete-time state-space model discussed by Berberidis and Giannakis (Berberidis and Giannakis 2017), we write:

$$x_k = M1_k x_{k-1} + v_k \tag{9}$$

$$y_k = M2_k x_k + w_k \tag{10}$$

where x_k is the vector of states for day k, $M1_k$ is the transition matrix linking consecutive market states, y_k is the index observation to be forecast, and $M2_k$ is the observation

matrix linking the market state to the index level. In the standard Kalman formulation, the noise terms v_k and w_k are usually modeled as white, Gaussian, mutually uncorrelated disturbances with known or estimated covariance matrices. In the proposed MSE framework, these terms connect the deterministic forecasting equations with classical state-space notation. Innovation-covariance estimation and covariance recursion are reserved for future probabilistic extensions. Accordingly, MSE is best understood as a Kalman-inspired deterministic linear forecasting framework that isolates the contribution of the observed market-state representation. The framework is tailored to our purpose through the following adaptations:

1. Market State as State Representation: We define x_k to be the cross-sectional vector of constituent prices (or its reduced-dimensionality projection) at time k. In the MSE framework, x_k is directly constructed from observed market data rather than inferred as an unobserved latent state. In the baseline specification, this state is defined in levels (normalized constituent prices) because the goal is to represent the contemporaneous cross-sectional structure of the market. Alternative encodings such as returns or factor-based states are possible, but they correspond to different modeling choices and are treated as extensions rather than replacements of the baseline formulation.
2. Estimating unknown matrices: In traditional applications, $M1_k$ and $M2_k$ are known, but in our case, they are unknown and must be estimated from the data. We estimate $M1_k$ and $M2_k$ using historical observations of x_k and y_k.

A synthesized model of the market state can therefore be used to summarize the economic data and substantially reduce the amount of initial data in a fully linear manner. This allows agile forecasts to be evaluated on a reduced dataset that is sufficiently representative of the initial set. Different tolerable information limits are set for these reductions and for assessing their impact on the prediction of a market value. Multivariate outputs are supported by letting $y_k \in \mathbb{R}^r$ (e.g., multiple indexes) and $M_2 \in \mathbb{R}^{r\times m}$, enabling joint forecasts within the same linear state.

For tractability, M_1 and M_2 are estimated from observed data using deterministic least squares pseudoinverse operations. This estimation step provides a transparent deterministic baseline, while innovation-covariance estimation, likelihood-based state-space identification, and Bayesian filtering define natural probabilistic extensions. We have:

$$M1_k = x_k x_{k-1}^{-1} \tag{11}$$

$$M2_k = y_k x_k^{-1} \tag{12}$$

where ${x^{-1}}_k$ and ${x^{-1}}_{k-1}$ are Moore–Penrose pseudoinverses of x_k and x_{k-1}, respectively. The high dimensionality and potential singularity of x_k requires the use of the pseudoinverse to ensure that the estimation is feasible, as standard inversion may not be possible.

Forecasting approaches

Predicting the index value y_{k+1} faces the challenge that x_{k+1} is unknown at time k. We propose the following three approaches:

1. MSE-1: constant matrices assumption. The state transition and observation matrices remain constant between time k and time $k+1$. We assume that both tomorrow's state transition matrix and tomorrow's measurement matrix will be the same as today's ($M1_{k+1}=M1_k$ and $M2_{k+1}=M2_k$).

$$x_{k+1} = M1_k x_k + v_k \Rightarrow x_{k+1} \simeq M1_k x_k \tag{13}$$

$$y_{k+1} = M2_k x_{k+1} + w_{k+1} \Rightarrow y_{k+1} \simeq M2_k x_{k+1} \tag{14}$$

This approach simplifies the prediction process and reduces computational complexity.

2. MSE-2: Forecasting Market State via Individual Stocks. Standard Kalman forecasting on each individual stock's historical price series is used to predict future prices p_{k+1}. This incorporates the temporal dynamics of each stock. The predicted market state vector $\hat{x}_{k+1}$ obtained by aggregating the forecasted stock prices and applying dimensionality reduction via random projections is used to obtain a reduced market state, if necessary.

$$M2_{k+1} = y_k \hat{x}_{k+1}^{-1}; \hat{y}_{k+1} = M2_{k+1}\hat{x}_{k+1} \tag{15}$$

Forecasting individual stock prices generates a more dynamic and potentially accurate prediction of the market state, incorporating the latest information available for each stock.

3. MSE-3: Forecasting Reduced Market State Directly. Standard Kalman forecasting is applied directly to the lower-dimensionality market state x_{k+1} to predict $\hat{x}_{k+1}$, on a similar basis to Eq. 15. This approach eliminates the need to forecast individual stock components, reducing the number of propagated variables and focusing on randomly projected linear combinations of the original market-state representation. These compressed directions should not be interpreted as principal components or economic factors, as they arise from random projections rather than covariance-based or factor-structured estimation.

Our reinterpretation of the Kalman framework treats the cross-sectional market state as the central predictive representation rather than as an external regressor. The transition matrix summarizes how this representation evolves between days, while the observation matrix links it to the index level to be forecast. In this sense, the framework embeds market structure into a linear forecasting scheme while preserving a transparent deterministic structure that can later be extended toward fully latent or Bayesian state-space identification.

Implementation and comparative analysis

As noted, the effectiveness of the MSE algorithms is evaluated by their implementation using data from the two US stock indexes considered here. This selection enables us to assess the algorithms across indexes with different compositions and market characteristics. Potential singularity issues when computing the inverses of ${x^{-1}}_k$ and ${x^{-1}}_{k-1}$ call for the use of the Moore–Penrose pseudoinverse (Bhandari et al. 2022), implemented using the *ginv* function in R introduced by Venables and Ripley (Venables et al. 2013). This ensures that the estimation of the matrices $M1_k$ and $M2_k$ is feasible even when x_k is not of full rank. We average the results over multiple iterations to mitigate the effects

of randomness in the dimensionality reduction process, obtaining reliable estimates of prediction errors and processing times. We compare the performance of the three MSE variants (MSE-1, MSE-2, and MSE-3) with a traditional Kalman filter applied directly to the historical time series of the index. The evaluation focuses on prediction accuracy, computational efficiency, and the impact of dimensionality reduction. The main metrics are log(MSE), percentage error, and computation time. Sect. "Additional diagnostic robustness checks" also reports the return MAE (mean absolute error) and directional hit ratio as diagnostic metrics.[4]

Finally, to ensure the reliability and stability of the predictive framework, we average the results of all experiments over at least 1,000 runs using different random projection matrices and seeds. We summarize performance by the sample mean and standard deviation (σ) of log(MSE) and of the per-prediction runtime. We report two-sided 95% confidence intervals (high and low) for the means, which coincide with the normal approximation when N is large. Confidence intervals for time are computed analogously: $\mathrm{CI95}_{\mathrm{low}}/\mathrm{CI95}_{\mathrm{high}}$ are the lower and upper bounds of the 95% confidence interval for those mean values where $halfwidth = 1.96 \times \sigma/\sqrt{(N)}$, $\mathrm{CI95}_{\mathrm{low}} = \bar{x} - halfwidth$, $\mathrm{CI95}_{\mathrm{high}} = \bar{x} + halfwidth$. This approach helps to mitigate the variance induced by randomness and improves the robustness of our findings.

MSE versus traditional Kalman time series analysis

Comparing the three MSE variants enables us to explore the trade-offs between modeling complexity, computational effort, and forecasting performance when using market state data. Each method reflects a different assumption about system dynamics, from static structures (MSE-1) to dynamic, fine-grained forecasts (MSE-2 and MSE-3), and we thus reveal the most effective configurations under different market conditions.

Standard Kalman analysis on the time series

A baseline for evaluating the performance of our MSE algorithms is established by first applying a traditional Kalman filter to the historical time series of the indexes under study. This standard approach provides a yardstick for comparing the predictive accuracy and computational efficiency of the MSE algorithms. We use the KalmanForecast function in R, which implements the Kalman filtering procedure as described by Durbin and Koopman (Durbin and Koopman 2012) and Gardner et al. (Gardner et al. 1980). The function applies a state-space model to the time series data, estimating unobserved states and generating forecasts based on observed index values. The steps involved in the standard Kalman analysis are as follows:

1. Data Preparation: We use the processed one-year daily closing price data for each index, as outlined in Sect. "Dataset and initial data processing", and normalize the time series to mitigate scale differences and enhance numerical stability.
2. Model Specification: We define a state-space model for the financial time series, using a local level or local trend specification depending on the series configuration

[4] Risk-adjusted performance, abnormal returns, portfolio efficiency, and transaction-cost-adjusted profitability require a separate implementation layer.

implemented by the KalmanForecast routine, and initialize the corresponding state and observation noise variances.

3. Parameter Estimation. The Kalman filter is used to recursively estimate the unobserved state variables and to optimize the model parameters by maximizing the likelihood function. Hyperparameters (including the process and measurement noise covariances as well as the initial state and covariance) are estimated by maximum likelihood on the training split (70%). We begin with a diffuse prior and use likelihood maximization to determine the values of Q and R without manual tuning. The resulting parameters remain fixed for one-step-ahead forecasting on the test portion of that window. Because the Kalman benchmark is included as a comparative forecasting baseline, the analysis focuses on forecast accuracy and runtime; a full residual-diagnostics study would provide a useful extension of the benchmark evaluation.
4. Forecasting: We generate one-step-ahead forecasts for the index values over the testing period and compare the predicted values with the values actually observed to assess performance.

Figure 3 summarizes the performance metrics of the standard Kalman filter applied to the DJ and Nasdaq-100 indexes.

The standard Kalman filter demonstrates competent predictive capabilities on the historical time series of both indexes, with notably high accuracy for the Nasdaq-100. The low MSE values indicate that the model effectively captures the underlying dynamics of the index prices. The computational efficiency is also high, with average processing times per iteration under 0.0040 s for both indexes. This efficiency is relevant for computationally lightweight forecasting settings, although the present study does not evaluate real-time system latency or live-trading deployment. These results serve as a benchmark for our MSE algorithms. Comparing the MSE algorithms to this standard approach allows us to assess the added value of incorporating market state data and dimensionality reduction techniques. Specifically, we aim to determine whether MSE algorithms achieve comparable or improved predictive accuracy while potentially providing computational advantages.

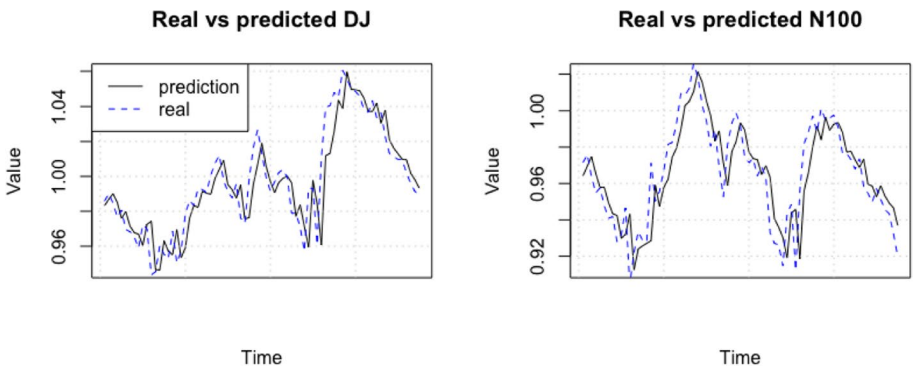


Fig. 3 Actual vs predicted value for the control series using Kalman forecasting on the one-year time series

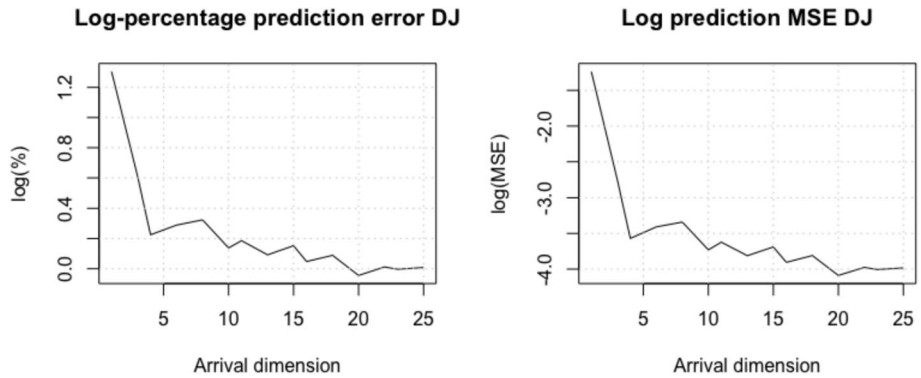


Fig. 4 Log(MSE) and percentage error of the DJ predictions using MSE-1 with data from the index's constituent companies

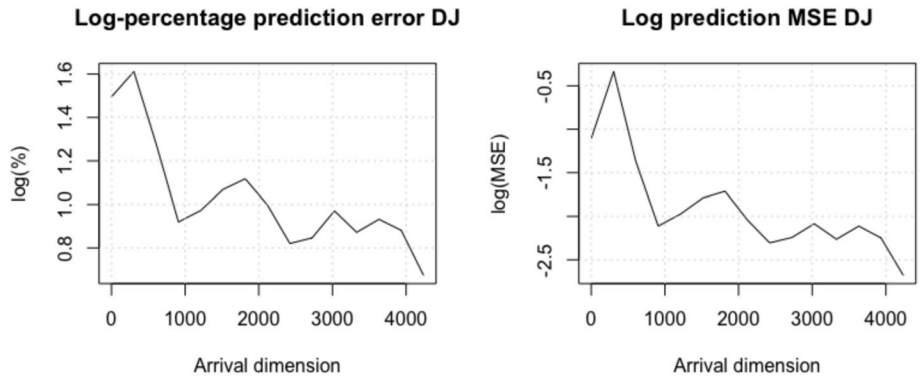


Fig. 5 Log(MSE) and percentage error of the DJ predictions using MSE-1 with data from all selected market companies

MSE-1

We apply MSE-1 to predict the DJ and Nasdaq-100 indexes using the market state data on their constituent companies. MSE-1 operates under the assumption that the state transition matrix $M1_k$ and the observation matrix $M2_k$ remain constant from time k to time $k+1$, as described in Eqs. (13) and (14). The impact of dimensionality reduction on prediction accuracy and computational efficiency is assessed by applying random projections to reduce the dimensionality of the market state xk. Various original-to-reduced dimension settings are considered, ranging from no reduction (1:1) to stronger reductions such as 25:3, which means that the original 25-dimensional DJ state is projected into three dimensions. Prediction accuracy is evaluated by using the logarithm of the mean squared error (log(MSE)) and the percentage error between the predicted and the actual index values. Figures 4 and 5 present the results for the DJ, while Figs. 6 and 7 show the results for the Nasdaq-100.

Figures 4, 5, 6 and 7 reveal that the measurement of the error in the prediction percentage and the MSE on the prediction percentage follow closely similar patterns, with minor variations between their shapes. Nevertheless, it does not appear to be necessary to input more companies to obtain better predictions. This finding aligns with those of Berberidis and Giannakis (Berberidis and Giannakis 2017), who reported that

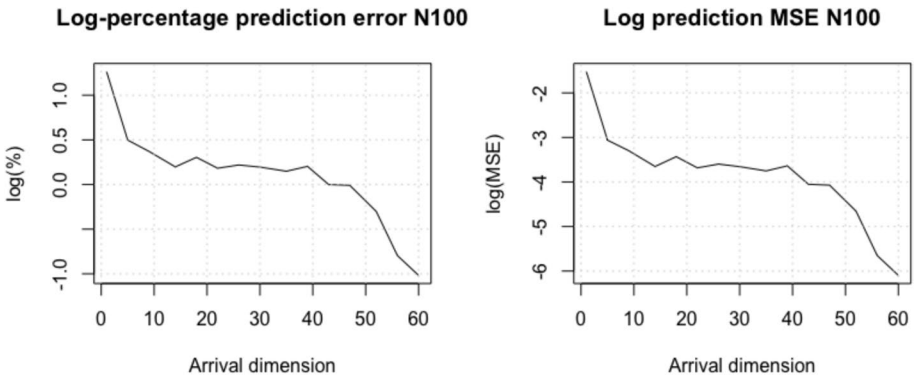


Fig. 6 Log(MSE) and percentage error of the Nasdaq-100 predictions using MSE-1 with data from the index's constituent companies

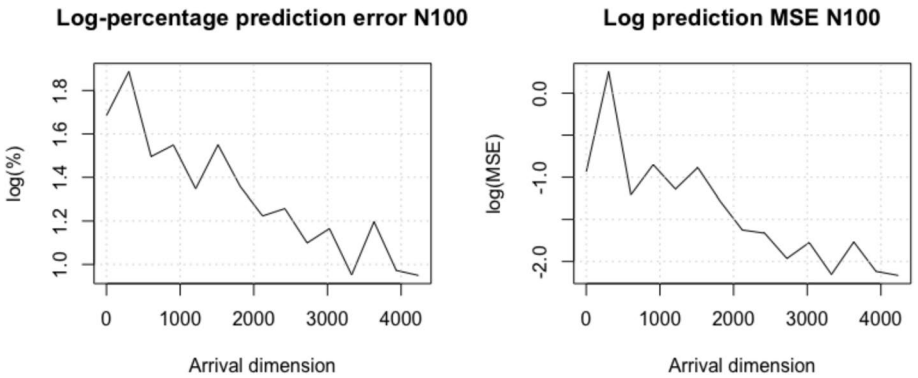


Fig. 7 Log(MSE) and percentage error of the Nasdaq-100 predictions using MSE-1 with data from all selected market companies

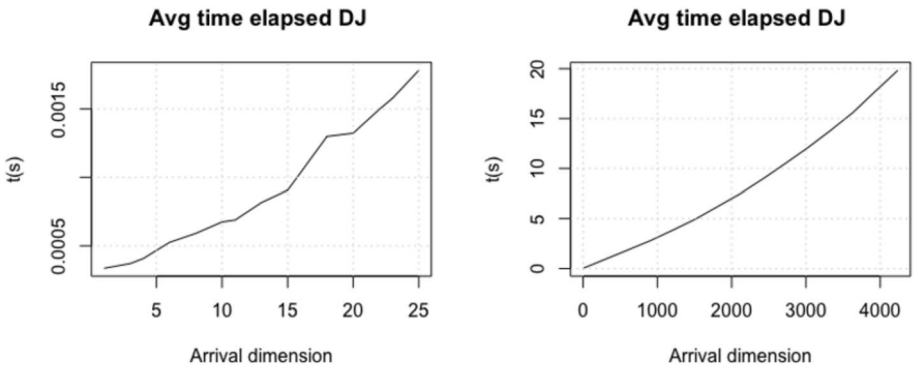


Fig. 8 Average execution time per prediction for the DJ using MSE-1: **a** with data from the index's constituent companies; **b** with data from all available companies

the performance of the Kalman filter may be enhanced by reducing data inputs through random projections. The average computation time for each simulation depends on the chosen level of dimensionality reduction, as shown in Figs. 8 and 9.

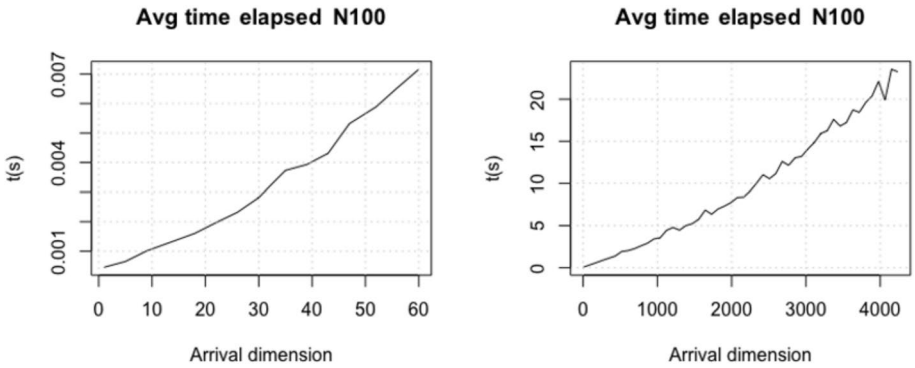
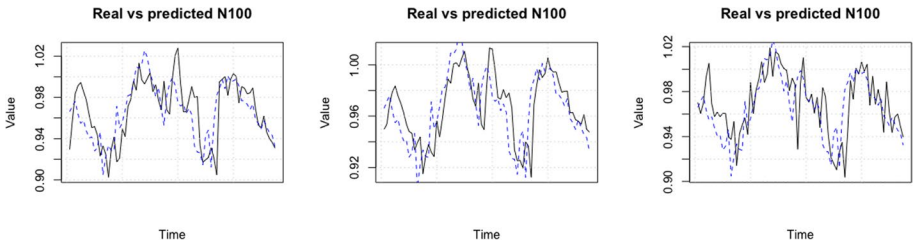


Fig. 9 Average execution time per prediction for Nasdaq-100 using MSE-1: **a** with data from the index's constituent companies; **b** with data from all available companies



Reduction 1:1
Log(MSE) = -8.91

Reduction 2:1
Log(MSE) = -5.69

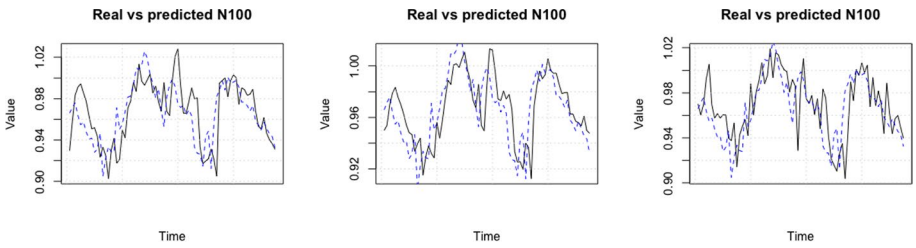
Reduction 25:3
Log(MSE) = -3.83

Fig. 10 Actual vs. predicted DJ values using MSE-1 with varying levels of dimensionality reduction on the constituent companies over a one-year time series

There is a directly proportional relationship between dimensionality and computational complexity—the fewer the dimensions, the faster the computation. The results show that restricting the data to the index constituents improves both accuracy and efficiency compared to using the full market. Therefore, we focus on using the respective datasets for each index for more efficient and reliable assessments. For example, when predicting the DJ value using all 4300 available companies, each prediction takes approximately 20 s and results in a log(MSE) of approximately -2.5. In contrast, using only the companies in the DJ and reducing the dimensionality to just three components leads to a similar error but with computation times reduced to mere milliseconds.

Figures 10 and 11 compare the actual versus predicted index values at various levels of dimensionality reduction. These figures reveal that as dimensionality is reduced, the predictions become less accurate and noisier, highlighting the conclusion that higher-dimensional data (i.e., with less of a reduction) lead to more accurate and stable predictions.

As illustrated in Figs. 10 and 11, the prediction error increases as the representation is compressed from 1:1 to stronger reductions, such as 25:3 for DJ and 12:1 for Nasdaq-100. Specifically, the prediction error for the DJ (Fig. 10) increases as dimensionality is reduced from 25 to three, with the log(MSE) deteriorating from -8.910 for 1:1 to -3.830 for 25:3. In turn, the log(MSE) for Nasdaq-100 (Fig. 11) deteriorates from -5.990 to -4.930 when reducing the dimensions from 1:1 to 12:1.



Reduction 1:1 Log(MSE) = -5.99 | Reduction 2:1 Log(MSE) = -5.85 | Reduction 12:1 Log(MSE) = -4.93

Fig. 11 Actual vs predicted Nasdaq-100 values using MSE-1 with varying levels of dimensionality reduction on the constituent companies over a one-year time series

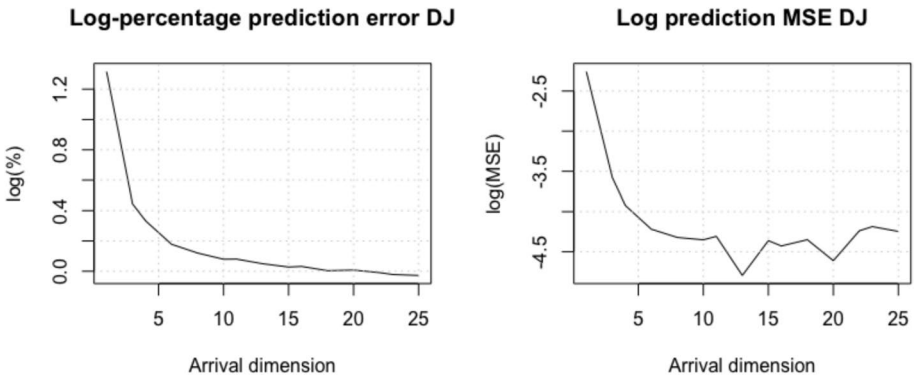


Fig. 12 Log(MSE) and percentage error of the DJ predictions using MSE-2 with data from the constituent companies

In sum, MSE-1 shows that accurate index predictions can be achieved using only the constituent market state data, even with a reduction in dimensionality. However, a trade-off emerges: lower dimensionality improves computational efficiency but may reduce predictive accuracy. The optimal balance depends on the specific forecasting requirements and acceptable error levels.

MSE-2

Unlike MSE-1, which assumes constant system dynamics, MSE-2 seeks to predict the future market state x_{k+1} by first forecasting the individual stock prices for the next day and then applying dimensionality reduction. By forecasting the individual stock prices, MSE-2 incorporates the temporal dynamics of each stock, potentially enhancing the prediction of the overall market index. Figures 12 and 13 show the prediction errors and mean squared errors for the DJ and Nasdaq-100 indexes, respectively, revealing how MSE-2 maintains or slightly improves the prediction accuracy compared to MSE-1 across various levels of dimensionality reduction.

Figure 12 shows that the prediction accuracy of MSE-2 is comparable to that of MSE-1. In certain cases, MSE-2 slightly outperforms MSE-1 in both relative error and MSE, suggesting that incorporating individual stock forecasts may enhance the prediction for the DJ. Similarly, Fig. 13 reveals that at certain levels of dimensionality reduction, MSE-2

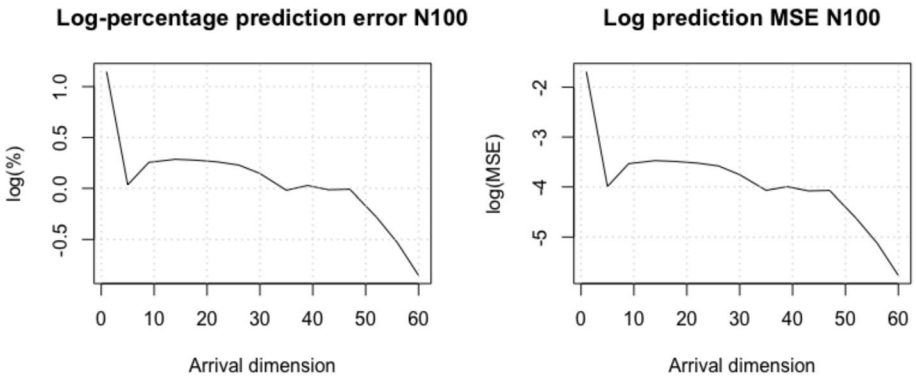


Fig. 13 Log(MSE) and percentage error of the Nasdaq-100 predictions using MSE-2 with data from the constituent companies

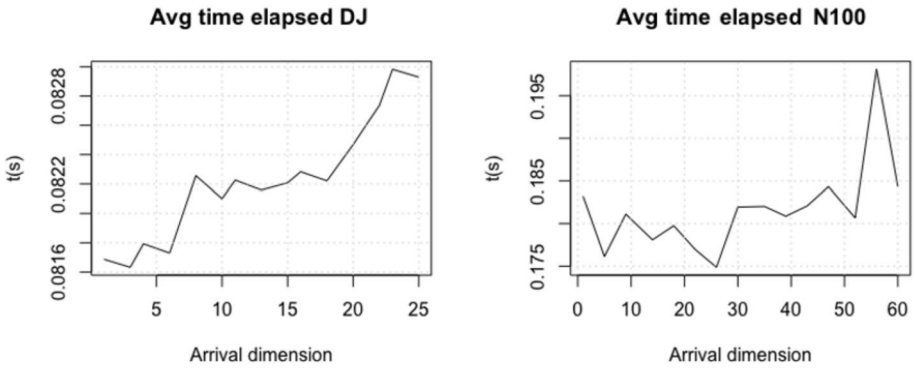


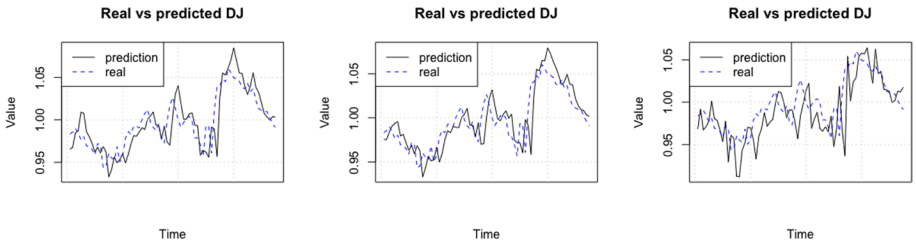
Fig. 14 Average execution time per prediction for both indexes using MSE-2 with data from the constituent companies

records a similar or slightly lower number of errors than MSE-1. Interestingly, intermediate levels of dimensionality reduction sometimes improve the accuracy compared to the use of all the dimensions. However, MSE-2 requires additional computational effort because it involves forecasting each company's stock price individually, thereby increasing processing time, as depicted in Fig. 14.

Figure 14 shows that the processing time for MSE-2 is higher than that for MSE-1, with the increase in computation time being more significant for the Nasdaq-100 index because it has more constituent companies. This additional time is attributed to the need for individual Kalman forecasts for each stock.

We analyze how different levels of dimensionality reduction affect the prediction accuracy and computation time. Figures 15 and 16 illustrate the actual versus predicted index values at various reduction levels for DJ and Nasdaq-100, respectively.

Figure 15 shows that reducing the dimensionality of the DJ index increases the prediction error. The best performance, indicated by the lowest log(MSE) of − 8.030, is achieved with no dimensionality reduction (1:1). However, even with a strong 25:3 original-to-reduced dimension setting, MSE-2 maintains reasonable accuracy with a log(MSE) of − 4.190. This suggests that while higher dimensionality increases prediction accuracy, MSE-2 can still perform adequately when the data are significantly

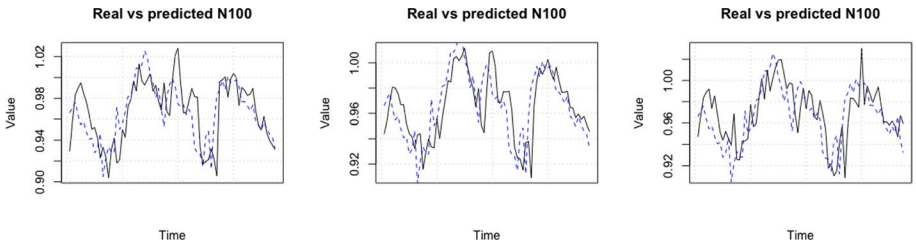


Reduction 1:1
Log(MSE) = -8.03

Reduction 2:1
Log(MSE) = -5.82

Reduction 25:3
Log(MSE) = -4.1 9

Fig. 15 Actual vs. predicted DJ values using MSE-2 with varying levels of dimensionality reduction over a one-year time series



Reduction 1:1
Log(MSE) = -5.96

Reduction 2:1
Log(MSE) = -7.59

Reduction 12:1
Log(MSE) = -4.95

Fig. 16 Actual vs predicted Nasdaq-100 values using MSE-2 with varying levels of dimensionality reduction over a one-year time series

compressed. Figure 16 reveals an interesting result for the Nasdaq-100 index: reducing the dimension by half (2:1) actually improves the log(MSE) from − 5.960 to − 7.590. This indicates that a moderate reduction enhances the prediction accuracy. However, a further reduction to a 12:1 setting decreases accuracy, with the log(MSE) deteriorating to − 4.950. This suggests that excessive reduction may discard information that is valuable for accurate predictions.

While MSE-2 may provide improved prediction accuracy, particularly at certain levels of dimensionality reduction, it has the drawback of increased computation time. The method requires each individual stock to be forecast, which is computationally intensive, especially for indexes with numerous constituent companies such as the Nasdaq-100. There is therefore a trade-off between prediction accuracy and computational efficiency that needs to be considered when choosing between MSE-1 and MSE-2. MSE-2 also shows that incorporating individual stock forecasts before aggregating them into the market state enhances index predictions. MSE-2 is as accurate as MSE-1 for the DJ, albeit with a higher computational cost due to the additional forecasting steps. In turn, MSE-2 outperforms MSE-1 for Nasdaq-100 at certain levels of dimensionality reduction, highlighting the crucial role of dimensionality in prediction accuracy. However, higher computational requirements mean that MSE-2 is best suited for applications in which prediction accuracy is prioritized over computation time or when sufficient computational resources are available.

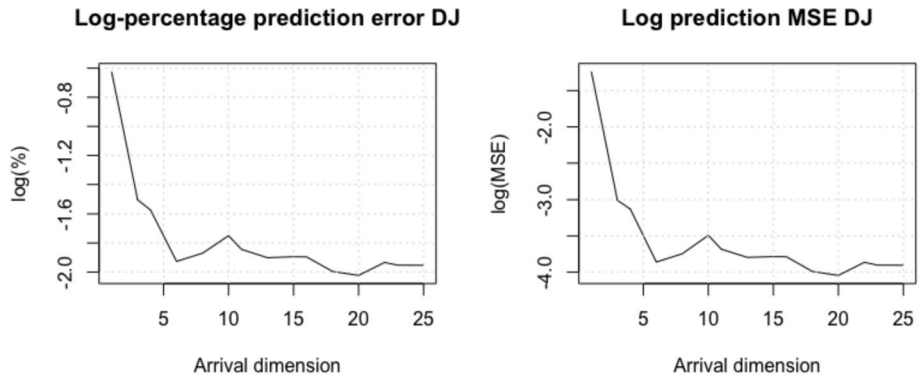


Fig. 17 Log(MSE) and percentage error of the DJ predictions using MSE-3 with data from the constituent companies

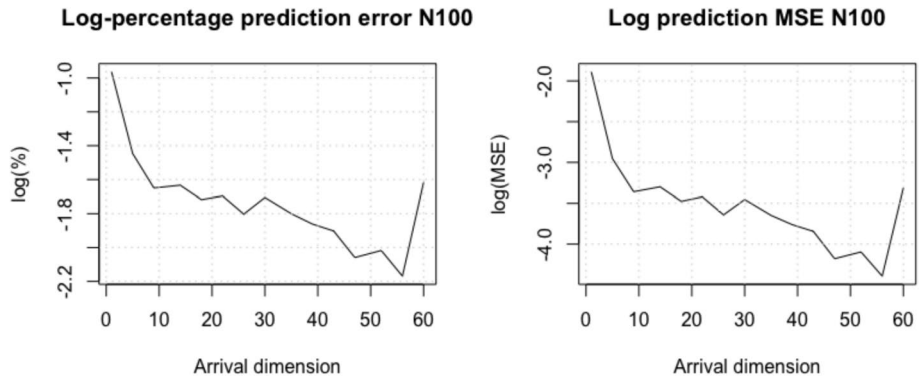


Fig. 18 Log(MSE) and percentage error of the Nasdaq-100 predictions using MSE-3 with data from the constituent companies

MSE-3

Unlike MSE-1 and MSE-2, which either assume constant system dynamics or forecast individual stock prices, MSE-3 aims to predict the future market state x_{k+1} directly within the lower-dimensional space. This approach bypasses the need to predict each company's stock price individually and instead focuses on the reduced market state as a whole, using Kalman forecasting. MSE-3 therefore propagates compressed linear combinations of the original market-state representation without interpreting these directions as principal components or estimated economic factors. Figures 17 and 18 present the prediction errors and mean squared errors for the DJ and Nasdaq-100 indexes, respectively, revealing how MSE-3 performs across various levels of dimensionality reduction.

Figures 17 and 18 show that MSE-3 records lower percentage errors than MSE-1 and MSE-2. This suggests that, in relative terms, MSE-3 provides better predictions of index movements. Despite the improvement in percentage errors, the log(MSE) values are higher (worse) for MSE-3 than for the preceding methods. This indicates that the absolute differences between the predicted and actual index values are larger.

Regarding computation time, MSE-3 requires more processing time than MSE-1 and MSE-2. Figure 19 illustrates the average execution time per prediction for both indexes using MSE-3.

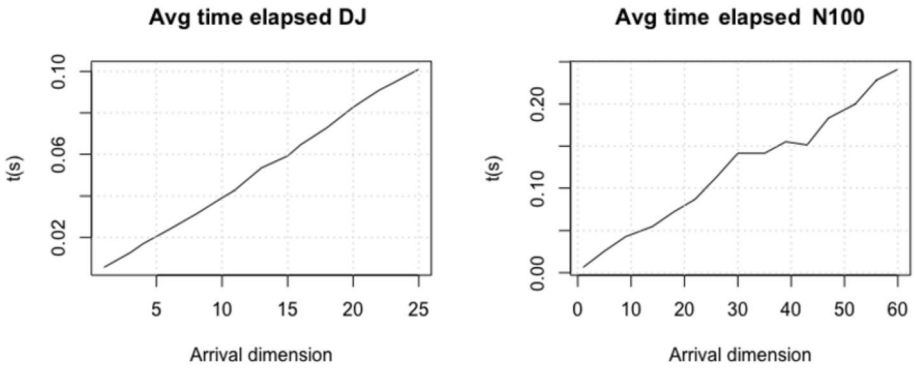
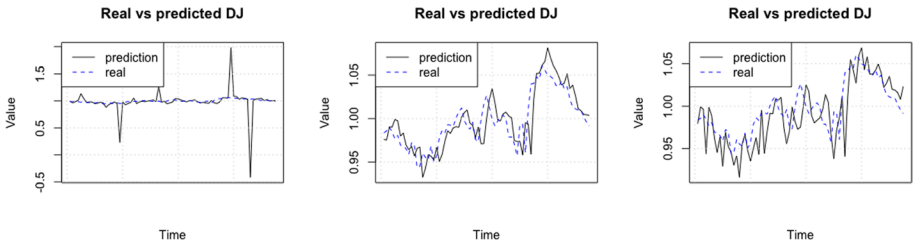


Fig. 19 Average execution time per prediction for both indexes using MSE-3 with data from the constituent companies



Reduction 1:1
Log(MSE) = -3.79

Reduction 2:1
Log(MSE) = -5.72

Reduction 25:3
Log(MSE) = -4.45

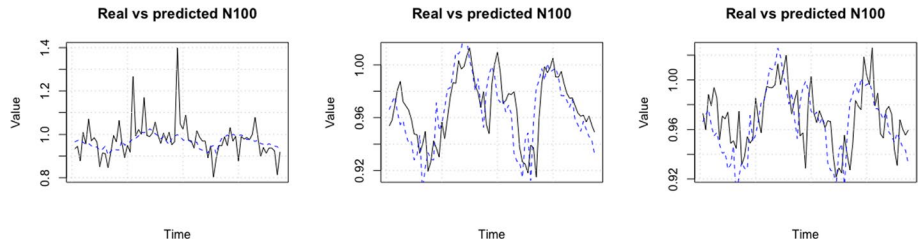
Fig. 20 Actual vs. predicted DJ values using MSE-3 with varying levels of dimensionality reduction over a one-year time series

MSE-3 is slower because it involves forecasting the reduced market state, which is less constant and has condensed information, making the Kalman calculations more complex.

Comparison with other methods: MSE-3 takes longer than MSE-1 (which assumes constant matrices) and longer than MSE-2 (which forecasts individual stock prices), indicating that the computational benefits of dimensionality reduction are offset by the increased complexity of forecasting in the reduced space. In any case, reducing the state size from m to k decreases the core linear algebra costs from $\mathcal{O}(m^3)$ to $\mathcal{O}(k^3)$, explaining the runtime differences reported.

We further analyze the effect of dimensionality reduction on prediction accuracy and computational efficiency. Figures 20 and 21 display the actual versus predicted index values for DJ and Nasdaq-100, respectively, at different levels of dimensionality reduction.

An interesting phenomenon arises when no dimensionality reduction is applied (reduction 1:1). In this scenario, the log(MSE) values are higher, indicating a weaker performance compared to those cases involving reduced dimensions. This outcome is counterintuitive because retaining all the dimensions would be expected to preserve more information and, consequently, yield better predictions. The increase in prediction error when no dimensionality reduction is applied may be attributed to a combination of the two factors. First, even when projecting onto a space of the same dimensionality, the



Reduction 1:1 Log(MSE) = -4.03 | Reduction 2:1 Log(MSE) = -5.59 | Reduction 12:1 Log(MSE) = -4.86

Fig. 21 Actual vs predicted Nasdaq-100 values using MSE-3 with varying levels of dimensionality reduction over a one-year time series

random projection introduces distortion due to the random matrix effect. This distortion impairs the quality of the data used for forecasting. Second, forecasting the projected market state adds additional error, especially in the absence of dimensionality reduction, when the variability is higher.

Dimensionality reduction (e.g., ratios of 2:1 or higher) generally improves log(MSE), indicating better absolute accuracy. However, excessive reduction (e.g., from 25 dimensions to 3) can increase errors because of information loss, suggesting an optimal reduction level. In comparative terms, MSE-3 achieves a lower percentage of errors than MSE-1 and MSE-2, indicating stronger relative predictions, but exhibits higher log(MSE), reflecting weaker absolute accuracy. This divergence suggests that MSE-3 captures proportional movements more effectively than exact index levels.

In terms of computation, MSE-3 is more demanding, even though it operates in a lower-dimensional space, because forecasting the compressed state is more complex and variable. This added cost does not translate into superior absolute accuracy, reducing its overall efficiency. Notably, moderate dimensionality reduction occasionally enhances performance—particularly for the Nasdaq-100—suggesting that random projections may filter noise and improve signal clarity under certain conditions.

Results and discussion

Main results

This section summarizes and analyses the performance of the standard Kalman filter and the three variants of our market state estimation forecasting framework (MSE-1, MSE-2, and MSE-3) applied to the DJ and Nasdaq-100 indexes. We compare prediction accuracies, computational efficiencies, and the impact of dimensionality reduction on forecasting performance. After presenting the results based on the baseline level and returns, we report additional diagnostic checks on dimensionality reduction, return-oriented metrics, and missing data treatment.

We begin by presenting the key results in Tables 1, 2, 3, 4, which compile the log(MSE) and the average computation time per iteration for each method. In the tables, the σ symbol indicates the standard deviation, and the $CI95_{low}/CI95_{high}$ notation shows the lower and upper bounds of the 95% confidence interval for those mean values, where $halfwidth = 1.96 \times \sigma/\sqrt{(N)}$, $CI95_{low} = \overline{x} - halfwidth$, $CI95_{high\_} = \overline{x} + halfwidth$; C is the compression ratio. The expressions with a bar over them, $\overline{log(MSE)}$ and $\overline{time}$, represent

Table 1 Standard kalman prediction for the 2019 dataset

Index	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(\text{err})_{\text{low}}$	$CI95(\text{err})_{\text{high}}$	$\overline{\text{time}}$	σ_{time}	$CI95(\text{time})_{\text{low}}$	$CI95(\text{time})_{\text{high}}$
DJ	−5.780	0.000	−5.780	−5.780	2.800	0.000	2.800	2.800
N100	−9.680	0.000	−9.680	−9.680	3.500	0.000	3.500	3.500

Table 2 MSE-1 performance for the 2019 dataset

Index	C	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(\text{err})_{\text{low}}$	$CI95(\text{err})_{\text{high}}$	$\overline{\text{time}}$	σ_{time}	$CI95(\text{time})_{\text{low}}$	$CI95(\text{time})_{\text{high}}$
DJ	1:1	−8.910	0.128	−8.920	−8.900	1.700	0.006	1.700	1.700
DJ	2:1	−5.690	0.343	−5.710	−5.670	0.700	0.000	0.700	0.700
DJ	25:3	−3.830	0.099	−3.840	−3.820	0.200	0.015	0.199	0.201
N100	1:1	−5.990	0.066	−5.994	−5.986	7.200	0.011	7.199	7.201
N100	2:1	−5.850	0.180	−5.860	−5.840	2.900	0.000	2.900	2.900
N100	12:1	−4.930	0.097	−4.940	−4.920	0.500	0.004	0.500	0.500

Table 3 MSE-2 performance for the 2019 dataset

Index	C	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(\text{err})_{\text{low}}$	$CI95(\text{err})_{\text{high}}$	$\overline{\text{time}}$	σ_{time}	$CI95(\text{time})_{\text{low}}$	$CI95(\text{time})_{\text{high}}$
DJ	1:1	−8.030	0.002	−8.030	−8.030	82.900	0.000	82.900	82.900
DJ	2:1	−5.820	0.003	−5.820	−5.820	82.200	0.002	82.200	82.200
DJ	25:3	−4.190	0.014	−4.191	−4.189	81.600	0.002	81.600	81.600
N100	1:1	−5.960	0.003	−5.960	−5.960	184.500	0.002	184.500	184.500
N100	2:1	−7.590	0.001	−7.590	−7.590	182.300	0.011	182.299	182.301
N100	12:1	−4.950	0.001	−4.950	−4.950	178.300	0.002	178.300	178.300

Table 4 MSE-3 performance for the 2019 dataset

Index	C	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(\text{err})_{\text{low}}$	$CI95(\text{err})_{\text{high}}$	$\overline{\text{time}}$	σ_{time}	$CI95(\text{time})_{\text{low}}$	$CI95(\text{time})_{\text{high}}$
DJ	1:1	−3.790	0.001	−3.790	−3.790	103.200	0.005	103.200	103.200
DJ	2:1	−5.720	0.002	−5.720	−5.720	43.100	0.015	43.099	43.101
DJ	25:3	−4.450	0.006	−4.450	−4.450	13.200	0.001	13.200	13.200
N100	1:1	−4.030	0.001	−4.030	−4.030	253.200	0.010	253.199	253.201
N100	2:1	−5.590	0.002	−5.590	−5.590	145.200	0.002	145.200	145.200
N100	12:1	−4.860	0.006	−4.860	−4.860	26.100	0.003	26.100	26.100

the mean (average) values for the log(MSE) and processing time over N repetitions. Times are measured in milliseconds and are calculated per prediction. Fixed pipelines can yield near-zero σ because the baseline Kalman configuration is deterministic and does not involve random projections.

From these results, we may infer the following. First, for the DJ index, the MSE-1 algorithm records the lowest log(MSE) of −8.91 without dimensionality reduction (1:1), outperforming the standard Kalman filter, which has a log(MSE) of −5.78. This indicates that MSE-1 provides more accurate predictions for the DJ when using the market state data on its constituent companies. These comparisons should be interpreted as forecast-comparison results. A lower log(MSE) indicates a more accurate index-level forecast

in the statistical sense and provides a useful input for future trading rule or portfolio strategy studies. The next step for establishing economic value would be an implementation study using tradable instruments, explicit portfolio rules, transaction costs, liquidity constraints, and risk-adjusted performance metrics. Second, for the Nasdaq-100 index, the standard Kalman filter yields the lowest log(MSE) of −9.68, surpassing all MSE variants. This suggests that historical price data alone are more effective for predicting this index than incorporating market state data through MSE algorithms. Third, the MSE-2 algorithm performs well, particularly for the Nasdaq-100 index, with a reduction ratio of 2:1 and a log(MSE) of -7.59, which is better than the log(MSE) results for MSE-1 and MSE-3 but still does not surpass the standard Kalman filter. Finally, the MSE-3 algorithm generally produces higher log(MSE) values than MSE-1 and MSE-2, reflecting a lower prediction accuracy in absolute terms. This is consistent across both indexes and various levels of dimensionality reduction.

The varying performance of the algorithms across the two indexes is consistent with the general NFL intuition that no single predictive method is uniformly best across all forecasting problems. We do not interpret these findings as a formal test of the NFL theorem. Rather, the results provide an empirical illustration that forecasting performance depends on the match between the predictive representation, the index structure, and the market regime. The main claim is that the effectiveness of forecasting methods for stock indexes depends on the index's structural characteristics, particularly volatility, sector composition, and balance between cross-sectional and time series information. For the Dow Jones Index, composed of 30 large-cap firms, incorporating market-state information with MSE-1—using the contemporaneous market state without forecasting individual stocks—outperforms the standard Kalman filter. This suggests that for indexes characterized by established firms and relatively stable price dynamics, cross-sectional market structure can enhance predictive performance. Conversely, Nasdaq-100, which is technology intensive and more volatile, is better predicted by the standard Kalman filter. Here, historical price dynamics capture information that is not fully reflected in the contemporaneous market state, making temporal dependencies more prominent in forecasting. Overall, these findings highlight the critical influence of the characteristics of the index on forecasting success.

While the comparative analysis focuses on descriptive performance metrics such as MSE and prediction error, the observed differences across the models and configurations are consistent and robust over multiple simulation runs, mitigating the effects of randomness and supporting the reliability of our findings. Next, we show the replication of our analyses for 2024 in Tables 5, 6, 7, 8.

The 2024 replication yields generally less negative log(MSE) values than 2019, indicating a more challenging post-COVID environment. Differences in interest rates, sector concentration, market cycle phase, and inflation-related beta instability may affect

Table 5 Standard kalman prediction for the 2024 dataset

Index	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(\text{err})_{\text{low}}$	$CI95(\text{err})_{\text{high}}$	$\overline{\text{time}}$	σ_{time}	$CI95(\text{time})_{\text{low}}$	$CI95(\text{time})_{\text{high}}$
DJ	−4.269	0.000	−4.269	−4.269	1.090	0.000	1.090	1.090
N100	−4.031	0.000	−4.031	−4.031	1.610	0.000	1.610	1.610

Table 6 MSE-1 performance for the 2024 dataset

Index	C	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(err)_{low}$	$CI95(err)_{high}$	$\overline{time}$	σ_{time}	$CI95(time)_{low}$	$CI95(time)_{high}$
DJ	1:1	−4.682	0.205	−4.695	−4.669	0.047	0.002	0.047	0.047
DJ	2:1	−4.468	0.210	−4.481	−4.455	0.045	0.000	0.045	0.045
DJ	25:3	−3.711	0.770	−3.759	−3.663	0.044	0.000	0.044	0.044
N100	1:1	−4.789	0.054	−4.792	−4.786	0.129	0.055	0.126	0.132
N100	2:1	−4.683	0.046	−4.686	−4.680	0.049	0.001	0.049	0.049
N100	12:1	−4.148	0.269	−4.165	−4.131	0.044	0.001	0.044	0.044

Table 7 MSE-2 performance for the 2024 dataset

Index	C	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(err)_{low}$	$CI95(err)_{high}$	$\overline{time}$	σ_{time}	$CI95(time)_{low}$	$CI95(time)_{high}$
DJ	1:1	−4.267	0.005	−4.267	−4.267	0.029	0.000	0.029	0.029
DJ	2:1	−4.261	0.007	−4.262	−4.261	0.028	0.000	0.028	0.028
DJ	25:3	−4.154	0.172	−4.165	−4.143	0.027	0.000	0.027	0.027
N100	1:1	−4.032	0.003	−4.032	−4.031	0.057	0.001	0.057	0.057
N100	2:1	−4.032	0.002	−4.032	−4.032	0.030	0.000	0.030	0.030
N100	12:1	−4.022	0.009	−4.023	−4.022	0.027	0.001	0.027	0.027

Table 8 MSE-3 performance for the 2024 dataset

Index	C	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(err)_{low}$	$CI95(err)_{high}$	$\overline{time}$	σ_{time}	$CI95(time)_{low}$	$CI95(time)_{high}$
DJ	1:1	−4.269	0.002	−4.269	−4.269	0.018	0.001	0.018	0.018
DJ	2:1	−4.268	0.004	−4.268	−4.267	0.019	0.004	0.019	0.019
DJ	25:3	−4.058	0.446	−4.086	−4.030	0.018	0.001	0.018	0.018
N100	1:1	−4.030	0.002	−4.030	−4.030	0.026	0.007	0.026	0.026
N100	2:1	−4.030	0.003	−4.030	−4.030	0.018	0.001	0.018	0.018
N100	12:1	−4.028	0.007	−4.029	−4.028	0.017	0.000	0.017	0.017

cross-sectional relationships (Valadkhani 2025). The comparison between 2019 and 2024 therefore provides initial evidence of regime sensitivity and motivates broader multiyear validation. Across both years, most between-method differences exceed the CI95 intervals, supporting stable rankings under repeated random projections.

Additional results using return representations

To assess whether the main findings depend critically on the use of normalized raw prices in the market-state definition, we conduct an additional sensitivity analysis in which the same framework is rerun using return-based inputs instead of price levels. The main results of this auxiliary analysis are reported in Tables 9, 10, 11. The return-based specification serves as a robustness check. It shows that the MSE framework is not restricted to a single data encoding, while the main analysis continues to rely on normalized price levels because they are more closely aligned with our baseline notion of market state.

In the return-based specification for 2019, MSE-1 without dimensionality reduction (1:1) achieves the highest predictive accuracy among the MSE variants for both Dow

Table 9 MSE-1 performance for the 2019 dataset with returns

Index	C	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(\text{err})_{\text{low}}$	$CI95(\text{err})_{\text{high}}$	$\overline{\text{time}}$	σ_{time}	$CI95(\text{time})_{\text{low}}$	$CI95(\text{time})_{\text{high}}$
DJ	1:1	−0.195	0.035	−0.201	−0.188	0.2109	0.0288	0.2052	0.2165
DJ	2:1	−0.172	0.051	−0.182	−0.162	0.0499	0.0040	0.0492	0.0507
DJ	25:3	0.373	0.300	0.315	0.432	0.0489	0.0030	0.0483	0.0495
N100	1:1	−0.242	0.018	−0.245	−0.238	0.0898	0.0046	0.0889	0.0907
N100	2:1	−0.237	0.024	−0.241	−0.232	0.0733	0.0013	0.0730	0.0736
N100	12:1	−0.161	0.065	−0.174	−0.148	0.0486	0.0003	0.0486	0.0487

Table 10 MSE-2 performance for the 2019 dataset with returns

Index	C	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(\text{err})_{\text{low}}$	$CI95(\text{err})_{\text{high}}$	$\overline{\text{time}}$	σ_{time}	$CI95(\text{time})_{\text{low}}$	$CI95(\text{time})_{\text{high}}$
DJ	1:1	0.013	0.002	0.013	0.013	0.1314	0.0137	0.1287	0.1340
DJ	2:1	0.013	0.003	0.012	0.013	0.0312	0.0002	0.0311	0.0312
DJ	25:3	0.053	0.050	0.043	0.063	0.0303	0.0003	0.0302	0.0304
N100	1:1	0.015	0.001	0.015	0.015	0.0575	0.0061	0.0563	0.0587
N100	2:1	0.015	0.002	0.015	0.015	0.0408	0.0007	0.0406	0.0409
N100	12:1	0.017	0.003	0.017	0.018	0.0303	0.0002	0.0303	0.0304

Table 11 MSE-3 performance for the 2019 dataset with returns

Index	C	$\overline{log(MSE)}$	$\sigma_{log(MSE)}$	$CI95(\text{err})_{\text{low}}$	$CI95(\text{err})_{\text{high}}$	$\overline{\text{time}}$	σ_{time}	$CI95(\text{time})_{\text{low}}$	$CI95(\text{time})_{\text{high}}$
DJ	1:1	0.011	0.002	0.011	0.011	0.0210	0.0066	0.0197	0.0223
DJ	2:1	0.011	0.003	0.011	0.012	0.0197	0.0006	0.0196	0.0199
DJ	25:3	0.042	0.037	0.035	0.050	0.0195	0.0026	0.0190	0.0200
N100	1:1	0.012	0.001	0.012	0.012	0.0212	0.0007	0.0210	0.0213
N100	2:1	0.012	0.002	0.012	0.012	0.0210	0.0008	0.0209	0.0212
N100	12:1	0.013	0.004	0.013	0.014	0.0197	0.0005	0.0196	0.0198

Jones and Nasdaq-100. This result indicates that the proposed framework is applicable under both level-based and return-based state representations, although the relative performance of the MSE variants depends on the representation, the target index, and the evaluation setting.

The return-based analysis complements the baseline level-based results by examining the framework under an alternative market-state encoding. Given that the study's central question is whether the contemporaneous cross-sectional market state is predictive, the main analysis continues to rely on normalized price levels. Because level-based and return-based targets differ in scale, log(MSE) values are most informative compared within each representation. Accordingly, the return-based results are intended to benchmark MSE variants within this specification and to assess whether the predictive relevance of market-state representations persists under alternative encoding.

Additional diagnostic robustness checks

We also conduct compact diagnostic checks on two implementation choices: dimensionality reduction and missing data treatment. These checks complement Tables 1–11 and should be interpreted within the auxiliary diagnostic setting rather than as replacements for the baseline results. First, we compare random projections with PCA using the same forecasting framework and comparable component counts. PCA may preserve dominant cross-sectional comovements more effectively, whereas random projections remain basis-agnostic and computationally simple. Under forward filling, PCA-MSE-1 produces the strongest diagnostic results, in terms of log(MSE) and return MAE, for both indexes. For Nasdaq-100, PCA-MSE-1 with no compression achieves log(MSE) = −5.161, return MAE = 0.00162, and hit ratio = 67.3%. For DJ, PCA-MSE-1 with no compression achieves log(MSE) = -5.569 and return MAE = 0.00100, while the 2:1 PCA configuration achieves the highest DJ hit ratio, 70.9%. These results support the value of the market-state representation and show that different reduction schemes preserve different predictive structures. Second, we test sensitivity to missing data treatment using forward filling, linear interpolation, sample dropping, mean imputation, and median imputation. The aim is to assess whether the market-state advantage remains visible under alternative preprocessing choices (Table 12).

The results show that the main conclusion is not dependent on forward filling. Across the reported treatments, MSE-1 remains the strongest market-state specification and outperforms the corresponding Kalman benchmark for log(MSE), return MAE, and hit ratio. Under forward filling, random projection MSE-1 improves the Kalman benchmark for both indexes: for DJ, the 1:1 configuration reaches log(MSE) = −5.203, return MAE = 0.00157, and hit ratio = 63.6%, compared with −4.448, 0.00381, and 34.6% for Kalman; for Nasdaq-100, the 1:1 configuration reaches log(MSE) = −5.025, return MAE = 0.00193, and hit ratio = 63.7%, compared with −4.226, 0.00505, and 36.5% for Kalman. Under interpolation and sample dropping, the hit ratios are even higher, exceeding 83% for DJ and 84% for Nasdaq-100 in the reported MSE-1 configurations.

The sensitivity check also shows that preprocessing choices matter. Mean and median imputation produces weaker absolute forecasting performance than forward filling, interpolation, or sample dropping, especially in terms of log(MSE) and return MAE. Nevertheless, MSE-1 remains stronger than the corresponding Kalman benchmark even under these more conservative central-tendency alternatives. This suggests that simple central-tendency imputation may preserve less of the cross-sectional geometry used by the market-state vector, while interpolation, forward filling, and sample dropping retain more of the information needed by the forecasting framework. The results therefore support two complementary conclusions: the market-state signal is not merely an artifact of forward filling, as its empirical strength depends on the preprocessing pipeline.

Figure 22 illustrates this effect for the DJ 2019 diagnostic experiment. Forward filling, interpolation, and sample dropping remain closely aligned with the realized path, whereas mean and median imputation produce less stable trajectories, especially for MSE-1. This supports the interpretation that continuity-preserving or sample-filtering procedures retain more of the cross-sectional structure used by the market-state forecast than central-tendency imputation.

Table 12 Diagnostic robustness results for missing data treatment and selected dimensionality-reduction settings

Index	Robustness check	Missing−data treatment	Projection	Method	Reduction	Components	Mean log(MSE)	Mean return MAE	Mean hit ratio
DJ	Baseline benchmark	Forward fill	N/A	Kalman	N/A	1	−4.448	0.00381	34.60%
DJ	Missing-data sensitivity	Forward fill	Random projection	MSE-1	1:1	24	−5.203	0.00157	63.60%
DJ	Missing-data sensitivity	Forward fill	Random projection	MSE-1	2:1	12	−5.004	0.00206	59.90%
DJ	Missing-data sensitivity	Forward fill	Random projection	MSE-1	25:3	2	−4.522	0.00396	55.10%
DJ	Dimensionality-reduction check	Forward fill	PCA	MSE-1	1:1	24	−5.569	0.00100	66.50%
DJ	Dimensionality-reduction check	Forward fill	PCA	MSE-1	2:1	12	−5.561	0.00101	70.90%
DJ	Dimensionality-reduction check	Forward fill	PCA	MSE-1	25:3	2	−5.082	0.00168	63.20%
DJ	Baseline benchmark	Linear interpolation	N/A	Kalman	N/A	1	−4.522	0.00380	37.40%
DJ	Missing-data sensitivity	Linear interpolation	Random projection	MSE-1	1:1	24	−5.426	0.00141	83.10%
DJ	Baseline benchmark	Sample dropping	N/A	Kalman	N/A	1	−4.287	0.00552	52.60%
DJ	Missing-data sensitivity	Sample dropping	Random projection	MSE-1	1:1	24	−5.337	0.00166	88.80%
DJ	Baseline benchmark	Mean imputation	N/A	Kalman	N/A	1	−3.232	0.01710	54.70%
DJ	Missing-data sensitivity	Mean imputation	Random projection	MSE-1	1:1	24	−4.103	0.00474	72.40%
DJ	Baseline benchmark	Median imputation	N/A	Kalman	N/A	1	−3.231	0.01720	52.20%
DJ	Missing-data sensitivity	Median imputation	Random projection	MSE-1	1:1	24	−4.180	0.00470	75.50%
NAS-DAQ-100	Baseline benchmark	Forward fill	N/A	Kalman	N/A	1	−4.226	0.00505	36.50%
NAS-DAQ-100	Missing-data sensitivity	Forward fill	Random projection	MSE-1	1:1	66	−5.025	0.00193	63.70%
NAS-DAQ-100	Missing-data sensitivity	Forward fill	Random projection	MSE-1	2:1	33	−4.893	0.00225	63.20%
NAS-DAQ-100	Missing-data sensitivity	Forward fill	Random projection	MSE-1	25:3	7	−4.516	0.00346	60.30%
NAS-DAQ-100	Dimensionality-reduction check	Forward fill	PCA	MSE-1	1:1	66	−5.161	0.00162	67.30%
NAS-DAQ-100	Dimensionality-reduction check	Forward fill	PCA	MSE-1	2:1	33	−5.143	0.00165	64.80%
NAS-DAQ-100	Dimensionality-reduction check	Forward fill	PCA	MSE-1	25:3	7	−4.943	0.00218	66.20%
NAS-DAQ-100	Baseline benchmark	Linear interpolation	N/A	Kalman	N/A	1	−4.308	0.00503	38.20%
NAS-DAQ-100	Missing-data sensitivity	Linear interpolation	Random projection	MSE-1	1:1	66	−4.897	0.00241	84.80%
NAS-DAQ-100	Missing-data sensitivity	Linear interpolation	Random projection	MSE-1	2:1	33	−5.054	0.00194	85.30%
NAS-DAQ-100	Baseline benchmark	Sample dropping	N/A	Kalman	N/A	1	−4.064	0.00725	53.80%

Table 12 (continued)

Index	Robustness check	Missing–data treatment	Projection	Method	Reduction	Components	Mean log(MSE)	Mean return MAE	Mean hit ratio
NAS-DAQ-100	Missing-data sensitivity	Sample dropping	Random projection	MSE-1	1:1	66	−4.800	0.00307	87.60%
NAS-DAQ-100	Baseline benchmark	Mean imputation	N/A	Kalman	N/A	1	−3.066	0.01997	53.00%
NAS-DAQ-100	Missing-data sensitivity	Mean imputation	Random projection	MSE-1	1:1	66	−3.916	0.00621	81.50%
NAS-DAQ-100	Baseline benchmark	Median imputation	N/A	Kalman	N/A	1	−3.063	0.01981	53.60%
NAS-DAQ-100	Missing-data sensitivity	Median imputation	Random projection	MSE-1	1:1	66	−3.968	0.00634	75.80%

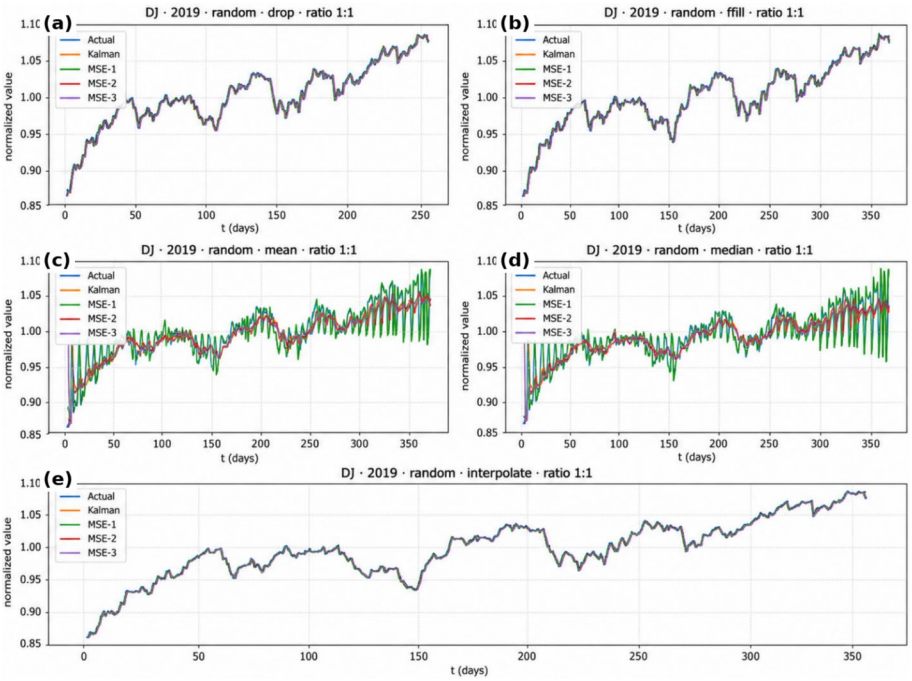


Fig. 22 Visual comparison of missing-data treatments in the DJ 2019 diagnostic experiment using random projection and a 1:1 ratio. The panels show actual normalized index values and forecasts from Kalman, MSE-1, MSE-2, and MSE-3 under **a** sample dropping, **b** forward filling, **c** mean imputation, **d** median imputation and **e** linear interpolation

Overall, the diagnostic checks reinforce the role of MSE-1, which uses the contemporaneous market state directly. MSE-2 and MSE-3 remain closer to the Kalman benchmark, suggesting that the predictive value mainly comes from the current cross-sectional configuration rather than from additional forecasting steps. The checks also show that moderate compression remains viable, while aggressive compression and weaker imputation choices can reduce stability.

Conclusions

This study tests whether cross-sectional market-state representations can improve daily one-step-ahead index forecasts relative to a univariate Kalman benchmark. Across the Dow Jones and Nasdaq-100 experiments, their contribution is inherently conditional—shaped by the target index, the state encoding, and the evaluation window—thereby strengthening the paper's comparative forecasting message while positioning asset-pricing validation, liquidity-adjusted trading performance, and portfolio implementation as natural next layers of inquiry.

Overall, the evidence supports normalized constituent prices as a reliable level-based market-state representation. Return-based experiments further indicate that the framework is not tied to a single encoding, and the diagnostic checks suggest that the market-state signal is not simply a byproduct of forward filling. In addition, PCA-based reductions can retain a stronger predictive structure than random projections in selected configurations, highlighting the relevance of data-dependent compression.

The main conclusions are as follows. First, the most informative input varies across indexes and regimes. For the Dow Jones—dominated by large, relatively stable firms with more coordinated movements—incorporating market-state information via MSE-1 yields more accurate forecasts than relying solely on historical index prices. For the Nasdaq-100, the main 2019 level-based random projection experiment favors the univariate Kalman benchmark, and yet the 2024 replication and diagnostics show that rankings can shift across regimes, encodings, and reduction methods. This pattern reinforces the conclusion that the value of market-state information is dependent on the target, period, and specification rather than being universal. Second, additional forecasting layers do not automatically translate into gains. Although it is feasible to combine market-state and historical index dynamics into richer schemes, the results provide limited evidence of systematic synergy, consistent with the possibility that the two sources overlap in the predictive content they deliver. Third, dimensionality reduction is practically effective, but its impact depends on the method and compression level. Random projections offer a simple, basis-agnostic mechanism for scaling market-state models, whereas PCA diagnostics suggest that data-adaptive reductions may preserve more predictive structure in certain settings. This scalability makes it feasible to exploit richer cross-sectional states without prohibitive runtimes and opens the door to broader markets or other asset classes. Finally, MSE-3 illustrates that combining historical forecasting with reduced market states can introduce nontrivial interactions among compression, state propagation, and index dynamics. Even when not advantageous in predictive accuracy, these patterns are informative in their own right and motivate further theoretical investigation into how a compressed cross-sectional structure propagates into aggregate index behavior.

Practical implications

Our findings highlight several considerations for researchers and practitioners in daily index forecasting. First, the choice of predictive method should be aligned with the characteristics of the forecasting target and the available information set. Historical index prices and cross-sectional market-state representations do not perform uniformly across indexes or periods. Accordingly, the proposed framework should be viewed as a

comparative forecasting tool to support model selection or signal-generation workflows, with potential extensions to trading or portfolio allocation through a dedicated implementation layer. Second, the results indicate that market-state information is more effective for indexes with relatively stable and coordinated constituent behavior, as observed for the Dow Jones. In contrast, for the more volatile Nasdaq-100, the univariate Kalman benchmark performs better in the main 2019 level-based random projection setting, suggesting that historical dynamics capture information that is not fully reflected in that specific market-state encoding. However, Sect. "Additional diagnostic robustness checks" shows that this result can vary under PCA-based reductions and return-based specifications, reinforcing the idea that the value of market-state information is dependent on the target, representation, and specification. Third, dimensionality reduction introduces a practical trade-off between predictive accuracy and computational efficiency. While random projections provide a simple and basis-agnostic compression mechanism, PCA diagnostics indicate that data-dependent reductions may retain additional predictive structure in some settings. Moderate compression reduces the computational burden while preserving relevant information, whereas excessive reduction may degrade performance.

From a financial perspective, these results are most informative within the statistical forecasting layer. Improved forecast accuracy can support portfolio analysis, abnormal return evaluation, and risk-adjusted performance assessment, but such applications require explicit implementation assumptions. Liquidity, in particular, is not merely a technical detail: trading costs, bid–ask spreads, and market depth determine whether forecast gains are economically exploitable. For index applications, ETF proxies offer a natural implementation route, yet their liquidity depends on both secondary-market conditions and the underlying asset basket (Agrrawal and Clark 2009; Amihud 2002; Amihud et al. 2005; Marshall et al. 2018). Liquidity-adjusted evaluation should therefore be treated as a distinct empirical layer.

Caution is also required when generalizing the results across macroeconomic regimes. Changes in inflation, interest rates, or volatility can alter systematic exposures and sector sensitivities, affecting the mapping between market-state representations and future index levels. The MSE framework should thus be interpreted as a recalibrated forecasting tool rather than a stable structural relationship, implying the need for periodic retraining, regime monitoring, and macroconditioned validation in practice.

Translating the framework into tradable systems would require time-varying index weights or divisors to be incorporated and market microstructure elements such as rebalancing rules, transaction costs, liquidity constraints, and execution latency to be addressed. These considerations define a separate implementation layer that would be necessary for robust portfolio or intraday applications.

More broadly, while the framework is linear and transparent, its deployment should be evaluated not only in terms of accuracy and efficiency but also with regard to accountability, market stability, and responsible use. The growing reliance on predictive algorithms in finance raises governance concerns related to transparency, access, and potential systemic effects. At the same time, such approaches offer opportunities for the development of more robust and interpretable decision-support tools. Future research

should therefore link forecasting performance to investment outcomes, risk management, governance practices, and realized economic results.

Limitations and directions for future research

The focused design of this study opens several future directions. First, the forecasting framework can be connected to asset-pricing and portfolio-efficiency tests by transforming forecasts into excess-return signals and evaluating them with GRS-type tests,[5] mean–variance, or stochastic discount factor diagnostics (Jobson and Korkie 1982; Gibbons et al. 1989; Hansen and Jagannathan 1997). Conditional beta models could also interact market-state features with inflation, interest rates, realized volatility, or recession indicators to test macroregime dependence (Valadkhani 2025). Dummy-variable and interaction-term specifications could further be used to assess whether market-state effects differ across market regimes, signals, or asset groups, rather than assuming a single homogeneous forecasting relationship (Waggle et al. 2001). Future specifications could also incorporate calendar dummies and interactions between market-state variables and month-end, quarter-end, holiday, or event-period indicators to test whether predictive content is concentrated in specific seasonal or regime-dependent conditions. Second, future work should extend the two single-year experiments to rolling multiyear validation across pre-COVID, COVID-crisis, post-COVID, inflation/rate-hike, bull, bear, and high-volatility regimes (Hanna 2018; Yang and Chuang 2023; Koukorini et al. 2025).

Third, future work should compare alternative market-state encodings, including log-returns, detrended prices, sector-normalized prices, volatility-adjusted prices, principal components, factor states, and sector aggregates (Wang and Wang 2016; Nobi and Lee 2016; Fama and French 1993). It should also compare random projections, PCA, eigenvector decompositions, factor models, and sector-based reductions using both forecast metrics and finance-specific diagnostics such as correlations, covariance structure, sector exposures, factor loadings, tail dependence, skewness, and higher moments (Jia et al. 2022; Reddy et al. 2020; Berberidis and Giannakis 2017; Donoho and Tanner 2009a; Vempala 2005). Fourth, the missing data treatment deserves broader analysis. The diagnostic checks show that MSE-1 is not driven solely by forward filling, but further research could compare Kalman smoothing, model-based imputation, multiple imputation, and missingness-ratio sensitivity, especially for broader universes, less liquid assets, and international markets (Little and Rubin 2019; Schafer and Graham 2002; Collier et al. 2024).

Fifth, the deterministic MSE framework can be extended into probabilistic state-space models with innovation-covariance estimation, likelihood-based identification, Bayesian filtering, and explicit uncertainty quantification (Bai et al. 2023; Khodarahmi and Maihami 2023; Durbin and Koopman 2012; Shumway and Stoffer 2017). Robust specifications with Huberised innovations or heavy-tailed likelihoods could improve resilience to non-Gaussian behavior (Yañez et al. 2024), while nonlinear and machine learning alternatives provide additional benchmarks (Luc and Wu 2025; Song et al. 2024; Wang

[5] GRS stands for the Gibbons, Ross, and Shanken (1989) test, which checks whether the pricing errors (alphas) are all zero and tests whether a model works and is efficient.

2024). Sixth, implementation-oriented work should add trading volumes and liquidity variables, including turnover, Amihud-type illiquidity, bid–ask spreads, volume-weighted average price (VWAP)-based prices, ETF liquidity, tracking error, transaction costs, and market impact (State Street Global Advisors 2026; Invesco 2026; Agrrawal and Clark 2009; Amihud 2002; Amihud et al. 2005; Marshall et al. 2018).

Finally, the forecasts could be converted into explicit trading or allocation rules by specifying tradable instruments, position sizing, rebalancing, transaction costs, liquidity filters, risk constraints, and benchmark portfolios (Song et al. 2025). Event-window CAR analysis could also evaluate whether market-state forecasts are associated with abnormal return responses around information releases (Agrrawal and Agarwal 2025). Such studies should report net and excess returns, Sharpe and Sortino ratios, maximum drawdown, turnover, abnormal returns, transaction-cost-adjusted performance, and formal predictive-accuracy tests (Sukma and Namahoot 2025; Diebold and Mariano 1995; Pesaran and Timmermann 1992).

Acknowledgements
We thank the editor and three anonymous reviewers for their valuable comments and suggestions, which have significantly improved the quality of this paper. The authors are also grateful to Javier Perote and participants in a seminar in the Universidad Carlos III de Madrid for their constructive comments and suggestions on previous versions of this study. Any errors are our own responsibility.

Author contributions
All authors made substantial contributions to the conceptualization, design, and execution of the study, as well as to the preparation and revision of the manuscript. Specifically: Josué Bustarviejo and Carlos Bousoño-Calzón led the development, implementation, and testing of the predictive algorithms. They also worked in the main interpretation of the results, refinement of the manuscript, literature review, data collection, preprocessing, statistical analysis, model validation, and the formatting and finalization of the submission. Pedro Aceituno-Aceituno and Jose A. Zuñiga-Vicente focused primarily on the literature review, statistical analysis, and model validation. Additionally, they contributed to formatting, proofreading, language editing, and finalizing the submission. All authors read and approved the final manuscript.

Funding
This study was supported by the Spanish Ministry of Science and Innovation and Universities (project number: PID2021-124641NBI00) and the Department of Education, Science and Universities (Community of Madrid, Spain) (Reference: PHS-2024/PH-HUM-294).

Data availability
Historical price data supporting the findings of this study, including data on the Dow Jones and Nasdaq-100 indices, have been deposited in Zenodo with https://doi.org/10.5281/zenodo.10957146 (https://zenodo.org/records/10957146). These datasets are openly available, ensuring adherence to the FAIR Guiding Principles of being findable, accessible, interoperable, and reusable. The datasets were generated as part of this research study and were processed using custom R and python software. Given the public interest and potential for further research, we opted for Zenodo due to its commitment to long-term archiving and its provision of a DOI for each dataset, facilitating easy citation and retrieval. As this research did not involve proprietary or sensitive information, there are no legal or ethical restrictions on the data. The datasets are thus fully accessible without any embargo. Researchers interested in exploring or extending our findings can access the data directly through Zenodo with the provided DOI. Consistent with our commitment to supporting the research community, we believe that sharing our datasets not only enhances the utility and reliability of our work but also increases its impact. We encourage other researchers to utilize these datasets in their work, and we are open to collaboration opportunities that these shared resources might inspire. For further correspondence, please contact the corresponding author, Josué Bustarviejo, at jbmunoz@tsc.uc3m.es. Please, let us know if further clarifications or materials are required.

Declarations

Competing interests
The authors declare no conflicts of interest in this paper.

Received: 5 January 2025 Revised: 10 June 2026 Accepted: 14 June 2026
Published online: 27 July 2026

References

Achlioptas D (2003) Database-friendly random projections: Johnson-Lindenstrauss with binary coins. J Comput Syst Sci 66(4):671–687

Agrrawal P, Agarwal R (2025) A longer-term evaluation of information releases by influential market agents and the semi-strong market efficiency. J Behav Financ 26(1):20–45. https://doi.org/10.1080/15427560.2023.2227303

Agrrawal P, Clark JM (2009) A multivariate liquidity score and ranking device for ETFs. Academy of Financial Services Proceedings.

Alexander C (2008) Market risk analysis, volume II: Practical financial econometrics. Wiley, Chichester.

Aloud ME, Alkhamees N (2021) Intelligent algorithmic trading strategy using reinforcement learning and directional change. IEEE Access 9:114659–114671. https://doi.org/10.1109/ACCESS.2021.3105007

Alp ÖS, Özbek L, Canbaloglu B (2023) An analysis of stock market prices by using extended Kalman filter: the US and China cases. Invest Anal J 52(1):67–82. https://doi.org/10.1080/10293523.2022.2166306

Amihud Y (2002) Illiquidity and stock returns: cross-section and time-series effects. J Financ Mark 5(1):31–56

Amihud Y, Mendelson H, Pedersen LH (2005) Liquidity and asset prices. Found Trends Financ 1(4):269–364

Bai Y, Yan B, Zhou C, Su T, Jin X (2023) State of art on state estimation: Kalman filter driven by machine learning. Annu Rev Control 56:100909

Balasubramanian P, Chinthan P, Baradudeen S, Sriraman H (2024) A systematic literature survey on recent trends in stock market prediction. PeerJ Comput Sci 10:e1700. https://doi.org/10.7717/peerj-cs.1700

Behrendt S, Schmidt A (2021) Nonlinearity matters: the stock price–trading volume relation revisited. Econ Model 98:371–385. https://doi.org/10.1016/j.econmod.2021.03.011

Berberidis D, Giannakis GB (2017) Data sketching for large-scale Kalman filtering. IEEE Trans Signal Process 65(14):3688–3701. https://doi.org/10.1109/TSP.2017.2701846

Bhandari HN, Rimal B, Pokhrel NR, Rimal R, Dahal KR, Khatri RK (2022) Predicting stock market index using LSTM. Mach Learn Appl 9:100320

Blazquez D, Domenech J (2018) Big data sources and methods for social and economic analyses. Technol Forecast Soc Change 130:99–113

Bustos O, Pomares-Quimbaya A (2020) Stock market movement forecast: a systematic review. Expert Syst Appl 156:113464

Campbell JY, Lo AW, MacKinlay AC (1997) The econometrics of financial markets. Princeton University Press, Princeton

Campisi G, Muzzioli S, De Baets B (2024) A comparison of machine learning methods for predicting the direction of the US stock market on the basis of volatility indices. Int J Forecast 40(3):869–880. https://doi.org/10.1016/j.ijforecast.2023.11.002

Candes EJ, Tao T (2005) Decoding by linear programming. IEEE Trans Inf Theory 51(12):4203–4215. https://doi.org/10.1109/TIT.2005.858979

Candes EJ, Wakin MB (2008) An introduction to compressive sampling. IEEE Signal Process Mag 25(2):21–30. https://doi.org/10.1109/MSP.2007.914731

Cheng D, Yang F, Xiang S, Liu J (2022) Financial time series forecasting with multi-modality graph neural network. Pattern Recogn 121:108218. https://doi.org/10.1016/j.patcog.2021.108218

Chhajer P, Shah M, Kshirsagar A (2022) The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction. Decis Anal J 2:100015. https://doi.org/10.1016/j.daj.2022.100015

Chiang TC, Zheng D (2010) An empirical analysis of herd behavior in global stock markets. J Bank Financ 34(8):1911–1921

Chiong KX, Shum M (2019) Random projection estimation of discrete-choice models with large choice sets. Manag Sci 65(1):256–273. https://doi.org/10.1287/mnsc.2017.2907

Chu G, Zhang W, Sun G, Zhang X (2019) A new online portfolio selection algorithm based on Kalman filter and anti-correlation. Phys a Stat Mech Appl 536:120949. https://doi.org/10.1016/j.physa.2019.120949

Cooper MJ, Gutierrez RC Jr, Hameed A (2004) Market states and momentum. J Financ 59(3):1345–1365

Collier ZK, Chawla K, Soyoye O (2024) Optimizing imputation for educational data: Exploring training partition and missing data ratios. J Exp Educ

Dasgupta S, Gupta A (2003) An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct Algor 22(1):60–65. https://doi.org/10.1002/rsa.10073

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13(3):253–263

Donoho D, Tanner J (2009a) Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. Philos Trans Royal Soc A 367(1906):4273–4293. https://doi.org/10.1098/rsta.2009.0152

Donoho DL, Tanner J (2009b) Counting faces of randomly projected polytopes when the projection radically lowers dimension. J Am Math Soc 22(1):1–53. https://doi.org/10.1090/S0894-0347-08-00600-0

Durbin J, Koopman SJ (2012) Time series analysis by state space methods. Oxford University Press, Oxford

Eldar YC, Kutyniok G (eds) (2012) Compressed sensing: Theory and applications. Cambridge University Press, Cambridge

Elsegai H (2022) Improving the process of early-warning detection and identifying the most affected markets: evidence from subprime mortgage crisis and COVID-19 outbreak—application to american stock markets. Entropy 25(1):70. https://doi.org/10.3390/e25010070

Fama EF (1970) Efficient capital markets. J. Finance 25(2):383–417

Fama EF (1991) Efficient capital markets: II. J. Finance 46(5):1575–1617

Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. J Financ Econ 33(1):3–56

Fischer T, Krauss C (2018) Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res 270(2):654–669

Gao H, Kou G, Liang H et al (2024) Machine learning in business and finance: a literature review and research opportunities. Finan Innov 10(1):86. https://doi.org/10.1186/s40854-024-00629-z

Gao W, Aamir M, Shabri AB, Dewan R, Aslam A (2019) Forecasting crude oil price using Kalman filter based on the reconstruction of modes of decomposition ensemble model. IEEE Access 7:149908–149925. https://doi.org/10.1109/ACCESS.2019.2946992

Gardner G, Harvey AC, Phillips GD (1980) Algorithm AS 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. J R Stat Soc C: Appl. Stat 29(3):311–321
Gibbons MR, Ross SA, Shanken J (1989) A test of the efficiency of a given portfolio. Econometrica 57(5):1121–1152
Gottschlich J, Hinz O (2014) A decision support system for stock investment recommendations using collective wisdom. Decis Support Syst 59:52–62
Hanna AJ (2018) A top-down approach to identifying bull and bear market states. Int Rev Financ Anal 55:93–110
Hansen LP, Jagannathan R (1997) Assessing specification errors in stochastic discount factor models. J Finance 52(2):557–590
Invesco (2026) Invesco QQQ ETF: Nasdaq-100 tracking description.
Jia W, Sun M, Lian J, Hou S (2022) Feature dimensionality reduction: a review. Complex Intell Syst 8(3):2663–2693. https://doi.org/10.1007/s40747-021-00600-x
Jobson JD, Korkie B (1982) Potential performance and tests of portfolio efficiency. J Financ Econ 10(4):433–466
Karmiani D, Kazi R, Nambisan A, Shah A, Kamble V (2019, February) Comparison of predictive algorithms: backpropagation, SVM, LSTM and Kalman Filter for stock market. In: Amity Int Conf Artif Intell (AICAI) 2019:228–234. IEEE. https://doi.org/10.1109/AICAI.2019.8701335.
Khandani AE, Kim AJ, Lo AW (2010) Consumer credit-risk models via machine-learning algorithms. J Bank Financ 34(11):2767–2787
Khodarahmi M, Maihami V (2023) A review on Kalman filter models. Arch Comput Methods Eng 30(1):727–747. https://doi.org/10.1007/s11831-022-09777-y
Kou G, Peng Y, Wang G (2014) Evaluation of clustering algorithms for financial risk analysis using MCDM methods. Inf Sci 275:1–12. https://doi.org/10.1016/j.ins.2014.02.137
Koukorini A, Peters GW, Germano G (2025) Generative-discriminative machine learning models for high-frequency financial regime classification. Methodol Comput Appl 27(2):36. https://doi.org/10.1007/s11009-025-10148-8
Krauss C, Do XA, Huck N (2017) Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. Eur J Oper Res 259(2):689–702. https://doi.org/10.1016/j.ejor.2016.10.031
Kumar G, Jain S, Singh UP (2021) Stock market forecasting using computational intelligence: A survey. Arch Comput Methods Eng 28(3):1069–1101
Kumbure MM, Lohrmann C, Luukka P, Porras J (2022) Machine learning techniques and data for stock market forecasting: a literature review. Expert Syst Appl 197:116659. https://doi.org/10.1016/j.eswa.2022.116659
Kurani A, Doshi P, Vakharia A, Shah M (2023) A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Ann Data Sci 10(1):183–208. https://doi.org/10.1007/s40745-022-00430-8
LeRoy SF, Werner J (2014) Principles of financial economics. Cambridge University Press, New York
Li B, Ergu D, Kou G (2025) Editor's introduction: special issue on the anomie of AI in finance; financial markets & investments; economic & policy analysis; corporate governance & related market dynamics. Financ Innov 11:121. https://doi.org/10.1186/s40854-025-00792-x
Li J, Balasubramanian K, Ma S (2023) Stochastic zeroth-order Riemannian derivative estimation and optimization. Math Oper Res 48(2):1183–1211. https://doi.org/10.1287/moor.2022.1284
Lin CY, Marques JAL (2024) Stock market prediction using artificial intelligence: a systematic review of systematic reviews. Soc Sci Humanit Open 9:100864. https://doi.org/10.1016/j.ssaho.2024.100864
Little RJA, Rubin DB (2019) Statistical Analysis with Missing Data, 3rd edn. Wiley
Liu L, Chen J, Zhao G, Fieguth P, Chen X, Pietikäinen M (2019) Texture classification in extreme scale variations using GANet. IEEE Trans Image Process 28(8):3910–3922. https://doi.org/10.1109/TIP.2019.2905640
Lo AW (2004) The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. J Portf Manag 30(5):15–29
Lo AW (2005) Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis. J Invest Consult 7(2):21–44
Luc GB, Wu CC (2025) A dynamic asset allocation approach under technical analysis and machine learning classification. Appl Econ. https://doi.org/10.1080/00036846.2025.2526177
Malkiel BG (2003) The efficient market hypothesis and its critics. J Econ Perspect 17(1):59–82
Marshall BR, Nguyen NH, Visaltanachoti N (2018) Do liquidity proxies measure liquidity accurately in ETFs? J Int Financ Mark Inst Money 55:94–111
Meinhold RJ, Singpurwalla ND (1983) Understanding the kalman filter. Am Stat 37(2):123–127
Mendel J (1971) Computational requirements for a discrete Kalman filter. IEEE Trans Autom Control 16(6):748–758. https://doi.org/10.1109/TAC.1971.1099825
Münnix MC, Shimada T, Schäfer R, Leyvraz F, Seligman TH, Guhr T, Stanley HE (2012) Identifying states of a financial market. Sci Rep 2(1):644. https://doi.org/10.1038/srep00644
Nobi A, Lee JW (2016) State and group dynamics of world stock market by principal component analysis. Physica A Stat Mech Appl 450:85–94. https://doi.org/10.1016/j.physa.2016.01.071
Patel J, Shah S, Thakkar P, Kotecha K (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert Syst Appl 42(1):259–268. https://doi.org/10.1016/j.eswa.2014.07.040
Pesaran MH, Timmermann A (1995) Predictability of stock returns: robustness and economic significance. J Finance 50(4):1201–1228
Pesaran MH, Timmermann A (1992) A simple nonparametric test of predictive performance. J Bus Econ Stat 10(4):461–465
Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, Baker T (2020) Analysis of dimensionality reduction techniques on big data. IEEE Access 8:54776–54788. https://doi.org/10.1109/ACCESS.2020.2980950
Rundo F, Trenta F, Pirrone R, Battiato S (2019) Machine learning for quantitative finance applications: a survey. Appl Sci 9(24):5574. https://doi.org/10.3390/app9245574

Sai S, Arunakar K, Chamola V, Hussain A, Bisht P, Kumar S (2025) Generative AI for finance: applications, case studies and challenges. Expert Syst 42(3):e70018
Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7(2):147–177
Schurz G, Thorn P (2024) Escaping the no free lunch theorem: a priori advantages of regret-based meta-induction. J Exp Theor Artif Intell 36(1):87–119. https://doi.org/10.1080/0952813X.2023.2170512
Sezer OB, Gudelek MU, Ozbayoglu AM (2020) Financial time series forecasting with deep learning: a systematic literature review: 2005–2019. Appl Soft Comput 90:106181. https://doi.org/10.1016/j.asoc.2020.106181
Shao Z, Yao X, Chen F, Wang Z, Gao J (2025) Revisiting time-varying dynamics in stock market forecasting: a multi-source sentiment analysis approach with large language model. Decis Support Syst 190:114362. https://doi.org/10.1016/j.dss.2024.114362
Sharma K, Cerezo M, Holmes Z, Cincio L, Sornborger A, Coles PJ (2022) Reformulation of the no-free-lunch theorem for entangled datasets. Phys Rev Lett 128(7):070501
Shiller RJ (2003) From efficient markets theory to behavioral finance. J Econ Perspect 17(1):83–104
Shumway RH, Stoffer DS (2017) Time series analysis and its applications: with R examples, 4th edn. Springer, New York
Sirignano J, Cont R (2019) Universal features of price formation in financial markets: perspectives from deep learning. Quantit Finance 19(9):1449–1459. https://doi.org/10.48550/arXiv.1803.06917
Song G, Zhao T, Ma X, Lin P, Cui C (2025) Reinforcement learning-based portfolio optimization with deterministic state transition. Inf Sci 690:121538. https://doi.org/10.1016/j.ins.2024.121538
Song Y, Cai C, Ma D, Li C (2024) Modelling and forecasting high-frequency data with jumps based on a hybrid nonparametric regression and LSTM model. Expert Syst Appl 237:121527. https://doi.org/10.1016/j.eswa.2023.121527
State Street Global Advisors (2026) SPDR Dow Jones Industrial Average ETF Trust: fund objective and benchmark description.
Sukma N, Namahoot CS (2025) Enhancing trading strategies: a multi-indicator analysis for profitable algorithmic trading. Comput Econ 65(6):3807–3840. https://doi.org/10.1007/s10614-024-10669-3
Sun Y, Liu L, Xu Y et al (2024) Alternative data in finance and business: emerging applications and theory analysis. Finan Innov 10(1):127. https://doi.org/10.1186/s40854-024-00652-0
Sun W, Liu Z, Yuan C, Zhou X, Pei Y, Wei C (2025) RCSAN residual enhanced channel spatial attention network for stock price forecasting. Sci Rep 15(1):21800. https://doi.org/10.1038/s41598-025-06885-y
Thakkar A, Chaudhari K (2020) Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks. Appl Soft Comput 96:106684. https://doi.org/10.1016/j.asoc.2020.106684
Tsay RS (2010) Analysis of financial time series, 3rd edn. Wiley, Hoboken
Valadkhani A (2025) Inflation-driven instability in US sectoral betas. J Asset Manag 26(5):506–513. https://doi.org/10.1057/s41260-025-00413-3
Vempala S (2005) The random projection method. Am Math Soc, Providence
Venables WN, Ripley BD (2013) Modern applied statistics with S-Plus. Springer, New York
Waggle D, Agrrawal P, Johnson DT (2001) Interaction between Value Line's timeliness and safety ranks. J Invest 10(1):53–62
Wang Y, Wang J (2016) Forecasting stock market indexes using principal component analysis and stochastic time effective neural networks. Neurocomputing 205:66–74. https://doi.org/10.1016/j.neucom.2016.04.030
Wang Z, Zhao H, Zheng M, Niu S, Gao X, Li L (2023) A novel time series prediction method based on pooling compressed sensing echo state network and its application in stock market. Neural Netw 164:216–227. https://doi.org/10.1016/j.neunet.2023.03.022
Wang C (2024) Stock return prediction with multiple measures using neural network models. Finan Innov 10(1):72. https://doi.org/10.1186/s40854-023-00608-w
Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1(1):67–82
Worldbank (2024) Market capitalization of listed domestic companies, *2020*. Accessed 15-November-2024. https://data.worldbank.org/indicator/CM.MKT.LCAP.CD.
Xu Z, Wang Y, Feng X, Wang Y, Li Y, Lin H (2024) Quantum-enhanced forecasting: Leveraging quantum gramian angular field and CNNs for stock return predictions. Financ Res Lett 67:105840. https://doi.org/10.1016/j.frl.2024.105840
Yañez C, Kristjanpoller W, Minutolo MC (2024) Stock market index prediction using transformer neural network models and frequency decomposition. Neural Comput Appl 36(25):15777–15797
Yang WR, Chuang MC (2023) Do investors herd in a volatile market? Evidence of dynamic herding in Taiwan, China, and US stock markets. Financ Res Lett 52:103364. https://doi.org/10.1016/j.frl.2022.103364
Yaroslavsky LP (2015) Can compressed sensing beat the Nyquist sampling rate? Opt Eng 54(7):079701. https://doi.org/10.1117/1.OE.54.7.079701
Yue P, Xu HC, Chen W, Xiong X, Zhou WX (2017) Linear and nonlinear correlations in the order aggressiveness of Chinese stocks. Fractals 25(5):1750041
Zhang Q, Wang X, Zhou X, Chen Q (2020) Application of improved adaptive Kalman filter in China's interest rate market. Neural Comput Appl 32:5315–5327. https://doi.org/10.1007/s00521-019-04568-0

Publisher's Note