



Athens University of Economics and Business

Department of Economics

**A Comprehensive Survey of Economic and
Financial Time Series Forecasting: From
Econometric Models to PCA-Based Statistical
Arbitrage**

Pinelopi Vavoule

Athens, Greece
November 2025



Supervisor: Elias Tzavalis

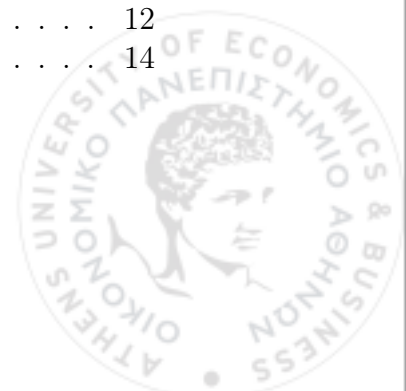
1st Examiner : Spyridon Pagratis

2nd Examiner: Ioannis Kospentaris



Contents

Abstract	VI
Περίληψη	1
1 Introduction	2
2 The Historical and Econometric Foundations of Financial Forecasting	4
2.1 Early Origins of Forecasting	4
2.2 The Keynesian Revolution and the Rise of Econometric Models	5
2.3 Crisis of Econometrics: The Lucas Critique and Rational Expectations	5
2.4 Modern Shifts: Data-Driven and Machine Learning Approaches	6
3 A Taxonomy of Univariate and Multivariate Forecasting Models	7
3.1 Classical Econometric Models	8
3.1.1 The ARIMA family: Modeling Univariate Dynamics . .	8
3.1.2 The GARCH family: Modeling Volatility	9
3.1.3 Vector Autoregression (VAR): Modeling Multivariate Systems	9
3.2 Machine Learning and Deep Learning Paradigms	10
3.2.1 Gradient Boosting Machines (GBM, XGBoost, Light-GBM)	10
3.2.2 Transformer Models: The State of the Art	10
3.3 Comparative Summary	11
4 Principal Component Analysis as a Cornerstone of Modern Financial Modeling	12
4.1 The Mathematical and Conceptual Framework of PCA	12
4.2 PCA for Factor Modeling and Risk Decomposition	14



4.3	Constructing and Trading Statistical Arbitrage Spreads with PCA	14
4.3.1	From Stock Returns to Eigen-Portfolios	14
4.3.2	Modeling Residuals and Mean Reversion	15
4.3.3	Generating Time Signals and Portfolio Construction	15
5	Advanced Topics and the Future of PCA-Based Forecasting	16
5.1	Frontiers in PCA: Advanced Variants and Applications	16
5.2	Hybrid Models: Fusing PCA with Deep Learning	18
6	Methodology and Empirical Implementation	20
6.1	Introduction	20
6.2	Empirical Data Description	21
6.3	Underlying Economic Theory	22
6.4	Extracting Market Factors via PCA	22
6.5	Lagged Features and Data Preparation	24
6.6	Incorporation of Macro and Market Variables	25
6.7	PCA-Based Regression Modeling	26
6.8	Post-LASSO OLS Analysis	28
6.9	Comparison with Time-Series Model	28
6.10	Feature Performance	30
6.11	Next-Day Forecast	31
6.12	Conclusions	32
7	Conclusions and Strategic Guidance	33
7.1	Comparative Overview of Forecasting Models	34
7.2	Strategic Recommendations	34
7.3	Concluding Remarks	35
	Bibliography	35



List of Tables

3.1	Comparative Summary of Forecasting Model Families.	11
6.1	Structure of the data set used for predictive modeling	24
6.2	Regression model performance comparison	25
6.3	Comparison of PCA-Based Regression Models	27
6.4	Post-LASSO OLS t-statistics for significant predictors	28
6.5	Forecasting Performance of the Models	29
6.6	Next-Day Forecast of T10Y2Y Spread using LASSO	31



List of Figures

6.1	Cumulative variance explained by principal components. The dashed red line shows the 90% threshold.	23
6.2	Actual vs Predicted Treasury spread for LASSO regression. . .	27
6.3	Comparison of forecasts from all models.	30
6.4	Feature Importance from LASSO regression.	31



Abstract

This thesis investigates the evolution and empirical implementation of forecasting methodologies for economic and financial time series. It examines the progression from classical econometric models—such as ARIMA, GARCH, and VAR— to modern machine learning frameworks including Gradient Boosting and Transformer architectures. The analysis highlights the trade-offs between interpretability and predictive performance and emphasizes the growing relevance of data-driven methods in high-dimensional financial settings.

A central contribution of the study is the integration of Principal **Components Analysis (PCA)** as a unifying statistical framework bridging econometric and computational approaches. PCA is applied to large panels of equity returns to extract latent factors that capture systematic market dynamics, which are then incorporated into regularized regression models for forecasting movements in the U.S Treasury yield curve.

The empirical analysis uses two datasets: daily stock prices for the 503 S&P 500 constituents from **Yahoo Finance**, and the 10-year minus 2-year U.S Treasury yield spread from **FRED**, covering the period September 2020 - September 2025. The results show that PCA-based regressions, particularly those employing **LASSO** regularization, outperform traditional econometric benchmarks in forecasting accuracy and stability. Overall, the findings support hybrid methodologies that combine econometric structure with data-driven modeling as an effective framework for capturing complex financial dynamics.



Περίληψη

Η παρούσα διατριβή διερευνά την εξέλιξη και την εμπειρική εφαρμογή των μεθοδολογιών πρόβλεψης σε οικονομικές και χρηματοοικονομικές χρονοσειρές. Εξετάζει τη μετάβαση από τα κλασικά οικονομετρικά υποδείγματα—όπως τα ARIMA, GARCH και VAR— προς τα σύγχρονα πλαίσια μηχανικής μάθησης, συμπεριλαμβανομένων των Gradient Boosting και Transformer. Η ανάλυση αναδεικνύει τους συμβιβασμούς μεταξύ ερμηνευσιμότητας και προγνωστικής ακρίβειας, καθώς και τον αυξανόμενο ρόλο των δεδομενοκεντρικών προσεγγίσεων σε περιβάλλοντα υψηλής διαστατικότητας.

Κεντρική συμβολή της μελέτης αποτελεί η ενσωμάτωση της **Ανάλυσης Κύριων Συνιστωσών (PCA)** ως ενοποιητικού στατιστικού πλαισίου που γεφυρώνει την οικονομετρία με τις υπολογιστικές μεθόδους. Η PCA εφαρμόζεται σε εκτεταμένα πάνελ αποδόσεων μετοχών για την εξαγωγή λανθάνοντων παραγόντων που αποτυπώνουν τη συστηματική δυναμική των αγορών, οι οποίοι στη συνέχεια ενσωματώνονται σε κανονικοποιημένα παλινδρομικά μοντέλα για την πρόβλεψη μεταβολών στην καμπύλη αποδόσεων των αμερικανικών κρατικών ομολόγων.

Η εμπειρική ανάλυση βασίζεται σε ημερήσια δεδομένα τιμών μετοχών των 503 εταιρειών του δείκτη S&P 500 από το **Yahoo Finance** και στη διαφορά αποδόσεων 10ετούς - 2ετούς διάρκειας των αμερικανικών ομολόγων από τη βάση **FRED** για την περίοδο Σεπτέμβριος 2020-Σεπτέμβριος 2025. Τα αποτελέσματα δείχνουν ότι τα υποδείγματα PCA, ιδιαίτερα εκείνα με κανονικοποίηση LASSO, υπερέχουν των κλασικών οικονομετρικών μοντέλων, ως προς την ακρίβεια και τη σταθερότητα πρόβλεψης. Συνολικά, τα ευρήματα υποστηρίζουν τις υβριδικές προσεγγίσεις που συνδυάζουν την οικονομετρική δομή με τις δεδομενο-κεντρικές μεθόδους ως αποτελεσματικό πλαίσιο για την κατανόηση και πρόβλεψη των πολύπλοκων δυναμικών των χρηματοοικονομικών αγορών.



Chapter 1

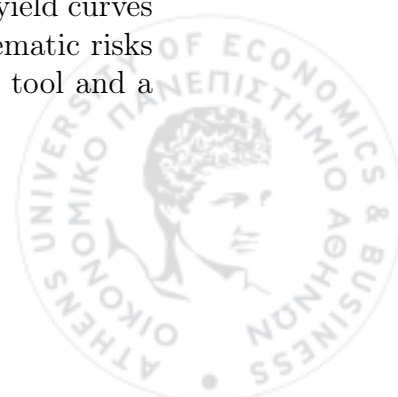
Introduction

Forecasting economic and financial time series has always been a central human pursuit. From ancient attempts to predict harvests to today's algorithmic trading models, the desire to anticipate the future reflects both necessity and ambition. Forecasts guide government policy, corporate strategy, and individual decisions; they shape expectations and often determine outcomes. When forecasts fail—as they did before the Great Depression or the Global Financial Crisis—the consequences are felt globally. When they succeed, they create stability, confidence, and opportunity.

The challenge lies in the nature of the economy itself. Economic systems are adaptive and complex, where human behavior, technology, and institutions interact in unpredictable ways. Models that once seemed powerful eventually reveal their limits. The Keynesian macro-economic systems of the mid-20th century could not explain stagflation in the 1970s. Rational expectations and DSGE frameworks offered elegance, but they largely failed to anticipate the 2008 crisis. Even today, advanced machine learning models—capable of learning non-linear patterns from massive datasets—remain vulnerable to regime shifts and unexpected shocks.

Yet each failure has sparked progress. Econometrics introduced rigor and statistical testing. Machine Learning expanded predictive power. And across both traditions, one recurring need stands out: the ability to make sense of high-dimensional data. Financial markets are noisy and interconnected, with thousands of variables moving simultaneously. Identifying the structure within this complexity is essential.

This is where Principal Components Analysis (PCA) becomes a cornerstone. The PCA distills vast amounts of correlated information into a small number of underlying factors. In bond markets, it reduces entire yield curves into level, slope, and curvature. In equities, it helps isolate systematic risks and reveals mispricing hidden in residuals. As both a statistical tool and a



bridge between classical and modern approaches, PCA allows the forecasters to cut through noise and capture the essence of financial dynamics.

This report is motivated by that perspective. It argues that forecasting matters not just for academic inquiry, but for practical decision-making in environments of uncertainty. It situates PCA within the long arc of forecasting history, examines its methodological role, and demonstrates its empirical applications in financial markets. It also looks forward, considering advanced variants of PCA and hybrid approaches with machine learning.

To situate PCA in context, however, we must first look back. The Chapter 2 traces the long arc of forecasting history—from ancient heuristics and early statistical models to the econometric revolutions of the 20th century and the rise of machine learning. These historical and econometric foundations provide the essential background for understanding both the strengths and the limitations of today’s forecasting toolkit.



Chapter 2

The Historical and Econometric Foundations of Financial Forecasting

The practice of forecasting economic and financial outcomes has evolved through cycles of innovation, application, and eventual reassessment. Each generation of methods has been shaped both by technological progress and by the inability of earlier approaches to anticipate major disruptions. Understanding this trajectory is essential: it highlights not only the foundations of modern econometric forecasting but also the motivations behind the shift toward data-driven methods such as principal component analysis (PCA) and machine learning.¹

2.1 Early Origins of Forecasting

The origins of financial forecasting can be traced to ancient civilizations, where agricultural and financial planning relied on observable indicators. For instance, records from Mesopotamia and Egypt document early attempts to predict harvests using the cyclical patterns of river flooding. While rudimentary, these practices established the principle of using data about the past to guide future decisions.¹

The Renaissance brought a major intellectual shift with the development of probability theory by figures such as Girolamo Cardano, which allowed uncertainty to be treated in a systematic way.



By the 17th and 18th centuries, forecasting had taken more recognizable forms: William Playfair’s graphical representations of data enabled time-series visualization, while indices developed by Laspeyres and Paasche provided the first systematic measures of aggregate prices. These innovations set the stage for the statistical and econometric methods that would define modern forecasting.¹

2.2 The Keynesian Revolution and the Rise of Econometric Models

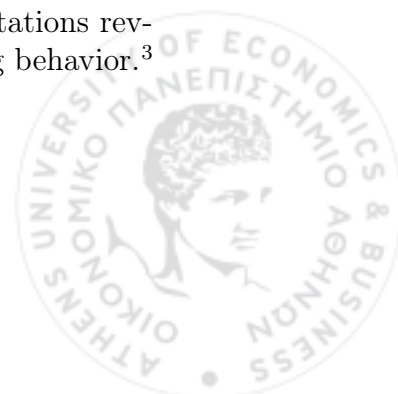
Modern macroeconomic forecasting emerged in response to the failures of early 20th business “business barometers“ that had failed to anticipate the Great Depression. The Keynesian revolution provided the theoretical framework for a new generation of models. Keynes’s *General Theory of Employment, Interest and Money* (1936) and Hicks’s IS-LM framework re-conceptualized the economy as a set of interdependent aggregate variables.²

This theoretical breakthrough coincided with the creation of National Income Accounts during World War II, which offered reliable measures of GDP and related aggregates. Building on these developments, Jan Tinbergen and Lawrence Klein constructed the first large-scale macroeconomic models. By the 1960’s, many governments and research institutions were producing regular forecasts using such systems. Early successes, such as predicting the effects of U.S tax cuts in the 1960’s solidified the status of econometric forecasting as a policy tool.³

2.3 Crisis of Econometrics: The Lucas Critique and Rational Expectations

The optimism surrounding large-scale econometric models was short-lived. The stagflation of the 1970s, triggered by oil shocks and the collapse of the Bretton Woods system, revealed the models’ limitations. Relationships such as the Phillips curve, once thought stable, broke down in practice.³

Theoretical critique soon followed. In 1976, Robert Lucas argued that the parameters of econometric models were not structural but dependent on the policy environment. When policies changed, agents would adjust their expectations, undermining the predictive power of models estimated from historical data. This “Lucas critique“ spurred the rational expectations revolution, which emphasized microfoundations and forward-looking behavior.³



The Dynamic Stochastic General Equilibrium (DSGE) models became the paradigm, embedding rational expectations and optimization by households and firms. Yet, like their predecessors, these models faltered during the Global Financial Crisis of 2008-2009, largely because they failed to incorporate financial instability, systemic risk, and nonlinear feedback mechanisms.³

2.4 Modern Shifts: Data-Driven and Machine Learning Approaches

The shortcomings of both Keynesian and DSGE frameworks reinforced a longstanding lesson: no single theoretical structure can fully capture the complexity of economic and financial systems. Advances in computational power, coupled with the explosion of high-frequency and high-dimensional financial data, have driven a shift toward more flexible, data-driven approaches.³

Modern forecasting now employs a hybrid toolkit that combines classical econometric models with machine learning techniques. Methods such as random forests, support vector machines, and neural networks have been introduced to capture nonlinearities and complex interactions that traditional models overlook. Importantly, factor-based methods such as Principal Components Analysis (PCA) have become central in financial applications, where dimensionality reduction and latent factor extraction are critical for handling large panels of market and macroeconomic variables⁶.

This evolution reflects a broader reorientation of the field: from theory-first approaches that impose structure on data, toward empirically driven techniques that let the data reveal hidden patterns. Chapter 3 examines the PCA in greater detail, positioning it as a cornerstone of modern financial modeling and as a bridge between econometric traditions and statistical arbitrage strategies.



Chapter 3

A Taxonomy of Univariate and Multivariate Forecasting Models

The modern forecaster's toolkit spans a wide range of methodologies, from classical econometric approaches rooted in statistical theory to modern machine learning and deep learning techniques capable of uncovering complex, nonlinear patterns. These models differ not only in their mathematical structure but also in their assumptions, interpretability, and predictive performance.

A critical concept in time series modeling is **stationarity**, where statistical properties such as mean, variance, and autocorrelation remain constant over time. Most financial series, however, exhibit trends or structural breaks and are therefore non-stationary. Preprocessing techniques, such as differencing or detrending, are often required to achieve stationarity before model estimation.⁷

Another key to the methodological distinction is between **in-sample forecasting**—evaluating a model on the same data it is trained on—and **out-of-sample forecasting**, which tests predictive ability on unseen data. The latter is the gold standard for assessing real-world forecast accuracy.

This chapter provides a taxonomy of major forecasting models, spanning classical econometrics, volatility modeling, multivariate systems, and contemporary machine learning paradigms.



3.1 Classical Econometric Models

Classical econometric models remain indispensable as the foundation of time series forecasting. Their assumptions of linearity and stationarity limit their flexibility but make them interpretable and theoretically grounded. These models often serve as benchmarks against which more complex methods are evaluated.⁷

3.1.1 The ARIMA family: Modeling Univariate Dynamics

The **Autoregressive Integrated Moving Average (ARIMA)** model generalizes earlier models to handle both stationary and non-stationary series. It is expressed as ARIMA(p,d,q), where:

- **Autoregressive (AR)** component (order p): captures the dependence of the current value on its past values.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t \quad (3.1)$$

- **Integrated (I)** component (order d): applies differencing to remove non-stationarity. For example, first differencing is⁷:

$$y'_t = y_t - y_{t-1} \quad (3.2)$$

- **Moving Average (MA)** component (order q): models the current value as a function of past forecast errors⁷.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (3.3)$$

ARIMA is particularly effective for short-term univariate forecasting where the immediate past strongly predicts the near future. Its key weaknesses are its linearity assumption and its rapid loss of accuracy over longer horizons, especially in volatile financial contexts.⁷



3.1.2 The GARCH family: Modeling Volatility

Financial returns exhibit **volatility clustering**, where periods of high volatility tend to follow each other. The **Generalized Autoregressive Conditional Heteroskedasticity (GARCH)** family captures this feature by modeling the variance of the error term as a dynamic process.¹⁴ A standard **GARCH(1,1)** specification is¹⁴:

$$y_t = \mu + \varepsilon_t, \quad \varepsilon_t = \sigma_t z_t, \quad z_t \sim N(0, 1) \quad \sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (3.4),$$

where:

- σ_t^2 is the conditional variance at time t .
- The parameters satisfy $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$.
- Persistence is captured when $\alpha + \beta$ is close to 1.

GARCH models are central in risk management, portfolio optimization, and option pricing. Extensions such as **EGARCH** and **TGARCH** address asymmetries (leverage effects), where negative shock increase volatility more than positive ones of the same magnitude.¹⁴

3.1.3 Vector Autoregression (VAR): Modeling Multivariate Systems

Economic and financial variables often evolve jointly. Vector Autoregression (VAR) was introduced by Sims (1980), generalizes the AR framework to multiple variables, capturing feedback loops and dynamic interactions¹⁹. For a bivariate system with variables y_t and x_t , a VAR(1) can be written as:¹⁹

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + [\varepsilon_{y,t} \varepsilon_{x,t}] \quad (3.5).$$

VAR is widely used in macroeconomic forecasting, especially for policy analysis. In the key tools are included:

- **Granger causality tests**: to evaluate predictive relationships between variables¹⁹.
- **Impulse response functions (IRFs)**: to trace the dynamic effects of shocks through the system.¹⁹



3.2 Machine Learning and Deep Learning Paradigms

While classical econometric models prioritize interpretability, they struggle with nonlinear dynamics and high-dimensional datasets. Machine Learning (ML) and deep learning methods, by contrast, are data-driven and excel at pattern recognition in large, complex datasets.¹

3.2.1 Gradient Boosting Machines (GBM, XGBoost, LightGBM)

Gradient Boosting builds models sequentially, where each new model attempts to correct the errors of the ensemble built so far.²³

Formally, the prediction at iteration m is:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (3.6),$$

where $h_m(x)$ is a weak learner (e.g., decision tree), and γ_m is the learning rate.

They are widely used in credit scoring, algorithm trading, and portfolio construction. Also implementations such as XGBoost, and LightGBM improve efficiency and reduce overfitting.²⁷

3.2.2 Transformer Models: The State of the Art

Transformers replaces recurrence with self-attention, allowing simultaneous processing of entire sequences.³⁸ The core mechanism is:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (3.7),$$

where Q, K, V are query key, and value matrices, and d_k is the dimension of the keys. The transformer-based models (e.g., Informer, PatchTST) have achieved state-of-the-art results in time series forecasting, including high-frequency financial data. Their ability to capture long-range dependencies without sequential bottlenecks makes them a promising frontier.³⁸



3.3 Comparative Summary

Family Model	Core Concept	Strengths	Weaknesses	Primary Use Case
ARIMA	Linear dependence on past values and errors	Interpretable, handles non-stationarity	Linear assumption, weak at turning points	Short-term univariate forecasting
GARCH	Conditional variance depends on past shocks and variances	Captures volatility clustering, risk modeling	Only models volatility, not mean	VaR, option pricing, volatility forecasting
VAR	System of equations for multiple interdependent variables	Captures feedback loops, structural analysis	Over-parameterization, assumes linearity	Macroeconomic forecasting, policy analysis
GBM	Sequential ensemble of weak learners	Nonlinear modeling, robust, high accuracy	Risk of overfitting, less interpretable	Credit risk, algorithmic trading
Transformer	Self-attention mechanism	State-of-the-art performance, parallel processing	Complex, high computational cost	High-frequency forecasting, advanced financial models

Table 3.1: Comparative Summary of Forecasting Model Families.

This taxonomy illustrates the trade-off between interpretability and predictive performance. Classical econometric models are grouped in economic theory but often struggle with non linearity and high dimensional data. Machine learning and deep learning models, by contrast, offer superior predictive power at the cost of transparency.

The subsequent chapter turns to **Principal Component Analysis (PCA)**, a technique that bridges the domains by extracting latent factors from high-dimensional datasets—offering both interpretability and practical utility in statistical arbitrage.



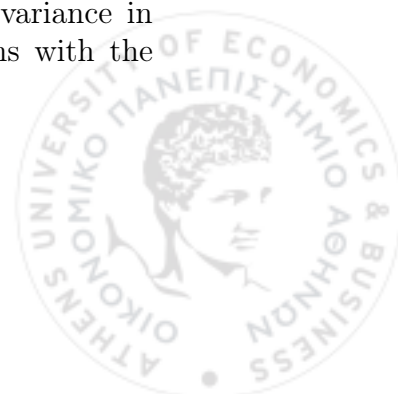
Chapter 4

Principal Component Analysis as a Cornerstone of Modern Financial Modeling

While the models discussed in the previous section provide frameworks for forecasting, they often require a well-defined set of input variables, or features. In modern finance, analysts are confronted with a high-dimensional problem: hundreds or thousands of potentially correlated shocks, economic indicators, and technical variables. The Principal Components Analysis (PCA) is a powerful statistical technique that serves as a cornerstone for addressing this dimensionality challenge. It provides a rigorous method for reducing a complex dataset to its most essential features, enabling more robust modeling, risk decomposition, and the construction of sophisticated trading strategies.

4.1 The Mathematical and Conceptual Framework of PCA

PCA was invented in 1901 by Karl Pearson and later independently developed by Harold Hotelling in the 1930s. At its core, PCA is a dimensionality reduction technique that transforms a dataset of possibly correlated variables into a smaller set of uncorrelated variables called principal components. This is achieved through an orthogonal linear transformation that reorients the data into a new coordinate system. In this system, the first axis, or first principal component, is aligned with the direction of greatest variance in the data. The second axis is orthogonal to the first and aligns with the second-greatest direction of variance, and so on.



By retaining only the first few principal components, one can capture most of the information (Variance) in the original data set with far fewer variables.⁴²

The mathematical engine behind PCA is the eigendecomposition of the data's covariance or correlation matrix. The process involves several steps. First, the data matrix is standardized so that each variable has a mean of zero and a standard deviation of one. This ensures that variables with larger variances do not disproportionately dominate the principal components. The standardized value is computed as:

$$Z = \frac{X - \mu}{\sigma} \quad (4.1),$$

where X is the original value, μ is the mean, and σ is the standard deviation of the variable.

Next, the covariance matrix of the standardized data is calculated. For a matrix of asset returns \mathbf{R} with T time periods and N assets, the covariance matrix is:⁴⁵

$$\Sigma = \frac{1}{T-1} (R - \bar{R})^\top (R - \bar{R}) \quad (4.2)$$

The core of PCA is the eigendecomposition of this covariance matrix:⁴²

$$\Sigma = V\Lambda V^\top \quad (4.3),$$

where V is the matrix of eigenvectors defining the directions of the principal components and Λ is the diagonal matrix of eigenvalues, representing the amount of variance captured by each component. The eigenvectors are ordered from largest to smallest eigenvalue, and selecting the top k eigenvectors allows for a lower-dimensional representation that preserves the most critical patterns in the data.



4.2 PCA for Factor Modeling and Risk Decomposition

In finance, PCA is widely used for factor modeling and risk management. A central concept in asset pricing is that individual asset returns are driven by a set of common underlying risk factors. PCA provides a systematic method to extract these factors empirically from the covariance structure of asset returns, without requiring pre-specified economic assumptions.

When applied to a matrix of stock returns, the first principal component (PC1) typically captures the largest source of common variation, which is often the overall market movement. Subsequent components, orthogonal to PC1, may represent other systematic risks such as sector-specific or style-based factors. For example, PCA has been extensively applied in yield curve analysis, where movements of many correlated interest rates across maturities can be decomposed into three components that explain over 95% of the variance: PC1 (level), PC2 (slope), and PC3 (curvature). This decomposition allows portfolio managers to understand and hedge interest rate risk effectively.⁵⁵

4.3 Constructing and Trading Statistical Arbitrage Spreads with PCA

PCA can also be applied to statistical arbitrage strategies, combining factor modeling with mean-reversion trading. The approach uses PCA to identify statistical risk factors (eigen-portfolios) and isolate the idiosyncratic component of each stock's return, which is then traded under the assumption that it will revert to its mean.

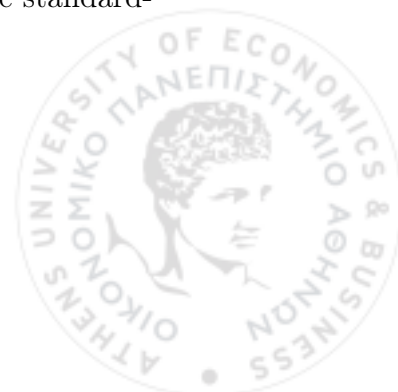
4.3.1 From Stock Returns to Eigen-Portfolios

The process begins with a matrix of historical returns, N stocks over M time periods. Each stock's return series is standardized to ensure high-volatility stocks do not dominate the analysis:

$$Z_{it} = \frac{R_{it} - \bar{R}_i}{\sigma_i} \quad (4.4)$$

PCA is then performed on the empirical correlation matrix of the standardized returns:

$$C = \frac{1}{T-1} Z^T Z \quad (4.5)$$



This yields eigenvalues λ_j and eigenvectors $v^{(j)}$. Each eigenvector defined an eigen-portfolio with weights:

$$Q_i^j = \frac{v_i^{(j)}}{\bar{\sigma}_i} \quad (4.6)$$

The returns of these eigen-portfolios, F_j , are computed and serve as the statistical factors.

4.3.2 Modeling Residuals and Mean Reversion

For each stock, the return is regressed against the eigen-portfolio returns to separate systematic and idiosyncratic components:

$$R_i = \beta_i F + \epsilon_i \quad (4.7),$$

where $\beta_i F$ represents the systematic component, and ϵ_i is the residual targeted for trading. The residual is modeled as a mean-reverting process using the Ornstein-Uhlenbeck stochastic differential equation:

$$dX_t = \kappa(\mu - X_t) dt + \sigma dW_t \quad (4.8),$$

where X_t is the residual at time t , μ is its long-term mean, k is the speed of mean reversion, σ is the volatility, and dW_t is a Brownian motion item.

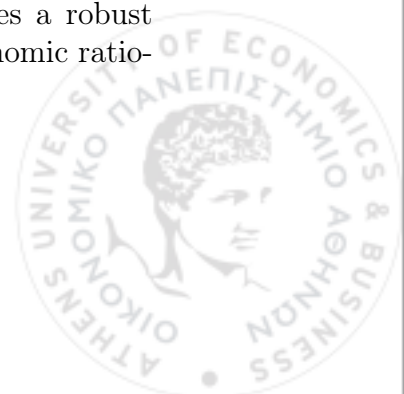
4.3.3 Generating Time Signals and Portfolio Construction

Trading signals are based on the standardized residual, or S-score:

$$s_i = \frac{X_t - \mu}{\sigma_{\text{eq}}} \quad (4.9)$$

A long position is opened when the residual is significantly below the mean (e.g., $s_i < -1.25$) and closed as it reverts toward zero. Short positions are opened when the residual is significantly above the mean (e.g., $s_i > 1.25$). To maintain market neutrality, positions are hedged against all eigen-portfolios based on the factor loadings $\beta_{i,j}$, ensuring the portfolio is exposed only to the idiosyncratic, mean-reverting component.

This methodology elegantly combines data-driven factor discovery via PCA with a model-based mean-reversion hypothesis. By deriving risk factors empirically rather than assuming them a priori, it provides a robust framework for statistical arbitrage while maintaining a clear economic rationale.¹



Chapter 5

Advanced Topics and the Future of PCA-Based Forecasting

While standard PCA provides a powerful foundation for financial modeling, its inherent limitations have spurred the development of more advanced variants and hybrid approaches. The future of financial forecasting lies not in a single master algorithm but in the intelligent combination of modular techniques. In this evolving landscape, PCA's role is shifting from a standalone analysis tool to a critical pre-processing and feature engineering component within larger, more sophisticated machine learning pipelines. This section explores these cutting-edge developments, from supervised extensions of PCA to its integration with deep learning architectures.

5.1 Frontiers in PCA: Advanced Variants and Applications

To understand the motivation for advanced PCA variants, it is essential to first summarize the strengths and weaknesses of classical PCA. Its primary strengths are the ability to reduce data dimensions, uncover hidden patterns by de-correlating features, and improve computational efficiency. However, its weaknesses include the assumption of linear relationships between variables, difficulty in interpreting principal components in terms of original economic variables, inevitable information loss during dimensionality reduction, and sensitivity to outliers.



Several advanced variants have been developed to address these limitations.

1. Scaled PCA (sPCA)

Scaled PCA is designed to overcome a critical shortcoming of standard PCA: its unsupervised nature. Classical PCA analyzes the covariance structure of predictors without considering the target variable, which means the components that explain the most variance may not be the most predictive.

The sPCA scales each predictor by its individual predictive slope on the target variable before performing eigendecomposition, giving more weight to predictors with stronger forecasting power and down-weighting noisy or irrelevant ones. Empirical research shows that sPCA can improve forecast accuracy, particularly when factors are weak or signal-to-noise ratios are low.⁶⁵

2. Sparse PCA

Sparse PCA addresses the interpretability problem inherent in standard PCA. Traditional principal components are linear combinations of all original variables, making them difficult to link to economic factors. Sparse PCA introduces a sparsity constraint, forcing many loadings to zero, which results in components that are combinations of only a small subset of variables, facilitating clearer economic interpretation.⁶⁶

3. Robust PCA

Robust PCA is developed to handle datasets with outliers or corrupted entries, which are common in financial data. It decomposes the data matrix into a low-rank component representing the true structure and a sparse component capturing outliers. This approach is less sensitive to extreme events, providing a more stable factor structure suitable for risk modeling in volatile markets.⁵⁸



5.2 Hybrid Models: Fusing PCA with Deep Learning

A significant frontier in financial forecasting involves hybrid models that combine PCA with deep learning techniques. Neural networks such as LSTMs excel at capturing non-linear dynamics but are sensitive to noisy, highly-dimensional, and correlated input data—a problem known as multicollinearity. PCA is ideally suited to address these challenges by denoising data, reducing dimensionality, and producing uncorrelated input features.³⁵

The process is synergistic: PCA transforms a large set of raw features, such as technical indicators or macroeconomic variables, into a smaller set of uncorrelated principal components. These components then serve as input to deep learning models, which can apply their non-linear modeling capabilities to a cleaner, more manageable feature space. Empirical studies have demonstrated that PCA-LSTM and PCA-BPNN (Backpropagation Neural Network) hybrid models often outperform models without PCA pre-processing. Dimensionality reduction simplifies network architecture, accelerates training, and can improve generalization to new data. However, the benefits are not universal, and in some cases, PCA pre-processing may filter out subtle but relevant information, highlighting the importance of rigorous validation before deployment.³⁵

A close comparison exists between PCA and autoencoders, a type of neural network for unsupervised dimensionality reduction. Autoencoders compress data into a low-dimensional latent space and reconstruct the original data from this encoding. While PCA performs a linear transformation, an autoencoder can learn complex non-linear mappings. Notably, a single-layer linear autoencoder is mathematically equivalent to PCA. The trade-off is that PCA is simple, computationally efficient, and interpretable as ordered variance directions, whereas autoencoders offer greater flexibility to capture non-linear structure but are more complex to train and less interpretable.

Ultimately, the choice between PCA and autoencoders depends on the nature of the data. For primarily linear relationships, PCA provides an efficient, robust solution. For complex, non-linear patterns, autoencoders offer more modeling power, albeit with added computational cost and reduced interpretability.



In either case, PCA remains a critical tool for pre-processing, feature engineering, and the reduction of high-dimensional financial data to its most informative structure, forming a bridge between classical statistical methods and modern machine learning techniques.



Chapter 6

Methodology and Empirical Implementation

6.1 Introduction

This chapter operationalizes the methodological framework developed earlier by applying it to real-world financial data, with a particular focus on forecasting the U.S Treasury 10-year minus 2-year spread (T10Y2Y)—a key indicator of the yield curve’s slope and a proxy for market expectations about future economic conditions. The empirical analysis relies on high-frequency financial data, including daily stock returns for S&P 500 constituents and the daily T10Y2Y spread. Principal components extracted from the cross-section of equity returns are used to summarize broad patterns in equity market behavior, which may contain predictive information for movements in the term structure.

While a rich set of macroeconomic indicators is available from the FRED database and could, in principle, be used to construct PCA-based predictors for the term structure, this analysis deliberately focuses on equity-based PCA factors. Equity prices incorporate forward-looking expectations about economic growth, inflation, and monetary policy, capturing real-time information often not fully reflected in individual macro variables. Moreover, macroeconomic data are typically released at lower frequency and may be subject to revisions, making them less suitable for daily forecasting. As shown in Section 6.6, incorporating macro variables alongside equity PCA factors adds limited incremental predictive power for short-term T10Y2Y movements, supporting the decision to rely primarily on market-based information while acknowledging that macro variables remain important for longer-term trends.



The economic rationale for this approach is that both equity returns and the yield curve are influenced by common macroeconomic expectations, such as anticipated economic growth, inflation, and monetary policy. The analysis is guided by three interrelated questions. First, to what extent can principal components derived from equity market returns provide meaningful insight into the dynamics of the T10Y2Y spread? Second, how do modern machine learning methods, including LASSO and Elastic Net, perform in forecasting compared with traditional economic models such as VAR, ARIMA, and GARCH? Third, which market-driven and macroeconomic variables exert the most significant influence on the behavior of the yield curve?

The following sections provide a detailed description of the data and the underlying economic theory, before presenting the empirical implementation of the forecasting methodology.

6.2 Empirical Data Description

The empirical analysis draws upon two primary datasets. The first dataset consists of daily stock prices for the 503 constituents of the S&P index. These data were obtained from *Yahoo Finance*, a publicly available financial database that provides comprehensive coverage of U.S. equity markets. From these price series, daily logarithmic returns were calculated, yielding a balanced panel of approximately 500 assets spanning the period of September 2020 to September 2025. This dataset captures broad cross-sectional variation in equity market behavior, including sectoral and idiosyncratic factors that are valuable for identifying systematic components through Principal Components Analysis (PCA).

The second dataset concerns the U.S. Treasury term structure, measured by the difference between the 10-year and 2-year **Treasury yields** (T10Y2Y). This variable was sourced from the **Federal Reserve Economic Data** (FRED) database, maintained by the Federal Reserve Bank of St. Louis. The 10Y2Y spread serves as a proxy for the slope of the yield curve, reflecting investors' expectations of future economic growth and monetary policy. Daily observations of the spread were merged with the equity return data to create a unified dataset suitable for high-frequency forecasting. Non-trading days and missing observations were handled through alignment and interpolation to ensure temporal consistency between the two series.



6.3 Underlying Economic Theory

The theoretical motivation for linking stock-return-based principal components to the term structure spread lies in their common dependence on macroeconomic expectations. According to the expectations hypothesis of the yield curve, long-term interest rates reflect the market's outlook for future short-term rates, inflation, and output growth. Similarly, stock returns encapsulate investors' expectations of future corporate earnings, which are themselves influenced by the same macroeconomic fundamentals.

When market participants anticipate economic expansion, both equity valuations and the yield curve slope tend to rise; conversely, periods of tightening financial conditions or declining growth expectations are typically associated with falling stock prices and a flattening or inversion of the yield curve. Consequently, principal components extracted from cross-sectional equity returns can be interpreted as latent economic factors that jointly influence both equity and bond markets. This theoretical linkage provides a sound economic basis for employing equity-derived PCA factors in forecasting movements of the T10Y2Y spread.

While a rich set of macroeconomic indicators is available from the FRED database and could in principle be used to construct PCA-based predictors for the term structure, we deliberately focus on equity based PCA factors. Equity prices incorporate forward-looking expectations about economic growth, inflation, and monetary policy, capturing real-time information often not fully reflected in individual macro variables. Moreover, macroeconomic data are typically released at lower frequency and may be subject to revisions, making them less suitable for daily forecasting. As shown in our results (Section 6.6), incorporating raw macro variables alongside equity PCA factors adds only limited incremental predictive power for short-term T10Y2Y movements, supporting the decision to rely primarily on market-based information while acknowledging that macro variables remain important for longer-term trends.

6.4 Extracting Market Factors via PCA

The first step of the analysis is the application of principal component analysis (PCA) to daily returns of S&P 500 stocks. By construction, PCA reduces the dimensionality of a large dataset, transforming many correlated stock returns into a smaller number of uncorrelated factors. These factors capture the dominant patterns in equity market behavior.



Stationarity of the T10Y2Y spread was confirmed using the *Augmented Dickey-Fuller (ADF)* test. A first difference transformation was applied to ensure stationarity, producing the dependent variable for the regression models.

To ensure that systematic influences are retained without including excessive noise, we select the number of components required to account for approximately 90% of the cumulative variance (Figure 6.1). This ensures that the most significant systematic influences are preserved without overfitting to minor fluctuations. The first few components explain the majority of variation, while additional components gradually capture more specific sources of variation, including sector-specific trends, style factors, and other systematic effects that may influence broader market dynamics. By focusing on these components, we distill complex market interactions into a manageable set of informative predictors suitable for forecasting financial spreads.

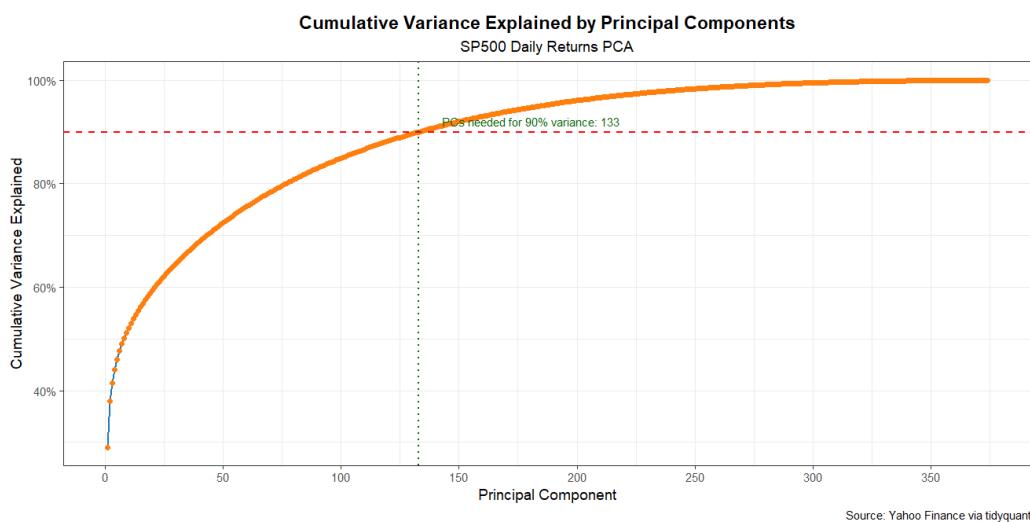


Figure 6.1: Cumulative variance explained by principal components. The dashed red line shows the 90% threshold.

In this study, we retained the first 133 principal components, which collectively capture approximately 90% of the variance in the 503-stock panel. This choice balances the need to preserve the dominant systematic market signals while avoiding overfitting to idiosyncratic noise, ensuring factor extraction for forecasting the T10Y2Y spread.



6.5 Lagged Features and Data Preparation

Following the extraction of principal components, these factors, were merged with the historical T10Y2Y spread data to form the foundation of the forecasting dataset. To capture short-term dynamics and temporal dependencies, lagged features were created for both the spread itself and the principal components, incorporating lags of one to three trading days.

The choice of lag length was based on a combination of economic reasoning and empirical inspection. Short lags (1 to 3 days) are appropriate for high frequency financial data because they allow the models to capture the immediate autocorrelation momentum effects, and short-term feedback mechanisms present in both equity market movements and the yield curve. Longer lags were not included, as preliminary analysis indicated they added limited predictive power for next-day movements and increased model complexity without improving performance.

For time-series benchmarks, lag lengths in ARIMA and VAR models were selected using Akaike and Bayesian Information Criteria (AIC/BIC), along with autocorrelation diagnostics, ensuring sufficient temporal structure without overfitting.

The resulting dataset provides the model with richer temporal context, enabling it to detect subtle relationships between past and current market conditions. The Table 6.1 below, illustrates the dataset structure, showing the inclusion of the spread, its lagged values, and the lagged principal components used as predictors.

Table 6.1: Structure of the data set used for predictive modeling

Date	T10Y2Y_diff	Spread _t	Spread _{t-1}	PC1 _{t-1}	PC2 _{t-1}	PC3 _{t-1}	...
2024-01-02	-0.05	1.25	1.28	0.012	-0.045	0.007	...
2024-01-03	0.00	1.27	1.25	0.015	-0.042	0.010	...
2024-01-04	0.00	1.30	1.27	0.017	-0.040	0.012	...
2024-01-05	0.00	1.28	1.30	0.013	-0.038	0.011	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

This preprocessing step ensures that the forecasting models have access to sufficient historical information to respond effectively to **daily market fluctuations** and short-run yield curve adjustments.



6.6 Incorporation of Macro and Market Variables

To broaden the scope of the analysis and examine potential **cross-market interactions**, additional macroeconomic and market-based variables were incorporated. These included **GDP growth**, **consumer price index (CPI)**, **unemployment rates**, and **Treasury yields**, along with key financial market indicators such as the VIX and sector-specific equity indices. The integration of these variables enables an assessment of whether macroeconomic conditions and broader market sentiment provide incremental predictive information beyond what is captured by equity-derived principal components.

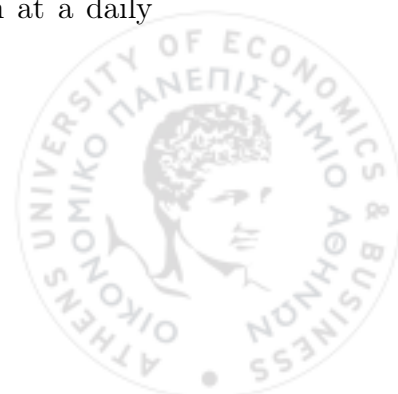
However, the empirical findings reveal that the inclusion of macroeconomic variables does not significantly enhance the model's short-term forecasting performance. Specifically, *LASSO PCA* models, with and without macroeconomic inputs produced almost identical results, as we can see from the Table 6.2 below. This suggests, that the principal components extracted from stock returns already embed much of the information relevant for explaining daily movements in the T10Y2Y spread.

Table 6.2: Regression model performance comparison

Model	RMSE	MAE
LASSO	0.03527	0.02777
LASSO + Macro	0.0331	0.0253
Ridge Regression	0.36742	0.36414
Elastic Net	0.03484	0.02711
Ordinary Least Squares (OLS)	0.55605	0.46349

This outcome highlights a key methodological point: while PCA on macro variables could theoretically summarize the term structure's information, equity-based PCA factors leverage the high-frequency, forward-looking signals embedded in market prices. Consequently, for short-term daily forecasting, equity market information provides superior responsiveness to market dynamics, whereas macro variables are more informative for medium-to long term trends.

We do not perform PCA on macroeconomic variables because they are released at lower frequencies, often subject to revisions, and provide limited incremental predictive power for daily spread movements. Equity-based PCA factors, by contrast, embed forward looking market information at a daily frequency, capturing short-term dynamics more effectively.



6.7 PCA-Based Regression Modeling

The core empirical model relates the daily changes in the T10Y2Y spread to lagged values of the spread and the PCA-derived factors. LASSO regression was chosen as the *main* modeling framework due to its dual capability for coefficient shrinkage and variable selection.

Formally, the LASSO estimator minimizes the residual sum of squares with an L1 penalty term:

$$\min_{\beta} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \quad (6.1),$$

where λ controls the level of regularization and determines the trade-off between model complexity and fit. The optimal penalty parameter λ was selected using *10-fold cross-validation*, ensuring that the model balances predictive accuracy with parsimony.

The LASSO framework is particularly suitable in this context because it simultaneously performs variable selection and shrinkage in a high-dimensional, correlated predictor space, allowing us to identify the principal components most relevant for forecasting the spread. Compared to a PCA regression using macro variables, LASSO applied to equity PCA factors captures richer, high-frequency information and isolates the factors with genuine short-term predictive power.

This approach yields a sparse and interpretable model by automatically excluding irrelevant predictors while stabilizing coefficient estimates – an essential feature in high-dimensional financial datasets where predictors are often correlated and noisy. After the LASSO estimation, a two-step post-estimation procedure was implemented: the retained predictors were re-estimated using Ordinary Least Squares (OLS) to compute valid t-statistics and assess statistical significance. Most coefficients exhibited absolute t-statistics above two, confirming their significance at 5% level and supporting the explanatory relevance of the selected principal components for short-term yield curve dynamics.

For comparison, Ridge regression—which applies an L_2 penalty to prevent excessively large coefficients without discarding variables—was also estimated. Similarly, the Elastic Net approach, combining the L_1 (LASSO) and L_2 (Ridge) penalties, provided a flexible compromise between variable selection and coefficient stabilization. Finally a standard Ordinary Least Squares (OLS) model was included as a benchmark to assess the relative gains from regularization techniques.

The results indicate that LASSO achieves the lowest forecast errors, highlighting its ability to handle high-dimensional and correlated predictors ef-

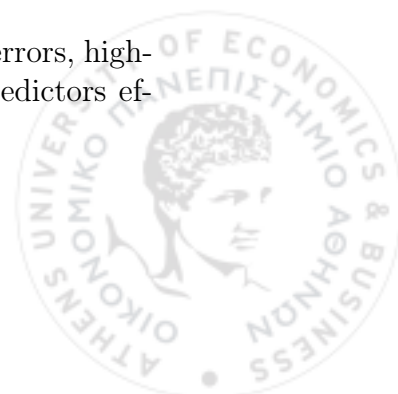


Table 6.3: Comparison of PCA-Based Regression Models

Model	RMSE	MAE
LASSO	0.0329	0.0253
Ridge	0.3679	0.3645
Elastic Net	0.0325	0.0252
Ordinary Least Squares (OLS)	3.6563	2.7602
LASSO + Macro	0.0331	0.0253
Ridge + Macro	0.3148	0.3115
Elastic Net + Macro	0.0337	0.0271
OLS + Macro	7.7809	5.5748

ficiently. Ridge and Elastic Net perform comparably well, while OLS exhibits considerably larger errors, reflecting the limitations of unregularized regression in this context. The residual diagnostics confirm that the selected LASSO model exhibits symmetric, near-zero residuals with no substantial autocorrelation, suggesting model stability and good predictive performance.



Figure 6.2: Actual vs Predicted Treasury spread for LASSO regression.

All regression methods capture the general direction of the spread series, but LASSO predictions track observed movements more closely, particularly during periods of sharp changes. The Ridge and Elastic Net provide intermediate accuracy, whereas OLS tends to deviate further. This highlights the advantage of regularization in high-dimensional financial datasets.



6.8 Post-LASSO OLS Analysis

To assess the statistical relevance of the predictors retained by the LASSO model, we performed a post-estimation OLS regression using only the variables selected in the LASSO step. This two-stage approach allows for conventional inference based on t-statistics, which are not directly available from penalized regression methods.

The table 6.4 below, reports the estimated coefficients, standard errors, and t-values for the significant predictors ($|t| < 2$) identified in the post-LASSO OLS model.

Table 6.4: Post-LASSO OLS t-statistics for significant predictors

Predictor	Estimate	Std.Error	t value	Pr(> t)
PC1	-0.00091	0.000016	-5.80	1.73e-08
PC112	-0.00644	0.00222	-2.90	4.04e-03
PC5_lag3	0.00181	0.00064	2.81	5.32e-03
PC64_lag1	-0.00479	0.00173	0.0253	5.88e-03
T10Y2Y_lag1	0.97721	0.01315	74.31	<2.2e-16

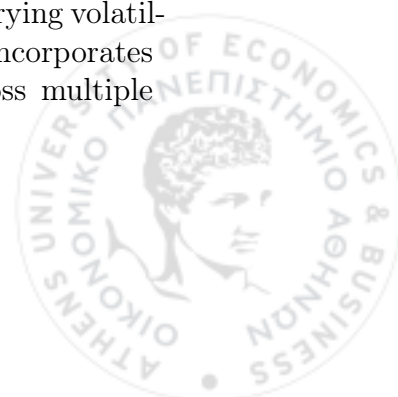
The results highlight the strong persistence of the T10Y2Y spread, as evidenced by the highly significant coefficient of its lagged value. In addition, several principal components—including PC1, PC112, PC5 (lag 3), and PC64(lag 1)—exhibit statistically significant effects.

This indicates that both dominant and less prominent equity market factors provide meaningful information about short-term variations in the yield curve spread.

These findings reinforce the interpretability of the LASSO-selected model and emphasize the importance of capturing cross-market interactions through latent factor structures derived from equity returns.

6.9 Comparison with Time-Series Model

To provide a comprehensive benchmark for the PCA-based regression framework, several classical time-series models were estimated directly on the T10Y2Y spread. These include the **Autoregressive Integrated Moving Average (ARIMA)** model, which captures serial dependence and moving-average effects; the **Generalized Autoregressive Conditional Heteroscedasticity (GARCH)** model, which explicitly models time-varying volatility; and the **Vector Autoregression (VAR)** model, which incorporates macroeconomic variables to account for interdependencies across multiple



economic indicators. We use a forecast horizon of 10-20 steps (approximately 2-4 weeks) to balance short-term accuracy with meaningful propagation of macroeconomic shocks: shorter horizons may not capture the lagged interactions among variables, while longer horizons introduce greater uncertainty due to accumulating forecast errors. This range, allows VAR forecasts to remain responsive to recent market conditions while providing informative near-term predictions for the 10Y2Y spread.

The results clearly indicate that PCA-based regression models—particularly those using LASSO regularization—outperform all classical time-series approaches in short-term forecasting accuracy.

Table 6.5: Forecasting Performance of the Models

Model	RMSE	MAE
LASSO	0.0329	0.0253
Ridge	0.3679	0.3645
Elastic Net	0.0325	0.0252
Ordinary Least Squares (OLS)	3.6563	2.7602
LASSO + Macro	0.0331	0.0253
Ridge + Macro	0.3148	0.3115
Elastic Net + Macro	0.0337	0.0271
OLS + Macro	7.7809	5.5748
VAR (Macro + Spread)	0.1697	0.1470
ARIMA	0.0828	0.0668
GARCH	0.0719	0.0559

The comparison in Figure 6.3 shows that most models drift away from the actual T10Y2Y spread, which remains close to zero during the test period. Among all forecasts, the PCA-only models (LASSO, Elastic Net) stay closest to the real values and generally avoid the large upward trend seen in the other approaches, consistent with their lowest RMSE and MAE in Table 6.5. When macroeconomic variables are added, the forecasts become noisier and tend to rise more sharply, suggesting that the macro data may be adding extra instability rather than improving accuracy, as reflected in the slightly higher errors. The VAR model performs particularly poorly, producing an unrealistically strong upward path, while both ARIMA and GARCH generate smooth but biased forecasts that fail to capture the small, frequent movements in the actual spread.



Overall, although none of the methods predict the level of the spread perfectly, the PCA-only models remain the most reasonable and least biased among the alternatives.

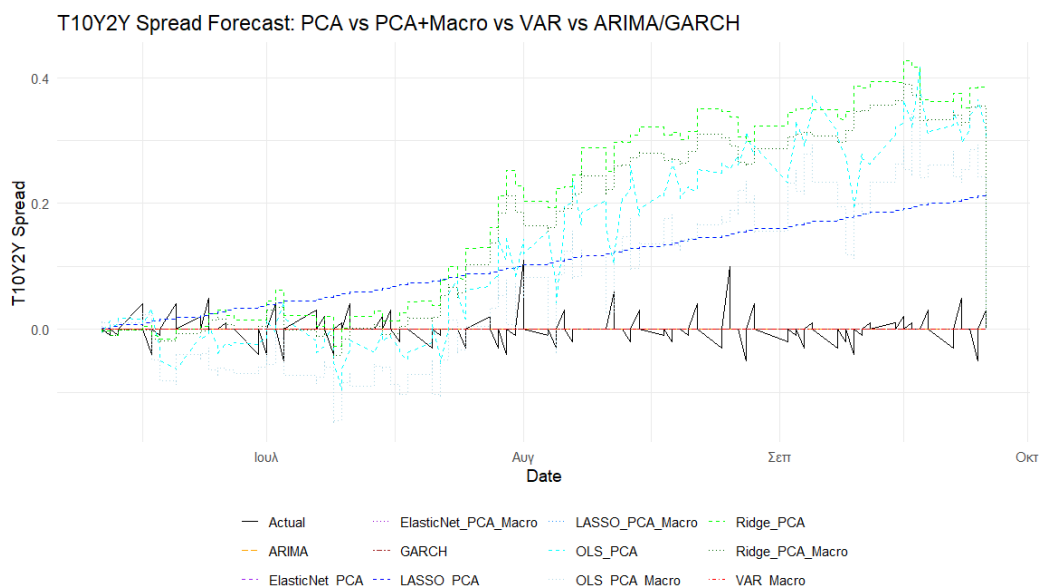


Figure 6.3: Comparison of forecasts from all models.

6.10 Feature Performance

LASSO's variable selection facilitates interpretability. By examining the coefficients retained in the final model, we identify which principal components are most relevant for forecasting. Interestingly, not all selected components correspond to those with the highest explained variance, indicating that sector-specific or less dominant factors contribute meaningfully to the prediction of yield curve dynamics. Positive coefficients imply that increases in the associated factor predict a widening of the spread, whereas negative coefficients indicate the opposite effect.



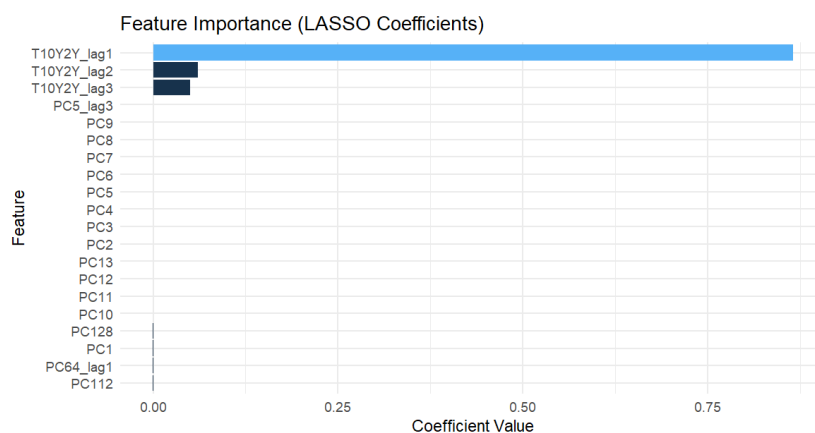


Figure 6.4: Feature Importance from LASSO regression.

This analysis provides insights not only into predictive performance but also into the underlying market mechanisms that influence the yield curve.

6.11 Next-Day Forecast

The best-performing LASSO model is applied to generate a next-day forecast for the Treasury spread.

Table 6.6: Next-Day Forecast of T10Y2Y Spread using LASSO

Model	Forecast Date	Predicted T10Y2Y Spread
LASSO (Top PCs + Lagged Spread)	2025-09-26	0.5289

This forecast demonstrates the model’s practical utility for producing actionable, forward-looking predictions. While point estimates are subject to uncertainty, the combination of equity market information and econometric modeling adds tangible value for real-time decision-making.



6.12 Conclusions

This analysis highlights the effectiveness of PCA as a dimensionality reduction tool and confirms the superiority of regularized regression models—particularly LASSO—in handling large sets of correlated predictors. Compared with classical time-series models such as ARIMA, GARCH, and VAR, PCA-based regressions yield lower forecast errors and better alignment with observed dynamics. Feature importance analysis reveals which components of stock market variation are most closely tied to yield curve movements, providing both predictive and interpretive insights.

While the inclusion of macroeconomic variables provides limited incremental benefit for daily forecasts, macro-based approaches remain valuable for understanding **medium-to long-term** trends and structural economic shifts. The results suggest that much of the information relevant to short-term yield curve dynamics is already embedded within equity market principal components, underscoring the interconnectedness of financial markets.

Moreover, the feature importance analysis offers meaningful insights into the specific market dimensions that influence yield curve movements, revealing the role of cross-asset linkages between equities and fixed income. Future research could extend this framework by incorporating additional macroeconomic and market indicators, implementing rolling-window forecasting to account for evolving market conditions, and exploring nonlinear or machine learning extensions to enhance predictive accuracy.

Overall, the proposed methodology provides a robust, interpretable and data-efficient foundation for both predictive modeling and analytical exploration of yield curve behavior in a high-dimensional financial environment.



Chapter 7

Conclusions and Strategic Guidance

This chapter presents a synthesis of the findings from this thesis on economic and financial forecasting, with particular attention to the role of Principal Component Analysis (PCA) in developing statistical arbitrage strategies. The previous chapters examined the evolution of forecasting methodologies, from classical econometric models to modern machine learning approaches, and explored how PCA can be leveraged to reduce dimensionality, extract latent factors, and improve the performance of trading strategies. By integrating insights from theoretical foundations, empirical analysis, and simulation results, this chapter aims to provide a coherent overview of the strengths and limitations of various forecasting models, and to propose strategic guidelines for both researchers and practitioners.

The objective of this chapter is twofold. First, it evaluates the comparative performance of different forecasting methods across key dimensions, including predictive accuracy, interpretability, data requirements, computational cost, and robustness to structural breaks. Second, it translates these insights into practical recommendations for implementing forecasting systems and PCA-based trading strategies, emphasizing adaptive and hybrid approaches that can respond to the dynamic nature of financial markets. In doing so, the chapter highlights the trade-offs inherent in model selection and the importance of combining complementary methodologies to achieve both predictive power and interpretability.



7.1 Comparative Overview of Forecasting Models

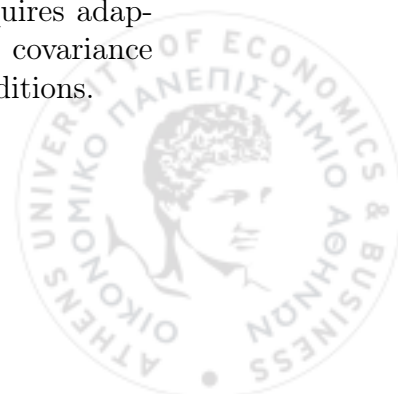
Forecasting models differ across several key dimensions, reflecting the trade-offs between classical econometric approaches and modern machine learning techniques. No single model is universally superior. Deep learning models, such as LSTM's and Transformers, are particularly effective at capturing complex short-term patterns, whereas simpler models like ARIMA or linear regressions can perform better over longer horizons. Gradient Boosting models, including XGBoost and LightGBM, often provide a balanced compromise, delivering strong predictive accuracy while remaining relatively interpretable and computationally efficient.

Interpretability remains a central differentiator among models. Econometric approaches such as VAR offer transparent, theory-driven insights suitable for causal analysis, whereas deep learning models tend to function as “black boxes“. PCA occupies an intermediate position: its components are mathematically well-defined but may be difficult to relate directly to economic factors. In terms of data and computational requirements, classical models are parsimonious and computationally light, whereas deep learning approaches require large datasets, extensive training, and specialized hardware. All models are challenged by non-stationarity and sudden shifts in markets regimes, but adaptive strategies such as rolling-window estimation can improve robustness to structural breaks.

7.2 Strategic Recommendations

For general economic and financial forecasting, a hierarchical approach is recommended. Research should begin with simple, robust models, such as ARIMA or VAR, to establish a baseline. Gradient Boosting models are particularly effective for structured, tabular datasets, offering high predictive performance with manageable complexity. Deep learning models should be used selectively, reserved for situations where long-range sequential dependencies are critical and sufficient high-quality data is available.

In the context of PCA-based statistical arbitrage, researchers should first implement the classical PCA framework and perform rigorous backtesting to establish a benchmark. More advanced PCA variants, such as scaled or supervised PCA, can enhance residual construction and improve the quality of trading signals. The dynamic nature of financial markets requires adaptive approaches: rolling-window PCA or exponentially weighted covariance matrices allow factor models to respond to changing market conditions.



Residual modeling can be further enhanced by combining the traditional Ornstein-Uhlenbeck process with GARCH or LSTM methods to capture time-varying volatility and non-linear mean-reversion patterns. All strategies should be rigorously evaluated using walk-forward, out-of-sample backtesting while accounting for transaction costs and market impact to ensure realistic performance assessment.

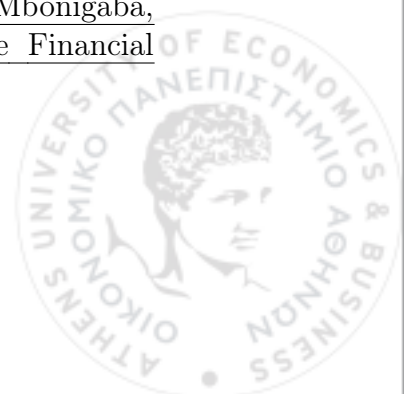
7.3 Concluding Remarks

The evolution of financial forecasting illustrates a clear trajectory from simple, theory-driven models to sophisticated hybrid approaches that integrate statistical, machine learning, and economic insights. The future does not lie in a single “super-model“ but in intelligent hybridization: combining PCA for dimensionality reduction, machine learning for non-linear pattern recognition, and econometric theory for interpretability. Success in forecasting depends on the thoughtful integration of these tools, adaptive methodologies, and rigorous evaluation to extract reliable signals from complex ever-changing markets.

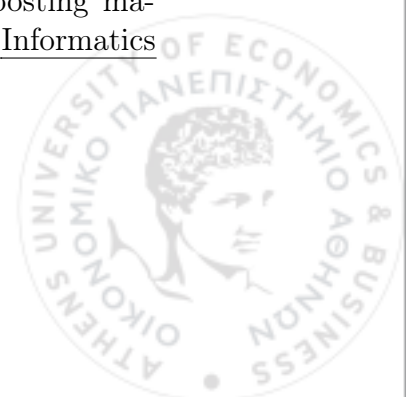


Bibliography

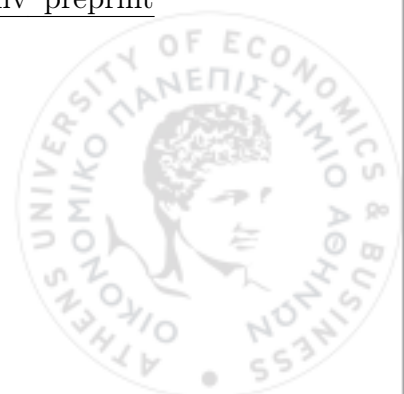
- [1] Abdi, H. and Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4):433–459.
- [2] Bao, W., Yue, J., and Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PloS one, 12(7):e0180944.
- [3] Basilevsky, A. T. (2009). Statistical factor analysis and related methods: theory and applications. John Wiley & Sons.
- [4] Biva, A. T. (2024). A Comprehensive Study of Machine Learning Approaches for Financial Time Series Forecasting. PhD thesis, University of Dhaka.
- [5] Bollerslev, T. (2008). Glossary to arch (garch). CREATES Research paper, 49.
- [6] Brundavani, P. (2024). A principal component analysis algorithm for seed enterprise financial performance and scientific and technological innovation. Journal of Computer Allied Intelligence (JCAI, ISSN: 2584-2676), 2(2):49–62.
- [7] Cao, D., Tian, Y., and Bai, D. (2015). Time series clustering method based on principal component analysis. In 5th International conference on information engineering for mechanics and materials, pages 888–895. Atlantis Press.
- [8] Castro-Iragorri, J. R. and Ramírez, J. (2021). Forecasting dynamic term structure models with autoencoders. Documentos De Trabajo, 19431.
- [9] Celestin, P. et al. (2025). Principal component analysis for simplifying multivariate financial data in portfolio risk analysis. Mbonigaba, Principal Component Analysis For Simplifying Multivariate Financial Data In Portfolio Risk Analysis (March 01, 2025).



- [10] Chakrabarty, B., Comerton-Forde, C., and Pascual, R. (2023). Identifying high frequency trading activity without proprietary data. Technical report, Working Paper.
- [11] Daly, K. (2008). Financial volatility: Issues and measuring techniques. Physica A: statistical mechanics and its applications, 387(11):2377–2393.
- [12] Dellaportas, P. and Pourahmadi, M. (2012). Cholesky-garch models with applications to finance. Statistics and Computing, 22(4):849–855.
- [13] Derbentsev, V., Matviychuk, A., Datsenko, N., Bezkorovainyi, V., and Azaryan, A. (2020). Machine learning approaches for financial time series forecasting. In Proceedings of the Selected Papers of the Special Edition of International Conference on Monitoring, Modeling & Management of Emergent Economy (M3E2-MLPEED 2020) Odessa, Ukraine, July 13-18, 2020, pages 434–450. CEUR Workshop Proceedings.
- [14] Diamantaras, K. (1996). Principal component neural networks theory and applications.
- [15] Dybvig, P. H. and Ross, S. A. (2003). Arbitrage, state prices and portfolio theory. Handbook of the Economics of Finance, 1:605–637.
- [16] Emami Gohari, H., Dang, X.-H., Shah, S. Y., and Zerfos, P. (2024). Modality-aware transformer for financial time series forecasting. In Proceedings of the 5th ACM International Conference on AI in Finance, pages 677–685.
- [17] Francq, C. and Zakoian, J.-M. (2019). GARCH models: structure, statistical inference and financial applications. John Wiley & Sons.
- [Frost] Frost, J. Principal component analysis guide & example. Statistics By Jim.
- [19] Guijarro-Ordóñez, J., Pelger, M., and Zanotti, G. (2021). Deep learning statistical arbitrage. arXiv preprint arXiv:2106.04028.
- [20] Hagerud, G. E. (1997). A new non-linear GARCH model. Stockholm School of Economics.
- [21] Hartanto, A. D., Kholik, Y. N., and Pristyanto, Y. (2023). Stock price time series data forecasting using the light gradient boosting machine (lightgbm) model. JOIV: International Journal on Informatics Visualization, 7(4):2270–2279.



- [22] Hawkins, J. (2005). Economic forecasting: history and procedures. Economic Round-up, (Autumn 2005):1–10.
- [23] Hendry, D. F. (2020). A short history of macro-econometric modelling. Journal of Banking, Finance and Sustainable Development, 1(1).
- [24] Ho, C.-T. B. and Wu, D. D. (2009). Online banking performance evaluation using data envelopment analysis and principal component analysis. Computers & Operations Research, 36(6):1835–1842.
- [25] Jaadi, Z. (2024). Principal component analysis (pca): A step-by-step explanation. Ανάκτηση, 9(06):2024.
- [26] Janićijević, S., Mizdraković, V., and Kljajić, M. (2022). Principal component analysis in financial data science. In Advances in principal component analysis. IntechOpen.
- [27] Jatana, V. (2024). Mastering time series forecasting (a concise guide to model selection and analysis). Available at SSRN 4730679.
- [28] Jeanne, M. and AKUMUNTU, J. (2025). Brainae journal of business, sciences and technology issn"2789-374x (print)" " 2789-3758 (online) volume 9, issue 02, february 2025.
- [29] Jindal, G. (2024). The role of finance tech in revolutionizing traditional banking systems through data science and ai. Journal Of Applied Sciences, 4(11):10–21.
- [30] Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065):20150202.
- [31] Joshi, S. (2025). Gradient boosting and explainable ai for financial risk management: A comprehensive review.
- [32] Kabir, M. R., Bhadra, D., Ridoy, M., and Milanova, M. (2025). Lstm-transformer-based robust hybrid deep learning model for financial time series forecasting. Sci, 7(1):7.
- [33] Krause, F. and Calliess, J.-P. (2024). End-to-end policy learning of a statistical arbitrage autoencoder architecture. arXiv preprint arXiv:2402.08233.



- [34] Lettau, M. (2024). 3d-pca: Factor models with restrictions. Technical report, National Bureau of Economic Research.
- [35] Li, H. (2019). Multivariate time series clustering based on common principal component analysis. Neurocomputing, 349:239–247.
- [36] Li, H., Huang, X., Luo, F., Zhou, D., Cao, A., and Guo, L. (2025). Revolutionizing agricultural stock volatility forecasting: a comparative study of machine learning and har-rv models. Journal of Applied Economics, 28(1):2454081.
- [37] Li, W. (2018). High frequency trading with speed hierarchies. Available at SSRN 2365121.
- [38] Liu, J. (2024). Navigating the financial landscape: the power and limitations of the arima model. Highlights in Science, Engineering and Technology, 88:747–752.
- [Liu et al.] Liu, N., Goh, Y. M., Du, S., and Chua, D. K. Data-driven construction project risk causal network: Integration of ensemble causal discovery and pls-sem validation. Available at SSRN 5295237.
- [40] Milne, R. and Bull, R. (2002). Back to basics: A componential analysis of the original cognitive interview mnemonics with three age groups. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 16(7):743–753.
- [41] Mishra, P., Al Khatib, A. M. G., Yadav, S., Ray, S., Lama, A., Kumari, B., Sharma, D., and Yadav, R. (2024). Modeling and forecasting rainfall patterns in india: a time series analysis with xgboost algorithm. Environmental Earth Sciences, 83(6):163.
- [42] Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2015). Introduction to time series analysis and forecasting. John Wiley & Sons.
- [43] Morton, C. (2025). Braving the elements: A time series analysis of e-scooter ridership assessing the impact of weather and seasonality across different climate regions. Case Studies on Transport Policy, 20:101431.
- [44] Mungai, F. (2016). Modeling and forecasting kenyan gdp using autoregressive integrated moving average (arima) models. Science Journal of Applied Mathematics and Statistics.



- [45] Nabi, R. M., Soran Ab M, S., and Harron, H. (2020). A novel approach for stock price prediction using gradient boosting machine with feature engineering (gbm-wfe). Kurdistan Journal of Applied Research, 5(1):28–48.
- [46] Naghshpour, S. and Iii, H. L. D. (2018). The impact of commercial banking development on economic growth: a principal component analysis of association between banking industry and economic growth in eastern europe. International Journal of Monetary Economics and Finance, 11(6):525–542.
- [47] Nelson, B. K. (1998). Time series analysis using autoregressive integrated moving average (arima) models. Academic emergency medicine, 5(7):739–744.
- [48] Oladapo, I. A. and Akinwale, Y. O. (2024). Islamic financial depth, inflation, interest rates, and economic growth in saudi arabia: An application of vector autoregression model. Banks and Bank Systems, 19(4):34.
- [49] Oprea, A. (2022). The use of principal component analysis (pca) in building yield curve scenarios and identifying relative-value trading opportunities on the romanian government bond market. Journal of Risk and Financial Management, 15(6):247.
- [50] Ozupek, O., Yilmaz, R., Ghasemkhani, B., Birant, D., and Kut, R. A. (2024). A novel hybrid model (emd-ti-lstm) for enhanced financial forecasting with machine learning. Mathematics, 12(17):2794.
- [51] Palmatier, R. W. and Sridhar, S. (2020). Marketing strategy: Based on first principles and data analytics. Bloomsbury Publishing.
- [52] Paul, J. (2024). Financial time series analysis with transformer models.
- [53] Pijarski, P., Kacejko, P., and Miller, P. (2023). Advanced optimisation and forecasting methods in power engineering—introduction to the special issue. Energies, 16(6):2804.
- [54] Polo, J., Martín-Chivelet, N., Alonso-Abella, M., Sanz-Saiz, C., Cuenca, J., and de la Cruz, M. (2023). Exploring the pv power forecasting at building façades using gradient boosting methods. Energies, 16(3):1495.
- [55] Roh, T. H. (2007). Forecasting the volatility of stock price index. Expert systems with applications, 33(4):916–922.



- [56] Saad, H., Samy, D., and Khalil, D. (2025). Integrating ai tools and var models for actuarial applications: an analysis of egp/usd exchange rate dynamics. Journal of Humanities and Applied Social Sciences.
- [57] Sakurada, M. and Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. In Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis, pages 4–11.
- [58] Satchell, S. and Knight, J. (2011). Forecasting volatility in the financial markets. Elsevier.
- [59] Selvin, S., Vinayakumar, R., Gopalakrishnan, E., Menon, V. K., and Soman, K. (2017). Stock price prediction using lstm, rnn and cnn-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci), pages 1643–1647. IEEE.
- [60] Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.
- [61] Smith, L. I. (2002). A tutorial on principal components analysis.
- [62] Srijiranon, K., Lertratanakham, Y., and Tanantong, T. (2022). A hybrid framework using pca, emd and lstm methods for stock market price prediction with sentiment analysis. Applied Sciences, 12(21):10823.
- [63] Tsay, R. S. (2005). Analysis of financial time series. John wiley & sons.
- [64] Uckan, T. (2024). Integrating pca with deep learning models for stock market forecasting: An analysis of turkish stocks markets. Journal of King Saud University-Computer and Information Sciences, 36(8):102162.
- [65] Wedel, M. and Kannan, P. (2016). Marketing analytics for data-rich environments. Journal of marketing, 80(6):97–121.
- [66] WU, L. (2025). The association between different succession methods and innovation investments in family businesses in china.
- [67] Xie, C., Zhang, Y., Wang, M., and Liu, Z. (2023). Quantamental trading: Fundamental and quantitative analysis with multi-factor regression model strategy. In International Conference on Business and Policy Studies, pages 1455–1470. Springer.



- [68] Xu, J., He, J., Gu, J., Wu, H., Wang, L., Zhu, Y., Wang, T., He, X., and Zhou, Z. (2022). Financial time series prediction based on xgboost and generative adversarial networks. International Journal of Circuits, Systems and Signal Processing, 16:637–645.
- [69] Yang, Y., Wu, Y., Wang, P., and Jiali, X. (2021). Stock price prediction based on xgboost and lightgbm. In E3s web of conferences, volume 275, page 01040. EDP Sciences.
- [70] Yap, B. C.-F., Mohamad, Z., and Chong, K.-R. (2013). The application of principal component analysis in the selection of industry specific financial ratios. British Journal of Economics, Management & Trade, 3(3):242–252.
- [71] Yongchareon, S. (2025). Ai-driven intelligent financial forecasting: A comparative study of advanced deep learning models for long-term stock market prediction. Machine Learning and Knowledge Extraction, 7(3):61.
- [72] Zhang, Z. et al. (2022). Research on stock price prediction based on pca-lstm model. Academic Journal of Business & Management, 4(3):42–47.
- [73] Zivot, E. and Wang, J. (2006). Vector autoregressive models for multivariate time series. In Modeling financial time series with S-PLUS®, pages 369–413. Springer.

