Mathematics in Stock Portfolio Management

Zhe Wang

July 2025

1 Introduction

This report presents a survey of mathematical methods applied throughout the design of stock trading strategies, following the standard quantitative investment process outlined in Grinold and Kahn [1]. Which starts from data preprocessing, then to signal generation, portfolio construction, and performance evaluation. In each of these stages, mathematics is used.

We focus on a structured methodology that incorporates stock selection, data collection, feature engineering, scoring and ranking, portfolio optimization, and backtesting. Classical methods such as mean-variance optimization [2], conditional value at risk (CVaR) [4], and the Black-Litterman model [3] are discussed alongside modern advancements like shrinkage estimators [5], entropy pooling [6], and Bayesian updating [33]. At each stage, we highlight the mathematical foundations, explain their mechanism, and discuss how they facilitate strategy design.

2 Stock Selection and Data Acquisition

The first step in designing a quantitative trading strategy is to define the investable stocks and acquire relevant data. This universe is typically filtered based on criteria such as market capitalization, liquidity, or sector classification. For the first two attributes, data acquired often includes large amount of noise, and statistical outlier detection using Mahalanobis distance [7] measures the multivariate distance of each data point from the distribution center, would remove the extreme observations that could distort parameter estimates. For classification of stocks, clustering algorithms (e.g., K-means) [8] partition assets into groups by minimizing within-cluster variance, thus creating homogeneous clusters that simplify subsequent analysis. Together, these steps ensure a valid, rigorous process of determine the stocks to trade.

3 Feature Engineering and Signal Construction

With stocks choosen, the next phase involves engineering predictive features. Momentum indicators—such as trailing twelve-month returns—capture trend persistence by measuring percentage change over a fixed lookback [9]. Volatility models like GARCH [10] dynamically estimate time-varying variance based on past shocks and residuals, providing risk forecasts. Valuation ratios (e.g., price-to-earnings) quantify relative cheapness or expensiveness versus fundamentals. Natural Language Processing (NLP) models such as FinBERT [11] apply transformer architectures to earnings-call transcripts to produce sentiment scores, extracting market sentiment signals. Time-series tools like ARIMA [12] fit autoregressive and moving-average terms to forecast returns from their own history. Factor analysis [13] and Independent Component Analysis (ICA) [14] decompose observed returns into latent factors or statistically independent sources, revealing underlying drivers. Finally, regularized regressions—Lasso (L1 penalty) [15] for factor selection and Ridge (L2 penalty) [16] for coefficient shrinkage—help build parsimonious, robust predictive models. By standardizing and linearly combining these features, one obtains a composite score aligned with documented factor-premium relationships of the choosen stocks.

4 Ranking and Selection

Once asset scores are computed, ranking transforms continuous signals into stable decisions. Z-scores convert raw feature outputs into standard deviations from the mean, enabling cross-factor comparability, while quantile ranks assign each asset a percentile position, reducing sensitivity to outliers. Bayesian ranking models [17] treat scores as random variables with priors that are updated via observed performance, thus accounting for estimation uncertainty. Robust statistics—such as trimmed means or M-estimators [18]—downweight extreme observations that could distort rankings. These techniques ensure that only the most promising assets (e.g., top-N) are selected for portfolio inclusion, reducing noise-driven turnover.

5 Portfolio Optimization

Selected assets are then allocated weights via portfolio optimization. Mean-Variance Optimization (MVO) [2] solves a quadratic program to minimize portfolio variance for a target expected return, formalizing the risk-return trade-off under a normal-returns assumption. Risk Parity [19] solves for weights that equalize each asset's contribution to total portfolio volatility ($w_i(\Sigma w)_i = \text{constant}$), thereby diversifying risk rather than capital. The Black-Litterman model [3] produces posterior expected returns by Bayesian blending of equilibrium market returns with investor views, leading to more stable weight estimates. Covariance matrix shrinkage (e.g., Ledoit-Wolf) [20] reduces estimation error in high dimensions by combining sample covariances with a structured target. Entropy pooling [6] integrates multiple return scenarios by maximizing entropy under moment constraints, yielding a broad distribution of plausible outcomes. Finally, Conditional Value-at-Risk (CVaR) optimization [4] solves a linear program to minimize expected losses in the tail beyond a specified confidence level. Collectively, these approaches generate weight vectors that balance return prospects against various risks and estimation uncertainties.

6 Backtesting and Evaluation

After weights are determined, backtesting simulates historical performance under realistic conditions. Time-series cross-validation (rolling windows) [21] retrains and tests models sequentially to mimic live updates. Bootstrapping [22] resamples historical returns to compute confidence intervals for performance metrics (e.g., return, volatility). Statistical hypothesis tests (e.g., t-tests on excess returns) [23] assess the significance of observed alpha. Drawdown analysis [24] measures peak-to-trough declines to characterize downside risk, and stochastic dominance tests [25] compare entire return distributions across strategies. These methods together provide a rigorous evaluation of strategy robustness and potential live-trading behavior.

7 Risk Management and Deployment

Robust deployment requires formal risk controls. Value-at-Risk (VaR) [26] calculates the loss at a specified confidence quantile, while CVaR [4] computes the expected loss conditional on exceeding that quantile, capturing tail risk. Scenario analysis [27] subjects portfolios to historical or hypothetical crises to evaluate resilience. Bayesian updating [33] continuously revises model parameters and return forecasts by combining prior beliefs with incoming market data, enabling adaptive responses. Markov regime-switching models [28] estimate latent market states and transition probabilities, triggering strategy adjustments when structural shifts are detected. These risk-management frameworks ensure strategies remain resilient under evolving market conditions.

8 Data Requirements

The strategy relies on multiple layers of data. Market data includes OHLCV (open, high, low, close, volume) series adjusted for corporate actions. From these, log or arithmetic returns are computed and used to derive the rolling covariance matrix for risk estimates. Fundamental data includes valuation metrics such as P/E ratios, earnings growth, and book value, while alternative data sources provide sentiment scores via NLP or macroeconomic indicators like interest rates and inflation. These data can be obtained from providers such as Yahoo Finance [29], Quandl [30], Alpha Vantage [31], or Bloomberg [32].

9 Conclusion

Mathematics provides the foundation for systematic stock trading strategies, from feature generation to portfolio construction and risk evaluation. By integrating tools such as mean-variance optimization [2], risk parity [19], shrinkage covariance estimators [5], and Bayesian modeling [3] in a unified framework, traders can build data-driven strategies that adapt to changing market conditions. Each stage in the pipeline is enhanced by modern mathematical methods, ensuring that the strategy is not only statistically robust but also practically viable. The inclusion of intuitive reasoning, empirical evidence, and rigorous validation creates a powerful, repeatable design process that serves as the cornerstone of contemporary quantitative investing.

References

[1] R. C. Grinold and R. N. Kahn, *Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk*, McGraw-Hill, 2nd edition, 1999.

[2] H. Markowitz, *Portfolio Selection*, The Journal of Finance, Vol. 7, No. 1, 1952.

[3] F. Black and R. Litterman, *Global Portfolio Optimization*, Financial Analysts Journal, 1992.

[4] R. T. Rockafellar and S. Uryasev, *Optimization of Conditional Value-at-Risk*, Journal of Risk, 2000.

[5] J. Fan, Y. Liao, and M. Mincheva, *Large Covariance Estimation by Thresholding Principal Orthogonal Complements*, Journal of the Royal Statistical Society, 2013.

[6] A. Meucci, *Beyond Black–Litterman: Views and Priors in Portfolio Allocation*, Social Science Research Network (SSRN), 2005.

[7] P. C. Mahalanobis, *On the Generalized Distance in Statistics*, Proceedings of the National Institute of Sciences of India, 1936.

[8] S. Lloyd, *Least Squares Quantization in PCM*, IEEE Transactions on Information Theory, 1982.

[9] N. Jegadeesh and S. Titman, *Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency*, The Journal of Finance, 1993.

[10] R. F. Engle, *Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation*, Econometrica, 1982.

[11] D. Araci, *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*, arXiv:1908.10063, 2019.

[12] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, 1976.

[13] C. Spearman, *General Intelligence Objectively Determined and Measured*, American Journal of Psychology, 1904.

[14] A. Hyvärinen and E. Oja, *Independent Component Analysis: Algorithms and Applications*, Neural Networks, 2000.

[15] R. Tibshirani, *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society, Series B, 1996.

[16] A. E. Hoerl and R. W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, 1970.

[17] "Bayesian Ranking," Wikipedia, `https://en.wikipedia.org/wiki/Bayesian_rating_system`, accessed July 2025.

[18] P. J. Huber, *Robust Statistics*, Wiley, 1981.

[19] E. Qian, *Risk Parity Portfolios: Efficient Portfolios through True Diversification*, Panagora Asset Management White Paper, 2005.

[20] L. Ledoit and M. Wolf, *A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices*, Journal of Multivariate Analysis, 2004.

[21] C. Bergmeir and J. M. Benítez, *On the Use of Cross-Validation for Time Series Predictor Evaluation*, Information Sciences, 2012.

[22] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, Annals of Statistics, 1979.

[23] W. S. Gosset ("Student"), *The Probable Error of a Mean*, Biometrika, 1908.

[24] P. Magdon-Ismail, A. Atiya, and I. Pratap, *Maximum Drawdown*, Journal of Risk, 2004.

[25] H. Levy and G. Hanoch, *Stochastic Dominance and Means of Order* $\leq k$, SIAM Journal on Applied Mathematics, 1969.

[26] J.P. Morgan/Reuters, *RiskMetrics—Technical Document*, 4th edition, 1996.

[27] Basel Committee on Banking Supervision, *International Convergence of Capital Measurement and Capital Standards*, 2006.

[28] J. D. Hamilton, *A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle*, Econometrica, 1989.

[29] Yahoo Finance, `https://finance.yahoo.com`, accessed July 2025.

[30] Quandl, `https://www.quandl.com`, accessed July 2025.

[31] Alpha Vantage, `https://www.alphavantage.co`, accessed July 2025.

[32] Bloomberg, `https://www.bloomberg.com`, accessed July 2025.

[33] "Bayesian Inference," Wikipedia, `https://en.wikipedia.org/wiki/Bayesian_inference`, accessed July 2025.