Predictive Modeling of Correlation and Volatility in Multivariate Financial Time Series

Filip Aleksić, Mate Papak, and Stjepan Begušić

University of Zagreb Faculty of Electrical Engineering and Computing
Laboratory for Financial and Risk Analytics
Zagreb, Croatia
stjepan.begusic@fer.unizg.hr

***Abstract*—This paper focuses on risk prediction in multivariate financial return time series by decomposing the asset covariance into its correlation and volatility components. Using a machine learning approach, we develop predictive models which ensure positive definiteness of the resulting matrices. The proposed approach allows us to separately study the contributions of correlation and volatility prediction to overall risk forecasting performance. To evaluate the proposed approach, we compare multiple performance metrics and assess the economic significance of our findings through a portfolio optimization application. The results provide insights into the relative predictability of correlation and volatility and their roles in portfolio optimization and risk management.**

***Index Terms*—Correlation, covariance, volatility, factor model, financial risk modelling**

I. Introduction

Risk management in finance relies primarily on estimating the covariance matrix of asset returns, which serves as a key input for portfolio risk estimation, portfolio optimization, and asset pricing [1]. Traditional approaches to risk modeling mostly focus on reducing estimation noise in a given look-back window of returns, while assuming that the estimated covariance will hold in the future holding period [2].

The sample estimators (sample variance, corelation and covariance), despite their simplicity, are prone to instability and noise, especially when the number of assets approaches or exceeds the number of observations, which is often the case in financial applications [3]. Shrinkage estimators, such as the class of Ledoit-Wolf methods, blend sample estimates with structured targets, but may impose overly simplistic assumptions that hinder predictive accuracy [4]. Similarly, unsupervised learning methods, such as clustering and factor models, provide useful insights into market structure but are not inherently designed for predictive performance [5], [6]

A significant drawback of these conventional approaches is their reliance on static, backward-looking information [7], [8]. By treating risk estimation primarily as a statistical inference problem, they fail to leverage any predictive potential of modern machine learning techniques. This paper proposes a shift in perspective: instead of focusing solely on estimation, we frame the risk estimation problem as a supervised learning task. To do so, we utilitze a decomposition of the return covariance matrix into its correlation and volatility components, and treat the predictive modeling of each component as a separate task. By maintaining positivity of the predicted volatilities and positive definiteness of the predicted correlation structures, we ensure the positive definiteness of the covariance predictions [9]. Positive definiteness in the correlation predictions is addressed using a factor model, which provides a robust and interpretable structure [10]. By explicitly incorporating these constraints, our framework generates more stable and realistic risk predictions. Moreover, this proposed approach allows us to study the separate contributions of the volatility and correlation predictions to the overall predictive performance of the model.

To evaluate the effectiveness of the proposed approach, we conduct a comprehensive empirical study using multiple performance metrics. Additionally, we assess the economic significance of our results through a portfolio optimization application. Our findings offer insights into the relative predictability of correlation and volatility, as well as their distinct contributions to overall risk management. Through this predictive lens, we demonstrate the advantages of a supervised learning approach in enhancing the robustness and accuracy of financial risk models.

II. Model

A. Decoupling correlation and volatility

Let Y denote the p-dimensional random vector of daily asset returns[1] for p assets. The asset return covariance is most commonly estimated from the data sample $\mathbf{Y} \in \mathbb{R}^{p \times T}$ (where T is the sample size), using the sample estimator on a historical window of length T and believed to hold in the future (e.g. during the holding period of a portfolio). However, in this paper we aim to formulate a predictive approach for the covariance matrix, with the goal of studying the predictive potential and impact separately for correlations and variances. To do so, we consider the decomposition of the covariance between any two return time series:

$$\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j, \tag{1}$$

where σ_i and σ_j are the standard deviations (volatilities) of the individual returns and ρ_{ij} is the correlation.

[1]In this paper we use arithmetic returns $Y_t = S_t/S_{t-1} - 1$, where S_t is the price of the instrument at time step (day) t.

To assure that the predicted correlations form a positive-definite correlation matrix for a large number of assets (even when $p >> T$), we use a factor-based approach to model the return correlations. Even though the academic literature has recently focused on expanding the selection of factors which explain the dynamics of asset returns, often resulting in redundant factors [11], it is well known that a single dominant factor describing the market return accounts for the vast majority of the market dynamics and the asset return cross-correlations, even for extreme market events [12]. To focus on the core aspects of the approach proposed in this paper, we adopt a single factor model, in line with a wide body of recent research focusing on the improvement of estimates of the single factor model coefficients, known as betas [13], [14]. However, in this paper we apply the model to standardized returns, so that the covariance structure implied by the model corresponds to the asset return correlations. Let $Z_{it} = (Y_{it} - \mu_i)/\sigma_i$ denote the standardized returns of stock i, and let X_t denote the standardized returns of a broad market index (the market factor), both with mean 0 and variance 1. The model for each stock is given by:

$$Z_{it} = \alpha_i + \beta_i X_t + \epsilon_{it}, \tag{2}$$

where ϵ_{it} is the so-called idiosyncratic component. More specificaly, in this case it is a random variable with mean 0 and variance $\psi_i = 1 - \beta_i^2$ (which ensures that the variance of Z_{it} is 1). Also note that for this particular setting with standardized variables, the coefficient β_i (often called beta) is in fact limited in value to $\beta_i \in (-1, 1)$. A one-factor linear model explains the correlations between stocks throught their similarity to the overall market. Under some simple assumptions (notable the orthogonality of the idiosyncratic components for all stocks and the market factor), the correlation $\mathrm{Corr}(Z_i, Z_j) = R_{ij}$ between any pair of stocks i and j is calculated simply as:

$$R_{ij} = \beta_i \beta_j. \tag{3}$$

Finally, the entire correlation matrix for all asset pairs is constructed as:

$$\mathbf{R} = \begin{bmatrix} 1 & \beta_2\beta_1 & \cdots & \beta_n\beta_1 \\ \beta_1\beta_2 & 1 & \cdots & \beta_n\beta_2 \\ \vdots & \vdots & \ddots & \vdots \\ \beta_1\beta_n & \beta_2\beta_n & \cdots & 1 \end{bmatrix}, \tag{4}$$

which is positive definite as long as $\psi_i > 0 \ \forall i = 1, \ldots, n$. This procedure allows us to predict positive definite correlation matrices of any dimension relying only on a p-dimensional vector of betas: $[\beta_1, \ldots, \beta_p]^\top$. Therefore the correlation prediction problem boils down to the problem of formulating a predictive model for betas.

On the other hand, the problem of predicting volatilities themselves is straightforward. Once the volatility predictions $\hat{\sigma}_i$ are obtained, we construct the diagonal matrix D:

$$\hat{\mathbf{D}} = \begin{bmatrix} \hat{\sigma}_1 & 0 & \cdots & 0 \\ 0 & \hat{\sigma}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\sigma}_n \end{bmatrix}. \tag{5}$$

The beta predictions $\hat{\beta}_i$ are used to construct the predicted correlation matrix $\hat{\mathbf{R}}$, and the final covariance matrix prediction is then obtained as:

$$\hat{\Sigma} = \hat{\mathbf{D}}\hat{\mathbf{R}}\hat{\mathbf{D}}. \tag{6}$$

The above described approach ensures positive definiteness of the predictions even in very high dimensions. This also allows us to formulate supervised machine learning models for the prediction of volatilities and correlations. Finally this approach also allows us to separately study the contribution of correlations and volatility in assessing the predictive potential of the considered models.

B. Machine learning approach

The factor model betas and the individual asset volatilities are generally estimated using historical data and sample estimators. The sample variance is:

$$\hat{\sigma}_i^2 = \frac{1}{T-1}(Y_i - \overline{Y}_i)(Y_i - \overline{Y}_i)^\top, \tag{7}$$

where Y_i denotes the vector of returns for asset i and $\overline{Y_i}$ denotes its sample mean. The sample betas are estimated using the OLS approach:

$$\hat{\beta}_i = \frac{\sum_{t=1}^{T}(Z_{it} - \bar{Z}_i)(X_t - \bar{X})}{\sum_{t=1}^{T}(X_t - \bar{X})^2}. \tag{8}$$

These sample estimators will lead to positive definite estimates of the asset return covariance which will generally be somewhat reliable. However, in this paper our goal is to formulate predictive models for σ_i and β_i. While doing so, we need to ensure that the volatility predictions are positive, which can be easily imposed using either log volatilities (the log volatility predictions may be $\in \mathbb{R}$) or an activation function at the output of our model. A challenge in volatility prediction is the presence of extreme values, which may lead to overfitting. Applying a logarithmic transformation to volatility reduces the relative impact of extreme values, leading to a more balanced optimization process and improved predictive performance. Regarding betas, ensuring that $\beta_i \in (-1, 1)$ will guarantee positive definiteness of the correlation matrix estimates. This we do using a $\tanh()$ activation function at the output layer of our model.

The forecasting targets in the machine learning approach are therefore logarithmic volatilities $\ln(\sigma_i)$ and beta coefficients. The goal is to minimize mean squared error (MSE) toward targets $\ln(\sigma_i')$ and β_i' estimated via sample-based estimation on the look-ahead window from t to $t + T'$, where $T' = 21$ days. Instead of training separate models for each stock, a uniform approach is used, allowing patterns across different stocks to be captured effectively. Therefore, we formulate a

number of input-output pairs in the training set equal to the number of historical time steps (windows) times the number of assets. To evaluate the best approach, four machine learning models are trained for volatility prediction and another four with the same hyperparameters for beta prediction.

1) Linear Regression: In this context, the linear regression model is used to predict logarithmic volatilities and beta coefficients from features calculated on past data. The model assumes a linear relationship between the input features and the target variable, fitted using the OLS approach, and allowing for an intercept term.

2) Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees during training and merges their outputs to improve predictive accuracy and control overfitting. Each tree in the forest is trained on a random subset of data and features, providing diversity among the trees. To assess the performance for different hyperparameters, key values were varied within reasonable bounds: 200–400 estimators, maximum tree depth 25–35, the minimum samples required to split a node were varied between 2 and 3, and the minimum number of samples per leaf between 1 and 2. Evaluation on a random subset of 100 stocks showed consistent performance across configurations, suggesting robustness and likely good generalization to the full dataset. The final model is configured with 300 estimators, a maximum tree depth of 30, and selects features using the square root heuristic. Additionally, the minimum samples required to split a node is set to 2, and the minimum number of samples per leaf is 1, ensuring balanced complexity and generalization.

3) XGBoost: XGBoost, or extreme gradient boosting, is a highly efficient and scalable implementation of the gradient boosting framework. It sequentially builds models to correct errors made by previous models. This iterative approach allows XGBoost to effectively capture complex relationships in data, making it suitable for predicting both volatility and beta coefficients. The model is tuned with a learning rate of 0.1, a maximum depth of 2, and 200 estimators. Regularization techniques include subsampling at 80% and column sampling per tree at 70%, helping to prevent overfitting while ensuring robust predictions. Regarding hyperparameters sensitivity, consistent performance was observed across the following ranges: learning rate 0.05-0.15, maximum depth 1-3, 150-250 estimators, subsampling 70-90%, and column sampling 60-80%.

4) Feedforward Neural Network: The feedforward neural network model consists of an input layer, hidden layers, and an output layer, where each layer is fully connected to the next. The network is designed with 16 input neurons and two hidden layers, each containing 16 neurons with ReLU activation to introduce non-linearity. For beta coefficient estimation the output layer applies the $\tanh()$ activation to ensure the values remain between -1 and 1. To enhance generalization, dropout regularization is set at 0.5, and early stopping is employed to mitigate overfitting. The model is trained using the Adam optimizer with a mean squared error loss function over 50 epochs, using a batch size of 32 to balance computational efficiency and convergence stability. Neural network performance remained stable when varying neurons per layer in the range 8-32, dropout 0.4-0.6, epochs 40-60, and batch size 8-32.

III. Data and features

The dataset consists of historical prices and returns of 848 constituent stocks which were continuously present in the U.S. Russell 3000 index, spanning the period from 2000 to 2020. Historical values of the VIX index are also included, providing a widely used measure of expected market volatility based on S&P 500 index option data. Model training is performed on data from 2000 to 2015 using a random subset comprising 3/4 of the stocks, while evaluation is conducted on all 848 stocks over the 2015–2020 test period. To represent the market, a broad market index is constructed as an equal-weighted portfolio of all stocks. The following features are calculated on historical windows:

- Average return $\overline{Y}_i$,
- Standard deviation $\hat{\sigma}_i$,
- Beta coefficient $\hat{\beta}_i$,
- Average VIX coefficient $\overline{VIX}$.

For each time step t and stock i, features are calculated on multiple historical window lengths $T = [5, 21, 63, 126]$ (measured in days). Each of the four features is computed over all four historical windows, resulting in a total of 16 features used to predict the target variables. The target for time step t and stock i is calculated as the sample (for volatility) or OLS (for beta) estimate on a look-ahead (future) window of length $T' = 21$ days. The historical sample estimator, used as a benchmark, is also computed over a 21-day window for both volatility and beta. This choice of features is intended to effectively capture return dynamics and enable reliable forecasts of future volatilities and betas.

IV. Performance evaluation

A. Performance measures

The model performance is evaluated using two primary measures: mean squared error (MSE) and average log-likelihood (LL). MSE was calculated for both the training and test sets to assess the prediction of logarithmic volatility and beta coefficients.

Log-likelihood measures the likelihood of observing the given data (multivariate return time series) under a particular statistical model. It provides insight into how well the model fits the data: higher values of LL indicate that the model better represents the underlying distribution of the returns. For this analysis, LL was calculated only on the test set, for each window where the covariance matrix was predicted. As the most commonly employed distribution in many machine learning applications, we use the joint multivariate Gaussian log-likelihood $\mathcal{L}_{\mathcal{N}}(y_{t+1}, \ldots, y_{t+T'} | \hat{\mu}, \hat{\Sigma})$, measured for standardized returns across all test windows of length T'. Moreover, to account for the heavy tails in empirical return distributions and the well-known extreme events in financial markets, we also calculate the likelihood of the data given the multivariate Student's t-distribution $\mathcal{L}_t(y_{t+1}, \ldots, y_{t+T'} | \hat{\mu}, \hat{\Sigma}, \nu)$ [15].

Since most econometric literature dealing with factor models emphasises the importance of the finite 4^{th} moment in asset return data, we use the degrees-of-freedom parameter $\nu = 5$ for which kurtosis is finite.

B. Application to portfolio optimization

The predicted covariance matrices were used in a portfolio optimization scenario. We use the global minimum variance portfolio [16], which relies only on the estimated covariance matrix, and is therefore ideal for evaluating the economic impact of different predictive models. The goal is to find portfolio weights $w = [w_1, \ldots, w_p]^\top$ which minimize the variance:

$$\begin{aligned} \min_{w} \; & w^\top \Sigma w \\ \text{s.t.} \quad & \sum_{i=1}^{N} w_i = 1. \end{aligned} \tag{9}$$

We calculate the minimum variance portfolio for each test window and reconstruct the portfolio returns, which are then use to calculate the portfolio volatility, mean return, and the Sharpe ratio.

V. Results

First, we inspect the ability of the developed models to accurately predict future values of betas and volatilities. Table I shows the MSE on train and test sets for the log-volatility prediction, for different methods: historical sample estimates and predictive models (feedforward neural network – FF, linear model, random forest – RF, and xgboost – XGB). The predictive models notably outperform the historical sample estimates, and the test MSE results demonstrate that they generalize well out of sample. The random forest model seems to show the most significant improvement on the train set, which is expected due to the way random forests fit the training data. However, they still generalize fairly well out-of-sample and show the best results, although not to the same extent.

TABLE I
Train and test MSE for predictive models of individual stock log-volatilities $\ln(\sigma_i)$, calculated over all stocks and test windows from 2015 to 2020.

Model	Train MSE	Test MSE
historical sample	0.3454	0.2749
FF	0.1333	0.1349
linear	0.1230	0.1238
RF	0.0100	0.1163
XGB	0.0888	0.1200

Table II shows the MSE on train and test sets for the beta parameter prediction, for different methods (historical OLS estimates and predictive models). In analogy with the predictive performance for the volatility models, the predictive models for beta also outperform the historical sample estimates, and the test MSE results demonstrate that they generalize well out of sample. Again, the random forest model shows the best in-sample performance, and although it does generalize well, the linear regression model seems to slightly outperform other predictive methods out-of-sample.

TABLE II
Train and test MSE for predictive models of stock betas (β_i), calculated over all stocks and test windows from 2015 to 2020.

Model	Train MSE	Test MSE
historical OLS	0.03779	0.10234
FF	0.02960	0.07265
linear	0.01819	0.05527
RF	0.00160	0.05945
XGB	0.01567	0.05801

The results above demonstrate that the models do indeed perform the predictive tasks they were trained to do, and they all are show improved performance over the benchmark methods using historical estimates.

We also study the ability of the predictive models to explain the observed future returns. Table III shows the log-likelihood values for the correlation matrices given future standardized returns on the test set, for the historical OLS estimator and the considered predictive models.

TABLE III
Average Gaussian log-likelihood $\mathcal{L}_{\mathcal{N}}$ and Student's t log-likelihood $\mathcal{L}_t$ for look-ahead standardized returns on the test set windows of length T'.

Model	$\mathcal{L}_{\mathcal{N}}$	$\mathcal{L}_t$
historical OLS	-22256	-21509
FF	-22493	-21740
linear	-22183	-21471
RF	-22238	-21497
XGB	-22219	-21486

Interestingly, although all of the considered predictive models were shown to outperform the historical OLS estimators on the MSE measure, not all of them outperform regarding the return log-likelihood, most notably the feedforward neural network model. Nevertheless, the other considered predictive models do indeed outperform the historical OLS method.

Finally, we use the predicted covariances in the global minimum portfolio optimization process, and calculate the returns of the obtained portfolio on the test dataset. We do this for different combinations of models for the betas and volatilities, and report the results for all pairs in Table IV.

Thes results suggest that the historical sample and OLS methods yield global minimum variance which exhibit the lowes out-of-sample variances overall. However, these methods were notably outperformed by the predictive models, as measured by MSE and log-likelihood. This is particularly striking, since the results suggest that the predictive models which may indeed produce more accurate predictions of market betas and volatilities, do not necessarily produce the best outcomes when these predictions are used in portfolio optimization. This is a striking result, which may be explained by the fact that commonly used loss functions in machine

TABLE IV
ANNUALIZED VOLATILITIES OF THE GMV PORTFOLIOS OPTIMIZED USING PREDICTED COVARIANCES OBTAINED WITH DIFFERENT COMBINATIONS OF MODELLING APPROACHES FOR THE VOLATILITIES AND BETAS.

	Beta model				
Vol. model	hist. OLS	FF	linear	RF	XGB
hist. sample	10.62%	11.36%	11.09%	11.36%	11.23%
FF	13.02%	16.64%	16.50%	17.38%	17.17%
linear	12.78%	13.72%	14.76%	15.33%	15.18%
RF	12.19%	13.65%	13.88%	14.62%	14.27%
XGB	12.22%	12.97%	13.55%	14.22%	13.93%

learning simply disregard the economic utility of the predictive models. This aligns with recent findings on labeling approaches, which demonstrate that incorporating economic objectives into the labeling process – such as aligning labels with optimal portfolio decisions – can substantially enhance the real-world performance of such models in financial applications [17]. Although our choice of the MSE loss function is motivated by its widespread use across various domains and supported by prior research [14], [18], the somewhat unexpected results presented here offer fresh insights into the challenges of applying machine learning in financial contexts. We hope these findings encourage further research into the design of task-specific target functions tailored to the unique objectives of financial applications.

VI. CONCLUSION

In this paper we consider the problem of predictive risk modelling from multivariate time series of asset returns. We propose an approach based on the decomposition of the asset return covariance matrix into its volatilities and correlations. To produce positive definite predictions of correlations in high dimensions, we adopt a single factor model and pose the correlation prediction problem as a supervised task of predicting the factor loadings. We develop several models following the proposed approach and test their performance on historical asset data. We measure the performance of the developed models using MSE and the log-likelihood of the look-ahead return data. In addition, we construct the predicted covariance matrices following different combinations of models for the volatilities and betas, and use these covariances in a portfolio optimization setting. The results demonstrate that the developed models outperform the benchmark methods when considering the performance measures, but not in the portfolio optimization application. Moreover, further research should also incorporate multiple factors and enrich the representations of correlations, and perhaps include additional fundamental and macroeconomic data.

REFERENCES

[1] T. Roncalli, *Handbook of Financial Risk Management*. Chapman & Hall/CRC Financial Mathematics Series, 2020.

[2] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, vol. 19, pp. C1–C32, 2 2016.

[3] J. Bun, J. P. Bouchaud, and M. Potters, "Cleaning large correlation matrices: Tools from random matrix theory," *Physics Reports*, vol. 666, pp. 1–109, 2017.

[4] S. Begušić and Z. Kostanjčar, "Cluster-based shrinkage of correlation matrices for portfolio optimization," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 301–305, IEEE, 9 2019.

[5] J. Fan, Y. Liao, and M. Mincheva, "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, pp. 603–680, 9 2013.

[6] S. Begušić and Z. Kostanjčar, "Cluster-specific latent factor estimation in high-dimensional financial time series," *IEEE Access*, vol. 8, pp. 164365–164379, 9 2020.

[7] R. F. Engle, O. Ledoit, and M. Wolf, "Large dynamic covariance matrices," *Journal of Business and Economic Statistics*, vol. 37, pp. 363–375, 4 2019.

[8] S. Deshmukh and A. Dubey, "Improved covariance matrix estimation with an application in portfolio optimization," *IEEE Signal Processing Letters*, vol. 27, pp. 985–989, 2020.

[9] C. Lam, "High-dimensional covariance matrix estimation," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 12, 10 2020.

[10] J. Fan, Y. Liao, and M. Mincheva, "High-dimensional covariance matrix estimation in approximate factor models," *Annals of Statistics*, vol. 39, pp. 3320–3356, 12 2011.

[11] G. Feng, S. Giglio, and D. Xiu, "Taming the factor zoo: A test of new factors," *Journal of Finance*, vol. 75, pp. 1327–1370, 6 2020.

[12] P. Cizeau, M. Potters, and J. P. Bouchaud, "Correlation structure of extreme stock returns," *Quantitative Finance*, vol. 1, pp. 217–222, 2001.

[13] L. R. Goldberg and A. N. Kercheval, "James–stein for the leading eigenvector," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, 1 2023.

[14] W. Drobetz, F. Hollstein, T. Otto, and M. Prokopczuk, "Estimating stock market betas via machine learning," *Journal of Financial and Quantitative Analysis*, pp. 1–37, 2024.

[15] S. Kotz and S. Nadarajah, *Multivariate T-Distributions and Their Applications*. Cambridge University Press, 2004.

[16] R. Clarke, H. D. Silva, and S. Thorley, "Minimum-variance portfolio composition," *Journal of Portfolio Management*, vol. 37, pp. 31–45, 2011.

[17] T. Kovačević, A. Merćep, S. Begušić, and Z. Kostanjčar, "Optimal trend labeling in financial time series," *IEEE Access*, vol. 11, pp. 83822–83832, 2023.

[18] O. Ledoit and M. Wolf, "The power of (non-)linear shrinking: A review and guide to covariance matrix estimation," *Journal of Financial Econometrics*, vol. 20, pp. 187–218, 1 2022.