Graphical Bayesian Networks with Topic Modeling Priors for Predicting Asset Covariances

Minghan Li
Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA 15213
minghanl@andrew.cmu.edu

Fan Pu Zeng
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
fzeng@andrew.cmu.edu

1 Introduction

We investigate the effectiveness of Bayesian networks in predicting the covariance matrix of financial assets (specifically a subset of the S&P 500), evaluated against Heterogeneous Autoregressive (HAR) models. In particular, we consider both HAR-DRD, based on the DRD decomposition of the covariance matrix, and Graphical HAR (GHAR)-DRD, which is also based on DRD decomposition but also makes use of graphical relationships between the assets. To build the graph representing relationships between the assets, we apply Latent Dirichlet allocation (LDA) on the 10-K filings of each of the companies, and infer edges based on topic overlap. We show that this technique has limited usefulness in our setup, but provides recommendations on how it could be further improved based on our observations of its predictions.

2 Background

2.1 Covariance Matrix Estimation

Covariance matrix estimation is a century-old research topic in portfolio theory and quantitative finance [12]. A common goal in portfolio optimization has been maximizing expected return and minimizing variance of the resulting portfolio (Equation 1).

$$\max_{\mathbf{x}} \mu^T \mathbf{x} - \frac{\gamma}{2} \mathbf{x}^T \mathbf{V} \mathbf{x}, \qquad \text{where } \mathbf{x} \in \mathcal{X}. \tag{1}$$

where $\mathbf{x}$ is a vector of weights assigned to each asset subject to certain constraint, μ is a vector of expected return, $\mathbf{V}$ is the covariance matrix of asset return, and γ is penalty term.

While providing sound theoretical guarantees, portfolio optimization's success in the real-world portfolios is limited. In particular, the reliance on accurate estimates of the two major inputs to the models, expected return and covariance matrix of individual assets, makes the problem difficult. Hedge funds and asset managers usually leverage their expertise to build good models for expected asset returns, whereas methods for covariance matrix estimation are relatively standard in the industry. We recognize that the ability to accurately forecast the covariance could generate huge value, such as by leveraging novel modeling techniques.

2.2 Form 10-K

The form 10-K is an annual report that is filed with the Securities and Exchange Commission (SEC) by public companies. The report provides a comprehensive overview of the company's business and financial performance for the year [9]. The form 10-K is highly suitable for use in topic modeling since the format of the document is standardized and is publicly accessible for all listed companies.

10-708 Class Project (2022).

We will be making use of the "Business Description" section of the report, as it contains information about the main business of the company, its subsidiaries, the markets that it is involved in, etc.

To motivate why the form 10-K might be informative, we plotted the word cloud of three publicly traded stocks in different industries, which can be found in Appendix A.1.

2.3 Topic Modeling with Latent Dirichlet Allocation models

Topic modeling is a technique for discovering latent topics within a collection of documents. In our paper, we will use Latent Dirichlet Allocation (LDA) models [8]. Note that topic modeling is not the focus of this paper and therefore LDA performs sufficiently well for the purposes of extracting a graphical relationship.

LDA is a statistical model that estimates the probability of each document in a collection of documents belonging to each topic. It does this by first assigning each word in the document to one of a predefined number of topics, and then estimating the probability of each document belonging to each topic [5].

2.4 Bayesian Networks

Bayesian modeling is a framework for representing and reasoning about uncertain concepts. It is based on the idea that probability represents the degree of belief in an event or proposition.

Bayesian networks are a type of probabilistic graphical model that represent the relationships between random variables as a directed acyclic graph (DAG). Each node in the graph represents a random variable, and the edges between nodes represent the dependencies between those variables. Bayesian networks can be used to model financial phenomenon and make predictions about future events. In this paper, each node in the network would represent the price of a stock, and the edges between nodes would represent the dependencies between these variables, such as whether two stocks share common topics.

3 Related Work

The literature contains a variety of proposed methods to address the shortcomings of sample covariance matrix as the estimated covariance matrix for a large universe of assets. One family of solutions is to use factor models to perform dimensionality reduction [1]. While factor models have empirically shown to be successful in improving the accuracy in covariance estimation, the identification of appropriate factors remains a major bottleneck. Another family of methods use shrinkage, a process that applies a transformation to the sample covariance matrix. One of the most famous shrinkage methods is proposed by Ledoit & Wolf [10]. Yet, the optimal choice of the shrinkage target still remain ambiguous.

More recent work focused on modeling decomposition of covariance matrix using Heterogeneous Autoregressive (HAR) model [11]. [13] introduced HAR-DRD model, which forecasts individual individual elemments in DRD decomposition of the covariance matrix HAR model, which has been shown to yield promising results in high-frequency datasets. Most recently, Zhang [15] extended the HAR-DRD framework by introducing an graphical prior into the model and allow for dynamic construction of graphical relationships.

4 Methods

We investigate whether using Bayesian networks as a novel modeling paradigm could be applied to covariance matrix estimation. We first construct baselines using a modern framework for covariance estimation, called the Heterogeneous Autoregressive (HAR)-DRD model. We then improve upon the existing framework by applying it to graphical HAR-DRD models, which augments the autoregressive framework with the ability to make use of the relationships between different stocks. Finally, we develop a modeling framework based entirely on graphical Bayesian networks.

4.1 Dataset

The 10-K data was taken from the U.S. Securities and Exchange Commission's EDGAR online database using the `secedgar` Python library. The list of S&P 500 historical components from 2010-01-01 to 2022-01-01 is obtained from Wharton Research Data Services along with their ticker symbols and full company name. The 10-K raw HTML data was first queried using company tickers; in cases where a company's ticker cannot uniquely identify it, we re-query using the company's list of full names adopted during the period. The final 10-K raw data consists of filings from a total of 726 companies.

Historical price data was obtained from Yahoo Finance using the `yfinance` library [14]. We used adjusted closing price data from 2011-01-01 to 2022-12-01, which accounts for stock splits and dividends.

4.2 Data Cleaning

To prepare the 10-K documents, we extracted text from the business description section, stripped all HTML tags and numerical tokens, and removed stop words using the `nltk` library [2]. We used `pyLDAvis` [3] to perform visualization of the top terms found by a preliminary run of LDA on the corpus as a sanity check. This revealed that many of the top terms are very generic in the context of the "Business Description" section (i.e business, company, etc) (see Appendix B.1 for a visual ranking), and therefore we also decided to remove these terms. A full listing of removed terms is given in Appendix B.2. Finally, 10-K documents whose "Business Description" section had less than 1000 words are also considered malformed and are removed. Overall 341 tickers from S&P500 passed the data cleaning step.

4.3 Latent Dirichlet Allocation

We used all tokens from the cleaned 10-K data to form a dictionary, and then performed LDA on this corpus. LDA was performed using the online variational Bayes (VB) algorithm as proposed in [6] with the `Gensim` library [4]. LDA was run with 30 topics. Such a large number was chosen because it allows us to more finely differentiate between different industries, allowing us to have a sparser networks afterwards.

Due to computational limitations, we sampled 86 random stocks after LDA, and used them as all the assets that we will consider in the remainder of the paper. The full list of tickers used is given in Appendix B.4.

4.4 HAR-DRD and GHAR-DRD Models

We introduce the following notation used throughout the paper:

- N is the number of assets in consideration,
- $N^{\#}$ is the number of correlation coefficients between all assets, $N^{\#} = \frac{N(N-1)}{2}$,
- $\mathbf{p}_t \in \mathbb{R}^N$ is vector of the adjusted closing price of all stocks at time t,
- $\mathbf{r}_t \in \mathbb{R}^N$ is the vector of log-returns at time t, given by $\mathbf{r}_t = \log \mathbf{p}_t - \log \mathbf{p}_{t-1}$,
- $\mathbf{H}_t \in \mathbb{R}^{N \times N}$ is the realized covariance (RC) matrix based on the log-returns of the last 21 trading days (roughly corresponds to a month of physical days). The RC estimator is given by $\mathbf{H}_t \coloneqq \sum_{k=1}^{21} \mathbf{r}_{t-k} \mathbf{r}_{t-k}^\top$,
- $\mathbf{RV}_t \in \mathbb{R}^N$ is the daily realized volatility for all assets at time t. It is given by the diagonal entries of $\mathbf{H}_t$, i.e $\mathbf{RV}_t[i] = \mathbf{H}_t[i,i]$. We use square brackets for denoting accessing vector or matrix entries to avoid overloading notation of the subscripts used to denote time.
- $\mathbf{D}_t \in \mathbb{R}^{N \times N}$ is the diagonal matrix formed from $\mathbf{RV}_t$, i.e $\mathbf{D}_t[i,i] = \mathbf{RV}_t[i]$,
- $\mathbf{R}_t \in \mathbb{R}^{N \times N}$ is the correlation matrix of the log-returns at time t,
- $\mathbf{x}_t \in \mathbb{R}^{N(N-1)/2}$ is the vectorization of the lower triangular entries of $\mathbf{R}_t$, i.e $\mathbf{x}_t = \text{vech}(\mathbf{R}_t)$,

- $\mathbf{A} \in \{0,1\}^{N \times N}$ is the symmetric adjacency matrix of the assets, where $\mathbf{A}[i,j] = 1$ if asset i is related to asset j, and $\mathbf{A}[i,j] = 0$ otherwise. Whether two assets are related can be determined in various ways. [15] uses the Global Industry Classification Standard (GICS) to construct such relations. The GICS assigns companies to economic sectors, and therefore companies in the same sector share edges in the graph of their stocks.
- $\mathbf{O} \in \mathbb{N}^{N \times N}$ is the diagonal matrix with diagonal entries representing the degrees of each vertex, as given by $\mathbf{A}$,
- $\mathbf{W} \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix of the assets, given by $\mathbf{W} = \mathbf{O}^{-1/2}\mathbf{A}\mathbf{O}^{-1/2}$.

The HAR-DRD model on the decomposition of the realized covariance matrix $\mathbf{H}_t$ into the product of the diagonal matrix of realized volatilities and the correlation matrix [13], i.e

$$\mathbf{H}_t = \mathbf{D}_t \mathbf{R}_t \mathbf{D}_t.$$

In the HAR-DRD model, we represent $\mathbf{RV}_t$ and $\mathbf{x}_t$ as linear equations that depend on their past values, given by

$$\mathbf{RV}_t = \alpha_0 + \beta_d \mathbf{RV}_{t-1} + \beta_w \mathbf{RV}_{t-5:t-2} + \beta_m \mathbf{RV}_{t-22:t-6} + \mathbf{u}_t \tag{2}$$

$$\mathbf{x}_t = \alpha_0' + \beta_d' \mathbf{x}_{t-1} + \beta_w' \mathbf{x}_{t-5:t-2} + \beta_m' \mathbf{x}_{t-22:t-6} + \mathbf{u}_t', \tag{3}$$

where $\mathbf{u}_t$ is an error term with mean 0, and $\alpha_0, \alpha_0', \beta_d, \beta_d', \beta_w, \beta_w', \beta_m, \beta_m'$ are coefficients to be found from regression. The d, w, m subscripts denote days, weeks, and months respectively. The range indexing of $\mathbf{RV}$ and $\mathbf{x}$ corresponds to the averaged realized covariance over days in the last trading week and trading month, i.e

$$\mathbf{RV}_{t-5:t-2} = \frac{1}{4}\sum_{k=2}^{6} \mathbf{RV}_{t-k}, \qquad \mathbf{RV}_{t-6:t-22} = \frac{1}{17}\sum_{k=6}^{22} \mathbf{RV}_{t-k}, \tag{4}$$

and similarly for the case of $\mathbf{x}_{t-5:t-2}$ and $\mathbf{x}_{t-6:t-22}$.

GHAR-DRD (Graph HAR-DRD) model is very similar to HAR-DRD, except it adds more covariates to model the interdependence of asset variances. Specifically, we use the normalized adjacency matrix $\mathbf{W}$ to represent effects of neighborhood aggregation over different horizons.

We construct $\mathbf{W}$ from the results of topic modeling 10-K documents by taking the top 2 topics for each company, and defining that they share an edge if they share one topic in common. The graph $\mathbf{A}$ found by LDA is visualized in Figure 8 in the Appendix.

The GHAR-DRD framework allows us to model the relationships of $\mathbf{RV}_t$ and $\mathbf{x}_t$ in a more expressive way by being able to capture the impact of neighboring assets:

$$\begin{aligned}\mathbf{RV}_t =& \alpha_0 + \beta_d \mathbf{RV}_{t-1} + \beta_w \mathbf{RV}_{t-5:t-2} + \beta_m \mathbf{RV}_{t-22:t-6} \\ &+ \gamma_d \mathbf{W} \cdot \mathbf{RV}_{t-1} + \gamma_w \mathbf{W} \cdot \mathbf{RV}_{t-5:t-2} + \gamma_m \mathbf{W} \cdot \mathbf{RV}_{t-22:t-6} + \mathbf{u}_t, \end{aligned} \tag{5}$$

$$\begin{aligned}\mathbf{x}_t =& \alpha_0' + \beta_d' \mathbf{x}_{t-1} + \beta_w' \mathbf{x}_{t-5:t-2} + \beta_m' \mathbf{x}_{t-22:t-6} \\ &+ \gamma_d' \mathbf{x}_{t-1} + \gamma_w' \mathbf{x}_{t-5:t-2} + \gamma_m' \mathbf{x}_{t-22:t-6} + \mathbf{u}_t'. \end{aligned} \tag{6}$$

The first line of Equations 5 and 6 captures its historical relationship with itself and is the same as in Equations 2 and 3, while the second line captures the historical relationship of each asset with its neighbors as chosen by $\mathbf{W}$.

4.5 Bayesian Network Graphical Model

We can also formulate the forecasting of $\mathbf{RV}_t$ and $\mathbf{x}_t$ from a graphical model point of view. In this view, the realized covariance and the correlation view is the observation of a probabilistic process based on the realized covariances and correlations from previous days.

Our formulation for predicting $\mathbf{RV}_t$ and $\mathbf{x}_t$ using a graphical model generalizes that of GHAR-DRD given by Equations 5 and 6 as follows:

$$\mathbf{RV}_t = \alpha + \widetilde{\mathbf{W}}_d \mathbf{RV}_{t-1} + \widetilde{\mathbf{W}}_w \mathbf{RV}_{t-5:t-2} + \widetilde{\mathbf{W}}_w \mathbf{RV}_{t-22:t-6}, \tag{7}$$

$$\mathbf{x}_t = \alpha' + \widetilde{\mathbf{W}}_d' \mathbf{x}_{t-1} + \widetilde{\mathbf{W}}_w' \mathbf{x}_{t-5:t-2} + \widetilde{\mathbf{W}}_m' \mathbf{x}_{t-22:t-6}, \tag{8}$$

where our variables and distributional assumptions are

$$\widetilde{\mathbf{W}}_t[i,j] = \begin{cases} \beta_t^{i,j} & \text{if assets } i \text{ and } j \text{ are related, i.e } A[i,j]=1, \\ 0 & \text{otherwise,} \end{cases} \quad \text{for } t \in \{d,m,w\}, \tag{9}$$

$$\widetilde{\mathbf{W}}'_t[(i,j),(k,l)] = \begin{cases} \beta_t^{(i,j),(k,l)} & \text{if } \{i,j\} \cap \{k,l\} \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

$$\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha), \tag{11}$$

$$\beta_t^{i,j} \sim \mathcal{N}\left(\mu_t^{i,j}, \sigma_t^{i,j}\right), \quad \text{for } t \in \{d,m,w\}, \tag{12}$$

and $\alpha', \beta_t^{'(i,j),(k,l)}$ are defined analogously and omitted to save space. Note that we no longer require an error term since this can be modeled by the variance of $\mathbf{RV}_t$ and $\mathbf{x}_t$, and that $\widetilde{\mathbf{W}}_t \in \mathbb{R}^{N \times N}, \widetilde{\mathbf{W}}'_t \in \mathbb{R}^{N^\# \times N^\#}$.

Then to find our coefficients $\alpha, \alpha', \widetilde{\mathbf{W}}_t, \widetilde{\mathbf{W}}'_t$ for each $t \in \{d,m,w\}$, we can perform maximum likelihood estimation by assuming that both $\mathbf{RV}_t$ and $\mathbf{x}_t$ are also drawn from a normal distribution. Let $\widetilde{\mathbf{W}}$ be shorthand for the parameterization for $\widetilde{\mathbf{W}}_d, \widetilde{\mathbf{W}}_w, \widetilde{\mathbf{W}}_m, \alpha$, and analogously for $\widetilde{\mathbf{W}}'$. Then we can model the likelihood as the following, for some prior of the covariances of their distributions Σ, Σ':

$$\Sigma \sim \text{HalfCauchy}(\beta_\Sigma), \qquad \Sigma' \sim \text{HalfCauchy}(\beta_{\Sigma'}), \tag{13}$$

$$\mathbf{RV}_{t,obs} \sim \mathcal{N}(\mathbf{RV}_{t,\widetilde{\mathbf{W}}}, \Sigma), \qquad \mathbf{x}_{t,obs} \sim \mathcal{N}(\mathbf{x}_{t,\widetilde{\mathbf{W}}'}, \Sigma'), \tag{14}$$

where we use a Half Cauchy distribution for the prior since the covariance matrix is non-negative, and so our two separate maximum log-likelihood objectives are

$$\max_{\widetilde{\mathbf{W}}} \sum_{\mathbf{RV}_{t,obs}} \log p_{\mathbf{W}_t}(\mathbf{RV}_{t,obs}), \qquad \max_{\widetilde{\mathbf{W}}'} \sum_{\mathbf{x}_{t,obs}} \log p_{\mathbf{W}'_t}(\mathbf{x}_{t,obs}). \tag{15}$$

To perform training in practice, we do this using Markov Chain Monte Carlo methods. The specific sampling algorithm used is the No-U-Turn Sampler (NUTS), which is a self-tuning variant of the Hamiltonian Monte Carlo sampler [7]. This is achieved by adaptively constructing a piecewise linear trajectory that approximately follows the gradient of the Bayesian network.

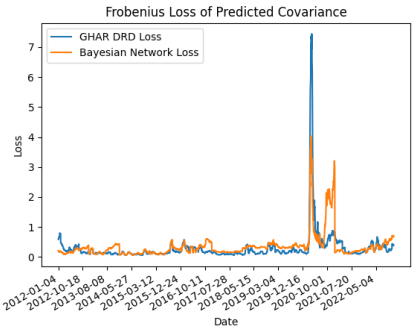
We used 500 draws, 200 tuning steps, and sampled a single chain when estimating each set of coefficients. If multiple chains were sampled, the NUTS sampler would also help to check if all chains are converging to the same stationary distribution and would report when that is not the case. However, in our case a single chain was sufficient for efficiency reasons as we ran prior experiments to see how many tune and draws were necessary for it to always converge.

Due to computation limitations, instead of actually using $\widetilde{\mathbf{W}}'_d, \widetilde{\mathbf{W}}'_w, \widetilde{\mathbf{W}}'_m$, we replace it with scalar variables following the normal distribution, $\widetilde{\mathbf{w}}'_d, \widetilde{\mathbf{w}}'_w, \widetilde{\mathbf{w}}'_m \sim \mathcal{N}(\cdot,\cdot)$. This is because $\widetilde{\mathbf{W}}'_t$ contains $O(N^3)$ parameters to optimize over, which becomes computationally intractable to perform sampling over so many iterations for our choice of $N = 86$. To get a sense of how dense $\widetilde{\mathbf{W}}_t$ can be, the degrees of the nodes in $\mathbf{A}$ are given as a histogram in Figure 9.

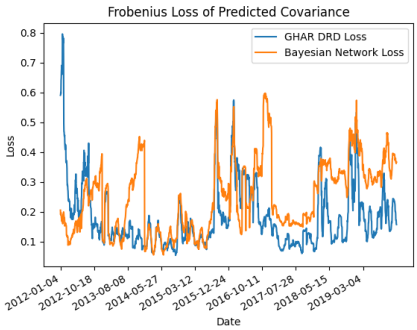
5 Results

5.1 Predicted Covariance Matrices

Figure 1 plots the loss as defined by the Frobenius norm of the difference between the ground truth realized covariance matrix, and the predicted realized covariance matrices using both GHAR-DRD and the Bayesian network. We provided plots over two time ranges, 2012-01-01 to 2022-12-01, and also 2012-01-01 to 2020-01-01. This is due to the large spike in loss experienced during the Covid-19 pandemic in 2022, which resulted in adverse market conditions that both models performed poorly under, and overshadows the datapoints from other periods of time.



(a) Frobenius loss of predictions between 2012-01-01 to 2022-12-01



(b) Frobenius loss of predictions between 2012-01-01 to 2020-01-01

Figure 1: Loss of GHAR-DRD vs Bayesian Network

5.2 Backtesting

In order to further assess the quality of the predicted realized covariance matrices, we designed a test that mimics a real-world problem faced by asset managers. In particular, asset managers often try to optimize the trade-off between expected return and variance in a portfolio. The goal is to construct a portfolio that would maximize expected return for a given variance limit and minimize variance for a given expected return.

Due to the difficulty of obtaining reliable expected returns, we decided to instead optimize over variance of the portfolio only. We define the problem of minimizing the return of a long-only fully-invested portfolio as follows:

$$\min_{\mathbf{x}} \mathbf{x}^\top \mathbf{V} \mathbf{x} \qquad \text{subject to } \mathbf{1}^\top \mathbf{x} = 1, \tag{16}$$

where $\mathbf{x}$ are the portfolio weights, and $\mathbf{V}$ is the covariance matrix, either realized or predicted.

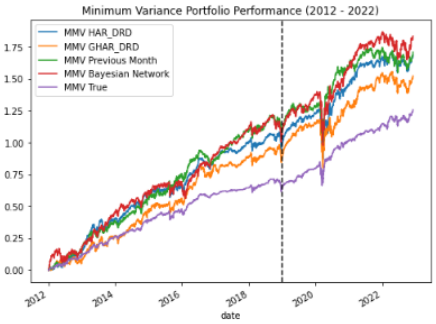
Suppose that we would like to invest in a strategy that mimics the return of the minimum-variance portfolio. In the beginning of each month, we construct the minimum-variance portfolio through an optimizer and observe its return over the month. The process continues from 2012-01-01 to 2022-12-01. We use different $\mathbf{V}$ from different prediction methods for comparison and try to compare the results with the result of true minimum-variance portfolio obtained using the realized covariance matrix. Notice that the result of true minimum-variance portfolio cannot be obtained in real trading. It is only a hypothetical scenario where an oracle could tell us the realized covariance matrix exactly. A predicted covariance matrix could potentially add value if its optimized portfolio match closely with the true minimum-variance portfolio.

Besides looking at the return curves of constructed portfolios, one could also look at the tracking error of the portfolios relative to the minimum-variance portfolio. The tracking error of a portfolio relative to a benchmark $\mathbf{x}_{\text{bm}}$ is defined as $(\mathbf{x} - \mathbf{x}_{\text{bm}})^\top \mathbf{V} (\mathbf{x} - \mathbf{x}_{\text{bm}})$. Tracking error is a more direct quantitative measure of closeness between portfolios.

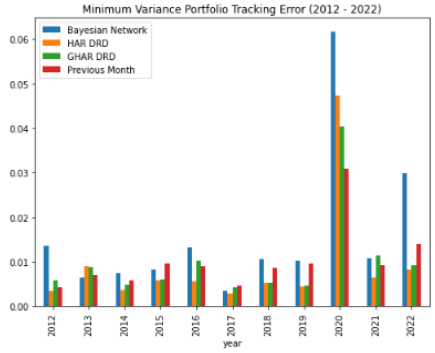
6 Discussion and Analysis

6.1 Analysis of Loss of Predicted Covariance Matrices

In Figure 1(b), we see that sometimes the Bayesian network has better loss than GHAR-DRD, and vice versa. There are also many periods of times where their predictions tend to agree in terms of loss, such as between 2014 to 2017. However, from Figure 1(a), we see that the peak loss incurred by the Bayesian network was only half that of GHAR-DRD, indicating that it might be a better model to capture uncertainty. This could be due to how it takes a Bayesian modeling approach and therefore can model uncertainty better since a conditional Bayesian model allows us to predict a distribution of distributions. Overall though, GHAR-DRD still incurred less overall loss during the entire test period.

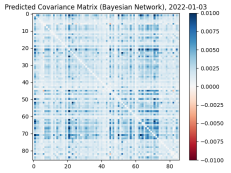


(a) Minimum Variance Portfolio Cumulative Log Return

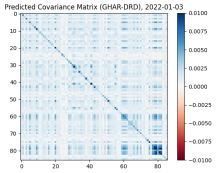


(b) Minimum Variance Portfolio Tracking Error

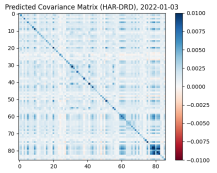
Figure 2: Minimum Variance Portfolio Backtest



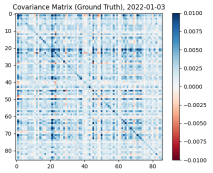
(a) Covariance matrix predicted by Bayesian network



(b) Covariance matrix predicted by GHAR-DRD model



(c) Covariance matrix predicted by HAR-DRD model



(d) Ground-truth covariance matrix

Figure 3: Comparison of covariance matrix predicted by the Bayesian network, GHAR-DRD, HAR-DRD, and ground truth.

6.2 Backtesting Result

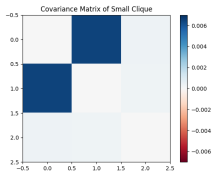
The plot of cumulative sum of the log return for minimum-variance portfolio constructed using different methods is shown in Figure 2. The return series that most closely track true minimum variance portfolio return series is GHAR DRD, followed by HAR DRD, Beysian Network, and then Previous Month. Looking at the total daily tracking error of each methods throughout the our testing period, portfolio constructed using HAR DRD and GHAR DRD consistently outperform those constructed using the other two methods by generating smaller tracking error to the minimum variance portfolio.

6.3 Visualization and Analysis of Covariance Matrix

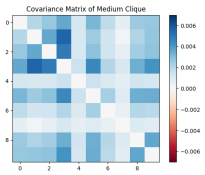
To further understand where the differences between the covariance matrix predicted by the Bayesian network, GHAR-DRD, and HAR-DRD differ from each other and the ground-truth, we visualized it on the same scale in Figure 3. The first thing that stands out in Figure 3 is the difference in the diagonal elements. The diagonal of the covariance matrix predicted by the Bayesian network is very weak, meaning that the Bayesian network believes that the inherent variance of the log returns of each stock is very low and is mostly explained by the behavior of other stocks. On the other hand, the predictions by GHAR-DRD and HAR-DRD puts most of the weights on the diagonal, which is the opposite behavior and is closer to the true covariance.

It is also noteworthy that all the predicted covariance matrix do not learn many negative correlations, as compared to the ground truth covariance matrices. This is expected, since we intentionally built the graph $\mathbf{A}$ based on stocks which are related to one another.

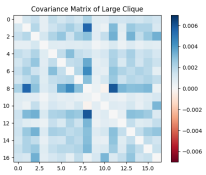
Finally, the predicted covariance matrices by GHAR-DRD and HAR-DRD are actually sparser than that of the Bayesian network, which is yet sparser still than the ground-truth covariance matrix. This could be due to the fact that GHAR-DRD and HAR-DRD are solved exactly by regression, which would sparsify the coefficients as those that do not contribute to optimality are ignored, whereas in



(a) Small 3-clique comprising the following tickers: AVB, NFLX, CTSH.



(b) Medium 10-clique comprising the following tickers: TTWO, CMG, SYY, CL, CRL, TER, MKC, BIO, GPC, WST.



(c) Large 17-clique comprising the following tickers: GIS, BWA, WST, ETN, PKG, MU, CAT, ROK, HON, IEX, AMD, QCOM, ITW, INTC, LIN, ANSS, MHK.

Figure 4: Predicted covariance matrices containing only assets from selected cliques in $\mathbf{A}$

Bayesian networks, the coefficients are sampled from a stationary distribution during Markov Chain Monte Carlo and are unlikely to be exactly 0. In addition, the sparsity of the diagonals in Bayesian networks means that the weights must be distributed to the neighbors of each asset in the covariance matrix.

In addition, we investigate the behavior of the predicted covariance matrix by Bayesian networks of randomly chosen cliques of different sizes (3, 10, 17), as shown in Figure 4. We observed that it tends to only form a few edges with very strong correlations in all clique sizes, but forms a lot of edges with weak correlations. This helps to show that the entire clique tends to move in the same direction, in line with our expectations as they all share the same common top topics.

6.4 Overall Analysis

Overall, our results show that our Bayesian network and GHAR-DRD models do not outperform HAR-DRD for covariance estimation for our setup, but we have identified limitations in our approach which can be addressed in future work to determine if this gap can be closed or surpassed.

First, predictions could be improved by incorporating both strong positive and negative relationships into the graph, as we showed that the predicted covariances are not making use of negative relationships. For future research, new ways of forming inductive bias about negative relationships should be developed. Our graphical prior could also be used as a complementary method to the former methods like HAR rather than a standalone method.

Second, due to the recent development in electronic trading, price relationship between stocks become increasingly volatile. The majority of the literature nowadays focus on high-frequency daily realized covariance constructed using intraday price data. Due to data limitations, we could only construct a monthly realized covariance matrix using daily data. Future work should aim at applying the same approach to higher frequency price series.

Lastly, we did not exploit the full generality of the Bayesian network for forecasting the correlations $\mathbf{x}_t$ due to computational limitations since we used scalar coefficients instead of $\widetilde{\mathbf{W}}_t$ (Equation 8). It could possibly perform better when it can harness its full range of expressiveness.

7 Teammates and Work Division

The team comprises Kevin Li and Fan Pu Zeng. Kevin worked on obtaining and processing the financial data, constructing HAR and GHAR baselines, and performing backtesting analysis. Fan Pu worked on LDA to extract the relationships between the assets, building and evaluating the Bayesian network, and performing analysis on the covariance matrices. Both members contributed to the report.

8 Access to our Code

All of our code can be accessed from the following public Github repository: `https://github.com/fanpu/10-708-project`.

References

[1] URL https://www.alacra.com/alacra/help/barra_handbook_US.pdf.

[2] URL https://www.nltk.org/.

[3] Pyldavis. URL https://pyldavis.readthedocs.io/en/latest/readme.html.

[4] Gensim: Topic modelling for humans, May 2022. URL https://radimrehurek.com/gensim/.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3 (null):993–1022, mar 2003. ISSN 1532-4435.

[6] M. Hoffman, F. Bach, and D. Blei. Online learning for latent dirichlet allocation. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/71f6278d140af599e06ad9bf1ba03cb0-Paper.pdf.

[7] M. D. Hoffman and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo, 2011. URL https://arxiv.org/abs/1111.4246.

[8] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, 2017. URL https://arxiv.org/abs/1711.04305.

[9] W. Kenton. 10-k: Definition, what's included, instructions, and where to find it, Oct 2022. URL https://www.investopedia.com/terms/1/10-k.asp.

[10] O. Ledoit and M. N. Wolf. Honey, i shrunk the sample covariance matrix. *SSRN Electronic Journal*, 2003. doi: 10.2139/ssrn.433840.

[11] T.-H. Lee and X. Long. Copula-based multivariate garch model with uncorrelated dependent errors. *Journal of Econometrics*, 150(2):207–218, 2009. doi: 10.1016/j.jeconom.2008.12.008.

[12] H. M. Markowitz. *Portfolio selection: Efficient Diversification of Investment*. Blackwell, 1991.

[13] D. H. Oh and A. J. Patton. High-dimensional copula-based distributions with mixed frequency data. *Journal of Econometrics*, 193(2):349–366, 2016. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2016.04.011. URL https://www.sciencedirect.com/science/article/pii/S0304407616300707. The Econometric Analysis of Mixed Frequency Data Sampling.

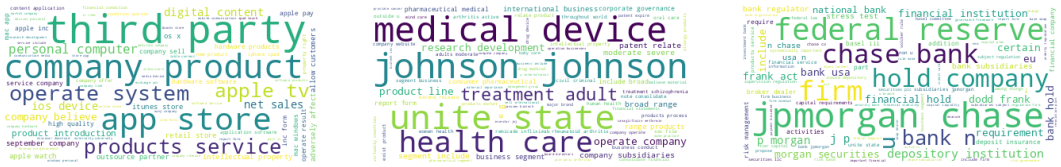
[14] Ranaroussi. Ranaroussi/yfinance: Download market data from yahoo! finance's api. URL https://github.com/ranaroussi/yfinance.

[15] C. Zhang, X. Pu, M. Cucuringu, and X. Dong. Graph-based methods for forecasting realized covariances. *SSRN Electronic Journal*, 2022. doi: 10.2139/ssrn.4274989.

Appendices

A Data Exploration

A.1 Word Clouds of 10-K reports



(a) Apple (AAPL) (b) Johnson and Johnson (JNJ) (c) JPMorgan Chase & Co. (JPM)

Figure 5: Word Cloud of the "Business Description" section of various stocks.

Figure A.1 shows the word clouds for the "Business Description" section of three stocks from different industries. This gives us some idea of the topics that we can expect from LDA.

B Data Cleaning

B.1 Preliminary Salient Terms

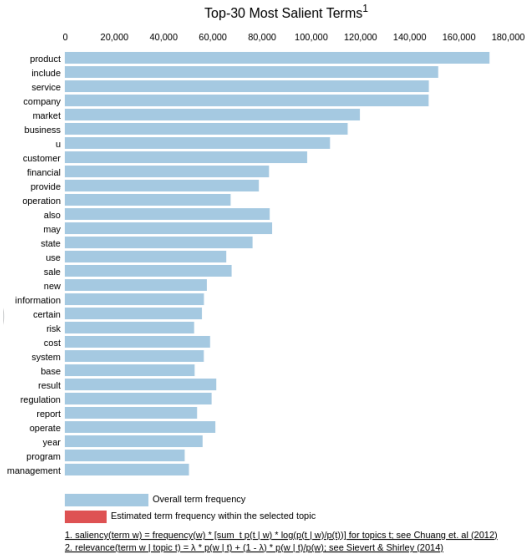


Figure 6: Top 30 Salient Terms

Figure 6 provides the list of the most salient terms originally found after only stripping generic stopwords and numerals. It can be seen that a lot of the top terms are very generic and has a high probability of appearing in the context of a 10-K document.

B.2 Additional Removed Stopwords

By human judgement, we decided that the following words are also stopwords in the context of the 10-K documents and are also removed:

PRODUCT COMPANY INCLUDE SERVICE MARKET BUSINESS U CUSTOMER FINANCIAL ALSO STATE MAY USE PROVIDE NEW YEAR RESULT COST CERTAIN

B.3 Topic Distribution and Distance Map

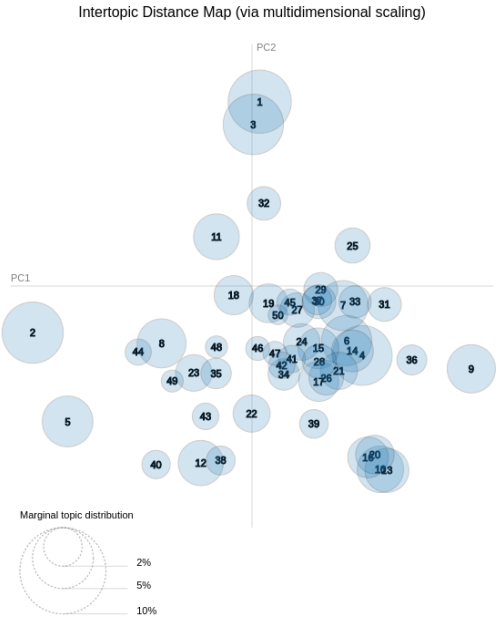


Figure 7: Topic distance and distribution

Figure 7 provides a visualization of the topic distribution and the distances between them.

B.4 Assets Used for Graphical Models

The full list of the 86 assets that we used for the graphical models are given below.

CTSH CI CMS ROST JNJ PFG PAYX AON FIS DXC ROK AMZN ISRG PNR GWW LNT CE NKE MCO AEP TPR UNH NVR XEL TTWO CMG GIS WST CAT XRAY ANSS QCOM INTC LIN MU GPC ITW EL DXCM COO PKG AMD ETN MHK IEX HON MKC MO BIO DIS NFLX RL CSX T WM TER CRL CL JKHY USB COF TFC GS AVB UDR ESS HIG PEAK SYY WAT BWA NEE PPL LLY INCY LVS COST XOM SLB DVN WMB PXD EOG PSA TSCO TAP

B.5 Visualization of Built Network

A visualization of the graph network recovered by LDA for the 86 stocks is given in Figure 8.

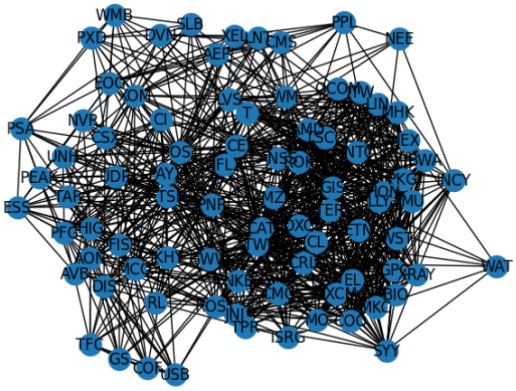


Figure 8: Graph showing the relationships between the 86 stocks found by LDA

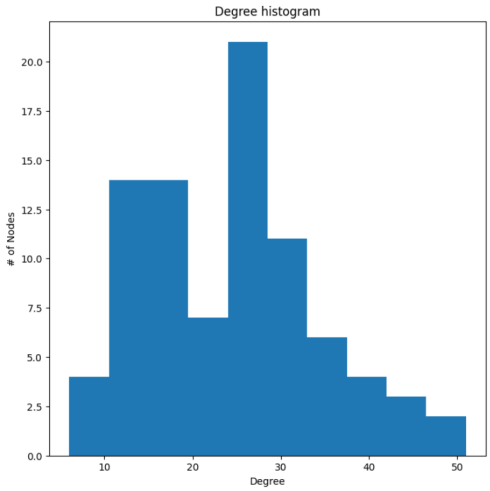


Figure 9: Histogram of Degrees in $\mathbf{A}$