STAT 5244 – Unsupervised Learning

Homework 3

Name: Chuyang Su UNI: cs4570

1 Graphical Models.

1.1 Data Processing.

The log return transformation was applied to the daily closing prices. The daily log return r_t for a stock price P_t was calculated as:

$$r_t = \ln(P_t) - \ln(P_{t-1})$$

This dataset of log returns, spanning 1,228 trading days, was used for all subsequent graphical model fitting.

1.1.1 Descriptive Statistics

The table below summarizes the descriptive statistics for the daily log returns.

Table 1: Descriptive Statistics of Daily Log Returns (Jan 2021 – Present)

	AAPL	AMZN	BAC	CVX	GOOGL	JNJ	JPM	KO	META	MSFT	NVDA	PFE	PG	WMT	XOM
Count	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00	1228.00
Mean ($\times 10^{-3}$)	0.63	0.27	0.53	0.64	1.02	0.33	0.81	0.38	0.65	0.66	**2.13**	-0.12	0.18	0.68	1.01
Std ($\times 10^{-2}$)	1.76	2.23	1.72	1.60	1.96	1.05	1.53	1.00	2.78	1.63	**3.29**	1.59	1.09	1.32	1.71
Min	-0.097	-0.151	-0.117	-0.086	-0.100	-0.079	-0.078	-0.072	**-0.306**	-0.080	-0.186	-0.070	-0.064	-0.121	-0.082
Max	0.143	0.127	0.081	0.085	0.097	0.060	0.109	0.046	0.209	0.097	**0.218**	0.103	0.042	0.091	0.062

The data clearly demonstrates the risk-return trade-off. The semiconductor stock `NVDA` shows the highest average daily return ($\sim 0.213\%$) but also the highest volatility (Standard Deviation: 3.29%) and largest maximum single-day return ($\sim 21.8\%$). Conversely, consumer staples stocks like `KO` (Coca-Cola) and `PG` (P&G) exhibit the lowest standard deviations ($\sim 1.0\%$), indicating high stability but lower returns. The largest single-day drop belongs to `META` (former FB) at -30.6%.

1.1.2 Time-Series Exploration

The cumulative returns plot (Figure 1) illustrates the differential performance across sectors over the analysis period.

1.1.3 Correlation Analysis

The correlation heatmap (Figure 2) reveals strong clustering of dependence among stocks within the same sector, which confirms the pervasive influence of systematic market risk.

Key Observations from the Heatmap:

- **Strong Correlation (0.6+):** High-tech stocks (`AAPL`, `MSFT`, `AMZN`, `GOOGL`, `NVDA`) are tightly coupled (e.g., `MSFT-AMZN` at 0.66, `MSFT-GOOGL` at 0.65). Financials (`JPM-BAC` at 0.82) and Energy stocks (`CVX-XOM` at 0.86) exhibit the highest correlations, reflecting their singular dependence on industry-specific factors (e.g., oil price, interest rates).
- **Weak/Low Correlation (0.0-0.3):** Healthcare stocks (`JNJ`, `PFE`) show low correlation with most other stocks (e.g., `JNJ` vs. Tech stocks often below 0.2), confirming their defensive, counter-cyclical nature.
- **Negative Correlation:** A notable weak negative correlation exists between the pharmaceutical stock `JNJ` and the high-growth technology stock `NVDA` (~ -0.09), suggesting an interesting divergence in their underlying risk drivers.

This preliminary analysis confirms the existence of strong, sector-specific dependencies, which the Graphical Lasso will aim to distill into a network of conditional dependencies.

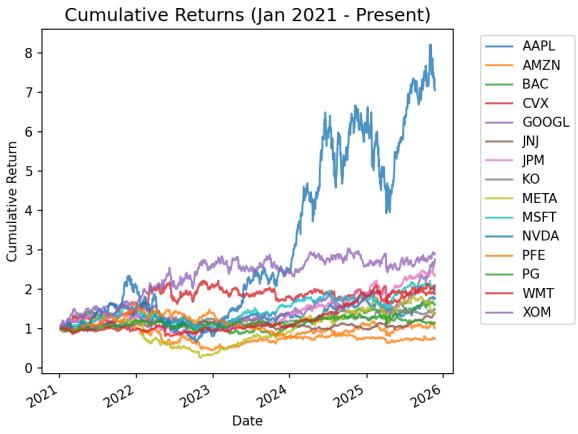

Figure 1: Cumulative Log Returns of Selected Stocks (Jan 2021 - Present)

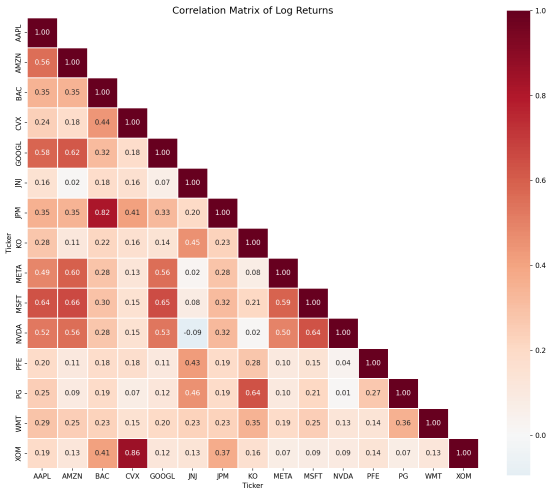

Figure 2: Correlation Heatmap of Daily Log Returns

1.2 Graphical Lasso.

Figure 3 shows the estimated precision matrices from both the Gaussian Graphical Lasso and the nonparanormal (rank-based) Graphical Lasso. The two heatmaps are almost identical, indicating that although individual stock returns are heavy-tailed, the dependence structure is well approximated by a Gaussian copula. Consequently, both methods recover essentially the same sparse conditional dependence network, suggesting that the underlying structure is stable, low-dimensional, and largely driven by sector-level factors.

The estimated graph highlights several strong conditional dependencies (e.g., AMZN–KO, JPM–GOOGL, WMT–BAC, NVDA–PFE) and a clear technology cluster consisting of AAPL, MSFT, GOOGL, AMZN, and META. NVIDIA does not join this cluster, likely due to its unusually strong and volatile performance during the sample period, which weakens

its partial correlations with the other technology stocks after conditioning on the full set of variables.

The regularization parameter α was selected via cross-validated Gaussian log-likelihood over a grid of 30 values spanning $\log_{10}(0.01)$ to $\log_{10}(0.8)$. This criterion is appropriate for unsupervised graphical models, as the validation likelihood measures generalization of the estimated precision matrix. The optimal values were

$$\alpha_{\text{Gaussian}} = 0.021287, \qquad \alpha_{\text{Nonparanormal}} = 0.013528.$$

For brevity, only the tuning curve for the Gaussian estimator is shown in Figure 4.

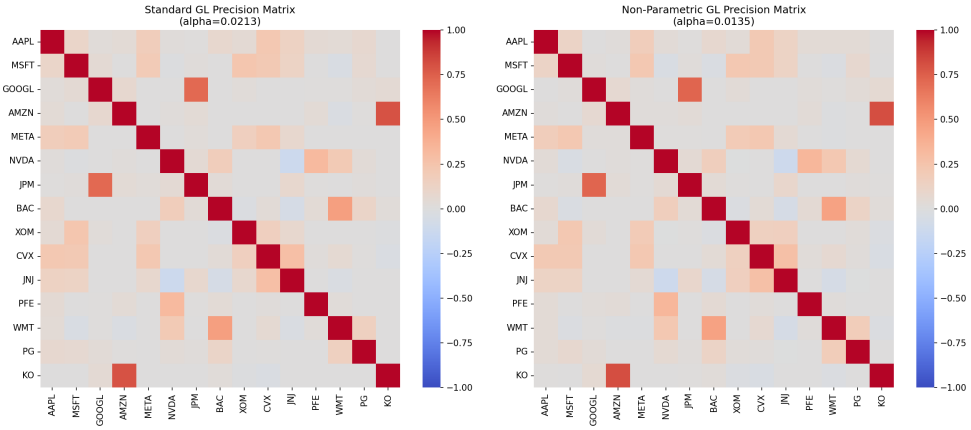

Figure 3: Gaphic Lasso Estimated Precision Matrices: Standard (Left) vs. Non-Parametric (Right)

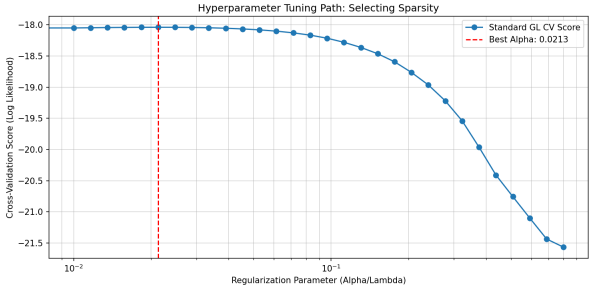

Figure 4: Cross-Validated Log-Likelihood Curve for Standard Graphical Lasso

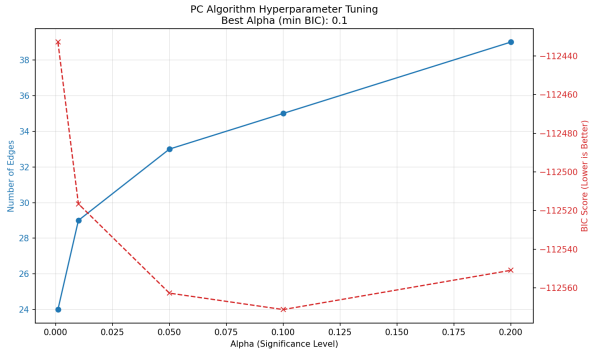

Figure 5: Best Alpha Selection for PC Algorithm via BIC Score

1.3 PC Algorithm.

To determine the optimal regularization level for the PC algorithm, we evaluated the Bayesian Information Criterion (BIC) across a range of significance thresholds

$$\alpha \in \{0.001,\ 0.01,\ 0.05,\ 0.1,\ 0.2\}.$$

The resulting BIC scores are shown in Figure 5. Although $\alpha = 0.05$ is often used as a conventional threshold, the BIC curve indicates that the model achieves its minimum score at $\alpha = 0.1$, implying that this level of sparsity provides the best balance between model fit and complexity.

Based on this criterion, we select $\alpha = 0.1$ and construct the final directed graph using the PC algorithm. The resulting structure is displayed in Figure 6, which represents the learned conditional independence relations and the corresponding Markov equivalence class under this optimal parameter choice.

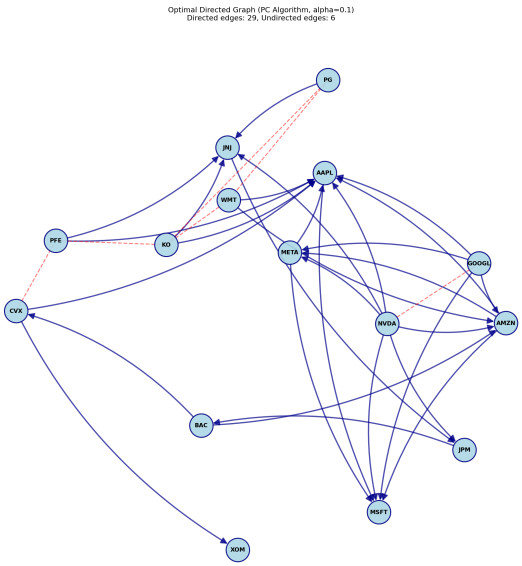

Figure 6: Best PC Algorithm Graph at $\alpha = 0.1$

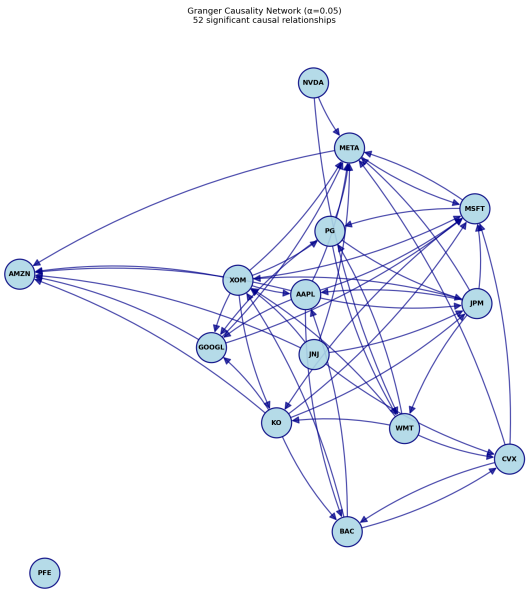

Figure 7: Granger Causal Graphical Model at $\alpha = 0.05$, Max Lag = 5

1.4 Comparison and Interpretation.

The learned structure from Figure 6 displays a clear pattern: technology stocks exhibit substantially richer and more directional dependency relationships compared to the other sectors. For example, MSFT appears as a child of all technology names, including AAPL, GOOGL, NVDA, AMZN and META, indicating that its conditional distribution depends structurally on multiple peers within the same sector. In contrast, NVDA and GOOGL share an undirected edge but act as parents to all other surrounding nodes, placing them in more central positions within the technology cluster. These findings are consistent with the results from part (b), in which the precision matrices also implied a tightly connected technology block.

Other sectors exhibit markedly more isolated behavior. For instance, in the energy sector, XOM is connected only to CVX, and once CVX is conditioned on, XOM becomes conditionally independent of all other stocks in the universe. This highlights both the internal coherence of the energy sector and its relative independence from the remaining market, which are consistent with common sense.

An interesting contrast emerges when comparing the PC graph with the precision matrices from part (b). Pairs such as KO–AMZN and JPM–GOOGL exhibit strong partial correlations under the Gaussian and nonparanormal graphical lasso, yet no direct edge appears between them in the PC graph. This difference reflects the distinct logics of the two models: graphical lasso identifies pairwise partial correlations, while the PC algorithm searches for a directed acyclic graph that satisfies a complete set of conditional independence relations. As a result, some associations are represented not by direct edges but by indirect paths—for example, KO connects to AMZN through AAPL or WMT. This is also consistent with economic intuition, as consumer staples and mega-cap technology firms often share indirect market linkages through broader macro or demand channels.

1.5 Granger Causal Graphical Model.

To evaluate the stability of the Granger causality model, we performed a hyperparameter sweep over significance levels $\alpha \in \{0.01, 0.05, 0.1\}$ and maximum lags $\{1, 3, 5, 7, 10\}$. The resulting edge counts and graph densities are summarized in Figure 8. Because interpretable causal networks should remain reasonably sparse, we selected $\alpha = 0.05$ and a maximum lag of 5.

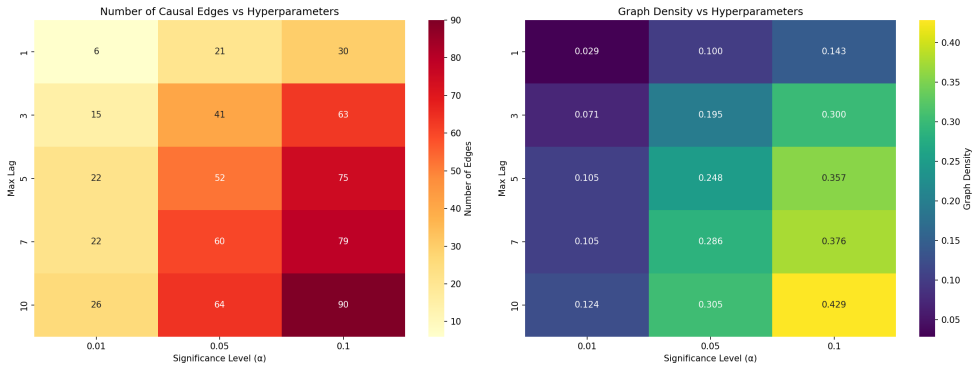

Figure 8: Granger Causality Model Tuning: Edge Counts (Left) and Graph Density (Right) Across Significance Levels and Maximum Lags

However, even under this moderate configuration, the resulting Granger graph still contains 52 significant edges (density = 0.248), which is far more dense than the sparse and interpretable structures obtained in parts (b) and (c). Figure 9 illustrates the adjacency matrix of significant Granger causal relationships, showing an unusually high concentration of edges across sectors. This density pattern suggests that the Granger framework is detecting a large number of spurious lead–lag relations rather than meaningful predictive structure.

The final inferred Granger network is shown in Figure 7. The resulting graph is visibly dense and lacks the sectoral organization and coherent clusters observed in the graphical lasso and PC algorithm results. Instead of revealing identifiable industry-level dependency patterns, the Granger graph displays widespread, cross-sector connections that contradict well-known market structure and are characteristic of noise-driven statistical artifacts in financial return data.

Together, these observations indicate that the Granger causality model does not fit the dataset well. Even after choosing a reasonably conservative hyperparameter setting, the method overfits the volatility and idiosyncratic noise inherent in daily stock returns, producing a dense and unstable network that fails to capture the stable structural dependencies recovered by the graphical lasso and PC models.

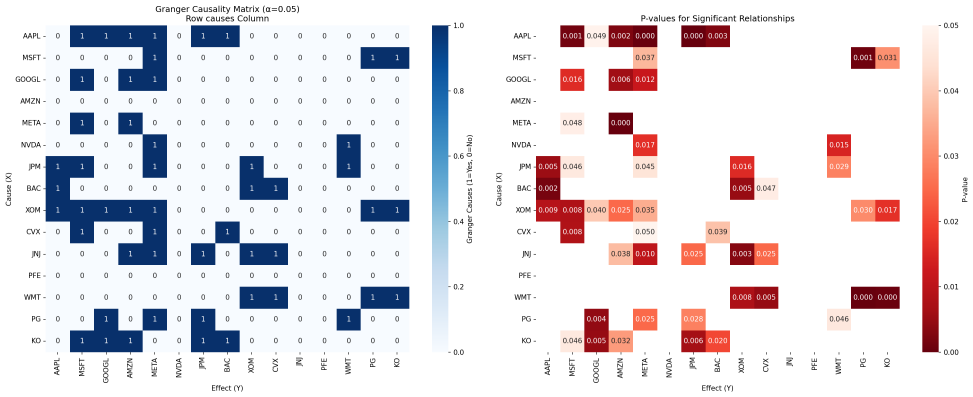

Figure 9: Granger Causality Adjacency Matrix at $\alpha = 0.05$, Max Lag $= 5$

2 Density estimation & Generative Models.

2.1 Kernel Density Estimation (KDE).

A grid search was performed on the bandwidth parameter h across three spaces (PCA-20/50 and original 64-dimensional space), with 5-fold cross-validation used to select the optimal value based on average log-likelihood, the result is shown in Figure 10. Since the CV score in the original space deteriorated significantly (-19606 vs -8081), confirming the curse of dimensionality, the PCA-20 space (optimal $h = 0.580$) was chosen for subsequent sampling. The pixel intensity distribution of generated samples closely matches the original data (mean 4.88 vs 4.95, standard deviation 6.02 vs 6.64), but spatial structure differs: generated samples exhibit only 23.91% sparsity compared to the original 48.93%, a 50% reduction, indicating KDE's tendency to produce denser pixel distributions. This results in some digits (e.g., 1, 0, 8) being clearly recognizable, though overall contrast remains insufficient(as shown in Figure 11).

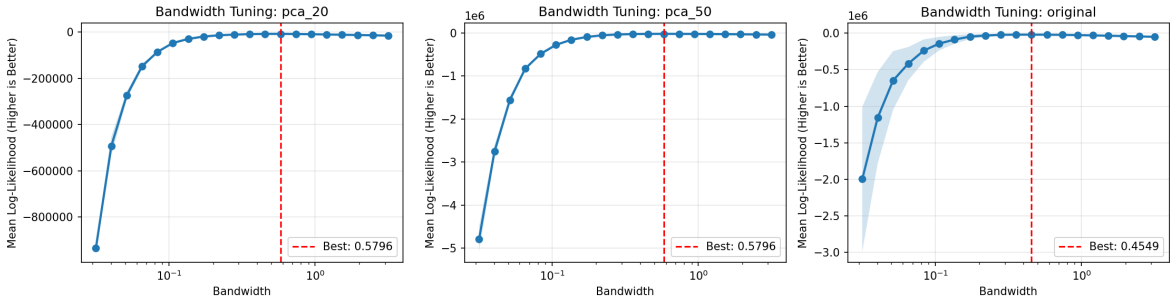

Figure 10: KDE Bandwidth Tuning via 5-Fold Cross-Validation in PCA-20, PCA-50, and Original Spaces

Figure 11: Generated Samples from KDE in PCA-20 Space