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 ($\times10^{-3}$)	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 ($\times10^{-2}$)	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.

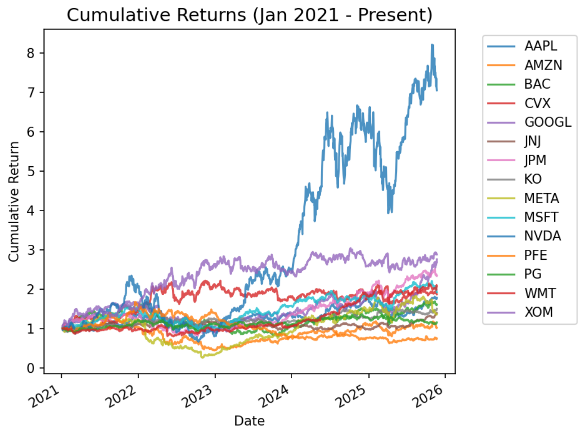

Figure 1: Cumulative Log Returns of Selected Stocks (Jan 2021 - Present)

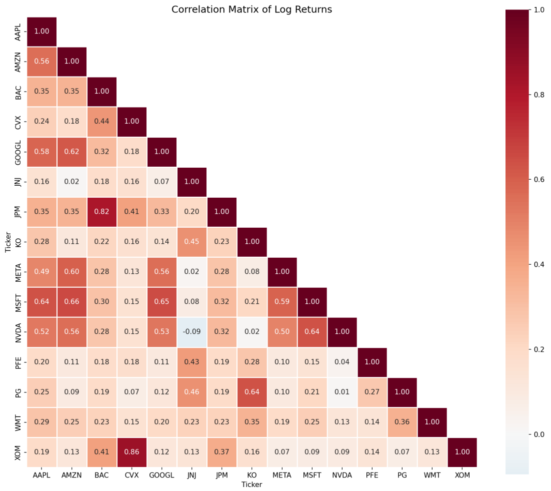

Figure 2: Correlation Heatmap of Daily Log Returns

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.

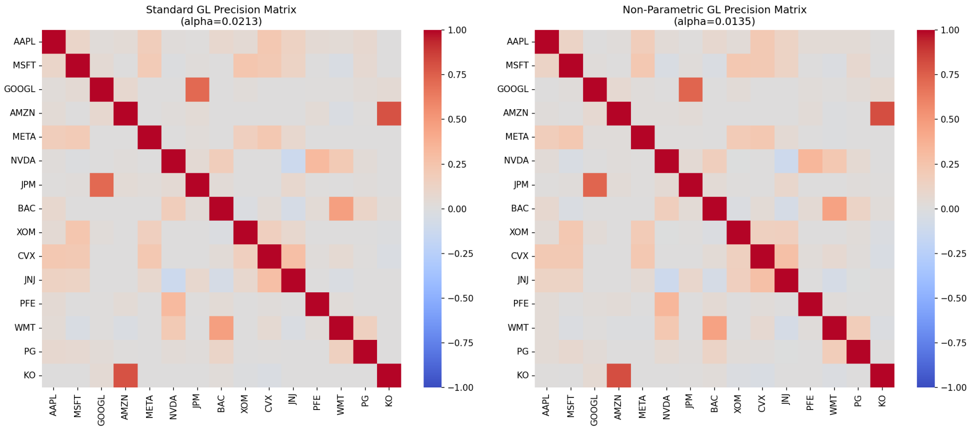

Figure 3: Gaphic Lasso Estimated Precision Matrices: Standard (Left) vs. Non-Parametric (Right)

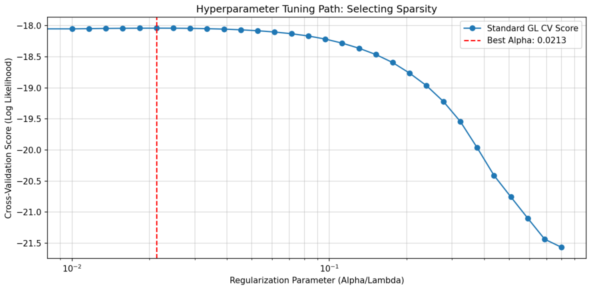

Figure 4: Cross-Validated Log-Likelihood Curve for Standard Graphical Lasso

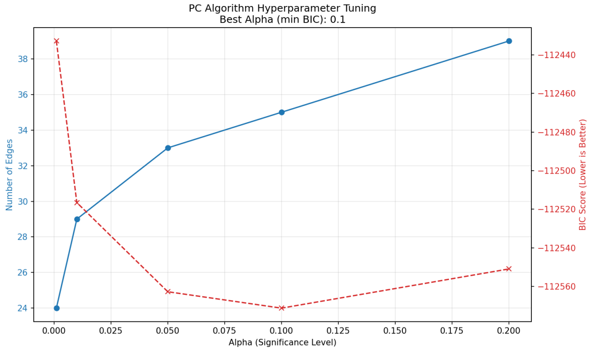

Figure 5: Best Alpha Selection for PC Algorithm via BIC Score

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.

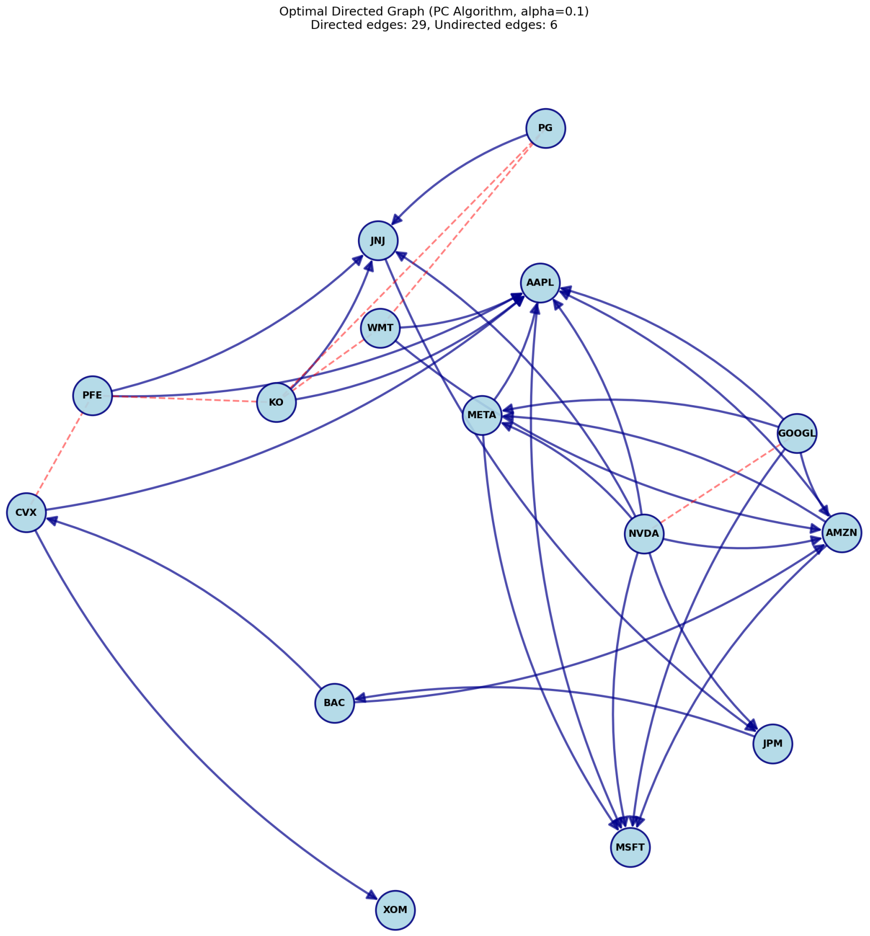
Optimal Directed Graph (PC Algorithm, alpha=0.1)
Directed edges: 29, Undirected edges: 6
PG
JNJ
AAPL
WMT
PFE
KO
META
GOOGL
CVX
NVDA
AMZN
BAC
JPM
MSFT
XOM

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