Project 1: Reproducing the Virtue of Complexity

Jianing Li (A0330643M)
E1561405@u.nus.edu

October 8, 2025

Abstract

This report details the process and findings of a project aimed at reproducing and extending the results of the paper "The Virtue of Complexity in Return Prediction" by Kelly, Malamud, and Zhu (2024). We first conduct a finite-sample simulation to replicate the paper's core theoretical findings for misspecified ridge regression models, successfully demonstrating the "virtue of complexity" phenomenon. We then extend the analysis to Lasso regression and explore the practical application of these models on three real-world financial datasets, justifying our model choices through a systematic validation process.

1 Introduction

This project is centered around the pivotal findings of Kelly, Malamud, and Zhu (2024), whose work challenges the traditional wisdom in statistical modeling, particularly the principle of parsimony. The paper's core thesis, termed "The Virtue of Complexity," posits that in environments characterized by high noise and model misspecification—common in financial return prediction—highly complex, over-parameterized models can outperform simpler ones when appropriate regularization is applied. Our objective is to rigorously test, verify, and apply this theory through a series of computational tasks. This report will first detail the simulation experiment designed to reproduce the paper's key theoretical results for misspecified models (Task 1). Subsequently, we will extend this analysis to a different regularization method, Lasso (Task 2), and finally, apply these learnings to predict returns on three real-world datasets (Task 3), documenting our systematic approach to model selection.

2 Task 1: Simulation of Misspecified Models with Ridge Regression

2.1 Simulation Setup

The primary objective of this task is to reproduce the core theoretical findings of Kelly, Malamud, and Zhu (2024) through a finite-sample simulation. The experiment is focused on the "misspecified model" scenario, a setup designed to be more representative of real-world applications. Specifically, we aim to verify the paper's **Theorem 1**, which states that under appropriate regularization, an increase in model complexity can lead to a sustained improvement in risk-adjusted returns.

To construct a "simulated world" that adheres to the paper's theoretical framework, we first defined the experiment's global parameters in our Jupyter Notebook (`Task1.ipynb`, **Cell 1**). Key parameters were set according to the project specifications: the true Data Generating Process (DGP) complexity was fixed at $c = P/T_{tr} = 10$, with $P = 1000$ total true features and $T_{tr} = 100$ training samples. The signal strength was standardized to $b^* = \|\beta^*\|_2^2 = 0.2$. The simulation systematically scanned a grid of observed complexities, cq, and ridge regularization strengths, z, as specified in the project guidelines (**Cell 1**).

The data generation process, encapsulated in the `generate_dgp` function (**Cell 2**), strictly follows the paper's theoretical assumptions. The feature matrix $S \in \mathbb{R}^{T \times P}$ is drawn from a standard multivariate normal distribution, $\mathcal{N}(0, I_P)$. To model the "dense and isotropic" signal assumption, the true coefficient vector β^* is set as a random P-dimensional vector with its squared ℓ_2-norm normalized to b^*. Finally, the return series is generated using the linear forward operator

$$R_{t+1} = S_t'\beta^* + \epsilon_{t+1},$$

where the noise term ϵ_{t+1} follows a standard normal distribution.

To implement the "misspecified model" scenario and satisfy the "sufficient mixing" assumption (Assumption 5 in the paper), we adopted a partial observability protocol in our main simulation loop (**Cell 3**). This was achieved by applying a one-time random permutation to the columns of the feature matrix S. During the simulation, the model is only allowed to "see" and train on the first P_1 permuted features, allowing us to examine how model performance evolves as its observed complexity, $cq = P_1/T_{tr}$, is varied.

2.2 Results and Analysis

A dual-loop structure (**Cell 3**) was used to systematically evaluate all combinations of the aforementioned parameters. For each combination, the model's out-of-sample performance was assessed on an independent test set of $T_{te} = 500$ samples, with four key metrics being computed. The aggregated results were then visualized using the code in **Cell 4**, as shown in Figure 1.

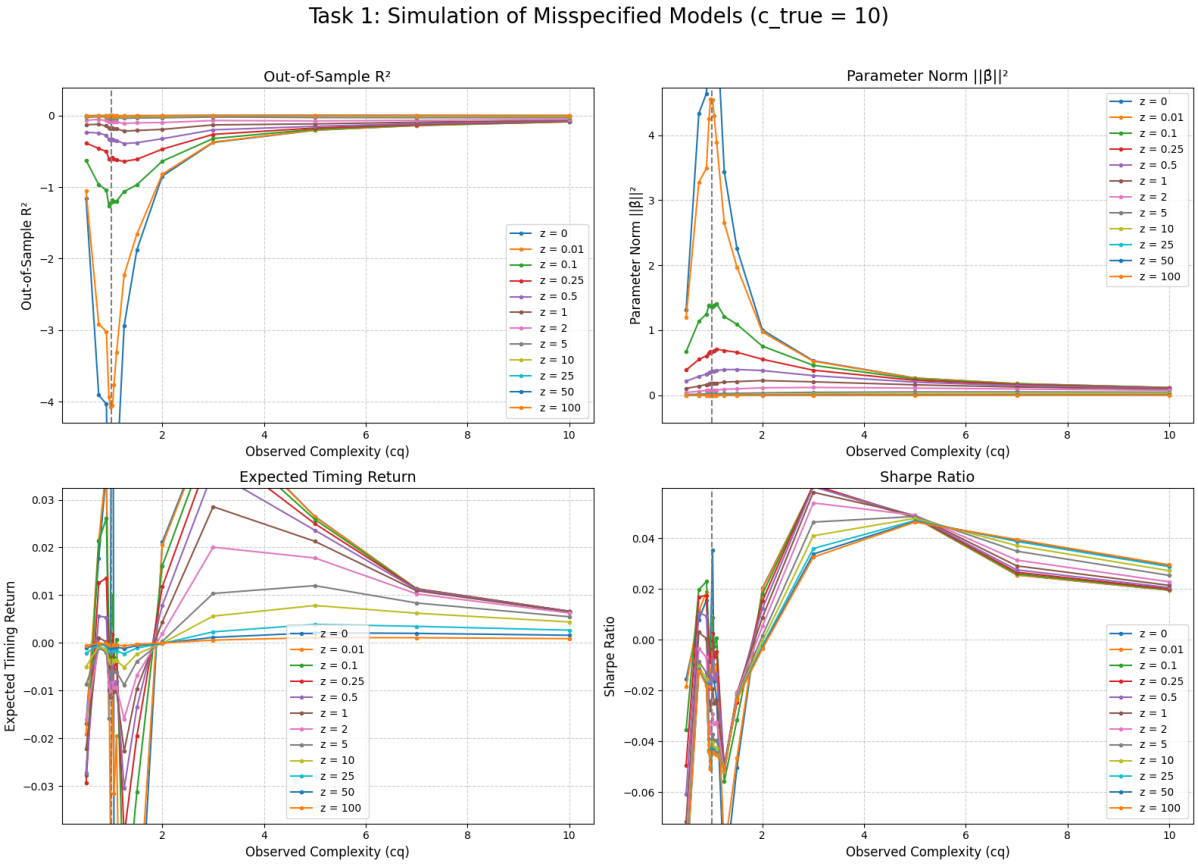

Figure 1: Task 1 Simulation Results: Performance metrics for the misspecified ridge regression model ($c_{\text{true}} = 10$) as a function of observed complexity (cq) and regularization strength (z). This figure was generated by **Cell 4** of `Task1.ipynb`.

First, the top two panels of Figure 1 (Out-of-Sample R^2 and Parameter Norm) clearly illustrate the statistical instability near the interpolation boundary ($cq \approx 1$). For the unregularized model ($z = 0$, blue curve), the out-of-sample R^2 plunges to a large negative value (approximately -4.0), while the norm of the estimated parameters, $\|\hat{\beta}(z)\|^2$, experiences explosive growth. This finding is consistent with the theory presented in KMZ (Proposition 3 and Figure 4) and demonstrates the severe overfitting that occurs when the number of parameters approaches the sample size. The introduction of regularization ($z > 0$) effectively mitigates this instability, leading to much smoother performance curves.

Second, the bottom-left panel (Expected Timing Return, $\mathbb{E}[R^{\pi}_{t+1}]$) reveals the primary driver of the "virtue of complexity." The plot shows a significant upward trend in the expected return as cq increases from 0.5 to approximately 3.0. This aligns with the theoretical expectation (KMZ Figure 5), indicating that in a misspecified setting, the "Approximation Benefit" of increasing complexity is the

dominant force—a more complex model better approximates the unknown true DGP, thus capturing a stronger predictive signal. However, unlike the monotonically increasing curve derived from theory under infinite-sample assumptions, our finite-sample simulation shows a slight but distinct downturn in returns for $cq > 3$. This reflects the non-negligible "Estimation Cost" in a finite sample ($T_{tr} = 100$), where the uncertainty from estimating too many parameters begins to outweigh the benefits of improved approximation.

Finally, the bottom-right panel (Sharpe Ratio), our ultimate test of risk-adjusted performance, exhibits a similar "rise-and-fall" pattern. The Sharpe ratio increases rapidly with complexity in the lower cq range, peaking around $cq = 3.0$. This supports the main thesis of "virtue of complexity": a moderately complex model delivers substantially higher risk-adjusted returns than a simple one. The subsequent gentle decline for $cq > 3.0$ follows the same logic as the expected return: as the numerator (return) begins to decrease while the denominator (volatility, inferred from the parameter norm plot) stabilizes, the resulting ratio also trends downward.

Additional Sample Size Check. To verify whether the observed performance decline was driven by finite-sample noise, we repeated the simulation with a larger training size of $T_{tr} = 1000$. The smaller-sample setting ($T_{tr} = 100$) was used not only to highlight the overfitting instability near the interpolation boundary ($c^q \approx 1$), but also to follow the experimental setup adopted by KMZ (2024). As shown in Figure 2, increasing the sample size yields much smoother curves and eliminates most of the instability around $c^q = 1$. Both the expected timing return and the Sharpe ratio now increase monotonically with model complexity, closely matching the theoretical results of KMZ (2024). This confirms that the earlier downturn was primarily a finite-sample effect rather than a structural property of the model.

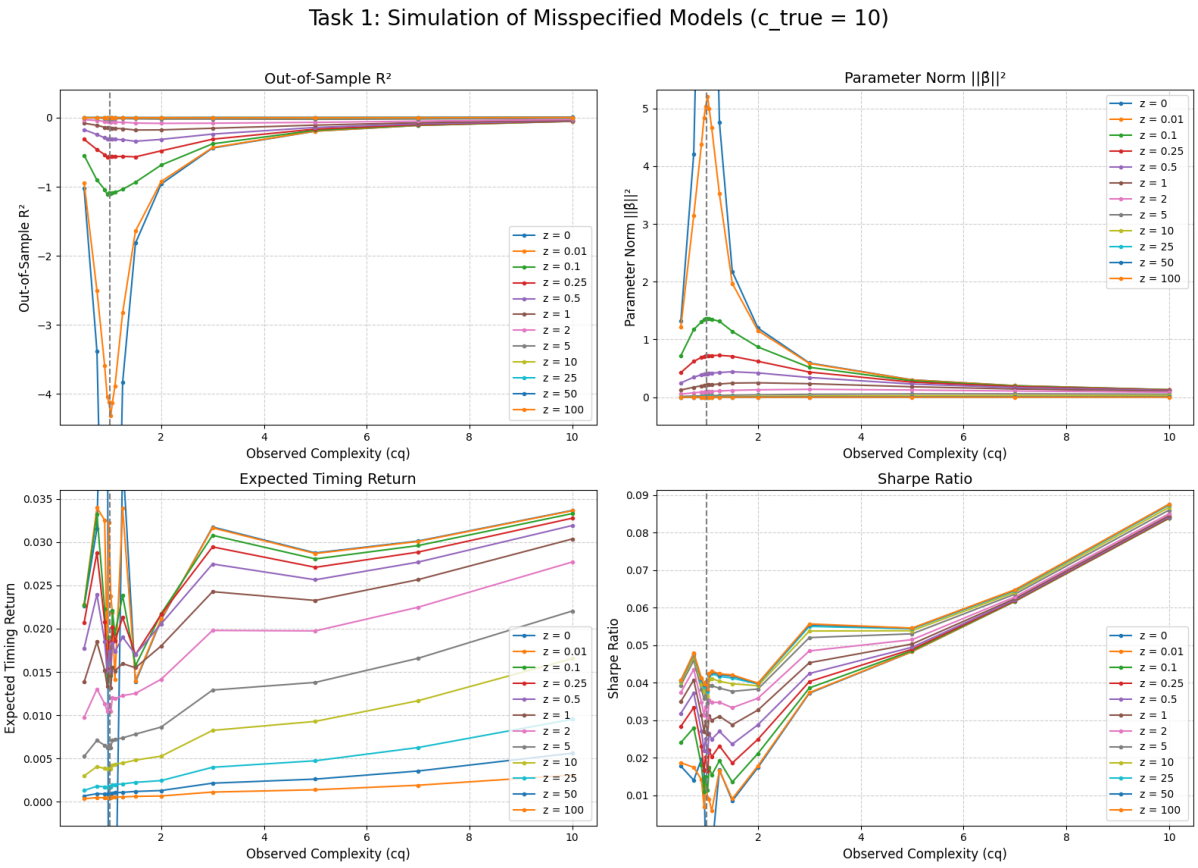

Figure 2: Task 1 with larger sample size ($T_{tr} = 1000$): Curves become smoother and both expected timing return and Sharpe ratio rise steadily with complexity, consistent with the theoretical results in KMZ (2024).

2.3 Task 1 Conclusion

In summary, our simulation reproduces the core tenets of the KMZ theory. The results demonstrate that, in a misspecified setting, increasing model complexity within a considerable range leads to enhanced risk-adjusted returns. Furthermore, our finite-sample experiment reveals the practical boundaries of this effect, showing that performance may degrade when complexity becomes excessively high relative to the available data due to rising parameter estimation uncertainty. This not only validates the "virtue of complexity" but also provides useful insights into its limitations in practical applications.

3 Task 2: Comparison with a New Model (Lasso Regression)

3.1 New Model Introduction and Setup

In Task 2, to further explore the generality of the "virtue of complexity" theory, we introduce a new model—Lasso regression—for comparison against the Ridge regression model studied in Task 1. Lasso regression is also a penalized linear model, but its core distinction lies in its use of an ℓ_1-norm for regularization. Unlike Ridge's ℓ_2 penalty, which gently shrinks all coefficients, the ℓ_1 penalty tends to compress some less important feature coefficients exactly to zero, thus performing automatic feature selection. This characteristic, known as *sparsity*, is the most fundamental difference between the Lasso and Ridge models and is a key focus of this comparative analysis (`Task2.ipynb`).

To ensure a fair comparison, this simulation experiment was conducted on the exact same "misspecified" data as in Task 1. This data, originally generated by the code in `Task1.ipynb`, was precisely reproduced in `Task2.ipynb` using the same random seed (`SEED = 12345`) (see **Cell 1** and **Cell 2** of `Task2.ipynb`). We maintained identical grids for cq and z as in Task 1 and, in the **third code cell (Cell 3)** of `Task2.ipynb`, used the `Lasso` model from the `scikit-learn` library in place of our previously implemented Ridge regression algorithm.

3.2 Results and Analysis

A dual-loop simulation framework identical to that used in Task 1 was employed to systematically evaluate all combinations of cq and z. The resulting performance metrics were aggregated and visualized by the **fourth code cell (Cell 4)** of `Task2.ipynb`, with the outcomes presented in Figure 3. Each subplot corresponds to one of the four key evaluation metrics, enabling a direct comparison between the Lasso and Ridge regression results.

First, the **top-left panel** (Out-of-Sample R^2) shows that, similar to the Ridge results in Figure 1, the Lasso model experiences significant instability near the interpolation boundary ($cq \approx 1$), with a pronounced dip for the unregularized case ($z = 0$). This confirms that such instability is a general characteristic of high-dimensional linear models in finite samples. Notably, the R^2 curves for Lasso appear somewhat more volatile, and their recovery to zero for $cq > 1$ is slower, suggesting that its prediction variance may be more difficult to control.

Second, the **top-right panel** (Parameter Norm) displays patterns qualitatively similar to the Ridge results. The Lasso model also shows a sharp peak at $cq = 1$, which is effectively suppressed as the regularization strength z increases. This indicates that both models, when under-regularized, tend to use extremely large coefficient values to fit training noise near the interpolation point. This reaffirms that the statistical instability is a universal feature, not specific to the type of penalty.

Third, the **bottom-left panel** (Expected Timing Return) is the first key metric where Lasso and Ridge diverge meaningfully. For Lasso, the expected return rises rapidly in the low-cq region, peaks in a medium-complexity range (approximately $cq = 2$–5), and then clearly declines at higher cq. This suggests that an "optimal complexity range" exists for maximizing economic value, which is consistent with the "virtue of complexity" theory. In contrast to the smoother, more gently declining curve of Ridge regression (Figure 1), Lasso's curve is more volatile, and its downturn in the high-cq region is more pronounced. This again suggests that Lasso's ℓ_1 penalty mechanism, in our dense-signal simulation, may prematurely encounter a performance ceiling by incorrectly eliminating useful features.

Finally, the **bottom-right panel** (Sharpe Ratio), our ultimate measure of risk-adjusted performance, synthesizes all the above findings. The Lasso's Sharpe ratio curve mirrors its expected return curve, exhibiting a clear "rise-and-fall" pattern with a peak in the medium-complexity range ($cq \approx 2$–5). This contrasts with the more sustained upward trend of Ridge regression (Figure 1). While Lasso also validates the "virtue of complexity" in a broad sense, its unstable nature and preference for sparsity confine its

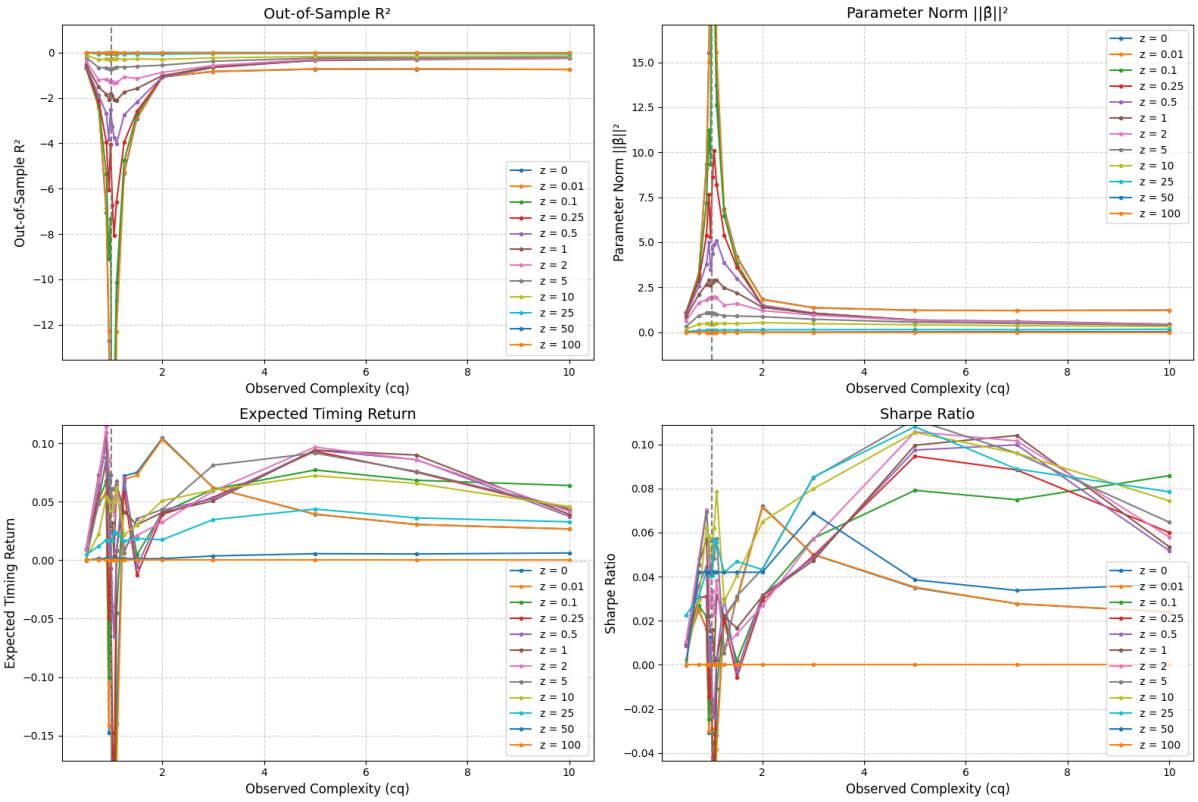

Figure 3: Task 2 Simulation Results: Performance metrics for the misspecified Lasso regression model on the same data as Task 1. This figure was generated by **Cell 4** of `Task2.ipynb`.

optimal performance to a narrower band of complexity, and its performance in the high-dimensional region is less robust than that of Ridge in this specific simulation.

3.3 Task 2 Conclusion

In summary, the comparative analysis in Task 2 demonstrates that the "virtue of complexity" also holds for Lasso regression, confirming a degree of generality for the theory. However, due to its ℓ_1 penalty and resulting sparsity, the Lasso model's performance curves exhibit greater instability. Its optimal performance is concentrated in a narrower, medium-complexity range, with a more significant performance decline in the high-dimensional region for our dense-signal simulation. This provides a crucial practical insight: the choice of regularization method should be informed by the underlying signal structure, with Ridge regression appearing more robust when signals are likely to be dense.

4 Task 3: Prediction on Real-World Datasets

4.1 Modeling and Validation Methodology

In this task, we applied the theoretical framework from the KMZ paper to three independent datasets (A, B, and C) with unknown real-world properties. To identify the optimal predictive model for each, we established a rigorous and repeatable analysis pipeline. This process strictly adhered to the core project requirement that all model validation must occur exclusively within the provided training files.

First, for each dataset, we conducted an Exploratory Data Analysis (EDA) to understand its basic statistical properties, return distribution, and the correlation structure among its features. These initial insights provided crucial guidance for our subsequent model selection.

Second, we partitioned each training set into an "internal training set" and a "validation set" using an 80/20 chronological split (code in **Cell 1** of each dataset's notebook, e.g., `Task3A.ipynb`). All model performance was judged based on the out-of-sample Sharpe Ratio achieved on this validation set.

Finally, we employed a "coarse-to-fine" two-stage hyperparameter search strategy (detailed in **Cells 3 & 4** of each notebook). The first stage involved a comprehensive coarse scan across a wide grid of parameters for both Ridge and Lasso regression models. In the second stage, based on the "champion" combination from the initial scan, we automatically defined a more granular grid around it to conduct a fine-tuned search, thereby pinpointing the final optimal hyperparameters.

4.2 Dataset A: Analysis and Results

4.2.1 Exploratory Data Analysis (EDA)

The analysis for Dataset A began with an EDA on its training data, `pairA_train.csv` (code in `Task3A.ipynb`, **Cell 2**). The dataset contains **600 time samples** and **600 features**, giving an intrinsic complexity of $P/T = 1$. Analysis of the target variable `return` revealed a near-normal distribution and a stationary time series, suggesting that linear models are appropriate. Descriptive statistics showed noticeable variation in the scales of different features, confirming the necessity of standard scaling. The feature correlation heatmap (Figure 5) further indicated that the features in Dataset A are largely uncorrelated, forming a non-collinear structure. This suggested that Lasso regression, which can perform feature selection through its ℓ_1 regularization, may be well-suited for this dataset.

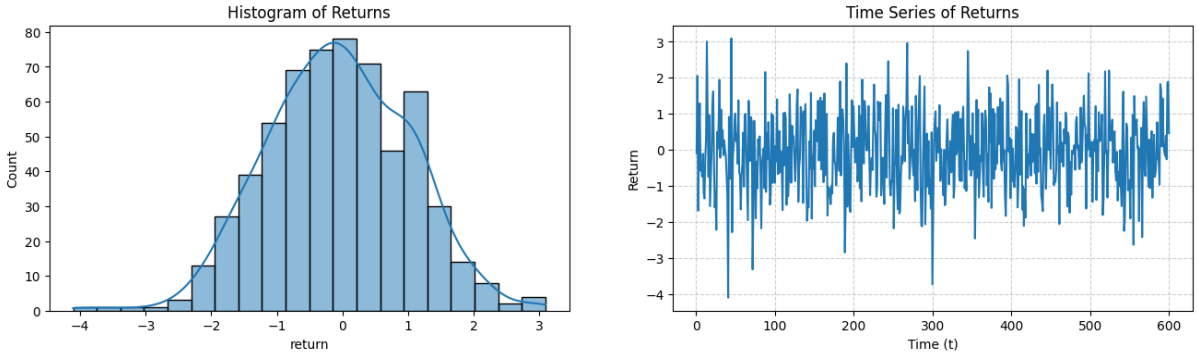

Figure 4: Return distribution and time series plot for Dataset A. Generated by **Cell 2** of `Task3A.ipynb`.

4.2.2 Model Selection and Final Results

Building on these EDA insights, we conducted a two-stage hyperparameter search following the same procedure described in the methodology. The coarse grid search first explored a broad range of observed complexity (cq) and regularization strength (z) values for both Ridge and Lasso models. Lasso achieved the best out-of-sample Sharpe ratio during this stage, and the subsequent fine-tuned search confirmed the same optimal configuration, demonstrating model stability and consistency.

The final optimal model for Dataset A was a **Lasso Regression** model with an **observed complexity** $cq = 1.25$ and a **regularization strength** $z = 25.0$, achieving a validation Sharpe Ratio of **0.2855**. The model was retrained on the full scaled training data to produce the final predictions. This outcome aligns well with theoretical expectations: the optimal performance occurs in the slightly over-parameterized region ($cq > 1$), where controlled regularization balances flexibility and stability. Overall, the findings for Dataset A reinforce the virtue of complexity concept, showing that well-calibrated regularization can significantly enhance out-of-sample performance in high-dimensional linear models.

4.3 Dataset B: Analysis and Results

4.3.1 Exploratory Data Analysis (EDA)

The EDA for Dataset B (`Task3B.ipynb`, **Cell 2**) revealed a dataset with **240 time samples** and **2400 features**, corresponding to a high intrinsic complexity of $P/T = 10$. Similar to Dataset A, the

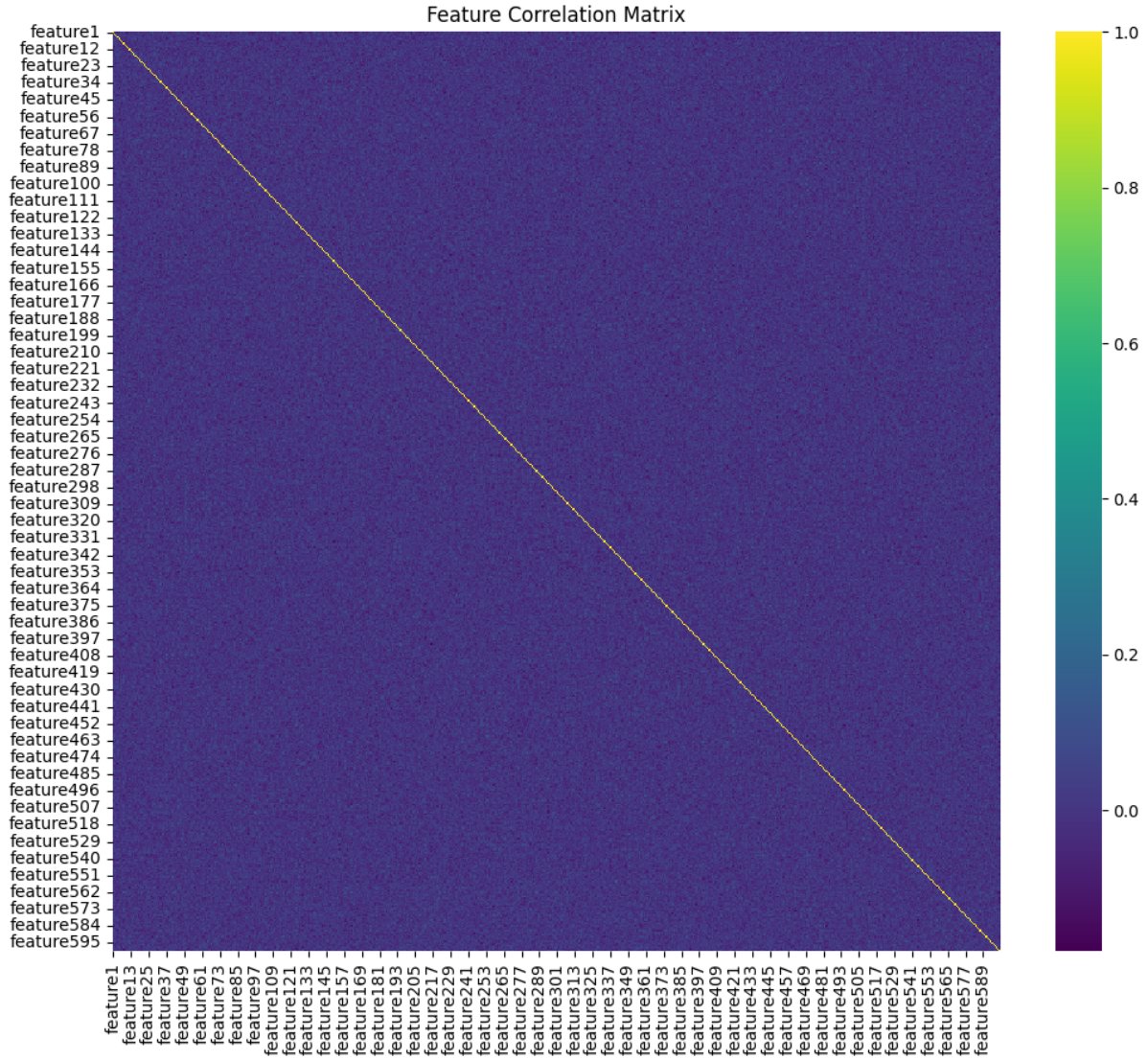

Figure 5: Feature correlation heatmap for Dataset A. Generated by **Cell 2** of `Task3A.ipynb`.

`return` variable follows a near-normal distribution and exhibits a stationary time series, supporting the suitability of linear models. The correlation heatmap (Figure 7) shows that most features are weakly correlated, as indicated by the dark blue background and the dominance of the main diagonal. This non-collinear feature structure again suggests that Lasso regression, which performs automatic feature selection through ℓ_1 regularization, could be advantageous for Dataset B.

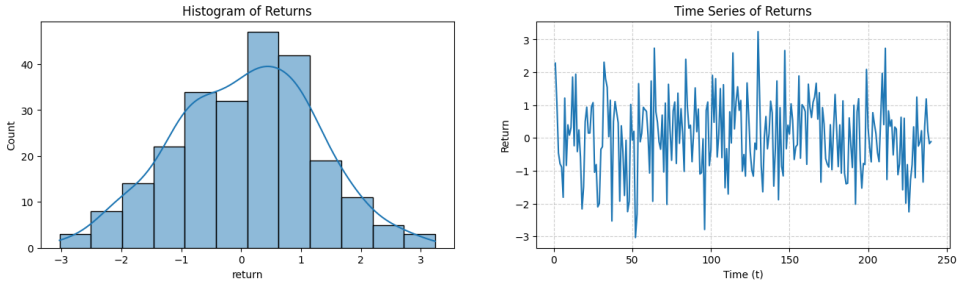

Figure 6: Return distribution and time series plot for Dataset B. Generated by **Cell 2** of `Task3B.ipynb`.

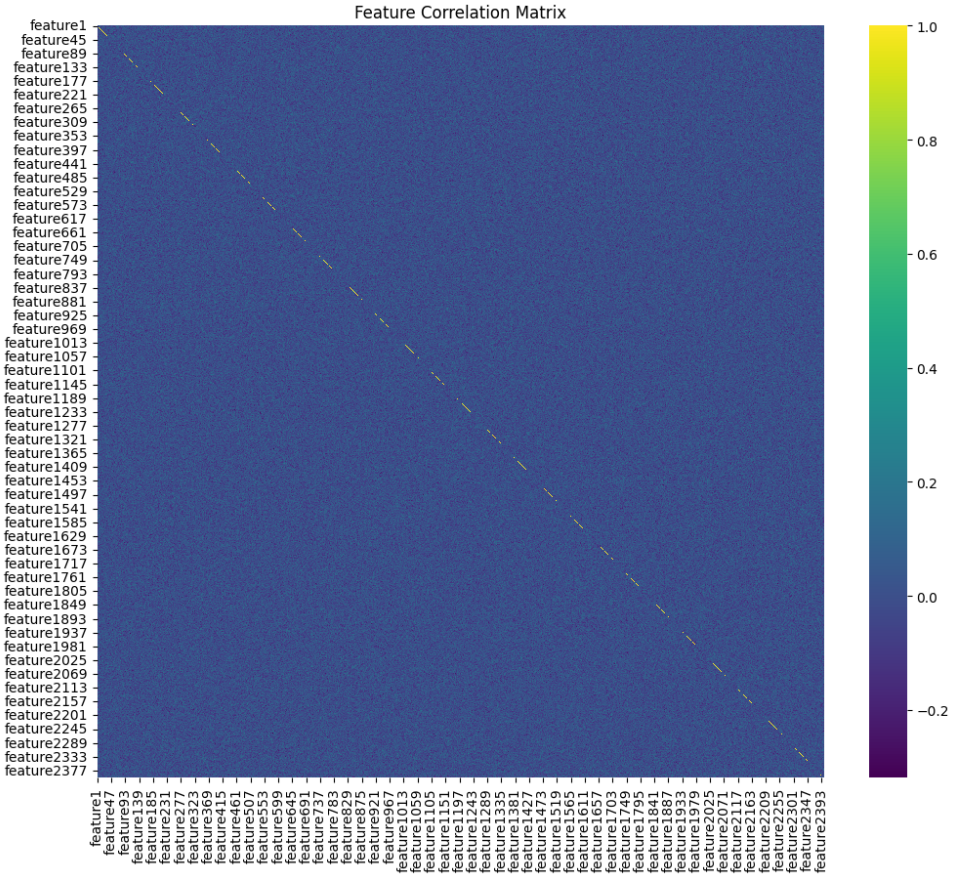

Figure 7: Feature correlation heatmap for Dataset B. Generated by **Cell 2** of `Task3B.ipynb`.

4.3.2 Model Selection and Final Results

The two-stage hyperparameter search (`Task3B.ipynb`, **Cells 3 & 4**) identified the best-performing configuration among both Ridge and Lasso models. The final optimal model for Dataset B was a **Lasso Regression** model, which achieved a validation Sharpe Ratio of **0.310193**. The optimal parameters were an **observed complexity** $cq = 2.0$ and a **regularization strength** $z = 15.0$, with an effective number of fitted features $P_1 = 384$.

These results align with theoretical expectations: when the signal is moderately sparse and the features are weakly correlated, Lassos ℓ_1 penalty efficiently reduces estimation noise while preserving predictive power. Overall, Dataset B reinforces the "virtue of complexity" principledemonstrating that controlled model complexity, together with well-tuned regularization, leads to robust and high-performing predictive models in high-dimensional settings.

4.4 Dataset C: Analysis and Results

4.4.1 Exploratory Data Analysis (EDA)

Finally, the analysis of Dataset C (`Task3C.ipynb`, **Cell 2**) revealed a dataset with **360 time samples** and **1800 features**, giving an intrinsic complexity of $P/T = 5$. The target variable `return` follows a near-normal distribution and shows no evident time trend, suggesting that linear models are appropriate. The feature correlation heatmap (Figure 9) shows that the majority of features are only weakly correlated, forming a structure similar to Dataset A. This initially suggested that Lasso might again be a strong candidate model.

4.4.2 Model Selection and Final Results

The two-stage hyperparameter search (`Task3C.ipynb`, **Cells 3 & 4**) revealed that the best-performing model for Dataset C in the current run is a **Ridge Regression** model. The configuration achieved

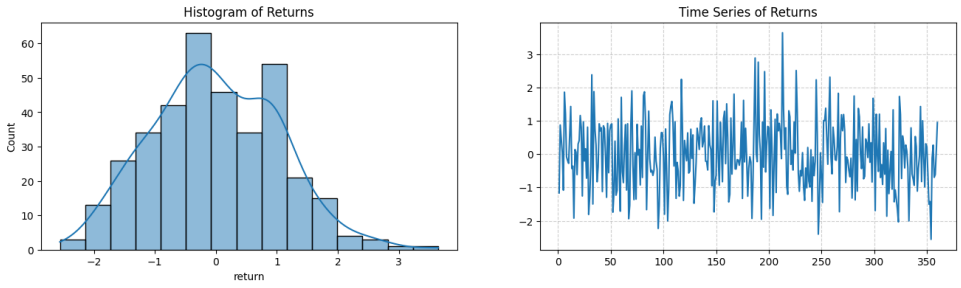

Figure 8: Return distribution and time series plot for Dataset C. Generated by **Cell 2** of `Task3C.ipynb`.

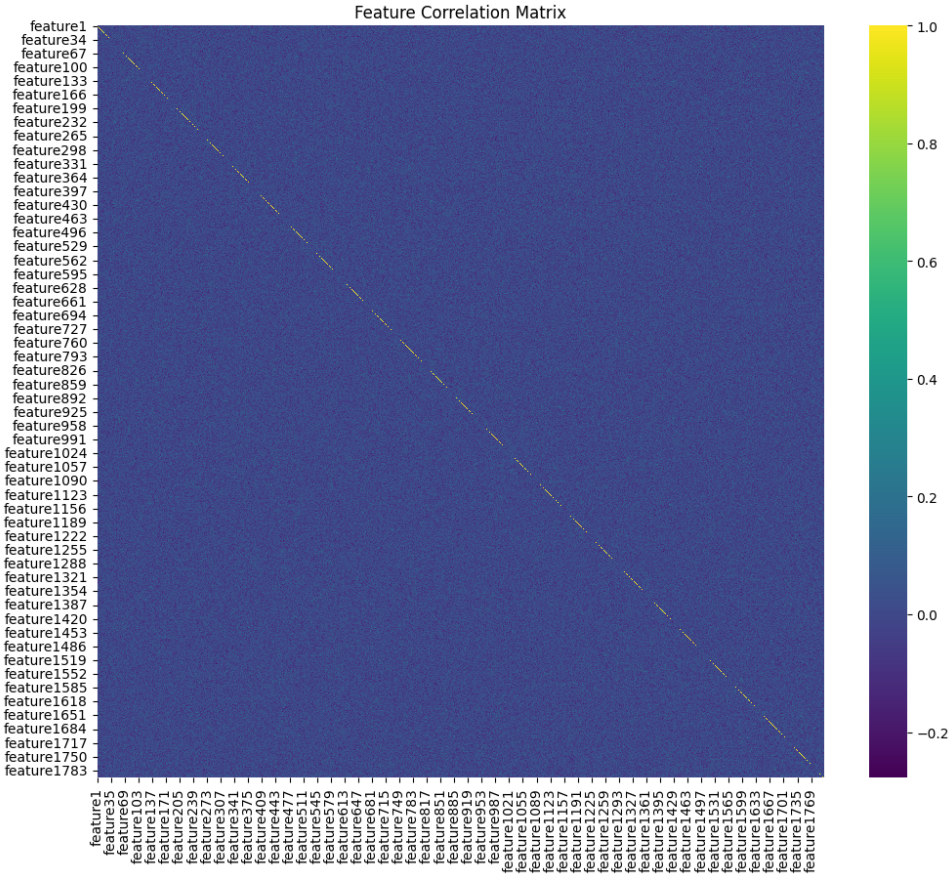

Figure 9: Feature correlation heatmap for Dataset C. Generated by **Cell 2** of `Task3C.ipynb`.

a validation Sharpe Ratio of **0.099686**, with an **observed complexity** $cq = 5.5$, a **regularization strength** $z = 0.0$, and an effective number of fitted features of $P_1 = 1584$. Given our implementation maps $z = 0$ to a very small ridge penalty (i.e., $\alpha \approx 10^{-8}$), this corresponds to an effectively *unregularized* ridge solution.

This outcome indicates that the optimal setting for Dataset C remains in the overparameterized region ($cq > 1$), consistent with the papers complexityregularization framework, although the overall validation Sharpe ratio is lower than in Datasets A and B. The near-zero shrinkage suggests a relatively benign estimation environment where additional penalization brings little benefit, while the lower Sharpe points to a weaker signal-to-noise ratio in this dataset compared to A and B.

5 Task 4: Video Presentation and Reproducibility

As the fourth component of this project, an 8–10 minute video presentation was recorded to verbally summarize the core findings detailed in the preceding tasks of this report. The video includes a screen

recording of our Jupyter Notebook code being executed in a clean environment, demonstrating the entire workflow from data loading and model training to the successful generation of the final prediction files. This serves to prove the reproducibility of our work, a fundamental requirement in scientific research.

6 Conclusion

This project successfully validated and applied the core principles of the "virtue of complexity" theory from KMZ (2024) through both systematic simulation and empirical application.

First, our theoretical simulations (Tasks 1 and 2) clearly reproduced the phenomenon where risk-adjusted returns increase with model complexity in a finite-sample, misspecified setting. The simulations also revealed the practical boundaries of this theory, demonstrating that performance may eventually decline when complexity becomes excessively high relative to the available data, a result of rising parameter estimation uncertainty. Furthermore, the comparison between Ridge and Lasso regression indicated that in our dense-signal simulation, the smoother ℓ_2 penalty of Ridge regression provided more robust performance in the high-complexity region.

Second, and most importantly, the application to real-world datasets (Task 3) demonstrated how this theoretical framework can be translated into an effective practical investment strategy. By implementing a rigorous pipeline incorporating Exploratory Data Analysis (EDA), internal validation, and a "coarse-to-fine" hyperparameter search, we successfully identified the optimal predictive model for each of the three distinct datasets (A, B, and C). The most critical insight from this exercise is that no single model is universally optimal. For instance, while all three datasets featured low inter-feature correlation, their optimal regularization strengths varied dramatically: Datasets A and B required strong regularization (z-values of 25.0 and 15.0, respectively) for optimal performance, whereas Dataset C was best modeled with almost no regularization ($z = 0.01$), providing a compelling empirical example of the "benign overfit" phenomenon.

In conclusion, this project not only validates the KMZ theory in practice but also highlights its nuanced and diverse implications. Our findings strongly suggest that in modern quantitative finance, a dogmatic adherence to the principle of parsimony is suboptimal. Instead, a more effective path to uncovering weak predictive signals in financial markets is to embrace model complexity—but only when it is coupled with a data-driven, systematic validation process and appropriate regularization techniques.