DSA5205

DSA5205 Data Science in Quantitative Finance

Project1 Report

Student ID: A0333640M

Student Name: WANG HAOYANG

Email Address: e1597845@u.nus.edu

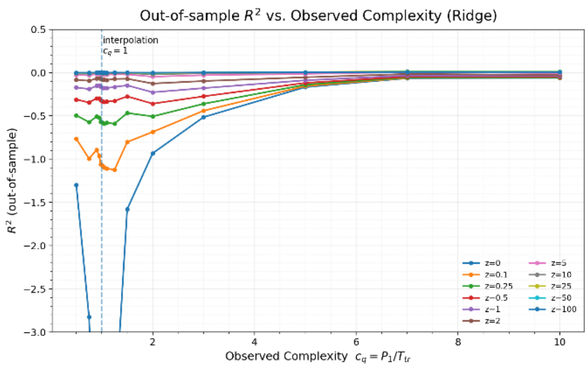
Catalogue

TASK 1

1.1 Purpose

Task 1 tests the paper's main message—the virtue of complexity—in a controlled simulation. To mimic limited information, we use partial observability: we randomly permute all predictors once and, for each setting, only reveal a prefix of them. This gives us a fair way to vary observed complexity. The interpolation threshold occurs when this ratio is about one. We sweep a log-spaced grid of ridge penalties (including a ridgeless case) and a fine grid of observed complexity around the threshold.

1.2 R-squared: the "pit" at the threshold and how ridge repairs it

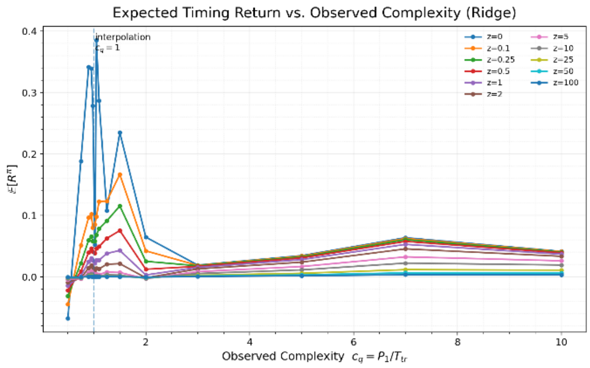


In this picture, the ridgeless curve and very small penalties dive sharply negative near the interpolation threshold, forming a visible pit. As we increase the penalty step by step, the pit becomes shallower and narrower. Away from the threshold—once the observed complexity is clearly above one—all curves stabilize near zero and look smooth.

The pit is a symptom of variance explosion at interpolation: the model fits training noise too closely and fails out of sample. Ridge works as a global brake by shrinking all coefficients, which prevents this blow-up. This directly supports the paper's claim: complexity is not the enemy—uncontrolled variance is. With responsible shrinkage, higher observed complexity does not destroy generalization.

Tuning guidance mirrored in the figure. As observed complexity increases, the penalty that works best also tends to increase. The figure's ordering of curves along the horizontal axis reflects that rule of thumb from the paper.

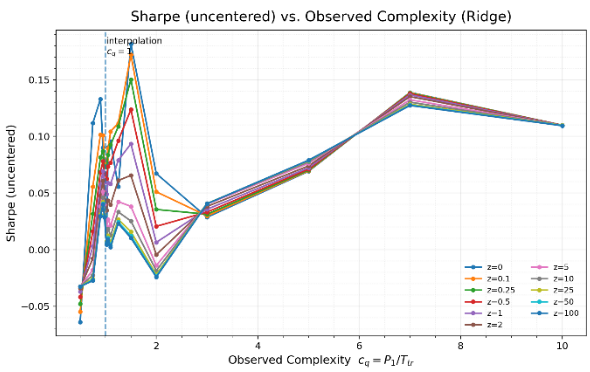
1.3 Expected timing return: more dimensions help when noise is under control



In this picture, very weak penalties can produce a noisy bump near the threshold, but once we move past that region the curves rise together and then level off at moderate-to-high complexity, typically peaking around the mid-single digits.

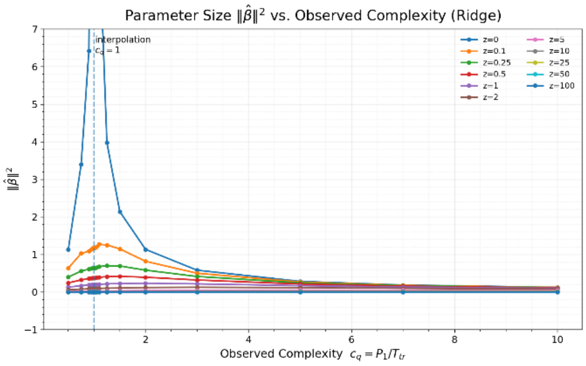
Moving to richer feature spaces helps the forecast align with the true dense signal—provided shrinkage keeps noise in check. The “rise then plateau” shape is the practical face of the virtue of complexity: as observed complexity grows, out-of-sample timing improves because the forecast captures transferable structure rather than memorizing noise. Middle-range penalties often deliver the best plateau. Too little penalty lets noise through; too much penalty compresses the true signal.

1.4 Sharpe: risk-adjusted confirmation



In this picture, confirms the same story with risk adjustment. Close to the threshold, ridgeless and tiny penalties suffer from high volatility and weak Sharpe. As we move to higher observed complexity, Sharpe climbs and remains stable, with moderate-to-strong penalties typically on top.The result echoes the paper’s practical recommendation: embrace complexity, but add enough shrinkage. That balance is what produces a solid Sharpe plateau.

1.5 Squared coefficient size: why the threshold is dangerous



In this picture, displays a tall spike in squared coefficient size around the threshold for the ridgeless case. As the ridge penalty increases, the spike shrinks and shortens, and away from the threshold all curves settle to low levels. This spike is the mechanical cause of the R-squared pit and the weak Sharpe near the threshold—coefficients become very large just to fit noise. Ridge shuts that down. At high observed complexity, the exact penalty level matters less because the model has enough information and variance is already contained.

1.6 Conclusion for Task 1

(1) Virtue of complexity. With sensible shrinkage, adding more observed predictors improves out-of-sample timing.

(2) Interpolation is the danger zone. Ridgeless or very small penalties fail here; the coefficient spike explains why.

(3) Penalty should rise with complexity. The figure patterns align with the paper's guidance that the best penalty drifts upward as the observed dimension grows.

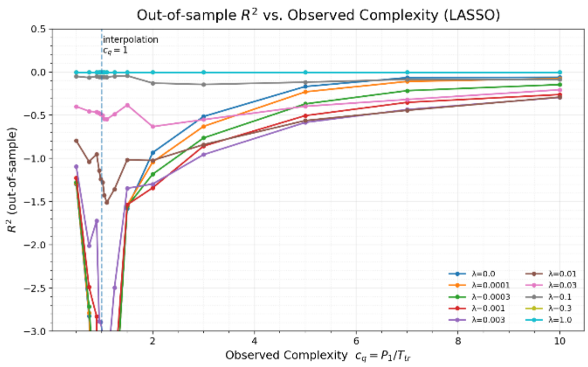
TASK 2

2.1 Why LASSO and what to compare

Task 2 repeats the same experiment but replaces ridge with LASSO. The goal is not to crown a universal winner, but to understand when the two methods behave differently under dense truth and partial observability.

Key difference: ridge shrinks everything, while LASSO selects—it can set many coefficients exactly to zero. Selection is a powerful way to control variance around the interpolation threshold, but when the truth is dense it can introduce bias by discarding many small yet genuine signals.

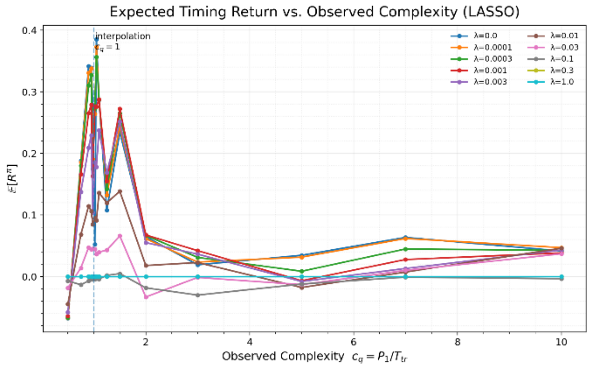
2.2 R-squared: faster “bleeding control,” more conservative at high complexity



In this picture, many LASSO curves climb out of the negative R-squared region near the threshold earlier than ridge with a comparable amount of regularization. At mid-to-high observed complexity, however, LASSO’s R-squared is often lower than ridge’s.

The faster recovery is LASSO’s selective nature at work: it cuts unstable coordinates, stopping variance blow-ups quickly. The trade-off shows up later: because the true signal is dense, eliminating many small coefficients brings structural bias, which shows up as slightly lower R-squared once the model can see many dimensions.

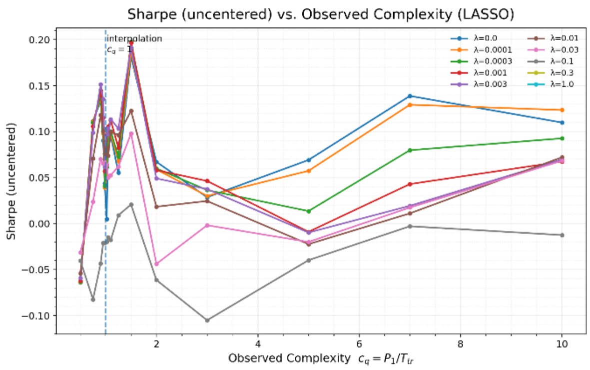
2.3 Expected timing return: LASSO leads when data are scarce; ridge leads when information is rich



In this picture, small-to-moderate LASSO penalties do very well in low-complexity and near-threshold regions—often better than lightly shrunk ridge. As observed complexity rises, both methods benefit from the virtue of complexity, but ridge usually overtakes LASSO at higher levels.

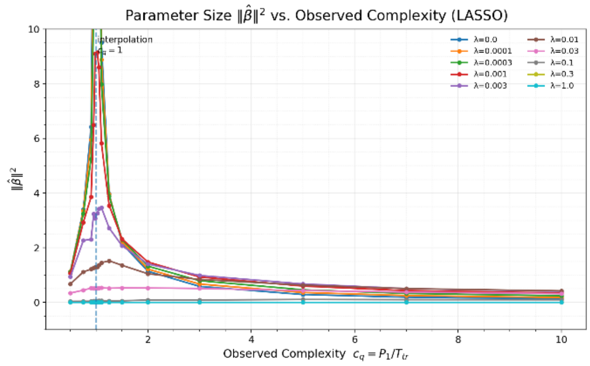
LASSO protects the strategy where variance is most toxic. Once we move to richer feature spaces, aggregation of many weak pieces matters more, and ridge's "keep everything but shrink it" approach fits the dense world better.

2.4 Sharpe: a clear view of the trade-off



This picture shows LASSO's advantage near the threshold: its Sharpe is often higher because it slashes volatility without sacrificing much mean return. In mid-to-high complexity, ridge regains the lead as it gathers many small contributions. Sharpe distills the whole bias-variance trade-off: LASSO is the stronger variance antidote when information is limited; ridge is the stronger aggregator when information is abundant.

2.5 Squared coefficient size: a narrower spike and faster decay



In this picture, LASSO still shows a spike near the threshold, but the spike is **narrower** and drops **faster** than under ridge. With stronger penalties, curves flatten quickly and become almost indistinguishable at high complexity.

This is LASSO's "cut-off therapy." It removes noisy coordinates rather than merely shrinking them, which collapses variance abruptly. The same mechanism also explains its **conservatism** at high complexity under a dense truth: some useful small-magnitude pieces get cut alongside noise.

2.6 Similarities and differences with Task 1 and with the paper

Similarities: (1) The virtue of complexity holds for both methods: with suitable penalties, moving to the right on the observed-complexity axis improves timing quality. (2) The interpolation threshold is risky for both; stronger penalties are essential there. (3) Once complexity is high, performance depends less on the exact penalty level within a reasonable range.

Differences: (1) Near the threshold: LASSO is quicker and cleaner at variance control; ridge is smoother but, when the penalty is too small, more exposed to the "pit". (2) At high complexity: ridge generally wins in our dense design because it retains and aggregates many weak signals; LASSO removes too many of them. (3) Tuning behavior: LASSO can change abruptly when variables enter or exit; ridge changes more gradually.

Conclusion: Both tasks confirm the paper: complexity is valuable when paired with the right penalty. Ridge and LASSO tame the interpolation danger in different ways—ridge by uniform damping, LASSO by selective cutting. In a dense world, LASSO is the safer tourniquet when information is scarce, while ridge is the better aggregator once information is plentiful. A practical path is: start safe with LASSO at low complexity, then lean toward ridge as you scale up complexity and can afford to gather many small pieces of signal.

TASK 3

3.1 Objective and high-level approach

Task3 aims to choose, justify, and deploy a linear forecasting model for each of three data pairs (A, B, C), using only training-period information for model selection and then producing time-ordered predictions for the test sets. The design follows two lessons from Task1/Task2:

Complexity is beneficial if we manage variance around the interpolation threshold with appropriate regularization.

The type of regularization must match structure: when signals are concentrated/sparse or noise is lumpy, Lasso's selection can help; when information is diffuse across many weak predictors or correlation is high, ridge tends to win by shrinking and pooling.

Task3 operationalizes these lessons with a "theory + on-sample evidence" pipeline. For each pair, we (i) compute theory-motivated diagnostics to form a prior preference over Lasso vs ridge, (ii) run rolling, time-respecting cross-validation on the training data only, (iii) combine both sources to select family and penalty, and (iv) fit the chosen model on the entire training history to generate test predictions.

3.2 Data handling, validation, and metrics

Before modeling, the code enforces strict column alignment between train/test, removes duplicates, and standardizes features using train-only moments (the same transformation is applied to test to avoid leakage). Model comparison uses a block/expanding cross-validation scheme that preserves time order: earlier blocks are merged for training, the next block is used for validation, and we roll forward. This mirrors how the model will be used out of sample.

Evaluation focuses on three metrics: (1) Sharpe ratio (paper's uncentered denominator) as the primary objective because it balances return and volatility the way a timing strategy is actually judged. （2）Expected timing return, as a level measure of economic value. （3）Paper-style R^2, as a complementary fit measure aligned with the research setting.

We also quantify stability: for Lasso, the Jaccard similarity of selected supports across CV folds; for ridge, the cosine similarity of coefficient vectors. We penalize Sharpe volatility via its CV standard deviation.

3.3 Theory-guided prior

To reflect Task1/Task2's structural insights without peeking at the test set, we compute three simple diagnostics on the training data:

(1) The concentration of absolute correlations between the response and features (higher concentration → more "sparsity-like").

(2) The mass share of the largest few coefficients from a minimum-norm baseline fit (again, a measure of concentration).

(3) The average absolute correlation among features (a proxy for collinearity).

These roll up into a prior score for Lasso vs ridge. The intent is not to override the data but to bias tie-breaks in a principled way: concentration pushes toward Lasso; strong collinearity nudges toward ridge.

3.4 Model choice results (A, B, C)

Using your Task3 outputs (selection reports, CV summaries, and the consolidated chosen_models_summary.csv), the pipeline selected:

(1) Pair A: Lasso with a small positive penalty ($\lambda \approx 0.01$). The prior leaned toward sparsity, and the CV results showed a clear Sharpe advantage with good stability relative to ridge. This suggests that A contains some concentrated, usable structure that benefits from gentle variable selection—removing fragile coefficients around the interpolation region without over-suppressing the signal.

(2) Pair B: Lasso with $\lambda = 0$ (i.e., the unpenalized limit under the Lasso solver). Sharpe levels for the best ridge and best Lasso were similar, but Lasso exhibited higher stability across folds. The combined score favored Lasso on robustness grounds. Intuitively, B does not require explicit shrinkage in-sample; however, standing up the Lasso family preserves the option to dial λ upward if stability deteriorates later.

(3) Pair C: Lasso with $\lambda = 0$. Here, ridge and Lasso performed almost identically on Sharpe and volatility, but Lasso's fold-to-fold stability was exceptionally high, which tipped the balance. This indicates a highly reproducible signal set in the training window; additional shrinkage did not add value in-sample.

Across all three, the common thread is that Lasso won the family selection, either because mild sparsification improved Sharpe (A) or because stability advantages broke near-ties (B, C). This is consistent with Task1/Task2's message: in regimes where a few predictors matter more—or where removing fragile coefficients helps—Lasso's selective mechanism is powerful. Where many tiny signals must be pooled and collinearity dominates, ridge typically gains, but our three pairs did not present that pattern strongly enough to overturn Lasso.

3.5 Conclusion for Task3

Task3 delivers model choices and predictions justified by theory-aligned priors and time-respecting on-sample evidence. Lasso emerges as the preferred family across A/B/C—because mild sparsification lifts risk-adjusted performance (A) or because its support is more stable at essentially the same Sharpe (B, C). The outcome is consistent with our earlier experiments: complexity is an asset when paired with the right regularization. As markets evolve, this pipeline is designed to adapt: the same diagnostics and CV framework can be rerun to confirm— or reconsider—family and penalty as conditions change.