DBA3803 Predictive Analytics in Business
Project 1

Group Members

Gao Jun	A0201822W
Carissa Ying Geok Teng	A0205190R
Celine Leo Ren Yi	A0173572B
Wu Jie Qi	A0203501A

Content Page

1. Overview

In portfolio optimization, diversification by investing in assets that move in different directions is important in reducing risk. With this in mind, some build portfolios with the minimum risk possible (MinVar). However, this tends to result in poor performance. Allocating various amounts of capital to each risky asset, aside from reducing risk, also affects the expected return of the portfolio. The main hurdle here is finding the best allocation of capital to each asset to maximise return. One common method is to distribute capital among all the assets equally, known as the equally-weighted (EW) portfolio. While EW performs well, it is not the best portfolio distribution.

This optimal distribution of weights can be found through modelling. By looking at past returns on our set of risky assets, we can use regression analyses to find the portfolio that would give the expected highest return. We can also compare the predicted return with that of EW to see if our prediction will perform better than the straight-forward EW portfolio.

Still, building a portfolio using different models comes with its own set of pros and cons, namely with regards to how complex the model should be. In this report, we will analyse the bias and variance tradeoff of model complexity using Lasso and Ridge built on 6, 24 and 48 industries respectively.

Using a random generator as illustrated in our code, the 6 and 24 industries that we have selected for our portfolio is as follows:

6: Mines, Whlsl, Gold, Banks, Ships, Paper
24: Mines, Whlsl, Gold, Banks, Ships, Paper, Drugs, FabPr, Smoke, Guns, Hshld, Beer, Fun, Trans, BusSv, Chems, Cnstr, RlEst, Txtls, Clths, BldMt, Oil, PerSv, Fin

2. Overfitting

There are many ways to build a regression model so as to estimate the optimal weights for a portfolio. Simple models tend to have fewer covariates while complex models have more. It is worth noting that while complex models have more covariates and hence fit well to training data, this does not necessarily translate to a better performance during testing. This phenomenon where the testing error is much larger than the training error is known as overfitting.

Overfitting is caused by the bias and variance tradeoff, in which a simple model has a high bias and low variance, whereas a complex model has a low bias but high variance. In other words, a highly complex model fits the training set too closely and is unable to generalise well to the test set, resulting in a large testing error. Under the general rule of thumb, for there to be no severe concern of overfitting, the number of observations ≥ 10 x covariates. This means that in a model where the number of observations are limited, having a large number of covariates will run the risk of overfitting.

$$\min_{\alpha, \vec{\beta}} f\left(\vec{y} - (\alpha + X\vec{\beta})\right) + \lambda_1 \|\vec{\beta}\|_1 \qquad \min_{\alpha, \vec{\beta}} f\left(\vec{y} - (\alpha + X\vec{\beta})\right) + \lambda_2 \|\vec{\beta}\|_2^2$$

Fig 1. Lasso definition

Fig 2. Ridge definition

The size of coefficients are closely tied to variance as a large coefficient means that a small change in one of the covariates will lead to a large change in the prediction. With many coefficients, this can lead to unpredictable results, causing overfitting. To mitigate this, a penalty(λ) is imposed on large coefficients. This is the concept of regularisation used in LASSO and Ridge regressions. The larger the λ, the higher the penalty on coefficients and the simpler the model. Meanwhile, the smaller the λ, the smaller the penalty on coefficients, giving a more complex model, which could in turn lead to overfitting.

Harkening back to the EW and MinVar portfolios, EW is the simplest model where $\lambda = \infty$ and MinVar is the most complex model where $\lambda = 0$. Hence, MinVar faces overfitting and future predicted returns become inaccurate when compared to the actual returns. To select a portfolio with optimal weights for each industry, we must strike a balance in model complexity. This can be done by adjusting λ to control the size of coefficients and by changing the number of covariates used in our model.

3. Lasso and Ridge in Portfolio Optimization

Using Lasso and Ridge regression, we created plots based on 63 observations for 6, 24 and 48 industries as the training set. Using 10-fold cross validation for Lasso and 63-fold cross validation for Ridge (RidgeCV only has "Leave-One-Out-CV"), we generated the mean-squared error (MSE) across different λ values. By comparing the MSEs, we will find the λ that results in the best model.

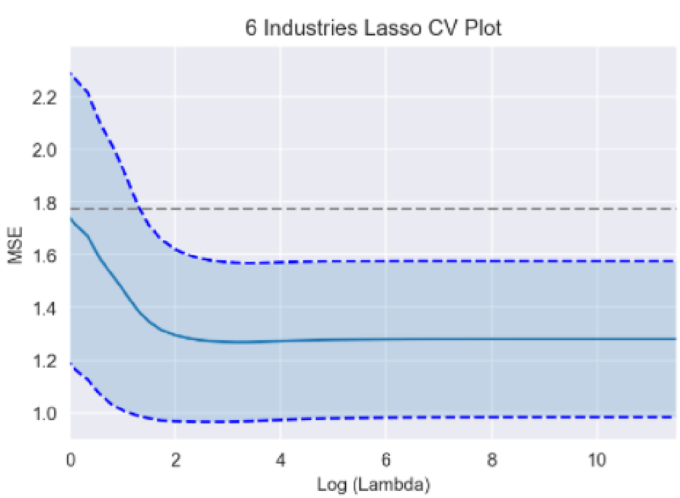

Fig 3. Lasso regression for 6 industries

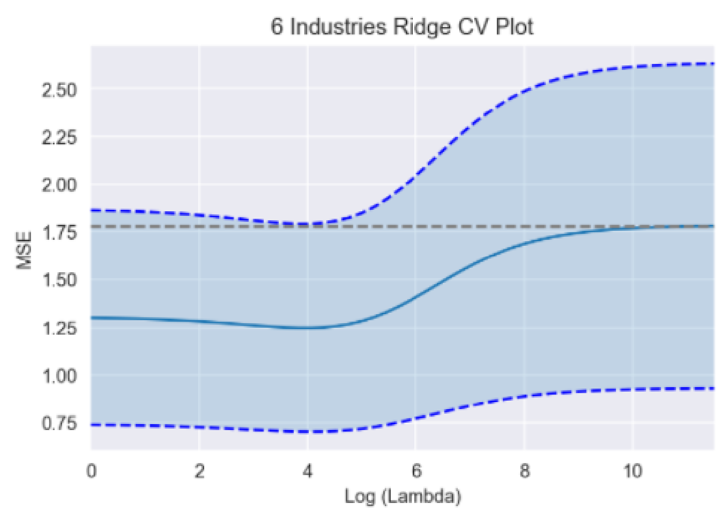

Fig 4. Ridge regression for 6 industries

As shown in the CV plots for 6 industries, the optimal λ is about 10^3 for LASSO and 10^4 for Ridge. As λ decreases beyond the optimal value, the penalty for large β decreases, increasing the degree of freedom of the coefficients. As a result, the model becomes more complex and overfits the training data. The model is hence unable to generalise well, giving a larger test error than expected.

However, in this case, the increase in complexity of the model does not lead to a significant increase in mean-squared error, as depicted by the relatively flat curve. This shows that there is no severe concern about overfitting when the portfolio only consists of 6 industries. This is because the rule of thumb is followed whereby 63 observations ≥ 10 x number of covariates. The model trained on the training set can be generalised to the test set to give optimal weights of each industry and relatively accurate predicted returns.

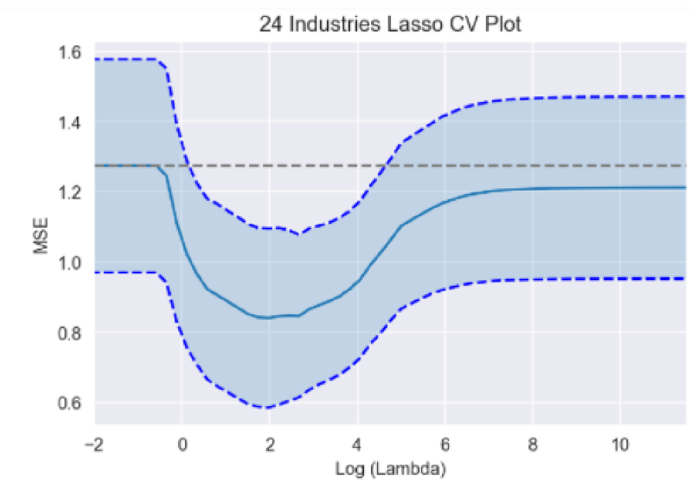

Fig 5. Lasso regression for 24 industries

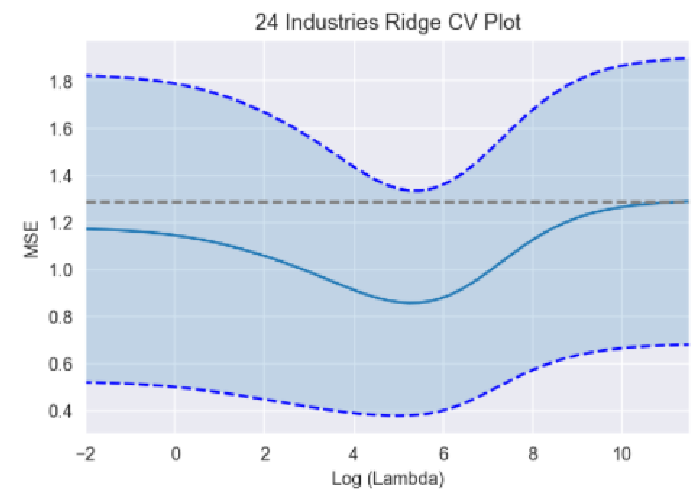

Fig 6. Ridge regression for 24 industries

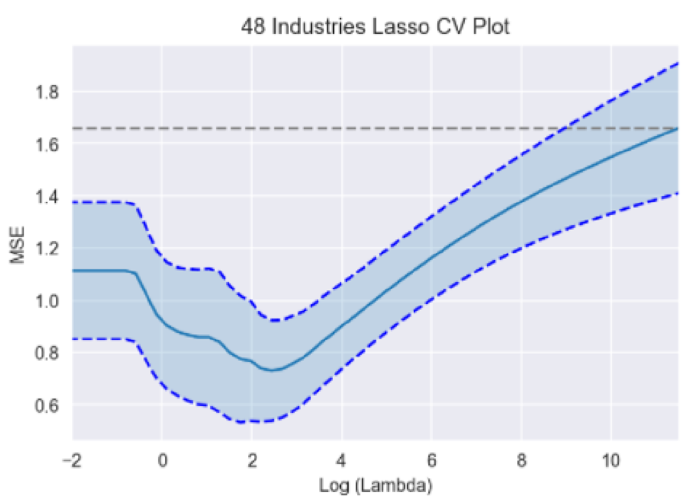

Fig 7. Lasso regression for 48 industries

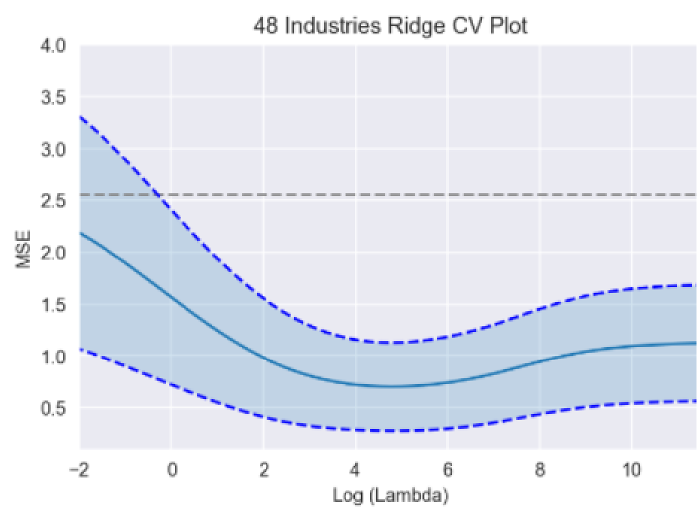

Fig 8. Ridge regression for 48 industries

From the CV plots for 24 and 48 industries, we see that a large λ gives a high prediction error in general. As λ increases, the penalty for large β increases, restricting the degree of freedom of coefficients. The coefficients are pulled towards zero, leading to a simpler model. However, in this case, the model becomes too simple, increasing its bias. As such, although the model has a low variance, it underfits the data and produces high prediction errors.

In these cases, we also observe that the effect of overfitting is stronger as depicted by the steeper curve with decreasing λ. For small λ, there is a greater degree of freedom for coefficients. The model becomes more complex, with potentially 24 or 48 covariates contributing significantly to the regression. This breaks the rule of thumb as the number of covariates x10 exceeds the number of observations (63). As a result, although the model is accurate in the training set, it is unable to give precise predictions in the test set. Hence, the model has low bias but high variance.

From our analysis, it is important to find λ such that the model neither overfits nor underfits. Applying this to portfolio optimization, it is important to find the right λ so that our model gives both accurate and precise predictions of the returns.

4. Wrong Alternative approaches

4.1 Using 42 observations instead of 63 as training

As mentioned, according to the rule of thumb, there is no severe concern about overfitting if the number of observations ≥ the number of covariates x 10. With fewer observations (42 instead of 63), there is a greater chance that the number of observations will fall below the number of covariates x 10. This will lead to overfitting, where a high variance results in inaccurate predictions as the model fits too closely with the training set and is unable to generalise to the test set. Hence, it is always better to have more observations so as to fulfil the general rule of thumb.

4.2 Using polynomial regression to create more features

In other regression models, it is possible to use polynomial regression by creating more covariates with second order or more terms. However, as discussed earlier, more covariates may lead to overfitting where the model is unable to provide accurate predictions in the test set caused by the high variance. In addition, for the portfolio optimization, it is given that it should be modelled by a linear regression as Y is defined as the demeaned return multiplied by the equal weighted matrix and X is the demeaned return multiplied by the modified identity matrix. Hence, it is unnecessary to create more features to increase the complexity of the model and model the portfolio using polynomial regression.

4.3 Using the sum of absolute deviations as a loss function because it is more robust

The sum of absolute values (LAD) minimises the absolute differences between the regression and the training data points, while OLS minimises the square difference. The performance is measured using variance, which is calculated as the sum of square differences between the predicted value and the test value. This is in line with how OLS is calculated. Hence, using OLS to minimise the square difference would help to achieve the minimum variance objective. On the other hand, LAD would place less weight on large deviations and would hence not minimise variance, giving a poorer performance.

4.4 Using PCR instead of Lasso and Ridge

PCR is a different regularisation method used to prevent overfitting and it typically regresses on only a subset of all principal components. While the accompanied low dimension structure could potentially generate some insights (especially in cases where the top principal components have a nice interpretation), it generally performs worse than both Lasso and Ridge in terms of prediction. This is because choosing principal components instead of choosing based on targets could result in the omission of important non-top principal components information. As such, broadly speaking, we would not recommend the use of PCR in place of Lasso and Ridge.

5. Implementing the right approach

Instead of using 63 observations as the training set, using 84 observations will likely provide a better model with more accurate predictions as the problem of overfitting is mitigated. Having a higher number of observations will reduce the likelihood of violating the general rule of thumb, thus, reducing the concern about overfitting.

Additionally, according to the learning curve, as the number of training points increase, the expected error of the test set will decrease and become closer to the expected error of the training set. This means that the performance variance decreases with more training data.

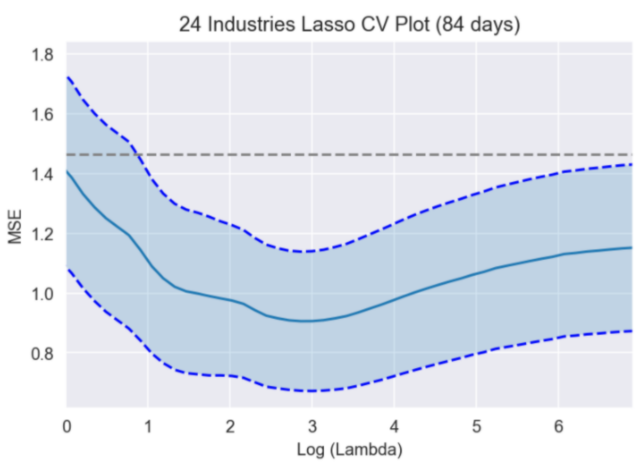

Fig 9. Lasso regression for 24 industries

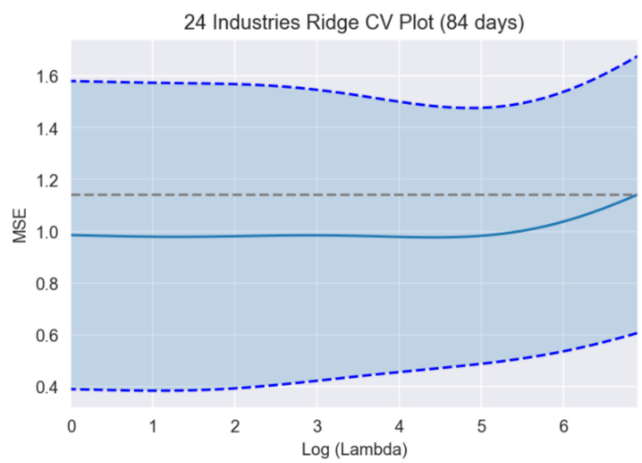

Fig 10. Ridge regression for 24 industries

As seen in the CV plots for 84 observations and 24 industries, as lambda decreases beyond the optimal point, the gradient of the curve is relatively flatter as compared to the CV plots for 63 observations and 24 industries. Although there is still overfitting, the concern of overfitting is reduced through the use of more observations. As such, despite the increase in model complexity, the trained model will still be able to better generalise to the test set with lower prediction error.

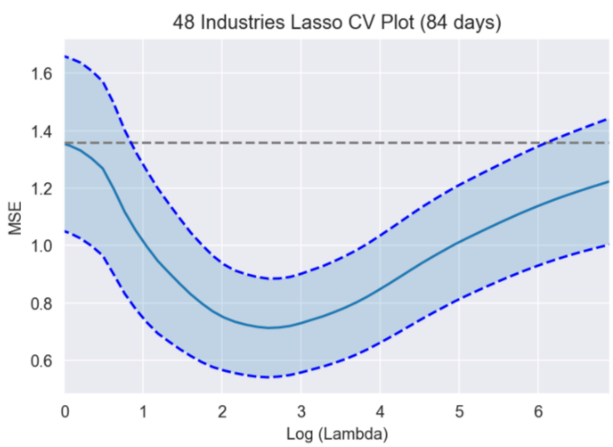

Fig 11. Lasso regression for 48 industries

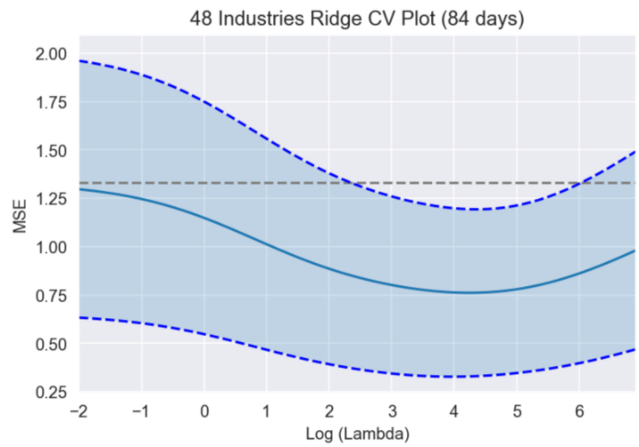

Fig 12. Ridge regression for 48 industries

However, when using 48 industries with 84 observations, we noticed that there is no significant difference between the CV plots using 84 observations as compared to using 63 observations. This is probably because overfitting is still a severe concern despite the use of more observations due to the large number of covariates. Hence, to overcome overfitting, it is necessary to increase the number of observations to ensure that the number of observations exceeds the number of covariates x 10, if a large number of covariates are being used.

Still, while we considered using 500 observations to fulfil the general rule of thumb for 48 industries, 500 observations stretches back years before the prediction date, which could be irrelevant. With more observations, the assumption that old data is representative of current returns might be too strong. While following the general rule of thumb is important, the relevance of the data is as important as the use of outdated data could result in poor prediction of the return.

6. Conclusion

The issue of overfitting is a serious concern when it comes to portfolio optimization as under-allocating or over-allocating capital to certain assets severely affect portfolio performances. Nonetheless, overfitting can be mitigated through the use of Lasso and Ridge models. While the use of other regularisation models such as PCR could generate alternative insights, our team feel that the use of Lasso and Ridge might be a better choice as they fare better in terms of performance. With that being said, different models suit different situations and the use of PCR could be more fitting in alternate scenarios. In trying to prevent overfitting, it is also worth noting that while more data can reduce the risk of overfitting, we need to ensure that our assumptions about the data are not too strong and that the data is relevant to our prediction model. Overall, regardless of the type of model used, it is important to find the right model complexity to ensure accurate and precise predictions.