WORLDQUANT UNIVERSITY

Submission Number: 1

Group Number: 15

Group Members:

Full Legal Name	**Location (Country)**	**E-Mail Address**	**Non-Contributing Member (X)**
Ewurama Fordjour	Ghana	ewurama.fordjour@gmail.com	
Deven N. Valecha	India	devenvalecha@gmail.com	
Ishaan Narula	India	ishaan.narula@outlook.com	

Statement of integrity: By typing the names of all group members in the text box below, you confirm that the assignment submitted is original work produced by the group (*excluding any non-contributing members identified with an "X" above*).

Ewurama Fordjour, Deven N. Valecha, Ishaan Narula

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.

N/A

** Note, you may be required to provide proof of your outreach to non-contributing members upon request.*

Work Split Report

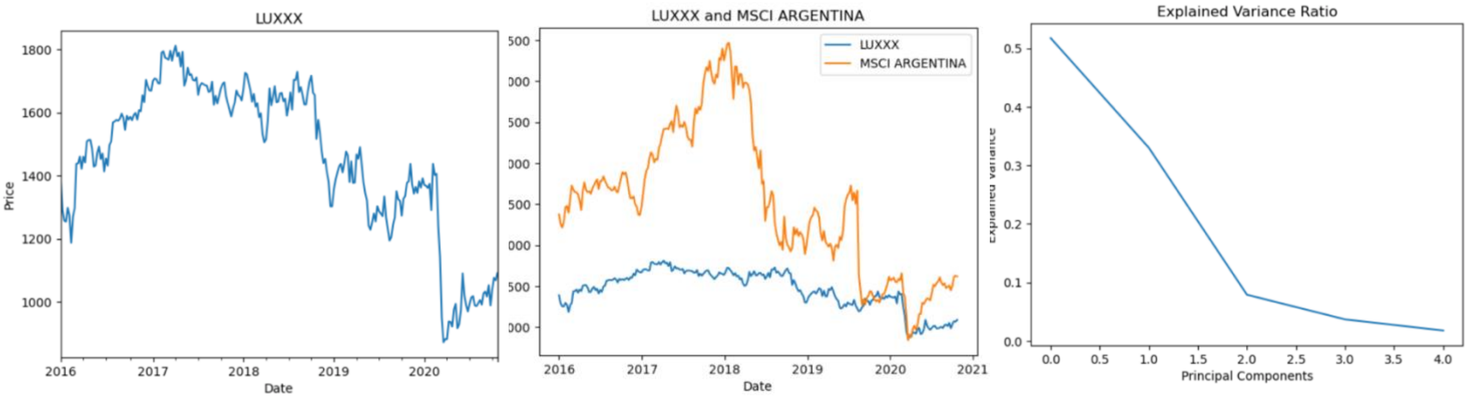
- Deven worked on questions 1 to 9
- Ishaan worked on questions 10 to 12
- Ewurama worked on questions 13 to 15
- Together we all had a look at each others part and made necessary suggestions for the successful completion of the assignment

MScFE 650 Machine Learning in Finance

Group #: 15 Submission #: 1

Technical Report

In our assignment we week to analysis our data using Principal component analysis and Lasso regression. We however incorporate other correlation models to help us know how best our data works. Per our graph of the time series of our LUXX data we can deduce that the data sample was a true representation of the dispersion. We also see a steady but progressive price strikes as the years go by. Though there was a drastic decline from the latter part of 2020 through to 2021.



Time Series Charts Showing Price Evolution of Various Funds over time; Explained Variance Contribution by PCs

The measure of the correlation coefficient for both data seemed to be having varying discrepancies. We see a very strong positive correlation with a time series data and itself when we used all the three correlation texts. They however tend to vary slightly when we did a particular data set with another data set. LUXX to Argentina showed 0.75 for the Pearson model, 0.76 for the Spearman model and 0.56 for the Kendall model.

Following the correlation analysis, we conduct a Principal Component Analysis to reduce the dimensionality of the 35-feature dataset into 5 features, each of which is essentially an eigenvector of the covariance matrix. Looking at the Explained Variance Ratio chart shown above, we see that the first 2 components explain more than 80% of the variation in the dataset. The breakdown of each component's contribution has also been tabulated below.

Principal Component	Contribution to Explained Variance
PC 1	55.58%
PC 2	27.31%
PC 3	5.21%
PC 4	3.68%
PC 5	2.25%

This is followed by a linear regression of the dependent variable LUXXX on the 5 principal components, which results in an R-squared value of 87.37%. This means that 87.37% of the total variation in LUXXX can be explained by its linear association with the 5 principal components.

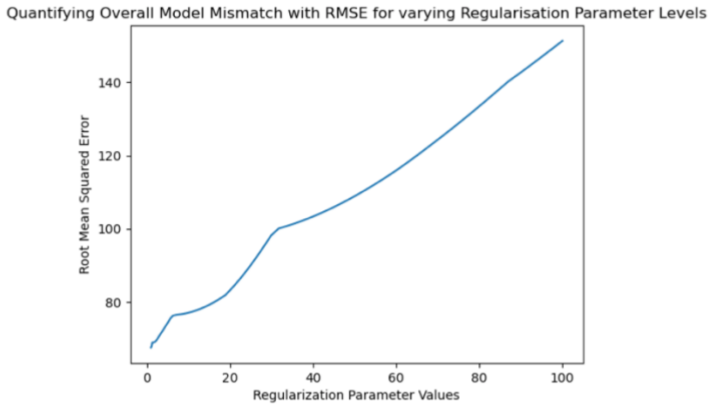
An alternative to PCA for achieving dimensionality reduction is Lasso Regression. This is basically a form of linear regression whereby the regressors which do not contribute much to explain the variation in the dependent variable are discarded. This is achieved by imposing by introducing a penalty parameter into the cost function, which takes the following form:

$$Cost(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\left|\beta_j\right|$$

Notice that the penalty is imposed by multiplying the sum of the absolute values of the beta coefficients with a parameter λ which controls the severity of the penalty. This is called L1 regularisation. It differs from L2 regularisation in that the beta parameters in the latter are squared.

We take 1000 different values of λ to test how quickly the regression coefficients for the various regressors reduce to zero as the penalty parameter increases. This gives us a sense of the relative importance of the various regressors. Notice that there is a trade-off between reducing the number of regressors and the resulting models' goodness of fit (captured by R-squared). However, a large number of regressors can lead to overfitting.

Upon fitting 1000 models, each corresponding to one value of λ, we plot the RMSE of each against increasing values of the penalty parameter, to capture extent of model mismatch.



We finally carry out an analysis to simplify our model to 7 predictors which do the most in explaining the variation in LUXXX. We also enlist all models (i.e. penalty parameter values, resulting regression coefficients and R-squared values) with have 7 or fewer non-zero predictors. We notice that a subset of models with 6 or 7 predictors have R-squared values exceeding 90%. Models with 4 or 5 predictors have R-squared less than 70%.

In conclusion, as elucidated above, both PCA and Lasso Regression follow different approaches to generate simple models with a handful of predictors without compromising on explanatory power. Although PCA results in a better fit because it preserves maximum variance, Lasso regression provides a better interpretation of results since the features therein can be meaningfully interpreted in the problem's context.

MScFE 650 Machine Learning in Finance

Group #: 15 Submission #: 1

Non-technical 1 Paragraph Email For Senior Management

Dear All,
Please see below our non-technical findings from the group assignment for your necessary action.
Principal component analysis (PCA) and Lasso regression both of which were used for dimensionality reduction helped us to somewhat reduce the number of variables used in our modelling. The PCA for our regression was 0.909 which was very indicative of the fact the model best fits the data under study. We however used 2 components to express 80% of the variation of our data, with the very first component accounting for more than half of the variance.

Thank you.