JP Morgan ALM R&S Internship
Quantitative Assignment

Ruoyu Lin

To be delivered Nov. 21 2022

1 Summary

1.1 Fred Query Script

For the JPM ALM R&S quantitative assignment, the standard Pythonic practice for API queries via built-in library `requests` is used, before which the Fred API key was manually downloaded. The query pipeline keeps the return type to be `pandas.DataFrame` for consistency. After querying all desired data series, all series are combined into one dataframe, and any row with *any* `NaN` value is dropped. This query process results in a data matrix of dimension $n \times d = 556 \times 6$ *i.e.* 556 observations and 6 features in total.

1.2 Model

Although several candidate models were attempted in the exploratory research, the final model selected is the most standard dimensionality reduction model – Principal Component Analysis (PCA). The standard implementation of `sklean.decomposition.PCA` is chosen[1]. The final parameter of `n_components`= 2 is used *i.e.* the dimension reduction from $\mathbf{6} \rightarrow \mathbf{2}$ is achieved. It is worth noting that as a standard practice of PCA, the returns are **de-meaned** before fitting to the model. Although 4 of 6 dimensions are eliminated, the first two principal components (PCs) still retain approximately **96.7**% of all variances in the data set, a more than reasonable loss of information for efficient features selection. The reduced data can be interpreted as orthogonal (mutual-independent), linear combinations of the original features. Final model:

```
sklearn.decomposition.PCA(n_components=2)
```

2 Model Selection

The model selection process is heavily influenced by a result presented at the 2022 Bloomberg-Columbia Machine Learning in Finance Workshop[2], namely a paper by Paul Bilokon and

[1] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[2] https://cfe.columbia.edu/content/8th-annual-bloomberg-columbia-machine-learning-finance-2022-workshop

David Finkelstein that introduces the iterated and the exponentially weighted extension to the traditional PCA[3]. The methods of Iterated Principal Component Analysis (IPCA) *i.e.* `xpca.IPCA` and Exponentially Weighted Moving Principal Component Analysis (EWMPCA) *i.e.* `xpca.EWMPCA` are both attempted in the exploratory research[4]. While both extended PCA demonstrate adequate performance on the queries data set (in terms of variance explained), the performance of their time stabilities in this *retrospect* study are not as good as the classical PCA. This thus resulted in the final choice of the classical PCA. However, it ought to be known by the reader that in a practical setting where one does not peek into the future, the IPCA and EWMPCA are robust models to be considered for dimensionality reduction. Both the exploratory research script `PCAs_comparison.py` and the Bilokon & Finkeltein paper are to be included in the submission for the interest of the reader.

3 Time Stability Analysis

To test for the time stability of the model, the Bresuch-Pagan test for heteroskedasticity is employed[5]. The assumption goes as follows:

Proposition 1. *If a dimensionality-reducing model is said to be* ***time-stable****, then the prediction problem of the original data set* $\boldsymbol{X}$ *using the reduced data set* $\widetilde{\boldsymbol{X}}$ *demonstrates* ***similar*** *performance throughout the time-span.*

And a direct consequence of this assumption is :

Remark 2. If a dimensionality-reducing model is said to be **time-stable**, then the OLS residual $\{\sigma_i^2\}_t$ of any time series $\mathbf{X}_i$ for $i = 1, \ldots, d$ demonstrates homoskedasticity.

Hence, holding this assumption, the Bresuch-Pagan test implemented by `statsmodels` library is used with an arbitrary critical value of $\alpha = 0.05$. The test results yielded p-values of $[15.61, 33.60, 9.83, 88.61, 6.08, 12.11]$ respectively for **GS1**, ..., **GS10**, which suffices to show that the model is **not time-stable** under above assumption.

Nonetheless, this result is expected under economic intuition, as there exist other external factors, whose information is not captured by the 6 time series, that induce outlier behavior in the time series, which consequently results in overfitting of the model. One prominent example is the sudden drop in returns during the March 2020 COVID-19 market crash, visible in both the original data and the reduced data, whose impact absolutely constitutes an outlier market incident.

Stability analysis summary: **Bresuch-Pagan** tests indicates that PCA(2) is not time-stable.

[3]Bilokon, Paul and Finkelstein, David, Iterated and Exponentially Weighted Moving Principal Component Analysis (August 30, 2021). Available at SSRN: https://ssrn.com/abstract=3913940 or http://dx.doi.org/10.2139/ssrn.3913940

[4]https://github.com/sydx/xpca

[5]Breusch, T. S., and A. R. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." Econometrica 47, no. 5 (1979): 1287–94. https://doi.org/10.2307/1911963.