From: **Gerry Windiarto Mohamad DUNDA** gwmdunda@connect.ust.hk
Subject: CSIC 5011: Project 1 Review
Date: 20 April 2023 at 3:16 PM
To: datascience.hw@gmail.com

Dear Professor,

Peer review 1: Ancestry Prediction via Dimensionality Reduction Techniques on SNPs Data. (9)

Summary
The study used dimensionality reduction methods on SNPs data to explore the relationship between genetic variation and geographical location. The results showed that PCA and MDS can effectively separate people from different regions based on essential principal components of SNPs, and an adequate number of SNPs with top importance can predict the region where people come from. The study also found that the similarity of SNPs principal components can reflect the relationship between geographical locations. Overall, the study supports the close relationship between human genes and geographic variation, and the effectiveness of dimensionality reduction methods in exploring this relationship using SNPs data.

The strengths

- Clear research goals and methods: The paper clearly outlines the research goals and methodology, which provides a clear direction for the study and establishes a strong foundation for the findings.

- Effective use of dimensionality reduction methods: The paper effectively applies dimensionality reduction methods, such as PCA and MDS, to explore the relationship between genetic variation and geographic location. This demonstrates the effectiveness of these methods in studying complex datasets, such as SNPs data.

- Relevant and significant findings: The paper's findings support the close relationship between human genes and geographic variation, which is a significant and relevant topic in the field of genetics. The paper's findings also demonstrate the potential of using SNPs data and dimensionality reduction methods in exploring this relationship further.

The weaknesses

- Lack of explanation on the description of SNP dataset. Perhaps, explain the statistics of SNPs/person. Also, the task for the model is not completely described. How many labels and what is proportion of train-valid-test and so on. and preprocessing techniques. Lastly, the data format is not clear such as what is the size of the vector

size of the vector.

- Lack of discussion on potential limitations: The paper does not discuss potential limitations of the study, such as sample size, data quality, or statistical assumptions. This could limit the interpretability of the findings and the ability to generalize the results to other populations or datasets. A discussion of potential limitations could help to provide a more comprehensive understanding of the research.

Evaluation on Clarity and quality of writing: 3
Evaluation on Technical Quality: 3
Overall rating: 3
Confidence on your assessment: 2

Peer review 2: Human Migration History: A Single Nucleotide Polymorphisms Perspective (7)
Summary
This paper uses genetic data and statistical methods to investigate human migration history, focusing on the African origin hypothesis of the geographic origin and early migration of anatomically modern humans. The author employs PCA to analyze SNP data and explore optimal model settings, as well as applying MDS and t-SNE as comparisons. The results show that all clustering coincides with the African origin hypothesis, providing valuable insights into human migration patterns and contributing to the discussion on the evolution of anatomically modern humans.

The strengths

- Comprehensive approach: The paper presents a comprehensive approach to studying human migration history, using SNP data and dimension reduction methods such as PCA, MDS, and t-SNE. This allows for a more detailed exploration of high-dimensional datasets and the identification of patterns that might not be immediately apparent.

- Clear and interesting results: The study has identified several migration patterns, including the migration from Africa to America and the possible migration of Europe and Central/South Asia populations from the Middle East. These findings are interesting and provide valuable insights into the genetic patterns of human migration.

- Suitable method: PCA has proven to be a suitable method for dimension reduction of SNP data, as it can effectively capture the

variation in the data and provide useful information for clustering and analysis. The authors also explore optimal model settings and compare the results obtained from PCA with those from MDS and t-SNE. This shows that the authors have put considerable effort into choosing the appropriate methods for their analysis.

The weakness

- The paper does not discuss the limitations of the statistical methods used for the analysis, such as PCA, MDS, and t-SNE. While these methods can be powerful tools for exploring high-dimensional data, they have certain assumptions and limitations that may affect their results.

Evaluation on Clarity and quality of writing: 5
Evaluation on Technical Quality: 5
Overall rating: 5
Confidence on your assessment: 2

Peer review 3: Visualization and Forecasting on Finance Data (11)

Summary
The report uses various data reduction techniques to analyze a collection of stock market indices of 500 companies and visualizes data with more than 1000 features on a 2D surface. PCA and SPCA are found to be effective in distinguishing different classes. For data forecasting, PCA, SPCA, and t-PCA methods are used, with t-PCA being more efficient for selecting a vital subgroup in high-dimensional data. The manifold learning results match the analysis well.

The strengths

- The paper uses a variety of data reduction techniques to analyze a large dataset of stock market indices, making it more comprehensive.

- The paper offers a detailed explanation of the forecasting results of several prediction methods, which adds to its credibility.

The weakness

- The justification of choosing the model is not clear. Providing more information about the selection of prediction methods and their specific parameters.

- Discussion of the results is lacking.

Evaluation on Clarity and quality of writing: 3

Evaluation on Clarity and quality of writing: 3
Evaluation on Technical Quality: 3
Overall rating: 3
Confidence on your assessment: 2

Peer review 4: Application of Shrinkage in Portfolio Optimization (10)

Summary
The study examines the effectiveness of different shrinkage approaches in improving the performance of mean-variance optimization portfolios by mitigating the effects of noisy covariance matrix estimation. The authors compare the performance of several shrinkage approaches on the S&P 500 index against various benchmarks. They find that directly applying shrinkage to the precision matrix leads to a significant improvement in out-of-sample portfolio performance. Additionally, the results suggest that shrinkage on the covariance matrix alone does not sufficiently address the problem of noisy estimation.

The strengths:

- The authors employ multiple shrinkage approaches and compare their effectiveness in improving portfolio performance on a well-known benchmark index (S&P 500).

- The authors provide useful insights that can guide portfolio managers in selecting appropriate methods for mitigating the effects of noisy estimation and improving portfolio performance.

- The study presents clear and concise results that demonstrate the superiority of shrinkage on the precision matrix in mitigating the effects of noisy estimation and improving portfolio performance.

The weaknesses:

- The study could benefit from a more detailed discussion of the underlying statistical concepts and assumptions, particularly around the use of shrinkage methods.

- More elaboration on the results is needed. For example, the authors could explore the impact of different shrinkage parameters on portfolio performance and provide guidance on optimal parameter selection.

Evaluation on Clarity and quality of writing: 4
Evaluation on Technical Quality: 4
Overall rating: 4
Confidence on your assessment: 3

Confidence on your assessment: 1

Peer review 5: Exploration of PCA family for handwritten digit classification (17)

Summary
The paper compares the performance of various PCA methods, including original PCA, SparsePCA, and KernelPCA, with different kernel functions on MNIST dataset. The data was reduced using these methods, and the classification task was performed using the random forest classifier. The results of the experiment indicate that PCA is effective in reducing dimensions, and KernelPCA with a linear kernel function performs the best.

The strengths:

- The paper clearly states its research goal of comparing the performance of different types of PCA, including original PCA, SparsePCA, and KernelPCA, with various kernel functions for dimension reduction.

- The paper likely used a systematic approach to compare the performance of different PCA methods, and utilized a standard evaluation metric, such as accuracy, to assess their effectiveness in reducing dimensions for classification tasks.

The weakness:

- No abstract

- No explanation on the significance of the problem

- While the study compared different types of PCA, it did not compare the performance of PCA with other dimensionality reduction methods, such as t-SNE, UMAP, or LLE. Comparing PCA with these methods could provide a more complete picture of the trade-offs and benefits of different dimensionality reduction approaches.

Evaluation on Clarity and quality of writing: 4
Evaluation on Technical Quality: 3
Overall rating: 3
Confidence on your assessment: 2

Best regards

Best regards,
Gerry