CSIC 5011 Mini-Project 1: Explore SNPs Data

FAN Junyi, XIAN Zhuozhi {jfanap, zxian}@connect.ust.hk

Department of Mathematics, HKUST

1. Introduction

In this project, we first applied some dimensionality reduction methods on the SNPs data to explore the relationship between the genetic variation of peoples and their geographical variation. Next we used random forest to predict the regions where people come from based on SNPs and compared the accuracy of predictions with or without dimensionality reduction.

2. SNPs Data

Single-nucleotide polymorphisms (SNPs) are a type of genetic variation that is present in a considerable large fraction of the population [5]. The SNPs data we used in this project contains a data matrix of $n = 488,890$ rows of autosomal SNPs and $p = 1043$ columns of peoples around the world. A file about the geographical information of each people is also used.

3. Methodology

To analysis the high dimensional SNPs data, the following methods are utilized for reducing dimension and extracting essential information.

- Principal Component Analysis (PCA) is widely used in data processing and dimensionality reduction. It uses orthogonal transformation to transform a set of possible variable correlation data into a set of linear uncorrelated variables called principal components [3].
- Multidimensional Scaling (MDS) is a methods that represents measurement of similarity among pairs of objects [1]. It uses the pairwise similarity of objects to construct a low-dimensional space, where the distance of each pair of objects is as consistent as possible in the original high dimensional space.
- t-Distributed Stochastic Neighbor Embedding (t-SNE) is a method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map, which is a variation of SNE which overcomes the crowding problem [4].

Moreover, since p is large, we used random projections for dimensionality reduction.

4. Dimensionality Reduction

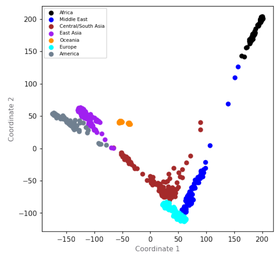

Fig. 1: PCA

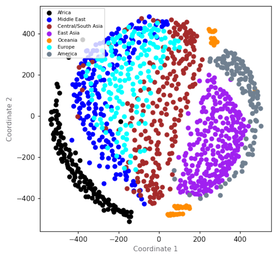

Fig. 2: MDS

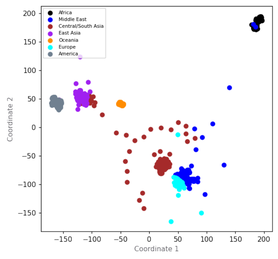

Fig. 3: t-SNE

Dimensionality reduction results for the SNPs data by PCA, MDS, t-SNE are shown in Fig. 1-3, respectively. All methods can separate peoples into the regions where they come from, while PCA has the best performance among them. It indicates that there is a relationship between human's genetic information and their geographical information.

We randomly selected $q = 100000, 10000, 1000$ rows of SNPs as random projection methods for PCA, and the corresponding results are shown in Fig. 4-6, respectively. When $q = 100000, 10000$, the results are similar to the original PCA result shown in Fig. 1, while when $q = 1000$, the result is not as good as the original one.

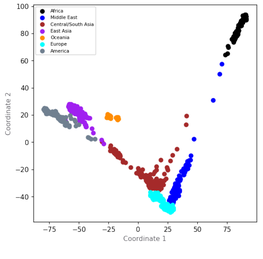

Fig. 4: $q = 100000$

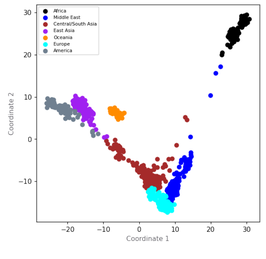

Fig. 5: $q = 10000$

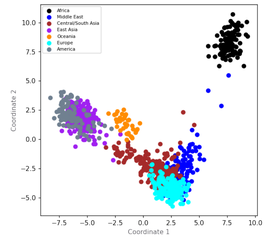

Fig. 6: $q = 1000$

5. Prediction

Method	Accuracy
Random Forest	96.5%
Random Forest + PCA	98.4%
Random Forest + MDS	82.7%
Random Forest + t-SNE	96.8%
Random Forest + PCA on 100000 randomly selected SNPs	97.8%
Random Forest + PCA on 10000 randomly selected SNPs	95.0%
Random Forest + PCA on 1000 randomly selected SNPs	85.6%

Tab. 1: Prediction by different methods

q	9000	8000	7000	6000	5000	4000	3000	2000
Accuracy	96.3%	95.9%	96.5%	95.3%	91.3%	93.2%	90.8%	88.4%

Tab. 2: Prediction by Random forest + PCA on randomly selected q SNPs

Since the SNPs data is related to geographical information, we used random forest with or without dimensionality reduction to predict the regions of peoples based on their SNPs. We keeped 2 principal components for PCA and used 2D MDS and t-SNE. The prediction accuracy from 5-fold cross-validation by different methods are listed in Tab. 1. Results show that predicting with PCA on all SNPs or $100000, 10000$ randomly selected SNPs has the same good performance as predicting directly.
Next we investigated the accuracy of prediction with PCA on different numbers of randomly selected SNPs. The corresponding results are listed in Tab. 2.

6. Further Analysis

In Fig. 1, except for those from Oceania, all the datapoints are laid in a 'V' shaped line where the order of regions from right to left is Africa, Middle East, Europe, Central/South Asia, East Asia and America, which is similar to the order of regions from left to right in Fig. 2. This order is consistent with the serial founder effect model for human settlement out of Africa in [2]. In particular, the proximity of the datapoints from East Asia and America in Fig. 1-2 is consistent with the hypothesis that native Americans migrated from Northeast Asia.

7. Conclusion and Future Work

For dimenisoality reduction part, we performed PCA, MDS and t-SNE to analysis and visualize the relevance between SNPs of peoples and their geographical information, while we randomly selected different numbers of SNPs to overcome the large dimensionality. For prediction part, we predicted the regions of peoples based on SNPs by the combinations of random forest and different dimenisoality reduction methods.
Both the results of dimenisoality reduction and prediction show that only a very small part of all SNPs can inform the regions where peoples come from, which leads to a problem about the lower bound of the number of peoples' SNPs required to identify their geographical information.

8. References

[1] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[2] Omkar Deshpande et al. "A serial founder effect model for human settlement out of Africa". In: *Proceedings of the Royal Society B: Biological Sciences* 276.1655 (2009), pp. 291–300.

[3] Sasan Karamizadeh et al. "An overview of principal component analysis". In: *Journal of Signal and Information Processing* 4.3B (2013), p. 173.

[4] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[5] David G Wang et al. "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome". In: *Science* 280.5366 (1998), pp. 1077–1082.

9. Contribution

- **Coding**: FAN Junyi
- **Poster**: XIAN Zhuozhi