AwesomePackage: A Valid R Package for Ancestry Inference

Jonathon Chow
Department of Mathematical Sciences
University of Science and Technology of China
Anhui, 230026, P.R.China
jonathonchow23@gmail.com

September 30, 2022

Abstract

Ancestry inference is an important topic in genetics. Its main task is to estimate population structure from genetic data. The PSD model has been adopted as the standard way of ancestry inference. Meanwhile, many statistical learning methods can be used to fit the PSD model, such as Markov chain Monte Carlo (MCMC), Expectation-Maximization (EM), sequential quadratic programming (SQP), variational inference (VI) and stochastic variational inference (SVI). Here, we implement the algorithms to fit the PSD model based on EM, SQP, VI and SVI respectively. We evaluate and compare these algorithms from multiple perspectives. From the perspective of algorithm accuracy, we show that all these algorithms perform well, among which the SQP algorithm performs best. From the perspective of algorithm efficiency, we show that the performance of the SVI algorithm is far better than other algorithms on large-scale data, and the performance of the VI algorithm is slightly better than other algorithms on small-scale data. From the perspective of data structure, we show that the VI algorithm and the SVI algorithm tend to reveal only the main features, while the EM algorithm and the SQP algorithm tend to reveal the fine structure. From the perspective of population number, the optimal value of the VI algorithm tends to appear earlier, while the optimal values of the EM algorithm and the SQP algorithm tend to appear at the largest population number. We test these algorithms on simulated data, TGP data and HGDP data. Our R package, AwesomePackage, is freely available online at `https://github.com/JONATHONCHOW/AwesomePackage`.

***K*eywords** PSD model · EM algorithm · SQP algorithm · VI algorithm · SVI algorithm

1 Introduction

Ancestry inference, which reveals the structure of a population from genotypic data, has become an essential task in genetics. It is relevant to many important topics in genetics, such as inheritable diseases (Francioli et al. 2014), conservation genetics (Pearse and Crandall 2004; Randi 2008), the ancestry and migration patterns of natural populations (Rosenberg et al. 2002; Reich et al. 2009), etc. With decreasing costs in sequencing and genotyping technologies, enormous amounts of genetic data about people and other organisms have become available. There is a growing need for fast and accurate tools to uncover the structure of populations from vast amounts of genetic data.

Model-based (likelihood and Bayesian) and non-model-based (PCA and K-means clustering) methods were developed to identify populations and assign individuals to the identified populations using marker genotype data. Model-based methods are favoured because they are based on a probabilistic model of population genetics with biologically meaningful parameters and thus produce results that are easily interpretable and applicable. Furthermore, they often yield more accurate structure inferences than non-model-based methods.

The probabilistic model of Pritchard, Stephens and Donnelly (Pritchard, Stephens, and Donnelly 2000), known as the PSD model, has become a standard tool for those model-based methods.

Many statistical learning methods can be used to fit the PSD model. From the perspective of maximum likelihood estimation, we can use Expectation-Maximization (EM), sequential quadratic programming (SQP), and sparse non-negative matrix factorization (SNMF). From the perspective of estimating the Bayesian posterior distribution, we can use Markov chain Monte Carlo (MCMC), variational inference (VI), and stochastic variational inference (SVI). Almost all of these algorithms have been developed into software, such as STRUCTURE (MCMC) (Pritchard, Stephens, and Donnelly 2000), FRAPPE (EM) (Tang et al. 2005), ADMIXTURE (SQP) (Alexander, Novembre, and Lange 2009), sNMF (SNMF) (Frichot et al. 2014), fastSTRUCTURE (VI) (Raj, Stephens, and Pritchard 2014), and TeraStructure (SVI) (Gopalan et al. 2016). There are also some recent developments, such as PopCluster (Wang 2022). These algorithms have different advantages. For example, STRUCTURE, FRAPPE and ADMIXTURE can parse fine structures, fastSTRUCTURE can highlight salient features, and TeraStructure can analyze large-scale data.

Meanwhile, the collection of human genetic data is proceeding apace. The two most typical projects are the 1000 Genomes Project (Abecasis et al. 2012) and the Human Genome Diversity Project (Cann et al. 2002; Cavalli-Sforza 2005), Known as TGP and HGDP.

In *Models and Methods*, we briefly describe the PSD model and the theoretical basis of the algorithms. We also briefly illustrate the relationship between the PSD model and some other models. We then describe the implementation details of the algorithms and the schemes to accelerate computation. Finally, we introduce some criteria for algorithm evaluation to help evaluate the accuracy of the results, choose population number, and compare the performance of different algorithms. In *Applications*, we compare the accuracy and time complexity of different algorithms on simulated genotype data sets. Then we demonstrate the results, especially the selection of population number, on TGP data set and HGDP data set.

2 Models and Methods

Todo

2.1 PSD model

observed variable: genotype matrix G

latent variable: matrix Z of the true origin of genes

parameters: population scale matrix P, gene scale matrix F

hyperparameter: population number K

2.2 EM algorithm

E-step: compute the expectation a_{ijk} and b_{ijk}

M-step: compute the maximization and update the parameters p_{ik} and f_{kj}

convergence criterion: the log-likelihood of incomplete data $\mathcal{L}(G|P, F)$ converges

2.3 SQP algorithm

update parameters: update P and F block by block alternately

convergence criterion: the log-likelihood of incomplete data $\mathcal{L}(G|P, F)$ converges

2.4 VI algorithm

update parameters: update variational parameters $\tilde{z_{ij}^a}$, $\tilde{p_i}$, $\tilde{f_{kj}^1}$, $\tilde{f_{kj}^2}$

convergence criterion: the ELBO converges

2.5 SVI algorithm

sample: sample a SNP

update parameters: iteratively update local parameter F_j at the SNP until it converges, then update global parameter P

convergence criterion: the log-likelihood at the validation set converges

2.6 Relationships with other models

Poisson NMF model (Carbonetto et al. 2021)

multinomial topic model

LDA model

3 Applications

We fit the PSD model on simulated data set, TGP data set and HGDP data set.

3.1 Simulated data set

To evaluate the performance of the different learning algorithms, we generated two groups of simulated genotype data sets. In simulated data set A, we focus on the influence of the strength of population structure and the choice of parameter `K` on the performance of different algorithms. In simulated data set B, we focus on the performance of different algorithms when the mixing ratio gap between different individuals is small.

3.1.1 Simulated Data Set A

We generated the simulated data set A (Raj, Stephens, and Pritchard 2014) in three steps. First, generate the population scale matrix P using the Dirichlet distribution; In the second step, the gene scale matrix F is generated using beta distribution. The third step is to generate the genotype matrix G using the binomial distribution. We set the number of individuals I to 600, the number of SNPs J to 2500, and the number of populations K to 3.

Step 1. The population scales for each sample are drawn from a symmetric Dirichlet distribution to simulate small amounts of gene flow between the three populations. Here we use $Dirichlet(\frac{1}{10}\mathbf{1}_3)$, of course, in the implementation code, we can adjust the parameters of the Dirichlet distribution.

Step 2. The ancestral allele frequencies $\bar{f}_j$ for each SNP are drawn from a natural data set to simulate allele frequencies in natural populations. Here we use the HGDP data set. First, $\bar{f}_j$ is equal to the total number of suballeles observed at the jth SNP divided by twice the number of individuals. Then, we assume that the samples are drawn from a three-population demographic model. The edge weights correspond to the parameter F_k (Wright 1949) [1] in the model that quantifies the genetic drift of each of the three current populations from an ancestral population. Here we choose $(F_1, F_2, F_3) = (0.1, 0.05, 0.01)$ to simulate strong structure and $(F_1, F_2, F_3) = 0.5 \times (0.1, 0.05, 0.01)$ to simulate weak structure. Thus, the allele frequency at a given locus for each population is drawn from a beta distribution (Balding and Nichols 1995) $f_{kj} \sim Beta\Big(\frac{1-F_k}{F_k}\bar{f}_j, \frac{1-F_k}{F_k}(1-\bar{f}_j)\Big)$.

Step 3. According to the PSD model, each element g_{ij} of the matrix G follows a binomial distribution with probability $(PF)_{ij} = \sum_{i=1}^{K} p_{ik} f_{kj}$ and number of trials 2.

3.1.2 Simulated Data Set B

We also use three steps to generate simulated data set B. In the second step, we set all F_k to 0.1. The third step is the same as for data set A. We just consider the first step. We set a Gaussian density for each ancestral population centered at its location and normalizing each individual such that all proportions sum

[1] In population genetics, F-statistics (also known as fixation indices) describe the statistically expected level of heterozygosity in a population; more specifically the expected degree of (usually) a reduction in heterozygosity when compared to Hardy–Weinberg expectation.

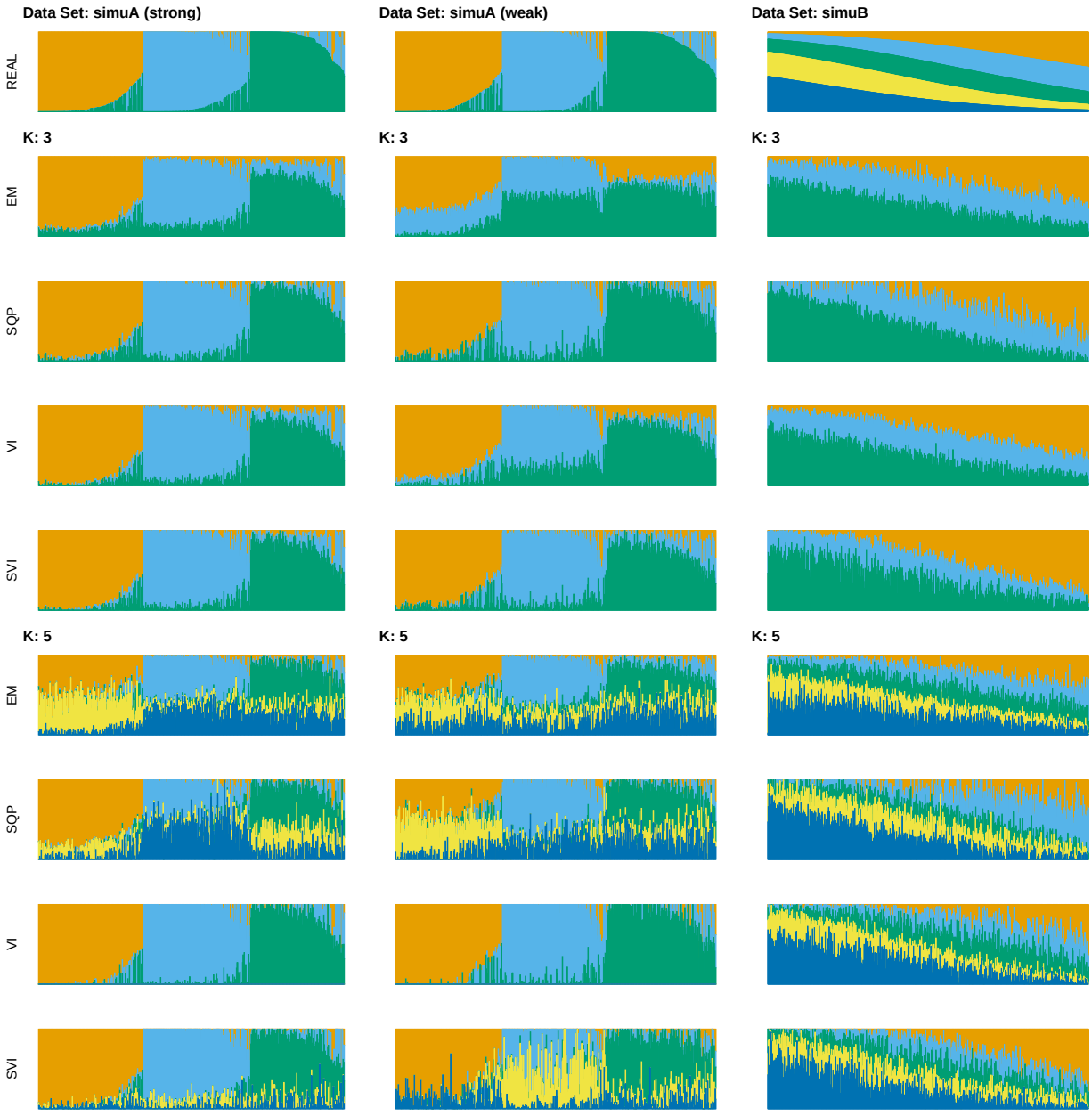

Figure 1: The structure plot of simulated data sets.

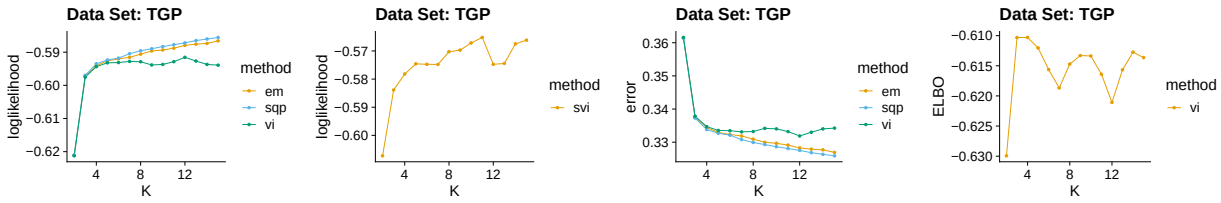

Figure 2: The evaluation indicators of TGP data set.

to 1 (Gopalan et al. 2016). In this case, each ancestral population is placed at a location evenly spaced along a line. Individuals are also positioned evenly on the line, and their proportions p_{ik} are a function of their proximity to each population's location. We set the number of individuals I to 1000, the number of SNPs J to 5000, and the number of populations K to 5.

3.1.3 Results

The main purpose of simulated data set A is to study the influence of the strength of population structure and the choice of parameter `K` on the performance of different algorithms.

For EM and SQP algorithms, they tend to reveal details, that is, they are sensitive to parameter `K` and structure strength. With the appropriate parameter `K`, this may be an advantage, as it reveals a finer structure. However, when the parameter `K` is too large, the phenomenon of overfitting is easy to occur. In addition, SQP algorithm is more accurate than EM algorithm.

For the VI algorithm, we notice that the results of VI are almost consistent for both parameter `K` and structure strength changes. This means that the VI algorithm only tends to reveal the main factors, thereby ignoring some smaller contributions. This is both a strength and a weakness.

The SVI algorithm is an ideal choice in many cases, because it can both highlight the main parts like VI, and react acutely when the structure is not obvious like EM and SQP.

The main purpose of simulated data set B is to study the performance of different algorithms when the mixing ratio gap between different individuals is small. In this case, EM algorithm and SQP algorithm can more faithfully reflect the structure of the dataset (the former is better), while VI algorithm and SVI algorithm will overemphasize some features (the former is worse).

3.2 TGP data set

Todo

3.2.1 Choose `K`

Criteria for choosing `K`: When there is no obvious gap in indicators, the smaller `K` is preferred. See Figure 2. We notice that the log-likelihood curves of EM and SQP have a continuous upward trend, which is due to the fact that there is no prior distribution constraint, which is prone to overfitting. The log-likelihood curves of EM and SQP slow down from `K` equals 4. The log-likelihood curve of VI flattens out from about `K` equals 4, and shows that the optimal `K` is 12. The log-likelihood curve of SVI is relatively irregular for three reasons. First, we use different training and validation sets to fit different `K`. Although we finally fixed the validation set when calculating the log-likelihood on the validation set, this was based on the assumption that the data are equivalent. Second, our convergence criterion may be relatively loose, resulting in some cases that do not really converge to the optimal value. Third, the sensitivity of the algorithm to the initial value leads to large errors in a single measurement. The log-likelihood curve of SVI shows that the optimal `K` is 11, and 8, 9, 10, and 11 are all good choices for `K`. The error curves of EM, SQP and VI are almost identical with the log-likelihood curves of EM, SQP and VI. The ELBO curve of VI shows the curve oscillating from `K` equals 3.

In conclusion, we note that when `K` is around 4, the fit is already doing very well. The optimal `K` should be reached around 11, but from the structure diagram, the populations appear redundant at this time.

For the **best** `K` (equals 4 and 11), we draw more subtle structures. We can compare with the results in the article (Gopalan et al. 2016), and the structure diagram is almost the same. See Figure 3.

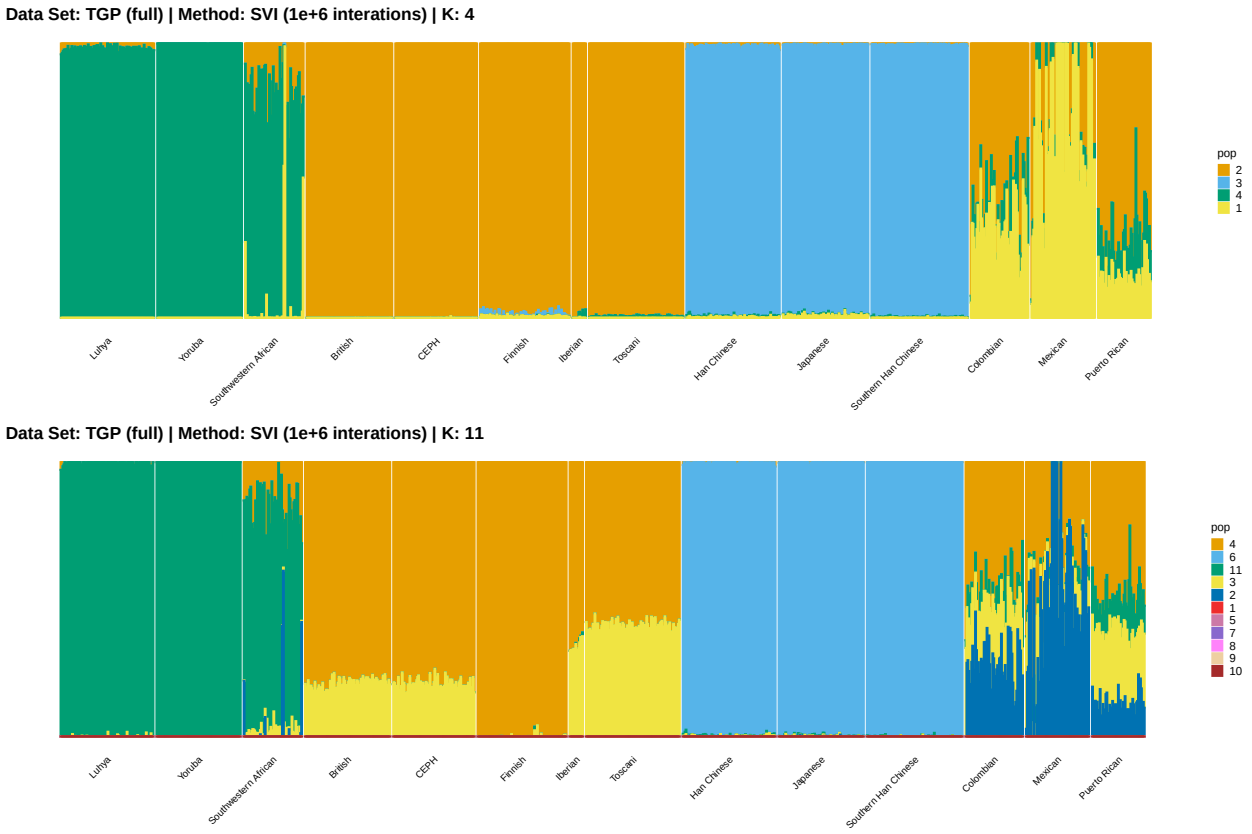

Figure 3: The structure plot of TGP data set.

3.3 HGDP data set

Todo

3.3.1 Choose `K`

Criteria for choosing `K`: When there is no obvious gap in indicators, the smaller `K` is preferred. See Figure 4. We notice that the log-likelihood curves of EM and SQP have a continuous upward trend, which is due to the fact that there is no prior distribution constraint, which is prone to overfitting. The log-likelihood curves of EM and SQP slow down from `K` equals 7. The log-likelihood curve of VI flattens out from about `K` equals 6, and shows that the optimal `K` is 8 and 11. The log-likelihood curve of SVI is relatively irregular for three reasons. First, we use different training and validation sets to fit different `K`. Although we finally fixed the validation set when calculating the log-likelihood on the validation set, this was based on the assumption that the data are equivalent. Second, our convergence criterion may be relatively loose, resulting in some cases that do not really converge to the optimal value. Third, the sensitivity of the algorithm to the initial value leads to large errors in a single measurement. The log-likelihood curve of SVI shows that the optimal `K` is 11 and 14, and 8, 9, 10, and 11 are all good choices for `K`. The error curves of EM, SQP and VI are almost identical with the log-likelihood curves of EM, SQP and VI. The ELBO curve of VI shows the curve oscillating from `K` equals 7.

In conclusion, we note that when `K` is around 7, the fit is already doing very well. The optimal `K` should be reached around 11, but from the structure diagram, the populations appear redundant at this time.

For the **best** `K` (equals 7 and 11), we draw more subtle structures. We can compare with the results in the articles (Li et al. 2008; Raj, Stephens, and Pritchard 2014; Gopalan et al. 2016), and the structure diagram is almost the same. See Figure 5.

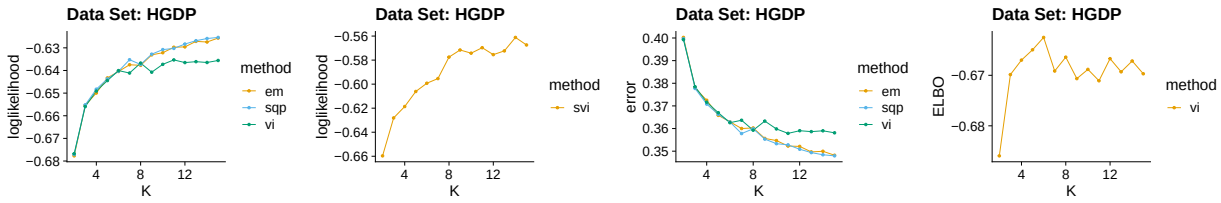

Figure 4: The evaluation indicators of HGDP data set.

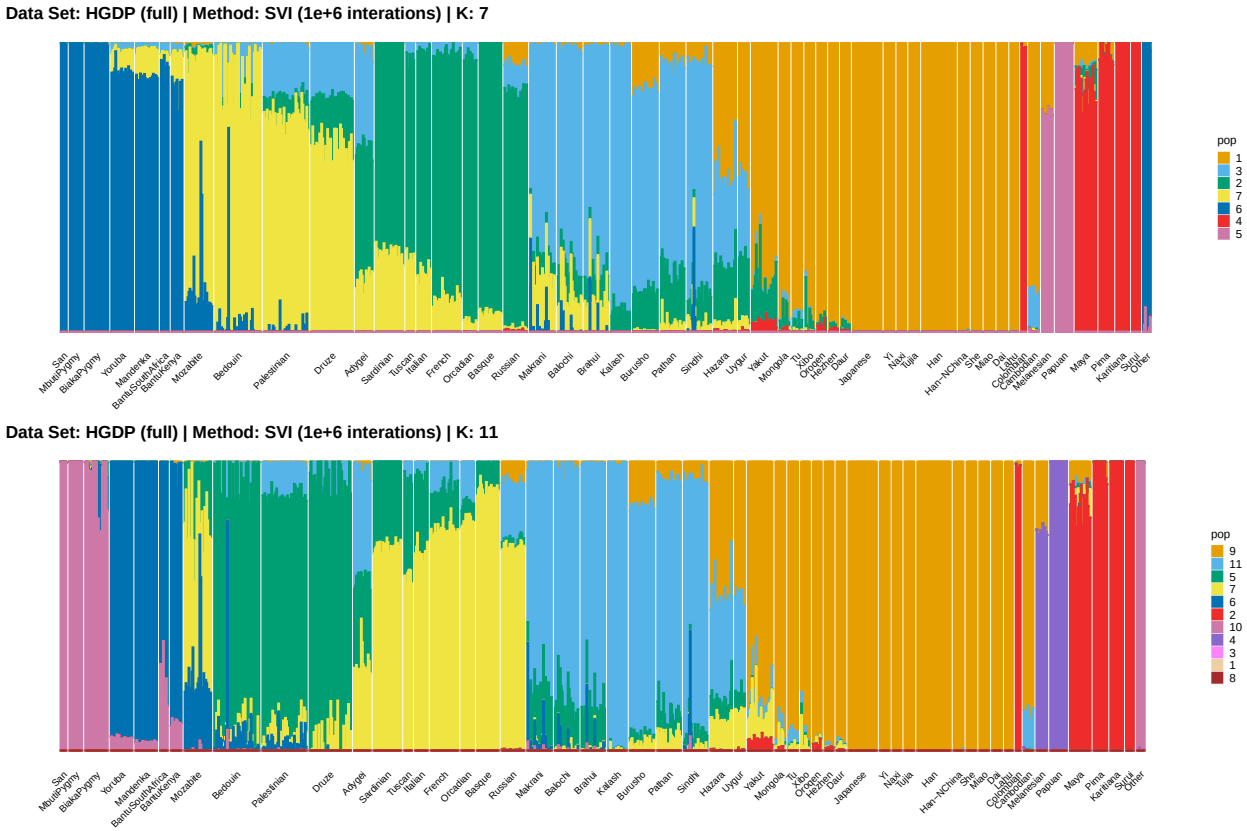

Figure 5: The structure plot of HGDP data set.

4 Discussion

Todo

4.1 Algorithm evaluation

Todo

4.1.1 Convergence accuracy

For suitable `K`, the SVI algorithm and SQP algorithm perform best in terms of convergence accuracy, followed by VI algorithm and finally EM algorithm. We can see this clearly in Section 3.1.

For the unknown `K`, due to the lack of prior constraints, the EM algorithm and SQP algorithm are prone to overfitting when the population number is redundant. Therefore, we had better use VI algorithm and SVI algorithm to select the appropriate `K`. See **Article** in AwesomePackage for details.

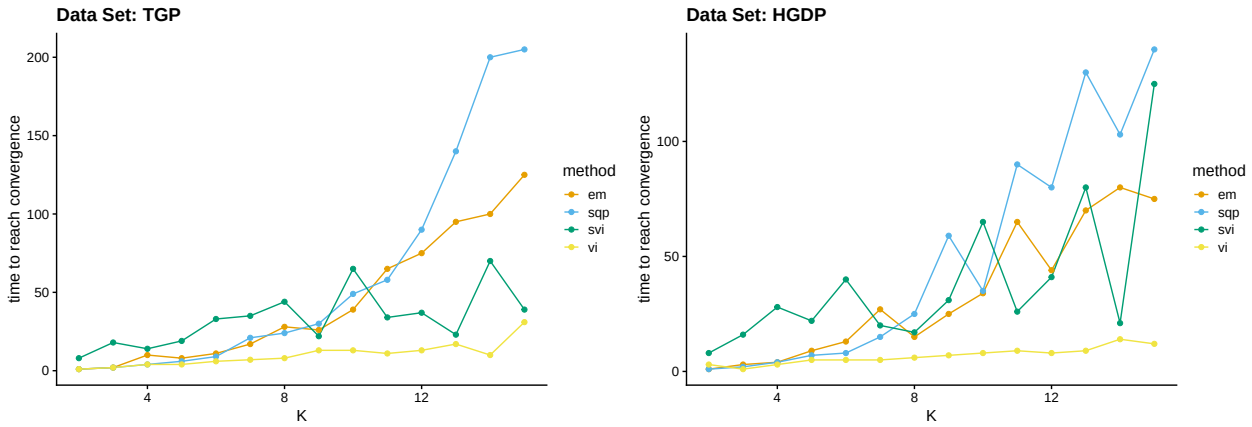

Figure 6: Convergence time. Since we only recorded integer values, we made a modest estimate of the value in integer hours.

4.1.2 Convergence efficiency

In addition to measuring the accuracy of convergence, we still need to consider the efficiency of convergence. We have two indicators to measure the convergence efficiency, which are convergence speed (the number of iterations required to achieve convergence) and convergence time (the time required for a single iteration). We can see the convergence time plots in Figure 6, and we can see the convergence speed plots in Appendix A.

EM algorithm is poor in terms of convergence time and convergence speed, and the convergence time increases rapidly with the increase of `K`.

Although SQP algorithm has a good performance in terms of convergence speed, the convergence time of the unaccelerated SQP algorithm is extremely slow, which increases rapidly with the increase of `K`.

VI algorithm has similar convergence speed with EM algorithm (both of them have poor performance), but in terms of convergence time, VI algorithm has excellent performance, especially with the increase of `K`, the required time increases slowly.

Due to different principles, we only consider the convergence time for the SVI algorithm. Although the performance of convergence time of SVI algorithm is poor on small data sets, the time of SVI algorithm is almost only related to the length of single sampling (the number of individuals), that is to say, for complete data sets, the convergence time of SVI is almost unchanged. This means that SVI has irreplaceable advantages for large data sets. Meanwhile, similar to VI algorithm, the change of convergence time of SVI algorithm is relatively insensitive to `K`. By the way, compared with other algorithms, the convergence time of SVI algorithm is irregular due to the randomness of sampling.

4.1.3 Algorithm selection criteria

In conclusion, we should consider both algorithm accuracy and algorithm efficiency. For small data sets, we can get good results by using VI directly. Or we can first use VI algorithm to reach the vicinity of the optimal value, and then use SQP algorithm to improve the convergence accuracy. The reason why the SQP algorithm is not directly used here is that the unaccelerated SQP algorithm is inefficient and the SQP algorithm is extremely easy to converge to local minima. For large data sets, we use the SVI algorithm without question. Of course, if `K` is unknown, we should pick `K` first, in the same way as above.

4.2 Others

Todo

Acknowledgments

Todo

Literature Cited

Abecasis, Goncalo R, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, et al. 2012. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 491 (7422): 56–65.

Alexander, David H, John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19 (9): 1655–64.

Balding, David J, and Richard A Nichols. 1995. "A Method for Quantifying Differentiation Between Populations at Multi-Allelic Loci and Its Implications for Investigating Identity and Paternity." *Genetica* 96 (1): 3–12.

Cann, Howard M, Claudia De Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, et al. 2002. "A Human Genome Diversity Cell Line Panel." *Science* 296 (5566): 261–62.

Carbonetto, Peter, Abhishek Sarkar, Zihao Wang, and Matthew Stephens. 2021. "Non-Negative Matrix Factorization Algorithms Greatly Improve Topic Model Fits." *arXiv Preprint arXiv:2105.13440*.

Cavalli-Sforza, L Luca. 2005. "The Human Genome Diversity Project: Past, Present and Future." *Nature Reviews Genetics* 6 (4): 333–40.

Francioli, Laurent C, Andronild Menelaou, Sara L Pulit, Freerk Van Dijk, Pier Francesco Palamara, Clara C Elbers, Pieter BT Neerincx, et al. 2014. "Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population." *Nature Genetics* 46 (8): 818–25.

Frichot, Eric, François Mathieu, Théo Trouillon, Guillaume Bouchard, and Olivier François. 2014. "Fast and Efficient Estimation of Individual Ancestry Coefficients." *Genetics* 196 (4): 973–83.

Gopalan, Prem, Wei Hao, David M Blei, and John D Storey. 2016. "Scaling Probabilistic Models of Genetic Variation to Millions of Humans." *Nature Genetics* 48 (12): 1587–90.

Li, Jun Z, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, et al. 2008. "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation." *Science* 319 (5866): 1100–1104.

Pearse, Devon E, and Keith A Crandall. 2004. "Beyond FST: Analysis of Population Genetic Data for Conservation." *Conservation Genetics* 5 (5): 585–602.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2): 945–59.

Raj, Anil, Matthew Stephens, and Jonathan K Pritchard. 2014. "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets." *Genetics* 197 (2): 573–89.

Randi, Ettore. 2008. "Detecting Hybridization Between Wild Species and Their Domesticated Relatives." *Molecular Ecology* 17 (1): 285–93.

Reich, David, Kumarasamy Thangaraj, Nick Patterson, Alkes L Price, and Lalji Singh. 2009. "Reconstructing Indian Population History." *Nature* 461 (7263): 489–94.

Rosenberg, Noah A, Jonathan K Pritchard, James L Weber, Howard M Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman. 2002. "Genetic Structure of Human Populations." *Science* 298 (5602): 2381–85.

Tang, Hua, Jie Peng, Pei Wang, and Neil J Risch. 2005. "Estimation of Individual Admixture: Analytical and Study Design Considerations." *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 28 (4): 289–301.

Wang, Jinliang. 2022. "Fast and Accurate Population Admixture Inference from Genotype Data from a Few Microsatellites to Millions of SNPs." *Heredity*, 1–14.

Wright, Sewall. 1949. "The Genetical Structure of Populations." *Annals of Eugenics* 15 (1): 323–54.

Appendix A

Todo

Appendix B

Todo