Topic Models & Ancestry Inference

MA Xiaoheng [1]

[1]School of Mathematics,
Harbin Institute of Technology
maxiaoheng7@hotmail.com

July 31, 2023

1. Topic Models
   - Topic Model Overview
2. Task1: A Simple Experiment
   - Algorithm
   - Experimental Results
3. Task2: Ancestry Inference
   - Ancestry Inference & PSD Model
   - Algorithm
   - Experimental Resluts

Topic Model Overview

Topic models are statistical models, which are often used in natural language processing and machine learning to discover abstract "topics" that occur in a collection of documents. Each "topic" is represented as a distribution over words.

Topic models diverge from Poisson models by assuming that documents are generated from K topics, where each topic is represented as a word distribution.

Not only are topic models useful in natural language processing, but they are also now widely applied in other fields, such as bioinformatics.

Data Generation Formula

The data generating procedure of a topic model can be described as:

$$x_{i1}, \ldots, x_{im} | L^*, F^*, t_i \sim \text{Multinomial}(t_i; , \pi_{i1}, \ldots, \pi_{im}),$$

where

$$\pi_{ij} = \left(L^* (F^*)^T \right)_{ij} = \sum_{k=1}^{K} l^*_{ik} f^*_{jk}.$$

Here, $t_i = \sum_{j=1}^{m} x_{ij}$ is the total count of words in document i, $\sum_{j=1}^{m} f^*_{jk} = 1$, $\sum_{k=1}^{K} l^*_{ik} = 1$. f^*_{jk} represents the probability of word j appearing in topic k, and l^*_{ik} is the topic composition of document i.

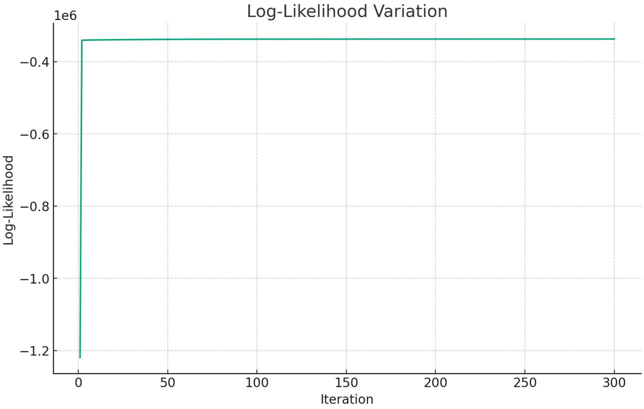
EM Algorithm for Topic Model Estimation

An EM (Expectation-Maximization) algorithm is developed to estimate L^* and F^* in the topic models. The following steps illustrate the implementation of the algorithm:

1. Initialize L and F randomly, normalize L by total word count t and normalize F by the sum across topics.
2. Perform EM iterations:
   - E-step: Compute the posterior distribution of topics for each word in each document, π_{ij}.
   - M-step: Update L and F by maximizing the expected log-likelihood, given π_{ij}.
3. Repeat the above steps until convergence, or until the maximum number of iterations (300) is reached.
4. Record the log-likelihood at each iteration to monitor the progress of the algorithm.

Likelihood Convergence Plot

Here is the log-likelihood value change across the EM algorithm iterations.



Ancestry Inference Overview

In recent years, accumulated genotype data has become a valuable resource for unravelling the complexities of human genetics. The genotype of an individual is known to be an admixture from different populations such as Asian, European, and African.

A key task in population genetics is to estimate the proportions of ancestry from each contributing population. This has led to the development of the PSD model, which is now a standard tool in this area.

PSD operates on a genotype matrix, $G \in \mathbb{R}^{n \times m}$, where n is the sample size and m is the number of biallelic genetic markers. An entry g_{ij} of G represents the observed number of the minor allele for individual i at marker j.

The PSD Model

Given the parameters p_{ik} and f_{kj}, representing the proportion of ancestry k for individual i and the frequency of the minor allele for marker j in each ancestry k, we model g_{ij} following a binomial distribution:

$$g_{ij} \sim \text{Bin}\left(2, \sum_{k=1}^{K} p_{ik} f_{kj}\right)$$

The probability density of g_{ij} is given by the following expression:

$$\Pr(g_{ij}) = \binom{2}{g_{ij}} \left(\sum_k p_{ij} f_{kj}\right)^{g_{ij}} \left(1 - \sum_k p_{ij} f_{kj}\right)^{2-g_{ij}}$$

The PSD model serves as a basis for various optimization methods like the Expectation-Maximization (EM) algorithm, the sequential quadratic programming (SQP) algorithm, and the stochastic variational inference (SVI) algorithm, all used for ancestry inference.

SVI for PSD model

1. **Data: Observed genotype for N individuals measured at L locations**
2. For all users $i \in 1, \cdots, N$, initialize the population proportions $\hat{\theta}_i$ randomly. Assume K ancestral populations.
3. **Repeat**
4. Sample a SNP location l and all observations $x_{1:N,\ell}$ at that location.
5. For $k \in 1, \cdots, K$, initialize $(\hat{\beta}_{k,\ell,0}, \hat{\beta}_{k,\ell,1})$ at SNP location ℓ to (a, b),
6. **(Allele frequency parameters)**
7. **Repeat:**
8. For $k \in 1, \cdots, K$ and $i \in 1, \cdots, N$ set

$$\begin{aligned} \phi_{i,\ell,k} &\propto \exp\left\{\mathrm{E}[\log \theta_{i,k}] + \mathrm{E}[\log \beta_{k,\ell}]\right\} \\ \xi_{i,\ell,k} &\propto \exp\left\{\mathrm{E}[\log \theta_{i,k}] + \mathrm{E}[\log(1-\beta_{k,\ell})]\right\} \end{aligned}$$

9. For $k \in 1, \cdots, K$ set the Beta parameters at SNP location l

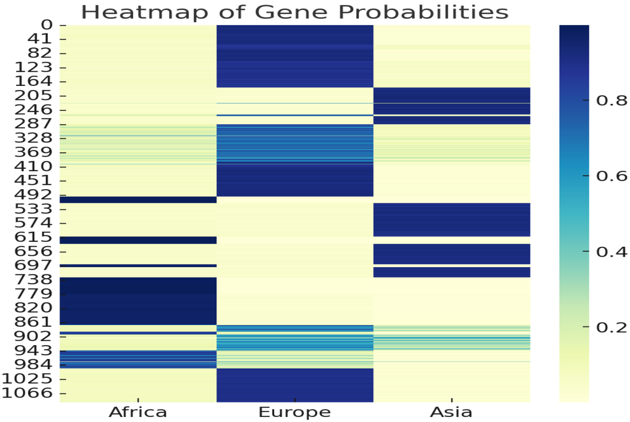
$$\begin{aligned} \hat{\beta}_{k,\ell,0} &= a + \textstyle\sum_{i=1}^{N} x_{i,\ell}\phi_{i,\ell,k} \\ \hat{\beta}_{k,\ell,1} &= b + \textstyle\sum_{i=1}^{N} (2 - x_{i,\ell})\xi_{i,\ell,k} \end{aligned}$$

10. **until** local parameters $\phi_{1:N,\ell}$, $\xi_{1:N,\ell}$ and $\hat{\beta}_{1:K,\ell}$ converge
11. **(Population proportions parameters)**
12. For $i \in \{1, \cdots, N\}, k \in \{1, \cdots, K\}$

$$\hat{\theta}^{t}_{i,k} = (1-\rho_t)\hat{\theta}^{(t-1)}_{i,k} + \rho_t L(c + x_{i,\ell}\phi_{i,\ell,k} + (2 - x_{i,\ell})\xi_{i,\ell,k})$$

13. Set the step-size $\rho_t = (\tau_0 + t)^{-\kappa}$ for iteration t
14. **until** convergence criteria are met

Results - Heat Map



- The determination of the continent types here is combined with the PCA plot of the dataset.

Results : 3D Scatter Plot and Correlation Plot

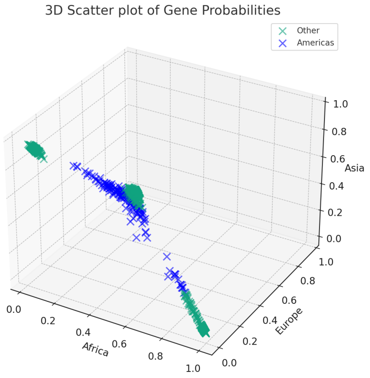


Figure: 3D Scatter Plot

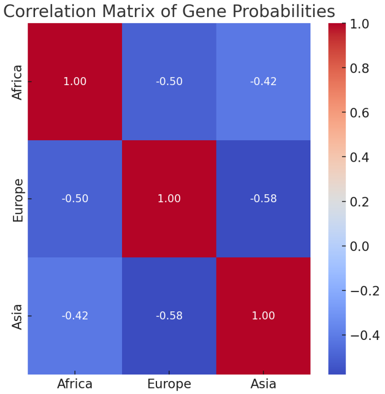


Figure: Correlation Plot

- The determination of the continent types here is combined with the PCA plot of the dataset.

Conclusions from Data Analysis

- Most samples in this dataset are pureblood from either Asia, Europe, or Africa.
- There are a few mixed-blood samples from America.
- Among the mixed-blood samples, those of European and Asian descent are the most common.
- Next are those of European and African descent.
- Mixed-blood samples of African and Asian descent are extremely rare.

References I

- Peter Carbonetto, Abhishek Sarkar, Zihao Wang, and Matthew Stephens. Non-negative matrix factorization algorithms greatly improve topic model fits. arXiv preprint arXiv:2105.13440, 2021.
- Prem Gopalan, Wei Hao, David M Blei, and John D Storey. Scaling probabilistic models of genetic variation to millions of humans. Nature genetics, 48(12):1587â1590, 2016.

Thank you !

MA Xiaoheng

School of Mathematics,
Harbin Institute of Technology
maxiaoheng7@hotmail.com

Project on GitHub: https://github.com/nosignalmxh/Topic_Models_Ancestry_Inference