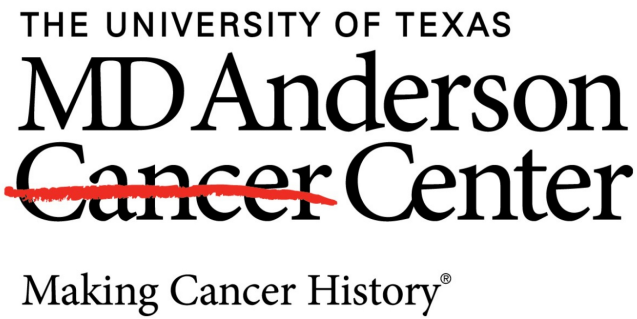
An Introduction to Mixture Models

Yu Ding

The University of Texas MD Anderson Cancer Center

Nov 11th, 2024



Outline

Outline

Finite Mixture Model

Let's note $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_N\}$ is a sample of size N, where Y_i is a P-dimensional random vector with probability density function $f(y_i)$ on $\mathbb{R}^P$, and y_i its realization.

$$f(y_i) = \sum_{k=1}^{K} \pi_k f_k(y_i),$$

where $f_k(y_i)$ is a component density of the mixture, and π_k the weight of population k subject to constraints $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$.

Finite Mixture Model

A new random variable is introduced, $\mathbf{Z} \in \{0,1\}^{N\times K}$. $z_{ik} = 1$ if y_i belongs to population k. $\{z_{i1}, \cdots, z_{iK}\}$ are assumed to be distributed according to a multinomial distribution:

$$\{z_{i1}, \cdots, z_{iK}\} \sim \mathcal{M}(1, \pi_1, \cdots, \pi_K).$$

The conditional, or posterior, distribution is

$$P\{z_{ik} = 1 | Y_i = y_i\} = \frac{\pi_k f_k(y_i|\theta_k)}{\sum_{k=1}^{K} \pi_k f_k(y_i|\theta_k)}.$$

Expectation-Maximization Algorithm

Let $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$ be the complete data sample space, where $\mathcal{Y}$ is the observed sample space and $\mathcal{Z}$ is the hidden sample space.
Define $\psi = \{\pi_1, \cdots, \pi_K, \theta_1, \cdots, \theta_K\}$. The complete-data log-likelihood function is

$$\log \mathcal{L}^c(\mathbf{X}; \psi) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log\{\pi_k f(y_i; \theta_k)\}.$$

In the Expectation step, we compute the expectation of the log-likelihood function given ψ'

$$Q(\psi, \psi') = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{\psi'}\{z_{ik} | Y_i = y_i\} \log\{\pi_k f(y_i; \theta_k)\}.$$

The EM algorithm consists of two steps:

- E-step: calculate $Q(\psi, \psi')$
- M-step: choose $\psi' = \arg\min_{\psi} Q(\psi, \psi')$

Outline

Negative binomial mixture model, and Binomial mixture model

In transcriptomic analysis, Li et al., 2023 [1], y_{ij} is the observed RNA counts for gene j in sample i.

$$y_{ij}|C_i = k \sim NB(\mu_{ijk}, \pi_j), \text{and } \log(\mu_{ijk}) = \log(s_i) + \beta_{jk},$$

where C_i is the cluster assignment for the ith sample, s_i is the normalization size factor of the ith sample.

In genomic analysis, Jiang et al., 2024 [2], let $\phi_i \in [0, 1]$ be the cellular prevalence (CP) of SNV i, we have the optimization problem as

$$\min_{\phi} \left\{ -\sum_{i=1}^{S} [r_i \log f(\phi_i) + (n_i - r_i)(1 - \log f(\phi_i))] + \sum_{1<j\leq S} p_\lambda(|\phi_i - \phi_j|) \right\}$$

Negative binomial mixture model, and Binomial mixture model

In transcriptomic analysis, Li et al., 2023 [1], y_{ij} is the observed RNA count for gene j in sample i.

$$y_{ij}|C_i = k \sim NB(\mu_{ijk}, \pi_j), \text{and } \log(\mu_{ijk}) = \log(s_i) + \beta_{jk},$$

where C_i is the cluster assignment for the ith sample, s_i is the normalization size factor of the ith sample.

In genomic analysis, Jiang et al., 2024 [2], SNVs from a common cancer cell population or subclone have identical CP.

$$r_i|C_i = k \sim Binomial(n_i, f_i(\phi_k))$$

Gaussian Mixture Model (GMM)

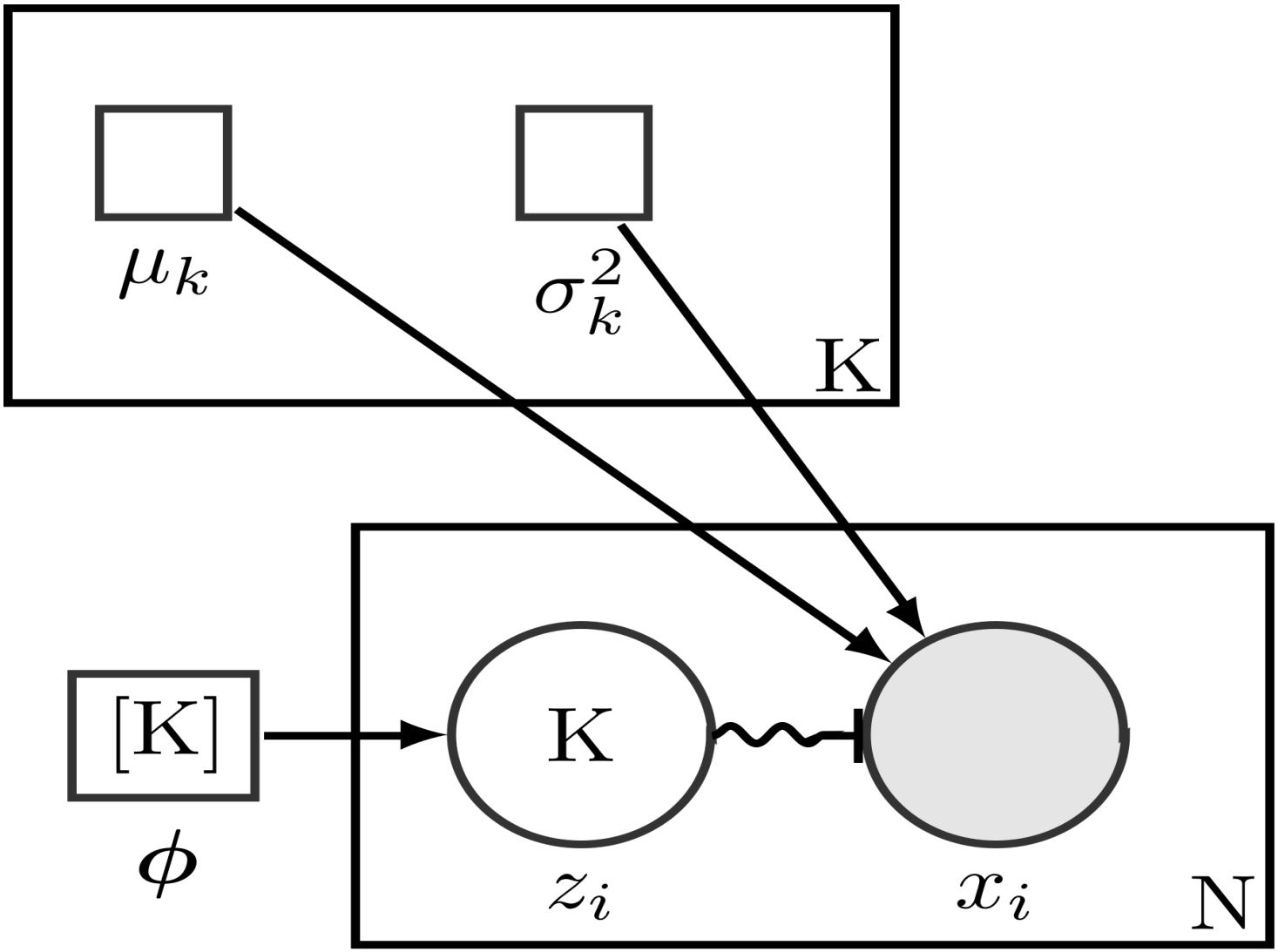
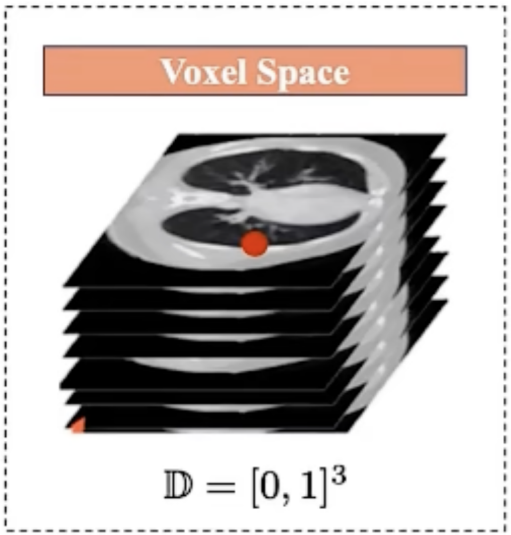
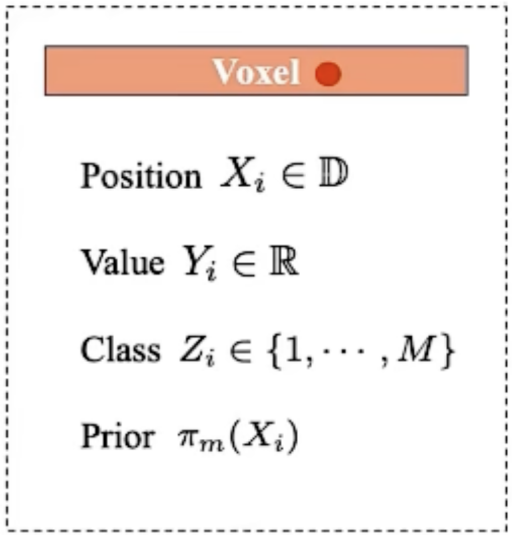


Figure: https://en.wikipedia.org/wiki/Mixture_model

A semiparametric Gaussian mixture model for chest CT-based 3D blood vessel reconstruction Zeng et al., 2024 [3]

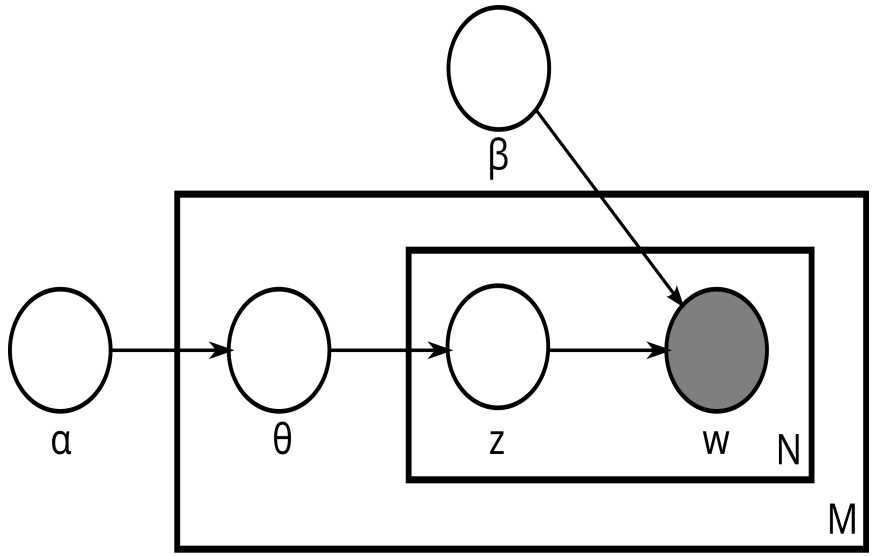




$$Y_i|\{X_i, Z_i = m\} \sim \mathcal{N}(\mu_m(X_i), \sigma_m^2(X_i))$$

Latent Dirichlet Allocation (LDA) Model

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.



Blei, Ng, and Jordan, 2003 [4]

LDA for Predicting Tissue-Specific Functional Effects of Noncoding Variation

For each variants i, there are K tissue-specific functional scores: $\boldsymbol{X}_i = \{x_{i1}, \cdots, x_{iK}\}$ of K functional annotations. There are two latent functional classes. Latent indicator $C_i = 1$ if variant i belongs to the first functional class.

(1) For each tissue j, choose $(1 - \pi_j, \pi_j) \sim \text{Dir}(\alpha_0, \alpha_1)$.
(2) Given π_j, for each variant i with $t_i = j$, choose a class $C_i \sim \text{Bern}(\pi_j)$.
(3) Given $C_1, \ldots, C_m, \mathbf{X}_1, \ldots, \mathbf{X}_m$ are independently generated such that each $\mathbf{X}_i$ is generated from the appropriate multivariate distribution: F_1 if $C_i = 1$ and F_0 otherwise.

Backenroth et al., 2018 [5]

Outline

Why mixture model?

- Capturing Heterogeneity
- Flexibility in Modeling Complex Distributions
- Probabilistic Interpretation

Outline

Variational Autoencoder (VAE)

In the multivariate form, GMM:
$\mathbf{y}_i|z_i = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

$\downarrow$

Assume the latent variable z is sampled from a high-dimensional space $\mathcal{Z}$, parameter θ is sample from $\boldsymbol{\Theta}$: $\mathbf{y}|\{z, \theta\} \sim \mathcal{N}(f(z, \theta), \sigma^2\mathbf{I})$.

$\downarrow$

How do we define z? What does function f represent?

$\downarrow$

We do not need to define z. Assume $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

$\downarrow$

Function f is a multi-layer neural network, which first maps z into some latent values of y, and then maps those latent values to observations.

VAE for Population Genetics

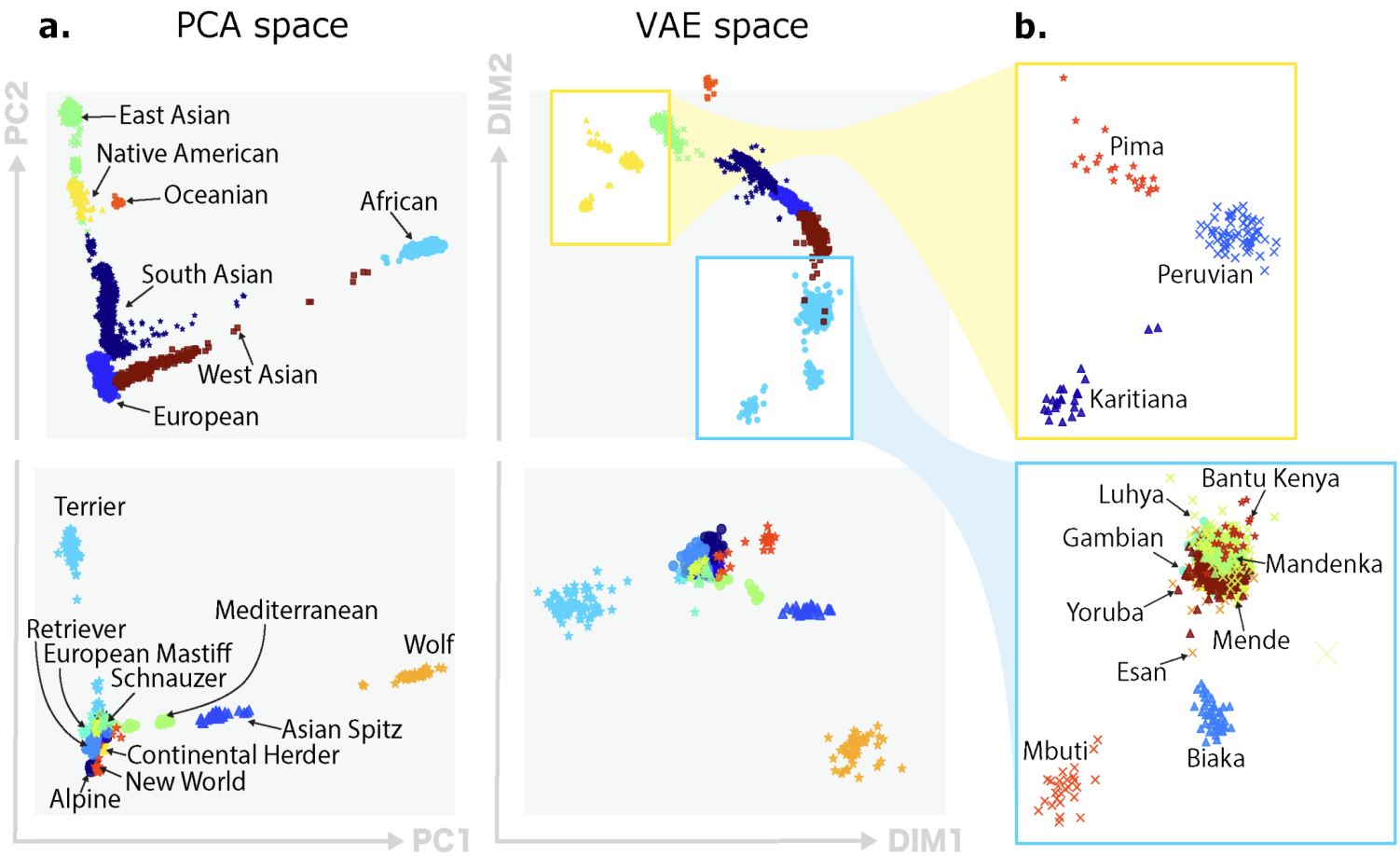


Fig. 1 Qualitative comparison of PCA and VAE projections. (a) The top row illustrates the projections generated by both PCA and VAE for 4,894 human samples using 839,629 SNPs. The second row displays projections of 489 canine samples using 198,473 SNP positions. **(b)** Focus of VAE projections of Native American subpopulations (in yellow), and African subpopulations (in blue).

Geleta et al., 2023 [6]

Thank you!

Questions?


Yujia Li, Tanbin Rahman, Tianzhou Ma, Lu Tang, and George C Tseng.
A sparse negative binomial mixture model for clustering rna-seq count data.
Biostatistics, 24(1):68–84, 2023.


Yujie Jiang, Matthew D Montierth, Kaixian Yu, Shuangxi Ji, Shuai Guo, Quang Tran, Xiaoqian Liu, Seung Jun Shin, Shaolong Cao, Ruonan Li, et al.
Pan-cancer subclonal mutation analysis of 7,827 tumors predicts clinical outcome.
bioRxiv, pages 2024–07, 2024.


Qianhan Zeng, Jing Zhou, Ying Ji, and Hansheng Wang.
A semiparametric gaussian mixture model for chest ct-based 3d blood vessel reconstruction.
Biostatistics, page kxae013, 2024.


David M Blei, Andrew Y Ng, and Michael I Jordan.
Latent dirichlet allocation.

Journal of machine Learning research, 3(Jan):993–1022, 2003.

Daniel Backenroth, Zihuai He, Krzysztof Kiryluk, Valentina Boeva, Lynn Petukhova, Ekta Khurana, Angela Christiano, Joseph D Buxbaum, and Iuliana Ionita-Laza.
Fun-lda: a latent dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications.
The American Journal of Human Genetics, 102(5):920–942, 2018.

Margarita Geleta, Daniel Mas Montserrat, Xavier Giro-i Nieto, and Alexander G Ioannidis.
Deep variational autoencoders for population genetics.
biorxiv, pages 2023–09, 2023.