L_2 distance between Gaussian Mixture Models

Kisung You
kyoustat@gmail.com

November 1, 2019

1 Introduction

Point set refers to a collection of points or observations. When we have multiple of such objects in a common ambient space, including $\mathbb{R}^d$, it is often convenient two represent each point set as a mixture of well-known probability distributions. This enables us to consider point sets as measures so that the analysis becomes operations on the space of measures.

In this note, we focus on a mixture of Gaussians, also known as Gaussiam mixture model (GMM), and a measure of distance in the sense of $L_2(\mathbb{R}^d)$ since it gives us explicit formula to compare two GMM-modeled measures P and Q. Existence of explicit expression makes it a fascinating option compared to other measures which may depend on Monte Carlo integration.

2 Problem Statement

Suppose we have two mixtures of Gaussians

$$P(x) = \sum_{n=1}^{N} \alpha_n N(x|\mu_n, \Sigma_n) \quad \text{and} \quad Q(x) = \sum_{m=1}^{M} \beta_m N(x|\eta_m, \Lambda_m)$$

where α_n and β_m are nonnegative (actually, positive) coefficients that sum to 1 respectively, i.e., $\sum_{n=1}^{N} \alpha_n = 1$ and $\sum_{m=1}^{M} \beta_m = 1$. $N(\cdot|\mu, \Sigma)$ and $N(\cdot|\eta, \Lambda)$ are Gaussian distributions in $\mathbb{R}^d$ with mean vectors μ, η and covariance matrices Σ, Λ respectively. As everyone reading this note is probably well aware, the density function for $N(x|\mu, \Sigma)$ is in form of

$$f(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

with $|\Sigma|$ being determinant of square covariance matrix Σ.

Two distributions are endowed with densities $p(x)$ and $q(x)$ for P and Q thanks to the above parametric forms and they are smooth functions over $\mathbb{R}^d$. This enables to define L_2 distance between two distributions,

$$L_2(P, Q) = \left\{ \int_{\mathbb{R}^d} (p(x) - q(x))^2 dx \right\}^{1/2}. \tag{1}$$

For notational simplicity, we will denote $p_n(x)$ and $q_m(x)$ for densities of P_n and Q_m respectively.

3 Derivation

We want to compute L_2 distance between two distributions P and Q. This can be written as

$$\begin{aligned}
L_2^2(P,Q) &= \int_{\mathbb{R}} (p(x) - q(x))^2 dx \\
&= \int \left(\sum_{n=1}^{N} \alpha_n p_n(x) - \sum_{m=1}^{M} \beta_m q_m(x) \right)^2 dx \\
&= \sum_{n,n'} \alpha_n \alpha_{n'} \int p_n(x) p_{n'}(x) dx + \sum_{m,m'} \beta_m \beta_{m'} \int q_m(x) q_{m'}(x) dx \\
&\quad - 2 \sum_{n,m} \alpha_n \beta_m \int p_n(x) q_m(x) dx
\end{aligned}$$

which requires to evaluate integral for a product of two Gaussian densities,

$$\int p_n(x) q_m(x). \tag{*}$$

From the formula to be introduced in the next section, we have following equalities,

$$\begin{aligned}
A_{n,n'} &= \int p_n(x) p_{n'}(x) dx = N(\mu_n | \mu_{n'}, \Sigma_n + \Sigma_{n'}) \\
B_{m,m'} &= \int q_m(x) q_{m'}(x) dx = N(\eta_m | \eta_{m'}, \Lambda_m + \Lambda_{m'}) \\
C_{n,m} &= \int p_n(x) q_m(x) dx = N(\mu_n | \eta_m, \Sigma_n + \Lambda_m).
\end{aligned}$$

Therefore,

$$L_2(P,Q) = \left\{ \sum_{n,n'} \alpha_n \alpha_{n'} A_{n,n'} + \sum_{m,m'} \beta_m \beta_{m'} B_{m,m'} - 2 \sum_{n,m} \alpha_n \beta_m C_{n,m} \right\}^{1/2}$$

Fact (*)

According to Section 8.1.8 of Matrix Cookbook [1], product of two gaussian densities is in form of following,

$$N(x | m_1, \Sigma_1) \cdot N(x | m_2, \Sigma_2) = c_c \cdot N(x | m_c, \Sigma_c)$$

where

$$\begin{aligned}
c_c &= N(m_1 | m_2, (\Sigma_1 + \Sigma_2)) \\
m_c &= (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} (\Sigma_1^{-1} m_1 + \Sigma_2^{-1} m_2) \\
\Sigma_c &= (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}.
\end{aligned}$$

From the fact that integral of density function is 1, we have

$$\int N(x | m_1, \Sigma_1) \cdot N(x | m_2, \Sigma_2) dx = c_c.$$

References

[1] K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. Version 20121115.