

Quantum Clustering and Gaussian Mixtures

Mahajabin Rahman

Davi Geiger

Abstract

The mixture of Gaussian distributions, a soft version of k -means ([2]), is considered a state-of-the-art clustering algorithm. It is widely used in computer vision for selecting classes, e.g., color[4, 1, 5], texture[1, 9], shapes [12, 10]. In this algorithm, each class is described by a Gaussian distribution, defined by its mean and covariance. The data is described by a weighted sum of these Gaussian distributions.

We propose a new method, inspired by quantum interference in physics [7]. Instead of modeling each class distribution directly, we model a **class wave function** such that its magnitude square is the class Gaussian distribution. We then mix the **class wave functions** to create the **mixture wave function**. The final mixture distribution is then the magnitude square of the **mixture wave function**. As a result, we observe the quantum class interference phenomena, not present in the Gaussian mixture model. We show that the quantum method outperforms the Gaussian mixture method in every aspect of the estimations. It provides more accurate estimations of all distribution parameters, with much less fluctuations, and it is also more robust to data deformations from the Gaussian assumptions. We illustrate our method for color segmentation as an example application.

1 Introduction

Given a large set of non labeled data, being able to separate it in clusters is of great interest in computer vision (e.g., [2, 4, 1, 5, 9, 12, 10].) Clustering may also be interpreted in a Bayesian method as the data likelihood model. When combined with the prior model on the data (such as encouraging neighbors to be labeled similarly), one forms the posterior probability.

This paper is solely focused on the clustering model and in particular we focus on data generated from multiple Gaussian distributions, each one characterized by mean and covariances. The Gaussian Mixture Model (GMM) is a standard model to describe such data and combined with the expectation-maximization algorithm is the standard approach to recover the Gaussian parameters as well as the number of points per class (see [2]). Inspired by quantum methods, e.g., [7], we reformulate this classical model (GMM) and produce a quantum model. Using the theory of expectation-maximization we recover the Gaussian parameters as well as the number of points per class.

The inspiration for our proposed method comes from the wave interference phenomena and previous work [6]. In the classical model, as one mixes distributions (such as a Gaussian mixture - GMM), no probabilities are canceled. By formulating the distributions in terms of wave functions, such that their magnitude square describe the probability distribution, we allow their mixtures to interfere. The effect is the cancellation of probabilities, which allows us to identify the clusters of data more accurately.

2 Classical Clustering

The input data is given by a set of N points $\{p_1, p_2, \dots, p_N\}$. If we are in two dimensions, our default point is, unless specified, $p_i = (x_i, y_i)$ and if we are in three dimensions $p_i^{3D} = (x_i, y_i, z_i)$.

Given K classes, each class data is produced by a Gaussian distribution with center μ_k and Covariance \mathbf{C}_k , with $k = 1, \dots, K$, i.e., thus the probability of p_i given class k is the following:

$$P(p_i|k, \mu_k, \mathbf{C}_k) = \frac{1}{Z_k} e^{-\frac{1}{2} (p_i - \mu_k)^T \mathbf{C}_k^{-1} (p_i - \mu_k)}.$$

where $Z_k = (2\pi)^{d/2} |\mathbf{C}_k|^{1/2}$ is a normalization factor with $|\mathbf{C}_k|^{1/2} = \sqrt{\det \mathbf{C}_k}$ and d is the data dimensionality.

Suppose from the set of N points originated from K classes, there are N_k points per class ($N = \sum_{k=1}^K N_k$). Since N_k is assumed to be unknown, according to the Gaussian mixture method, the final distribution is given by a weighted sum of each class distribution of all N points

$$\begin{aligned} P(p_i|\{\theta_k\}) &= \sum_{k=1}^K P(p_i, k|\{\theta_k\}) = \sum_{k=1}^K P(k|\{\theta_k\}) P(p_i|k, \{\theta_k\}) \\ &= \sum_{k=1}^K \left[P(k|\{\theta_k\}) \frac{1}{\sqrt{(2\pi)^d |\mathbf{C}_k|}} e^{-\frac{(p_i - \mu_k)^T \mathbf{C}_k^{-1} (p_i - \mu_k)}{2}} \right] \end{aligned} \quad (1)$$

where $P(k|\{\theta_k\})$ is the prior probability that a point belongs to class k , and $\theta_k = \{\mathbf{C}_k, \mu_k\}$ are the Gaussian parameters of class k . The clustering problem is to find $P(p_i|\{\theta_k\})$ from the data, i.e., to extract the parameters $\theta = \{P(k|\{\theta_k\}), \theta_k; k = 1, \dots, K\} = \{P(k|\{\theta_k\}), \mathbf{C}_k, \mu_k; k = 1, \dots, K\}$ from the data. A standard method to extract these parameters is based on the EM method.

2.1 EM Method

In appendix B, we show a brief derivation of the EM method that leads to the iteration procedure $t = 1, \dots, T$ with the following two steps

E-step:

$$Q_i^t(k) = P(k|p_i, \{\theta_k^{t-1}\}) = \frac{P(p_i, k|\{\theta_k^{t-1}\})}{P(p_i|\{\theta_k^{t-1}\})} = \frac{P(p_i, k|\{\theta_k^{t-1}\})}{\sum_l P(p_i, l|\{\theta_k^{t-1}\})} \quad (2)$$

For the Gaussian mixture, we get

$$Q_i^t(k) \equiv Q_{i,k}^t = \frac{P^{t-1}(k|\{\theta_k^{t-1}\}) \frac{1}{Z_k^{t-1}} e^{-\frac{(p_i - \mu_k^{t-1})^T (\mathbf{C}_k^{-1})^{t-1} (p_i - \mu_k^{t-1})}{2}}}{\sum_{l=1}^K P^{t-1}(l|\{\theta_k^{t-1}\}) \frac{1}{Z_l^{t-1}} e^{-\frac{(p_i - \mu_l^{t-1})^T (\mathbf{C}_l^{-1})^{t-1} (p_i - \mu_l^{t-1})}{2}}} \quad (3)$$

Followed by

M-Step: The maximization of

$$\begin{aligned} [P^t(k), \theta_k^t] &= \arg \max_{P(k), \{\theta_k\}} \sum_i \sum_k Q_i^t(k) \log \frac{P(p_i, k|\{\theta_k^{t-1}\})}{Q_i^t(k)} \\ &= \arg \max_{P(k), \{\theta_k\}} \sum_i \sum_k Q_i^t(k) \log P(p_i, k|\{\theta_k^{t-1}\}) \end{aligned} \quad (4)$$

For the Gaussian mixture ($\theta_k = (\mathbf{C}_k, \mu_k)$) we get

$$\begin{aligned} N_k^t &= \sum_i Q_{i,k}^t \\ \Downarrow & N_k^t - \text{estimated number of points, at step } t, \text{ associated to class } k \\ P^t(k|\{\theta_k^{t-1}\}) &= \langle P^t(k|p_i, \{\theta_k^{t-1}\}) \rangle_N = \frac{N_k^t}{\sum_{l=1}^K N_l^t} \\ \mu_k^t &= \frac{1}{N_k^t} \sum_{i=1}^N Q_{i,k}^t p_i \\ \mathbf{C}_k^t &= \frac{1}{N_k^t} \sum_{i=1}^N Q_{i,k}^t (p_i - \mu_k^t)(p_i - \mu_k^t)^T \end{aligned} \quad (5)$$

where $\langle f(p_i) \rangle_N$ indicates the expectation of the function f over all N points $\{p_i\}$ (with respect to $P(p_i, k|\{\theta_k^{t-1}\})$). The Gaussian normalization $Z_k = \sqrt{(2\pi)^d |\mathbf{C}_k|}$ is also required to be computed.

3 Quantum Clustering

We introduce a wave function to describe the system of points. It is such that the probability associated to the system of points is the magnitude square of the wave function. The wave function associated to a point p_i from a class k is given by

$$\psi_k(p_i|\theta_k) \equiv \frac{1}{\sqrt{Z_k}} e^{-\frac{1}{4}(p_i - \mu_k) \mathbf{C}_k^{-1} (p_i - \mu_k)} e^{-i \phi_k},$$

where $Z_k = (2\pi)^{d/2} |\mathbf{C}_k|^{1/2}$. From this definition we then recover the classical probability distribution for a given class $P(p_i|k, \theta_k) = |\psi_k(p_i|\theta_k)|^2 = \frac{1}{Z_k} e^{-\frac{1}{2}(p_i - \mu_k) \mathbf{C}_k (p_i - \mu_k)}$. The new phase factor, $e^{-i \phi_k}$, does not alter the class distribution, but will have its impact in the mixture case.

We extend the quantum method when the data is described by a mixture of classes. Analogously to the mixture of Gaussians, we now mix quantum waves to produce the final wave, i.e.,

$$\psi(p_i|\{\alpha_k, \theta_k\}) = \sum_{k=1}^K \alpha_k \psi_k(p_i|\theta_k) = \sum_{k=1}^K \left[\frac{\alpha_k}{\sqrt{Z_k}} e^{-\frac{1}{4}(p_i - \mu_k) \mathbf{C}_k^{-1} (p_i - \mu_k)} e^{-i \phi_k} \right] \quad (6)$$

The final probability $P(p_i)$ (short for $P(p_i|\{\alpha_k, \theta_k\})$, with $\theta_k = (\mu_k, \mathbf{C}_k, \phi_k)$), is given by

$$\begin{aligned} P(p_i) &= |\psi(p_i|\{\alpha_k, \theta_k\})|^2 \\ &= \sum_{k=1}^K \left[\frac{\alpha_k}{\sqrt{Z_k}} e^{-\frac{1}{4}(p_i - \mu_k) \mathbf{C}_k (p_i - \mu_k)} \sum_{l=1}^K \frac{\alpha_l^*}{\sqrt{Z_l}} \cos \phi_{l,k}(p_i) e^{-\frac{1}{4}(p_i - \mu_l) \mathbf{C}_l (p_i - \mu_l)} \right] \end{aligned} \quad (7)$$

where α_l^* is the conjugate of α_l , and

$$\phi_{l,k} = \phi_k - \phi_l \quad (8)$$

and we must require that the set $\{\alpha_k\}$ satisfy $1 = \sum_i P(p_i)$. The final probability $P(p_i|\{\alpha_k, \theta_k\})$ exhibits wave interference phenomena, not present in the classical clustering formulation, due to $\cos \phi_{l,k}$ for $l \neq k$.

Following the Bayesian interpretation, $P(p_i|\{\alpha_k, \theta_k\}) = \sum_k P(p_i, k|\{\alpha_k, \theta_k\})$, where $\theta_k = (\mu_k, \mathbf{C}_k, \phi_k)$, the join distribution $P(p_i, k|\{\alpha_k, \theta_k\})$ is derived from equation 8

$$\begin{aligned} P(p_i, k|\{\alpha_k, \theta_k\}) &= \frac{\alpha_k}{\sqrt{Z_k}} e^{-\frac{1}{4}(p_i - \mu_k) \mathbf{C}_k (p_i - \mu_k)} \sum_{l=1}^K \frac{\alpha_l^*}{\sqrt{Z_l}} \cos \phi_{l,k}(p_i) e^{-\frac{1}{4}(p_i - \mu_l) \mathbf{C}_l (p_i - \mu_l)} \\ &\text{Defining the unnormalized Gaussian model :} \\ G_{i,k} &= \frac{1}{\sqrt{Z_k}} e^{-\frac{1}{4}(p_i - \mu_k) \mathbf{C}_k^{-1} (p_i - \mu_k)} \quad (\text{where } G_{i,k}^2 \text{ is normalized}) \\ &= \alpha_k G_{i,k} \left(\sum_{l=1}^K \alpha_l^* \cos \phi_{l,k} G_{i,l} \right) \end{aligned} \quad (9)$$

with the constraint on $\{\alpha_k\}$ that $1 = \sum_i \sum_k P(p_i, k|\{\alpha_k, \theta_k\})$.

Classical Case: For $\phi_{l,k} = \frac{\pi}{2} \forall l \neq k$, we recover the classical case. This can only happen if we have just two classes, otherwise for at least one pair of classes the phase difference will not be $\frac{\pi}{2}$.

3.1 EM Method for Quantum Clustering

E-step: From equation 2

$$\begin{aligned} Q_i^t(k) &= P(k|p_i, \{\theta_k^{t-1}\}) = \frac{P(p_i, k|\{\theta_k^{t-1}\})}{P(p_i|\{\theta_k^{t-1}\})} = \frac{P(p_i, k|\{\theta_k^{t-1}\})}{\sum_m P(p_i, m|\{\theta_m^{t-1}\})} \\ &= \left[\frac{\alpha_k G_{i,k} \left(\sum_{l=1}^K \alpha_l^* \cos \phi_{l,k} G_{i,l} \right)}{\sum_m \alpha_m G_{i,m} \left(\sum_{l=1}^K \alpha_l^* \cos \phi_{l,m} G_{i,l} \right)} \right]^{t-1} \end{aligned} \quad (10)$$

Note that, like in the classical case, the estimated number of points, at step t , associated to class k is given by

$$N_k^t = \sum_{i=1}^N P(k|p_i, \{\theta_k^{t-1}\}) = \sum_{i=1}^N Q_{i,k}^t \quad (11)$$

M-Step: The maximization of

$$\begin{aligned} (\alpha_k^t, \theta_k^t) &= \arg \max_{\{\alpha_k, \theta_k\}} \sum_i \sum_k Q_i^t(k) \log \frac{P(p_i, k | \{\theta_k^{t-1}\})}{Q_i^t(k)} \\ &= \arg \max_{\{\alpha_k, \theta_k\}} \sum_i \sum_k Q_i^t(k) \log P(p_i, k | \{\theta_k^{t-1}\}) \end{aligned} \quad (12)$$

where $\theta_k = (\mu_k, \mathbf{C}_k, \phi_k)$.

3.2 Two Classes and Constraints

For a two class model, $k = 1, 2$, without loss of generality, we can consider $\alpha_1, \alpha_2 \in \mathbb{R}$ since their possible phase is already absorbed by ϕ_1 and ϕ_2 . Then,

$$\begin{aligned} P(p_i, 1 | \{\alpha_k, \theta_k\}) &= \alpha_1^2 G_{i,1}^2 + \alpha_1 \alpha_2 G_{i,1} G_{i,2} \cos \phi \\ P(p_i, 2 | \{\alpha_k, \theta_k\}) &= \alpha_2^2 G_{i,2}^2 + \alpha_1 \alpha_2 G_{i,1} G_{i,2} \cos \phi \end{aligned} \quad (13)$$

with $\cos \phi = \cos \phi_{1,2} = \cos \phi_{2,1}$

Constraints: The probability $P(p_i, k | \{\theta_k^{t-1}\})$ must satisfy

$$0 \leq P(p_i, k | \{\theta_k^{t-1}\}) \leq 1$$

and

$$1 = \sum_i \sum_k P(p_i, k | \{\theta_k^{t-1}\}) \Rightarrow 1 = \alpha_1^2 + \alpha_2^2 + 2 \alpha_1 \alpha_2 \cos \phi \sum_i G_{i,1} G_{i,2}.$$

E-step: Replacing $\cos \phi = \frac{1 - \alpha_1^2 - \alpha_2^2}{2 \alpha_1 \alpha_2 \sum_i G_{i,1} G_{i,2}}$ into equation 10 and the E-step becomes

$$\begin{aligned} Q_i^t(1) &= \left[\frac{\alpha_1^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i}{\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i} \right]^{t-1} \\ Q_i^t(2) &= \left[\frac{\alpha_2^2 G_{i,2}^2 + \frac{1}{2} (\alpha o)_i}{\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i} \right]^{t-1} \end{aligned} \quad (14)$$

where

$$o_i = \frac{G_{i,1} G_{i,2}}{\sum_i G_{i,1} G_{i,2}} \quad \text{and} \quad (\alpha o)_i = (1 - \alpha_1^2 - \alpha_2^2) o_i \quad (15)$$

gives measures of overlap of the distributions.

M-step: Replacing $\cos \phi = \frac{1 - \alpha_1^2 - \alpha_2^2}{2 \alpha_1 \alpha_2 \sum_i G_{i,1} G_{i,2}}$ into the M-step equation 12 gives us the objective function

$$O(\{\alpha_k, \theta_k\}) = \sum_i \sum_k Q_i(k) \log P(p_i, k | \{\alpha_k, \theta_k\}) \quad (16)$$

and so

$$(\alpha_k^t, \theta_k^t) = \arg \max_{\{\alpha_k, \theta_k\}} \sum_i \{Q_i^t(1) \log [2 \alpha_1^2 G_{i,1}^2 + (\alpha o)_i] + Q_i^t(2) \log [2 \alpha_2^2 G_{i,2}^2 + (\alpha o)_i]\} \quad (17)$$

where in the maximization we must satisfy the constraint

$$-1 \leq \cos \phi = \frac{1 - \alpha_1^2 - \alpha_2^2}{2 \alpha_1 \alpha_2 \sum_i G_{i,1} G_{i,2}} \leq 1. \quad (18)$$

Define the vectors

$$\vec{\alpha}_k = \alpha_k \hat{g}_k \quad \text{and} \quad (\hat{g}_k) = (G_{1,k}, G_{2,k}, \dots, G_{i,k}, \dots, G_{N,k})^T \quad (19)$$

where $\hat{g}_k \cdot \hat{g}_k = \sum_{i=1}^N G_{i,k} G_{i,k} = \sum_{i=1}^N G_{i,k}^2 = 1$.

Thus, $\vec{\alpha}_k \cdot \vec{\alpha}_k = \alpha_k^2$ and $\vec{\alpha}_1 \cdot \vec{\alpha}_2 = \alpha_1 \alpha_2 \cos \varphi_{1,2} = \alpha_1 \alpha_2 \sum_{i=1}^N G_{i,1} G_{i,2}$. Therefore the constraint 18 can be written as

$$\min\{(\vec{\alpha}_1 - \vec{\alpha}_2)^2, (\vec{\alpha}_1 + \vec{\alpha}_2)^2\} \leq 1 \leq \max\{(\vec{\alpha}_1 - \vec{\alpha}_2)^2, (\vec{\alpha}_1 + \vec{\alpha}_2)^2\} \quad (20)$$

3.3 Gradient Equations for the M-Step

The maximization step can be achieved either at the constraint boundaries or where the gradient with respect to the free parameters are zero. We now compute these gradients of the objective function $O(\{\alpha_k, \theta_k\})$ (given in equation 16) with respect to each of its parameters (α_k, θ_k) and investigate when the gradient vanishes. Some of the detail calculations are presented in Appendix A

$$\begin{aligned} \mu_k^t &\rightarrow \partial_{\mu_k} O(\{\alpha_k, \theta_k\}) = \sum_i \sum_k Q_i^t(k) \frac{\partial_{\mu_k} P(p_i, k | \{\theta_k^{t-1}\})}{P(p_i, k | \{\theta_k^{t-1}\})} = 0 \\ \mathbf{C}_k^t &\rightarrow \partial_{\mathbf{C}_k} O(\{\alpha_k, \theta_k\}) = \sum_i \sum_k Q_i^t(k) \frac{\partial_{\mathbf{C}_k} P(p_i, k | \{\theta_k^{t-1}\})}{P(p_i, k | \{\theta_k^{t-1}\})} = 0 \\ \alpha_k &\rightarrow \partial_{\alpha_k} O(\{\alpha_k, \theta_k\}) = \sum_i \sum_k Q_i^t(k) \frac{\partial_{\alpha_k} P(p_i, k | \{\theta_k^{t-1}\})}{P(p_i, k | \{\theta_k^{t-1}\})} = 0 \end{aligned} \quad (21)$$

μ_k : derived in 34 and 35

$$\mu_k = \frac{\sum_i F_{i,k} p_i}{\sum_{j=1}^N F_{i,k}} \quad (22)$$

where $F_{i,k} = Q_i(k) - o_i \sum_{j=1}^N \frac{\frac{1}{2}(\alpha o)_i}{\alpha_1^2 G_{j,1}^2 + \alpha_2^2 G_{j,1}^2 + \frac{1}{2}(\alpha o)_j}$

\mathbf{C}_k : derived in 41 and 42

$$\mathbf{C}_k = \frac{\sum_i R_{i,k}^t (p_i - \mu_k) (p_i - \mu_k)^T}{\sum_i R_{i,k}^t} \quad (23)$$

where $R_{i,k}^t = F_{i,k} + \frac{\alpha_k^2 G_{i,k}^2}{\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,1}^2 + \frac{1}{2}(\alpha o)_i}$

α_k :

for ∂_{α_1} derived in 31

$$\begin{aligned}
0 &= \sum_i^N Q_{i,1}^t \frac{(2 G_{i,1}^2 - o_i)}{[2 \alpha_1^2 G_{i,1}^2 + (\alpha o)_i]} - Q_{i,2}^t \frac{o_i}{[2 \alpha_2^2 G_{i,2}^2 + (\alpha o)_i]} \\
&\text{applying equation 14} \\
&= \sum_i^N \frac{(G_{i,1}^2 - o_i)}{[\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i]}
\end{aligned} \tag{24}$$

for ∂_{α_2} derived in 32

$$\begin{aligned}
0 &= \sum_i^N -Q_{i,1}^t \frac{o_i}{[2 \alpha_1^2 G_{i,1}^2 + (\alpha o)_i]} + Q_{i,2}^t \frac{(2 G_{i,2}^2 - o_i)}{[2 \alpha_2^2 G_{i,2}^2 + (\alpha o)_i]} \\
&\text{applying equation 14} \\
&= \sum_i^N \frac{(G_{i,2}^2 - o_i)}{[\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i]}
\end{aligned} \tag{25}$$

where o_i and $(\alpha o)_i$ are measures of overlap given by equations 15

3.4 The Classical Limit

For the case $\phi = \frac{\pi}{2}$, the probabilities (13) becomes

$$\begin{aligned}
P(p_i, 1 | \{\alpha_k, \theta_k\}) &= \alpha_1^2 G_{i,1}^2 \\
P(p_i, 2 | \{\alpha_k, \theta_k\}) &= \alpha_2^2 G_{i,2}^2
\end{aligned} \tag{26}$$

and $\alpha_1^2 + \alpha_2^2 = 1 \rightarrow (\alpha o)_i = 0$, $F_{i,k} = Q_i(k) \rightarrow \mu_k = \frac{\sum_i Q_{i,k} p_i}{\sum_i Q_{i,k}}$, and $R_{i,k}^t = 2 Q_{i,k} \rightarrow \mathbf{C}_k = \frac{\sum_i Q_{i,k} (p_i - \mu_k)(p_i - \mu_k)^T}{\sum_i Q_{i,k}}$. We recover the classical mixture model. The two class case allows us to compare the performance of the two models, the quantum and the classical.

4 Experiments and Analysis

We restrict the analysis to the two class cases. The main reason is to compare the performance (empirically as well as theoretically) to the classical case. We first devise experiments with 2D randomly controlled data sets to fully analyze the statistical properties of the quantum inspired method vs the classical mixture of Gaussians (state of the art EM method in clustering). We then devise color segmentation experiments with 3D data (see section 4.4). to illustrate how the technique can be applied and outperform state of the art techniques in a computer vision application.

4.1 Statistical Analysis of Controlled Data Sets

We generated random data $p_i^k = (x_i, y_i)$ with mean $\mu^k = [\mu_x^k \ \mu_y^k]^T$ and covariance $\mathbf{C} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} (\sigma_x^2)^k & 0 \\ 0 & (\sigma_y^2)^k \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$, where θ is the rotation of the x-axis where σ_x^k is the standard deviation along x for class k . In all the tests we considered $N_1 = 500$ and $N_2 = 500$ or 1000 points per class. We refer to these parameters μ 's, σ 's, N 's as the ground truth of the experiments, as the samples come from single Gaussians with these parameters and with that many samples. This is to say that we did not sample from the mixture of Gaussians model. The goal of each experiment was to compare the performance of the quantum model and the classical one, by estimating the number of points of each class, the center of each class and the eigenvalues of each covariance (the variances). The estimations allow us to understand which model best identifies the original Gaussian distribution parameters and number of points (we refer to these parameters as the ground truth).

To account for the random process, 50 trials were produced for each combination of variables. The estimated mean per trial $i = 1, \dots, 50$ is represented as $[\mu_{x,i}^k, \mu_{y,i}^k]^T$, where

$$\bar{\mu}_x^k = \frac{1}{50} \sum_{i=1}^{50} \mu_{x,i}^k \quad (27)$$

In order to compare the estimated mean per trial to the true value of the mean $[\mu_x^{k*}, \mu_y^{k*}]$, we show the error, calculated as:

$$Error(\mu_x^k) = \sqrt{(\mu_x^{k*} - \bar{\mu}_x^k)^2} \quad (28)$$

and we refer to the fluctuation of this error as

$$Fluctuation(\mu_x^k) = \bar{\eta}_x^k = \sqrt{\frac{1}{50} \sum_{i=1}^{50} (\mu_{x,i}^k - \bar{\mu}_x^k)^2} \quad (29)$$

The fluctuation does not depend on the ground truth, as it characterizes the variation of the statistical experiments.

This error and fluctuations were calculated for all the parameters, namely the means (or centers) $\mu_{x,y}^k$ (for x and y separately), variances $(\sigma_{x,y}^k)^2$ (for x and y separately), and the number of points per class (N1 and N2).

4.1.1 Tests

We arbitrarily fixed $N^1 = 500$ and $N^2 = 1000$ to generate the data sets, so the sizes of each class are different and large enough for statistical measurement. The Gaussian parameters to generate the $N^1 = 500$ data points for the first class were $\mu_{x,y}^1 = 0, \sigma_{x,y}^1 = 3$. The center of the second Gaussian distribution is varied starting with $\mu_{x,y}^2 = 10$ and gradually brought closer together to $\mu_{x,y}^1 = 0$ to create greater overlap. As we brought the second center towards the first center, we repeated experiments for $\sigma_{x,y}^2 = 4$ and 5 (see some examples of data sets in figure 1.)

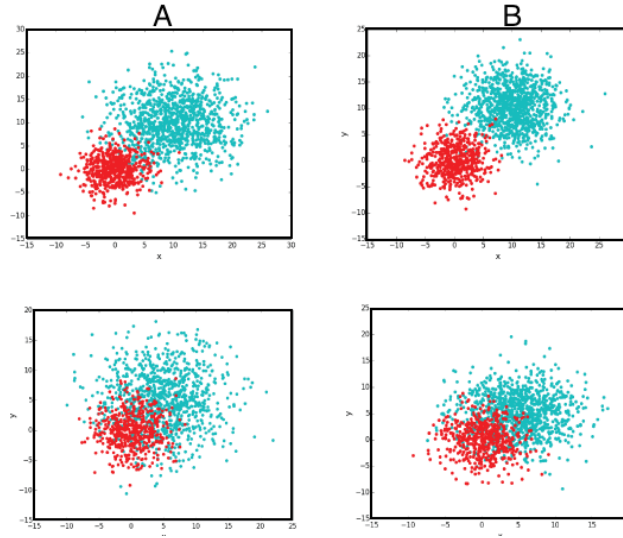


Figure 1: For all four images there are 500 red dots in Class 1 (N1), and 1000 blue-green dots in Class 2 (N2). Also, for all four images $\mu_{x,y}^1 = (0,0)$ and $\sigma_{x,y}^1 = 3$. First column, A: $\sigma_{x,y}^2 = 5$. Second column, B: $\sigma_{x,y}^2 = 4$. Top row: 'separated' distributions with $\mu_{x,y}^2 = 10$, and Bottom row: overlapping distributions with $\mu_{x,y}^2 = 5$.

For each experiment there were 50 trials (see figure 2). We then plot the errors (given by equation 28) and fluctuations (given by equation 29) for each experiment (see figures 3 and 4).

••• Classical Estimations
 ••• Quantum Estimations
 ••• Ground Truth

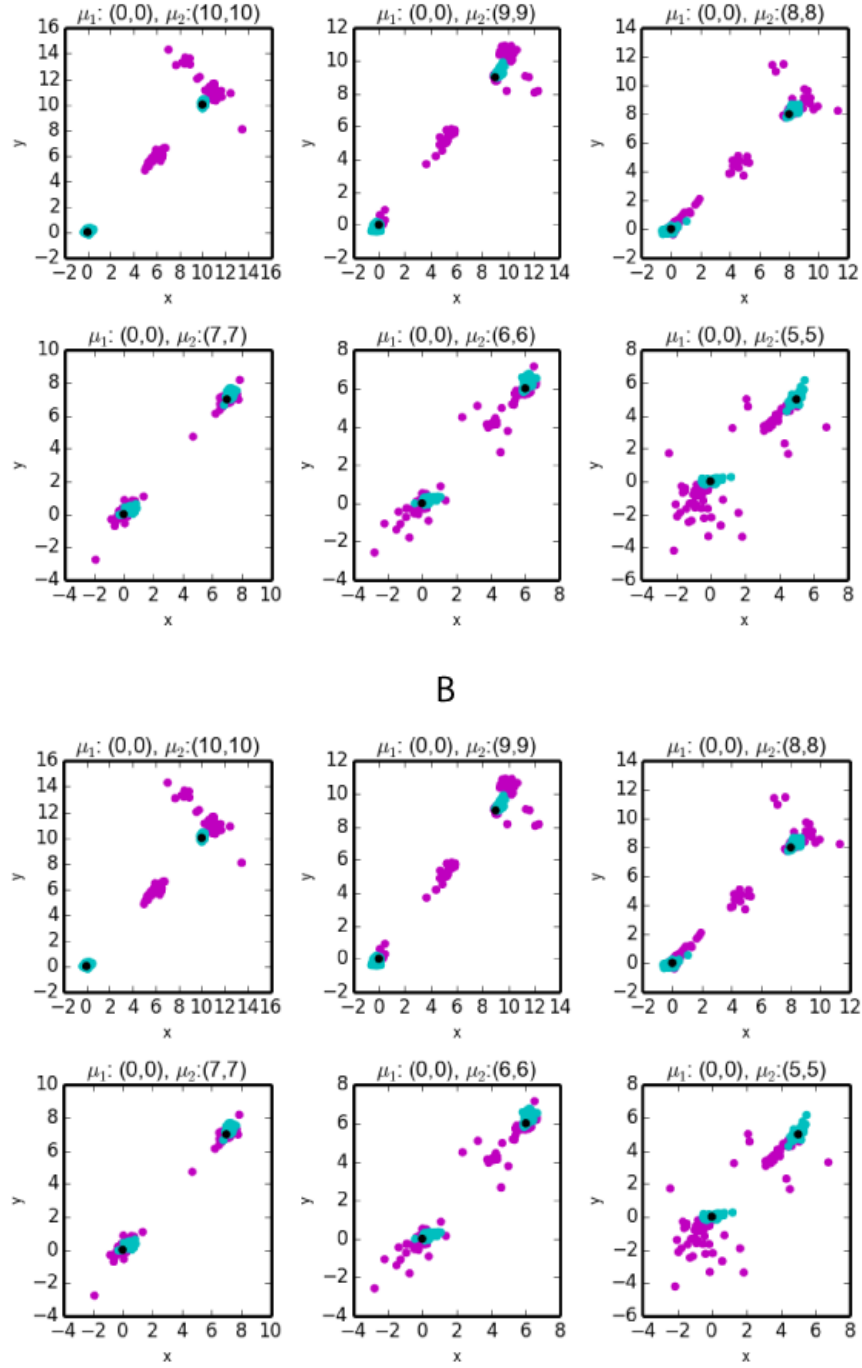
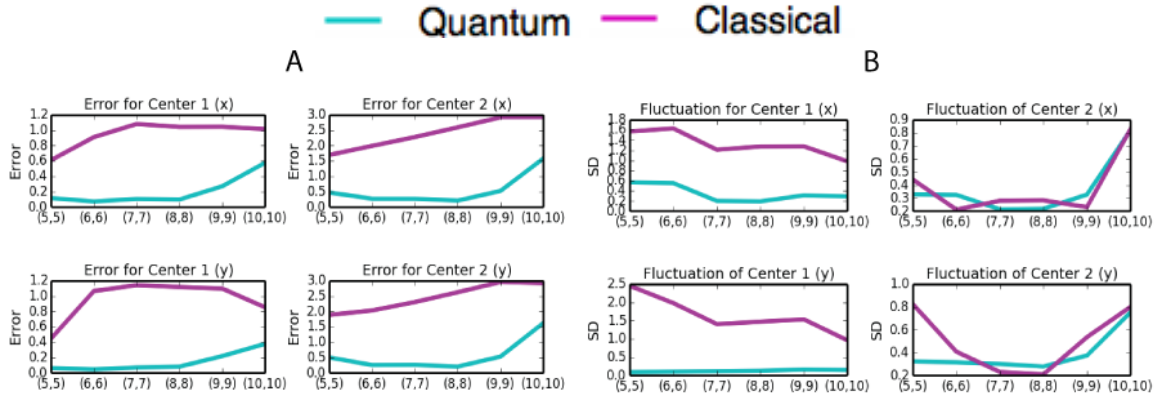
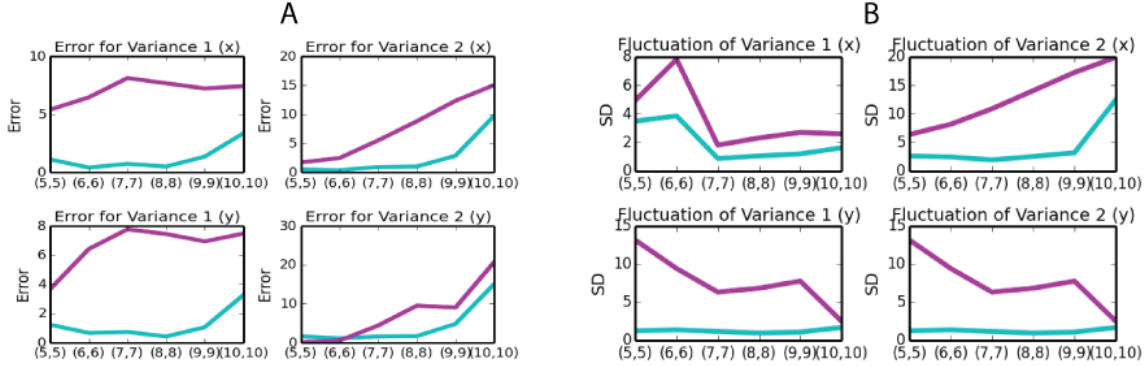


Figure 2: Each dot represents a center estimation for each trial. Purple dots represent estimation for the classical method while Cyan dots represents estimation from the Quantum method. The two Black dots per image represent the ground truth centers. For all 12 images, $\mu_{x,y}^1 = 0$, $\sigma_{x,y}^1 = 3$, $N_1 = 500$ and $N_2 = 1000$. A shows $\sigma_{x,y}^2 = 5$ and B shows $\sigma_{x,y}^2 = 4$. Both, A and B, show results as $\mu_{x,y}^2$ vary from 10 to 5. It is noticeable how the trials for the quantum method almost always provide good answers (and it is robust), while the classical method has less accuracy and much more variation. We quantify this assertion in figure 3

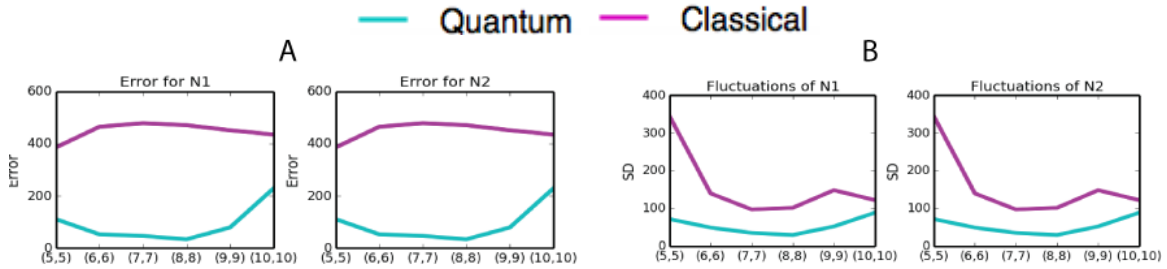


Results for centers



Results for variances

Figure 3: Y-axis: Column A.) Estimated error (given by equation 28) and Column B.) standard deviation (fluctuations)(given by equation 29). Top two rows are for both centers and x and y coordinates. Bottom two rows are for both variances and x and y coordinates. The X-axis for all graphs: varying $\mu_{x,y}^2$ values from (5,5) to (10,10). In these trials, the true $\mu_{x,y}^1 = 0$, and variances, $(\sigma_{x,y}^1)^2 = 9$ and $(\sigma_{x,y}^2)^2 = 25$. The quantum method proves to be more accurate and robust, as it gives a lower error and fluctuates less. As $\mu_{x,y}^2$ increases, there is less overlap between the two distributions. In the overlapping regimes, the highly fluctuating results of the classical case indicate that the method is unable to identify the parameters of the separate groups. We see the effects of this on the error curve, as the average of highly fluctuating estimated variables can be closer the ground truth.

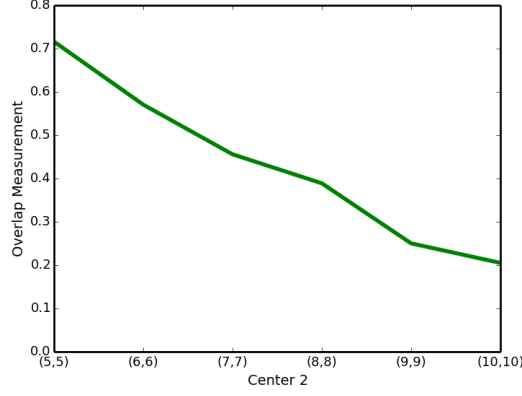


Results for number of points

Figure 4: Y-axis: Estimated error for N^1 , N^2 calculated using equation 28 and standard deviation calculated using equation 29 vs X-axis: $\mu_{x,y}^2$ is varied. Because the quantum method returns a lower error and standard deviation (fluctuations), it is the more accurate and robust method. As the data is more separated, the difference between the methods is reduced. Estimated error values for $N1$ and $N2$ are identical since $N^1 + N^2 = N$ is known/constant.

4.1.2 Overlap and Interference

We argue that the quantum method utilizes the phase parameter to help capture the overlapping of the two distributions. In the quantum method the phase differences between classes yields the interference phenomena, which is stronger as the data overlaps. We want to demonstrate that the interference better captures the overlapped data. For this, through out the experiments described above, we measure the overlap between the data as shown in figure 5 and we show the phase difference (measured by the cosine of it in the table of figure 5).



Overlap as distributions separate

Overlap	0.716	.571	0.456	0.389	0.250	.205
$\cos \phi$	3.59E-6	8.68E-6	7.01E-5	5.24E-5	5.93E-5	0.173

Figure 5: Y-axis: Overlap $\sum_i G_{i,1} G_{i,2}$, as $\mu_{x,y}^2$ varies on the X-axis for the fixed parameters $\mu_{x,y}^1 = 3$, $\sigma_{x,y}^1 = 3$, $\sigma_{x,y}^2 = 5$, $N^1 = 500$ and $N^2 = 1000$. When $\mu_{x,y}^2 = 10$, we see that there is still some overlap, measured approximately to be .205. This sheds light on our results in Figure 4, where the classical model still under performs compared to the quantum model. The classical method is highly sensitive to any overlaps while the quantum model chooses a phase difference between classes to accurately recover the Gaussian parameters and number of points per class. This is shown by the table above which calculates $\cos \phi$ experimentally using $\cos \phi = \frac{1 - \alpha_1^2 - \alpha_2^2}{2 \alpha_1 \alpha_2 \sum_i G_{i,1} G_{i,2}}$.

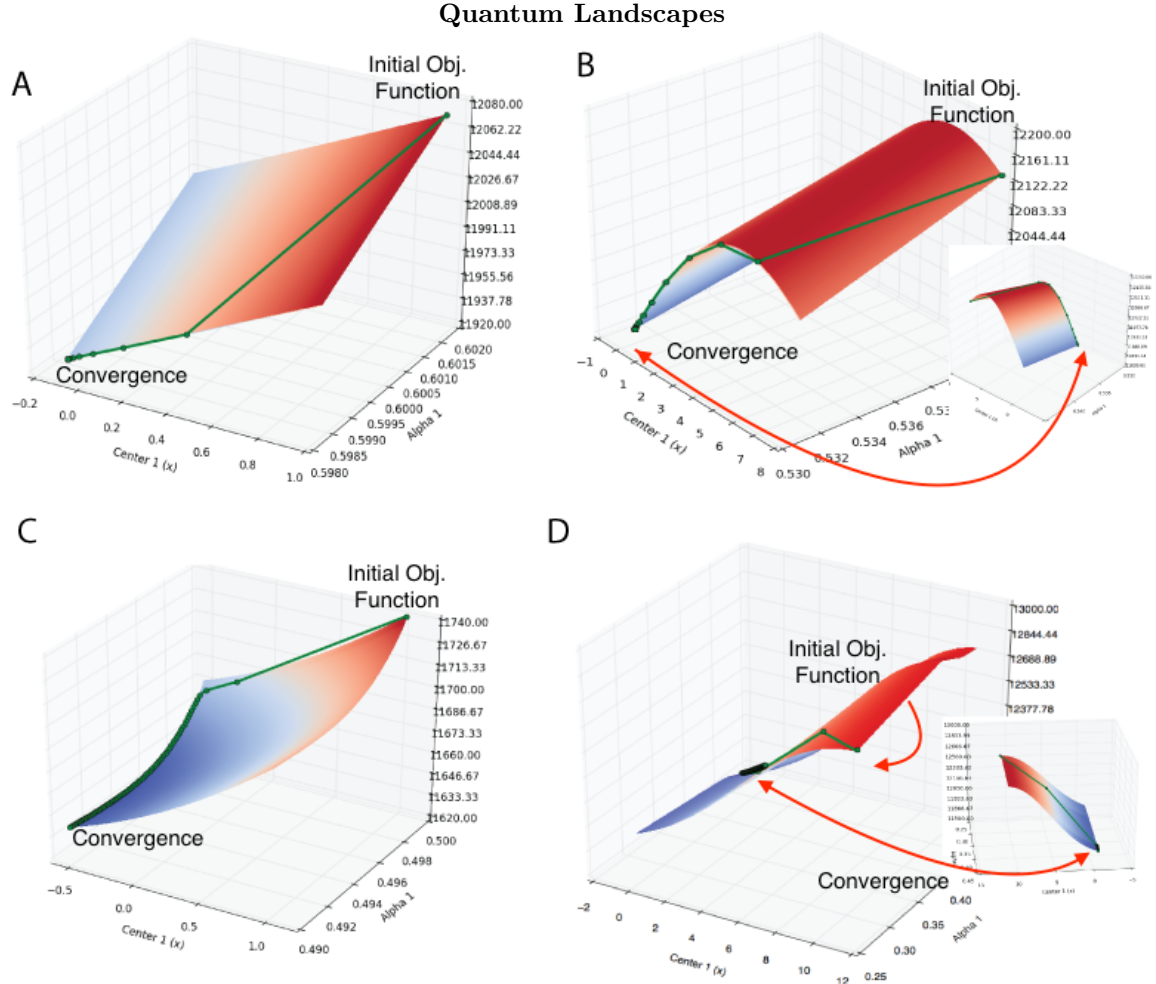
4.2 Assessment of Log Likelihood Landscapes

To better understand why the quantum method outperforms the classical method, we first assess the landscapes objective function, $O(\alpha_k^t, \theta_k^t) = \sum_{i=1}^N \sum_{k=1}^2 Q_i^t(k) \log P(p_i, k | \{\theta_k^{t-1}\})$, to examine how the values obtained fair with the desired solution. Note that the objective function is different for the quantum method (see 9 and 14) and the classical method (see ?? and 2), due to the calculation of $P(p_i, k | \{\theta_k\})$ and $Q_i(k)$ in each method. We wonder if a local minima is reached which might prevent variables at converging at the correct values. Second, we also asses the final solution's objective functions to see which method offers a better cost function (assessing the interference phenomena ability to lower the cost function). Third, in generating these landscapes, we also test to see whether initial values of the the centers have any affect on the estimated variables.

In order to create the landscapes we fix all parameters but μ_x^1 and α_1 (the parameter α_2 varies, but is determined from α_1). The sample over the μ_x^1 values is in the range between the minimum and maximum values from the EM iterations, and in increments of 0.01. For the classical method we do the same, we get the maximum and minimum values of α_1 and sample by increments of 0.01 in between these values. However, for the quantum method, to fully specify α_2 (and the objective function), we need the phase values as well as α_1 . Thus, for the quantum method, we sample over the α^1 values obtained from each iteration of the EM procedure. The number of iterations is not predetermined, but rather depends when there is convergence.

For both the classical and quantum methods, we note the objective function during the initialization of EM, and the during the ending of EM to ensure that the objective function did indeed

increase, and is close to the true objective function. For the classical method, the true objective function was obtained by fixing all the parameters in 12. For the quantum method, all the parameters except α_1 and α_2 were fixed. In order to obtain those two variables, we performed an exhaustive search over the $[0.0, 1.0]$ in increments of .001, to obtain the retrieve the lowest possible objective function defined by 16.

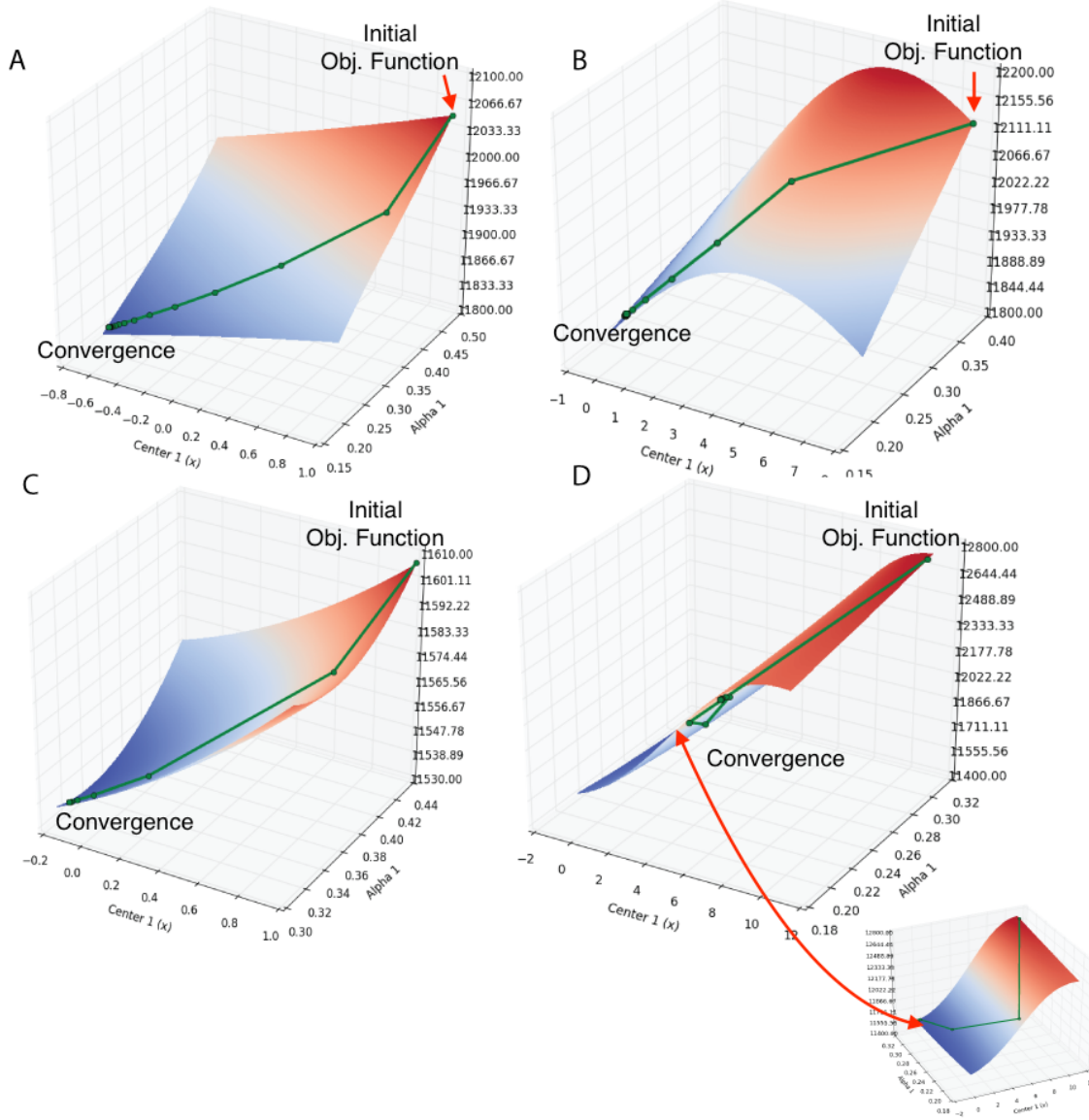


Estimated Values for Quantum Method

Image	Initial μ_x^1	Final μ_x^1	Estimated N1	Estimated N2	Initial Obj. Function	Final Obj. Function	True Obj. Function
A	1	-0.1289	438.572	1061.427	12045.968	11909.624	11941.578
B	8	-0.287	376.447	1123.553	12113.659	11895.729	11941.578
C	1	-0.073	460.572	1039.023	11620.084	11909.624	11535.664
D	11	-0.568	456.662	1043.337	12632.097	11628.296	11535.664

Figure 6: These landscapes from the quantum method show the objective function with changing μ_x^1 and α_1 . For A,B, C, and D, we fixed the parameters to $\mu_y^1 = 0$, $\sigma_{x,y}^1 = 3$, $\sigma_{x,y}^2 = 5$, N1 = 500, and N2 = 1000. For cases A and B, μ^2 was fixed to 5, and for cases C and D, μ^2 was fixed to 9. The table above specifies the original/true objective function, which is lower than the ones for the classical case, due to the interference phenomena.

Classical Landscapes



Estimated Values for Classical Method

Image	Initial μ_x^1	Final μ_x^1	Estimated N1	Estimated N2	Initial Obj. Function	Final Obj. Function	True Obj. Function
A	1	-0.551	326.469	1173.531	12047.035	11825.221	12176.206
B	8	-0.845	306.459	1193.540	12113.659	11814.813	12176.206
C	1	-0.279	462.194	1037.805	11586.998	11570.7837	11577.100
D	11	-0.120	463.622	1036.377	12476.823	11539.203	11577.100

Figure 7: These landscapes from the classical method show the objective function with changing μ_x^1 and α_1 . For A,B, C, and D, we fixed the parameters to $\mu_y^1 = 0$, $\sigma_{x,y}^1 = 3$, $\sigma_{x,y}^2 = (5,5)$, N1 = 500, and N2 = 1000. For cases A and B, μ^2 was fixed to 5, and for cases C and D, μ^2 was fixed to 9. The table above specifies the original/true objective function, which is higher than the ones for the quantum case, due to the interference phenomena. The objective function for the final values is lower than the original/true objective function, which means the method is performing adequately.

4.3 Deformations to the Gaussian Assumption

So far, the tested classes were generated from Gaussian distributions. Now, in order to deform the shape of the distribution, the data set is initially drawn from a Gaussian distribution and then, each point coordinates is displaced by a random number drawn from a uniform distribution in $[-\epsilon, \epsilon]$. In order to remove any biases that would cause the quantum method to outperform its classical counterpart, we start with distributions where both methods performed equally well. This was the case for which the distributions were separated with an overlap measurement of 6.82E-12, and both classes were equal in covariance and in the number of points. The parameters for both classes were fixed at $\mu_{x,y}^1 = 0$, $\mu_{x,y}^2 = 7$, and variances $(\sigma_{x,y}^1)^2 = (\sigma_{x,y}^2)^2 = 9$, $N^1 = N^2 = 500$. 50 trials were run for ϵ values of 0.00, 0.75, 1.50, 3.00, 3.75, 4.50, 5.25, and 6.00. (see figure 7, 8, 9). The quantum method outperforms the classical method, proving to have a more robust performance to distributions that are not "ideal Gaussian distributions".

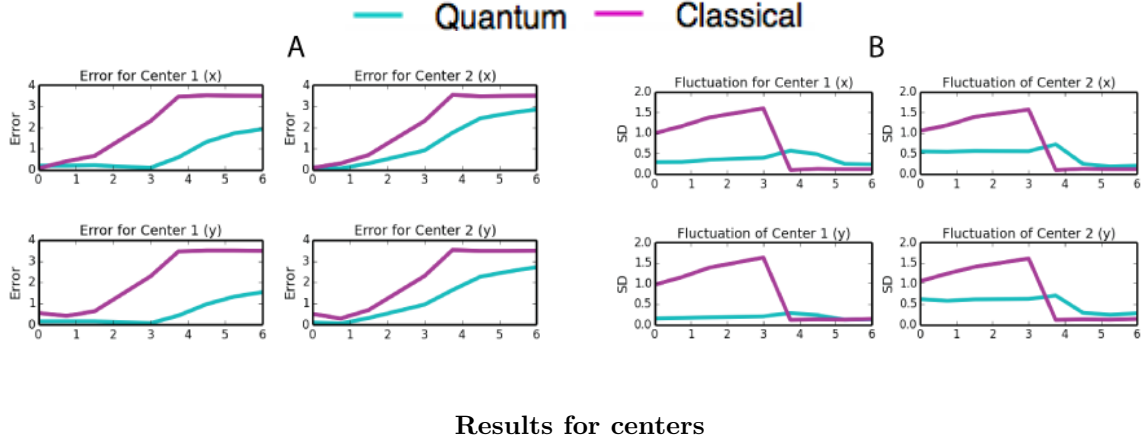


Figure 8: Y-axis: Estimated error (given by equation 28) and standard deviation (fluctuations)(given by equation 29) for both centers vs X-axis: the deformation ϵ was increased. Throughout these experiments, to avoid any advantage to the quantum method, we drew the data from $\mu_{1,2}^2 = \mu_{1,2}^1 = 3$, and variances $(\sigma_{x,y}^1)^2 = (\sigma_{x,y}^2)^2 = 9$, and $N^1 = N^2 = 500$. Without the deformations, both methods perform similarly, as evidenced by results at $\epsilon = 0$.

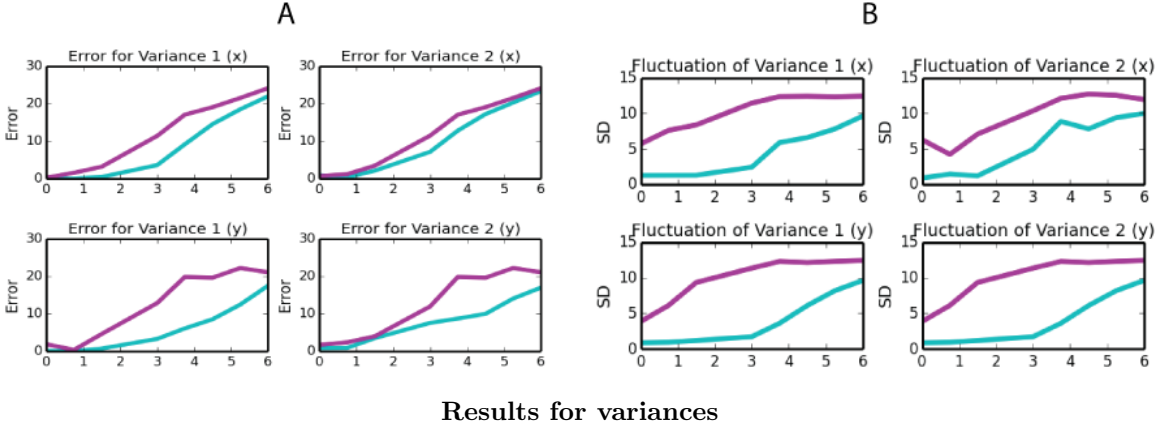


Figure 9: Y-axis: Estimated error (given by equation 28) and standard deviation (fluctuations) (given by equation 29) for both variances $(\sigma_{x,y}^1)^2$ and $(\sigma_{x,y}^2)^2$ vs X-axis: increasing $\epsilon, \mu_{x,y}^2$. In these trials, the true $(\sigma_{x,y}^1)^2 = 9$ and $(\sigma_{x,y}^2)^2 = 25$. The quantum method outperforms on both accuracy and robustness since it returns a lower error on the average variances and fluctuates less.

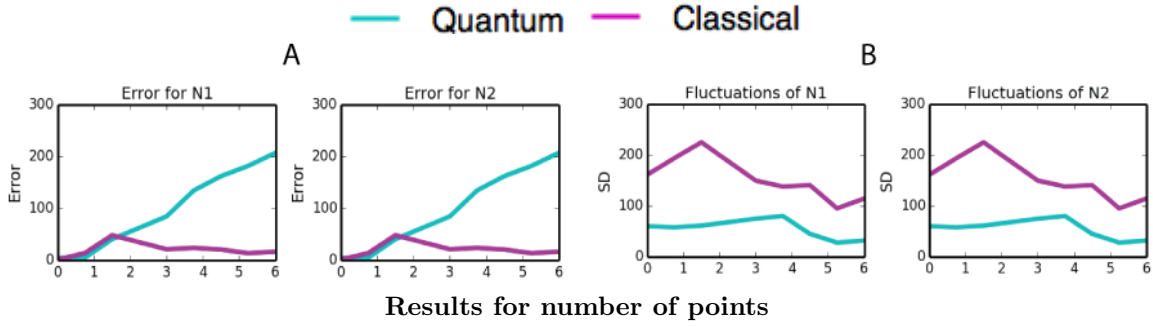


Figure 10: Y-axis: Estimated error (given by equation 28) and standard deviation (fluctuations)(given by equation 29) for N^1 (first column) and N^2 (second column) vs X-axis: the deformation ϵ was increased. Notice that after the classical system has completely failed to identify the other parameters accurately, the standard deviation decreases as does the error for identifying N^1 and N^2 . This suggests a bias towards $N^1 = N^2$, even if estimations are wrong.

4.4 Experiments in Color Segmentation: 3D Data

Color images have each pixel represented in RGB space, i.e., for each pixel three numbers are assigned representing the amount of red (R), green (G), and blue (B). The color segmentation problem consists of assigning to each pixel one class, foreground or background (1 or 0). From a clustering point of view, each pixel is one color data point, in 3D (in RGB space), and the task is to discover the two classes in color space and assign each pixel to a corresponding class.

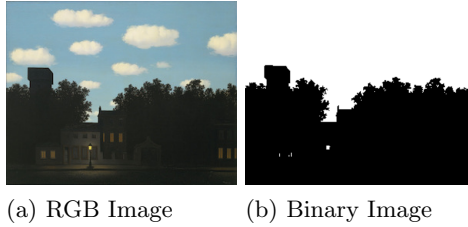


Figure 11: The RGB image of Magritte's *Empire of Light*. Each pixel was assigned either 0 or 1 (for black or white) by creating a threshold for each RGB channel. If red, green and blue levels in the original image were all above 100, then the pixel was assigned to 1, and 0 otherwise.

Of course, in computer vision, one does employ the notion of neighborhood, that nearest pixels are more likely to belong to the same class. Here, we only focus on the clustering problem, i.e., each pixel is treated as independent. To build a complete color segmentation problem, one could use the final clustering probabilities to a "neighborhood" prior and work on the posterior probability. This would lead to a graph partitioning problem that is known to be solved with the min cut algorithm [8, 3]. However, our goal here is to focus on data probability (likelihood) problem, i.e., on the clustering probabilities. To demonstrate that our quantum inspired method works significantly better than the state-of-the-art Gaussian mixture model.

L*a*b* color space: A Lab color space is a color-opponent space with dimensions L for lightness and a and b for the color-opponent dimensions, based on non-linearly compressed coordinates (e.g. CIE XYZ) ([11]). The lightness, L^* , represents the darkest black at $L^* = 0$, and the brightest white at $L^* = 100$. The color channels, a^* and b^* , will represent true neutral gray values at $a^* = 0$ and $b^* = 0$. The red/green opponent colors are represented along the a^* axis, with green at negative a^* values and red at positive a^* values. The yellow/blue opponent colors are represented along the b^* axis, with blue at negative b^* values and yellow at positive b^* values. It has been used in early computer vision color work precisely because it has good perceptual metric properties (e.g., [13, 11]).

Setting up the Clustering Problem in 3D $L^*a^*b^*$ space: We first colored binary images (black and white) according to two corresponding class Gaussian distributions, G^b and G^w generated in $L^*a^*b^*$ color space.

The experiment is then defined as follows. Each image pixel is mapped to a 3D point coordinate (L_i^k, a_i^k, b_i^k) . Then, we extract $\mu_{l,a,b}^k = (L^k, a^k, b^k)$ and the covariance, $\mathbf{C}^k = \begin{bmatrix} \sigma_L^k & 0 & 0 \\ 0 & \sigma_a^k & 0 \\ 0 & 0 & \sigma_b^k \end{bmatrix}$ from the set of points (as many as image pixels). Once these parameters are extracted (either from the Classical Gaussian of Mixture method or by the Quantum inspired method) we assign a class to each pixel.

Class Assignment: According to the clustering method (Classical or Quantum) the final expectation values $Q_i^b; i = 1, \dots, N$, and $Q_i^w; i = 1, \dots, N$ determine the class, where b indicates the class we assign the color "black" and w indicates the other class, that we assign the color "white". These distributions were used to compute the number of points per class by both methods. Thus, in order to produce the final segmentation, given a pixel, $p_i; i = 1, \dots, N$, if $Q_i^b - Q_i^w > 0$, then we assigned the "black" color, and otherwise, we assigned "white" color.

Results on Artificially Colored Images: For each clustering method, classical and quantum inspired, we ran 10 trials for each image, starting each time at a random initialization. To account for the variability in the classical case, we used the image which had estimated parameters closest to the average of its cohort.

To quantify the comparison of the methods, we calculated the estimated error for the variables $\mu_{L,a,b}^k$, $\sigma_{L,a,b}^k$, and N^k .

We found a combined error using

$$Error = \sqrt{(\mu_L^{k*} - \overline{mu_L^k})^2 + (\mu_a^{k*} - \overline{mu_a^k})^2 + (\mu_b^{k*} - \overline{mu_b^k})^2}, \quad (30)$$

where μ^{k*} denotes the ground truth, and $\overline{mu_L^k}$, the estimated value.

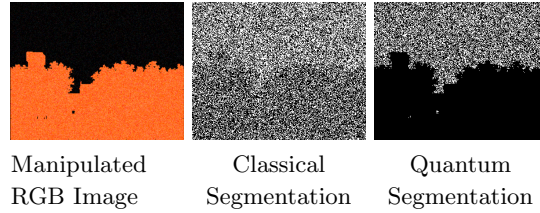


Figure 12: Every pixel in binary image seen in figure 10 was assigned a random point in $L^*a^*b^*$ space generated from corresponding class Gaussian distributions. The fixed parameters for this case, which defined the class Gaussian distributions were $\mu_{L,a,b}^w = (0,0,0)$, $\mu_{L,a,b}^b = (75,75,75)$, $\sigma_{L,a,b}^w = (4,4,4)$, $\sigma_{L,a,b}^b = (6,6,6)$, $N^w = 34034$ pixels, and $N^b = 37666$ pixels.

The table below summarizes the average of quantum and classical estimated errors using 30 and relative error is the ratio of the Quantum error to the Classical one.

Variables	Classical Error	Quantum Error	Relative Error
$\mu_{L,a,b}^w$	66.579	0.246	0.003
$\mu_{L,a,b}^b$	60.079	44.410	0.739
N^w	1015.121	19598.009	19.306
N^b	1015.121	19598.009	19.306
$(\sigma_{L,a,b}^w)^2$	64.721	61.488	0.950
$(\sigma_{L,a,b}^b)^2$	65.023	5.567	0.085

Note that although the error in number of point estimation of the classical case is reasonable, this is solely because this image has about the same number of points in both classes but the actual assignment to classes is very poor. So while the quantum method did poorly on the "black region" of the manipulated image, it accurately marked the "orange region" while classical performed poorly everywhere.

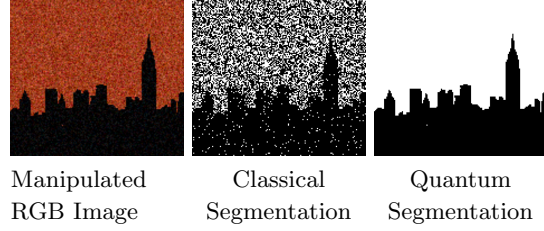


Figure 13: An original RGB image, available at <http://alexgoldblum.com/new-york/nyc-skyline/>, was binarized. We then colored it artificially randomly sampling each pixel from the $L^*a^*b^*$ with the following two Gaussian parameters $\mu_{L,a,b}^w = (40, 40, 40)$, $\mu_{L,a,b}^b = (0, 0, 0)$, $\sigma_{L,a,b}^b = (5, 5, 5)$, and $\sigma_{L,a,b}^w = (8, 8, 8)$. The number of pixels in $N^b = 8362$, and $N^w = 13478$. The table below summarizes the average of quantum and classical estimated errors using 30 and relative error is the ratio of the Quantum error to the Classical one.

Variables	Classical Error	Quantum Error	Relative Error
$\mu_{L,a,b}^w$	42.449	0.110	0.002
$\mu_{L,a,b}^b$	26.332	0.092	0.003
N^w	3418.877	31.000	.009
N^b	3418.877	31.000	.009
$(\sigma_{L,a,b}^w)^2$	34.117	3.948	0.115
$(\sigma_{L,a,b}^b)^2$	33.466	1.131	0.034

5 Conclusion

In this paper we solely investigated the clustering problem. The state of the art clustering method is EM as it gives close-form solutions for the (Classical) Gaussian Mixture Model (GMM) [2]. Inspired by quantum methods [7] in which interference phenomena causes probability cancellations, we reformulated this classical model in a quantum model and applied the EM method of parameter estimation.

We showed that the quantum method outperforms the classical (GMM) method in every aspect of the estimations, in cases where the two clusters have an uneven number of data points and different covariances. The Quantum method is able to recover more accurate estimations of all distribution parameters, with much less fluctuations. In regimes where the distributions are separated, we observe that the classical model is very sensitive to slight overlaps. We also show the Quantum method is robust to data deformations from the Gaussian assumptions and is able to segment colors more accurately in the color experiments.

We plotted the landscapes of each model and showed them to be smooth, leading often to global minima solutions. There were cases, when the data was very separated and the initial condition very far from the desired solution that the quantum method reached local minima and not global minima. We noted that the quantum method has an extra variable to be computed, the phase of each class.

The biggest difference between the two models is that the cost function for the quantum model returns a lower value in an overlapping case than the classical solution, suggesting that the interference phenomena drives the quantum system. This allowed the quantum method to produce much more accurate and robust solutions, compared to the classical method.

For working towards segmentation, clustering may also be interpreted in a Bayesian method as the data likelihood model and combined with the prior model on the data (such as encouraging neighbors to be labeled similarly) to form the posterior probability. In this case, for the future, once the clustering model is produced, techniques such as graph cuts[8, 3] could be employed to yield final and better results for color segmentation.

A Gradient Equations For the Parameters

α_k :

for ∂_{α_1}

$$\begin{aligned}
0 &= \sum_i^N Q_{i,1}^t \frac{\left(2G_{i,1}^2 - \frac{G_{i,1}G_{i,2}}{\sum_i G_{i,1}G_{i,2}}\right)}{\left[2\alpha_1^2 G_{i,1}^2 + (1 - \alpha_1^2 - \alpha_2^2) \frac{G_{i,1}G_{i,2}}{\sum_i G_{i,1}G_{i,2}}\right]} - Q_{i,2}^t \frac{\frac{G_{i,1}G_{i,2}}{\sum_i G_{i,1}G_{i,2}}}{\left[2\alpha_2^2 G_{i,2}^2 + (1 - \alpha_1^2 - \alpha_2^2) \frac{G_{i,1}G_{i,2}}{\sum_i G_{i,1}G_{i,2}}\right]} \\
&= \sum_i^N Q_{i,1}^t \frac{(2G_{i,1}^2 - o_i)}{[2\alpha_1^2 G_{i,1}^2 + (\alpha o)_i]} - Q_{i,2}^t \frac{o_i}{[2\alpha_2^2 G_{i,2}^2 + (\alpha o)_i]} \\
&\quad \text{applying equation 14} \\
&= \sum_i^N \frac{(G_{i,1}^2 - o_i)}{[\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i]} \tag{31}
\end{aligned}$$

for ∂_{α_2}

$$\begin{aligned}
0 &= \sum_i^N -Q_{i,1}^t \frac{\frac{G_{i,1}G_{i,2}}{\sum_i G_{i,1}G_{i,2}}}{\left[2\alpha_1^2 G_{i,1}^2 + (1 - \alpha_1^2 - \alpha_2^2) \frac{G_{i,1}G_{i,2}}{\sum_i G_{i,1}G_{i,2}}\right]} + Q_{i,2}^t \frac{\left(2G_{i,2}^2 - \frac{G_{i,1}G_{i,2}}{\sum_i G_{i,1}G_{i,2}}\right)}{\left[2\alpha_2^2 G_{i,2}^2 + (1 - \alpha_1^2 - \alpha_2^2) \frac{G_{i,1}G_{i,2}}{\sum_i G_{i,1}G_{i,2}}\right]} \\
&= \sum_i^N -Q_{i,1}^t \frac{o_i}{[2\alpha_1^2 G_{i,1}^2 + (\alpha o)_i]} + Q_{i,2}^t \frac{(2G_{i,2}^2 - o_i)}{[2\alpha_2^2 G_{i,2}^2 + (\alpha o)_i]} \\
&\quad \text{applying equation 14} \\
&= \sum_i^N \frac{(G_{i,2}^2 - o_i)}{[\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i]} \tag{32}
\end{aligned}$$

A.1 μ_k

For $k=1$:

$$\begin{aligned}
0 &= \sum_i \sum_k Q_i^t(k) \frac{\partial_{\mu_k} P(p_i, k | \{\theta_k^{t-1}\})}{P(p_i, k | \{\theta_k^{t-1}\})} \\
&= \sum_i Q_{i,1}^t \frac{\alpha_1^2 \mathbf{C}_1^{-1} (p_i - \mu_1) G_{i,1}^2 + \frac{1}{4} (1 - \alpha_1^2 - \alpha_2^2) \frac{G_{i,1} G_{i,2}}{\sum_i G_{i,1} G_{i,2}} \left(\mathbf{C}_1^{-1} (p_i - \mu_1) - \frac{\sum_i \mathbf{C}_1^{-1} (p_i - \mu_1) G_{i,1} G_{i,2}}{\sum_i G_{i,1} G_{i,2}} \right)}{\left[\alpha_1^2 G_{i,1}^2 + \frac{1}{2} (1 - \alpha_1^2 - \alpha_2^2) \frac{G_{i,1} G_{i,2}}{\sum_i G_{i,1} G_{i,2}} \right]} \\
&\quad + Q_{i,2}^t \frac{\frac{1}{4} (1 - \alpha_1^2 - \alpha_2^2) \frac{G_{i,1} G_{i,2}}{\sum_i G_{i,1} G_{i,2}} \left(\mathbf{C}_1^{-1} (p_i - \mu_1) - \frac{\sum_i \mathbf{C}_1^{-1} (p_i - \mu_1) G_{i,1} G_{i,2}}{\sum_i G_{i,1} G_{i,2}} \right)}{\left[\alpha_2^2 G_{i,2}^2 + \frac{1}{2} (1 - \alpha_1^2 - \alpha_2^2) \frac{G_{i,1} G_{i,2}}{\sum_i G_{i,1} G_{i,2}} \right]} \\
&= \sum_i Q_{i,1}^t \frac{\alpha_1^2 \mathbf{C}_1^{-1} (p_i - \mu_1) G_{i,1}^2 + \frac{1}{4} (\alpha o)_i \mathbf{C}_1^{-1} ((p_i - \mu_1) - \sum_i (p_i - \mu_1) o_i)}{\left[\alpha_1^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i \right]} \\
&\quad + Q_{i,2}^t \frac{\frac{1}{4} (\alpha o)_i \mathbf{C}_1^{-1} ((p_i - \mu_1) - \sum_i (p_i - \mu_1) o_i)}{\left[\alpha_2^2 G_{i,2}^2 + \frac{1}{2} (\alpha o)_i \right]} \\
&\quad \text{Replacing } Q_{i,1}^t \text{ and } Q_{i,2}^t \text{ from equations (14)} \\
&= \sum_i \frac{\alpha_1^2 \mathbf{C}_1^{-1} (p_i - \mu_1) G_{i,1}^2 + \frac{1}{2} (\alpha o)_i \mathbf{C}_1^{-1} ((p_i - \mu_1) - \sum_i (p_i - \mu_1) o_i)}{\left[\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i \right]} \\
&= \sum_i \mathbf{C}_1^{-1} (p_i - \mu_1) \frac{\alpha_1^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i}{\left[\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i \right]} \\
&\quad - \left(\sum_i \mathbf{C}_1^{-1} (p_i - \mu_1) o_i \right) \sum_i \frac{\frac{1}{2} (\alpha o)_i}{\left[\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i \right]} \\
&= \sum_i \mathbf{C}_1^{-1} (p_i - \mu_1) \left[\frac{\alpha_1^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i}{\left[\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i \right]} - o_i \sum_{j=1}^N \frac{\frac{1}{2} (\alpha o)_j}{\left[\alpha_1^2 G_{j,1}^2 + \alpha_2^2 G_{j,2}^2 + (\alpha o)_j \right]} \right] \tag{33}
\end{aligned}$$

Thus,

$$\mu_1 = \frac{\sum_i F_{i,1} p_i}{\sum_i F_{i,1}} \quad \text{and} \quad \mu_2 = \frac{\sum_i F_{i,2} p_i}{\sum_i F_{i,2}} \tag{34}$$

where

$$\begin{aligned}
F_{i,k} &= \frac{\alpha_k^2 G_{i,k}^2 + \frac{1}{2} (\alpha o)_i}{\left[\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,2}^2 + (\alpha o)_i \right]} - o_i \sum_{j=1}^N \frac{\frac{1}{2} (\alpha o)_j}{\left[\alpha_1^2 G_{j,1}^2 + \alpha_2^2 G_{j,2}^2 + (\alpha o)_j \right]} \\
&= Q_i(k) - o_i \sum_{j=1}^N \frac{\frac{1}{2} (\alpha o)_j}{\left[\alpha_1^2 G_{j,1}^2 + \alpha_2^2 G_{j,2}^2 + (\alpha o)_j \right]} \tag{35}
\end{aligned}$$

A.2 \mathbf{C}_k

We focus on $\Lambda_k = \mathbf{C}_k^{-1}$. Recall the maximum likelihood gives the equation

$$0 = \sum_i \sum_k Q_i^t(k) \frac{\partial_{\Lambda_k} P(p_i, k | \{\theta_k^{t-1}\})}{P(p_i, k | \{\theta_k^{t-1}\})} \tag{36}$$

Let us investigate $\partial_{\Lambda_k} P(p_i, k | \{\alpha_k^{t-1}, \theta_k^{t-1}\})$ and using equation (14) we conclude:

$$\begin{aligned}
\partial_{\Lambda_k} P(p_i, k | \{\alpha_k, \theta_k\}) &= \delta_{k',k} \left(2\alpha_k^2 G_{i,k} \partial_{\Lambda_k} G_{i,k} + \frac{1}{4} (1 - \alpha_k^2 - \alpha_{\not k}^2) \frac{G_{i,k} G_{i,\not k}}{\sum_i G_{i,k} G_{i,\not k}} \left[\frac{\partial_{\Lambda_k} G_{i,k}}{G_{i,k}} - \frac{\sum_i \partial_{\Lambda_k} G_{i,k} G_{i,\not k}}{\sum_i G_{i,k} G_{i,\not k}} \right] \right) \\
&+ \delta_{k',k} \left(\frac{1}{4} (1 - \alpha_k^2 - \alpha_{\not k}^2) \frac{G_{i,k} G_{i,\not k}}{\sum_i G_{i,k} G_{i,\not k}} \left[\frac{\partial_{\Lambda_k} G_{i,k}}{G_{i,k}} - \frac{\sum_i \partial_{\Lambda_k} G_{i,k} G_{i,\not k}}{\sum_i G_{i,k} G_{i,\not k}} \right] \right) \\
&= \delta_{k',k} \left(2\alpha_k^2 G_{i,k} \partial_{\Lambda_k} G_{i,k} + \frac{1}{4} (\alpha o)_i \left[\frac{\partial_{\Lambda_k} G_{i,k}}{G_{i,k}} - \frac{\sum_i \partial_{\Lambda_k} G_{i,k} G_{i,\not k}}{\sum_i G_{i,k} G_{i,\not k}} \right] \right) \\
&+ \delta_{k',k} \left(\frac{1}{4} (\alpha o)_i \left[\frac{\partial_{\Lambda_k} G_{i,k}}{G_{i,k}} - \frac{\sum_i \partial_{\Lambda_k} G_{i,k} G_{i,\not k}}{\sum_i G_{i,k} G_{i,\not k}} \right] \right) \tag{37}
\end{aligned}$$

Developing it for the different (k, k') cases, and replacing $\partial_{\Lambda_k} G_{i,k} = G_{i,k} \partial_{\Lambda_k} \log G_{i,k}$, we get for $k = 1$:

$$\begin{aligned}
0 &= \sum_i Q_{i,1}^t \frac{2\alpha_1^2 G_{i,1}^2 (\partial_{\Lambda_1} \log G_{i,1}) + \frac{1}{4} (\alpha o)_i [(\partial_{\Lambda_1} \log G_{i,1}) - \frac{\sum_i G_{i,1} (\partial_{\Lambda_1} \log G_{i,1}) G_{i,2}}{\sum_i G_{i,1} G_{i,2}}]}{\alpha_1^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i} \\
&+ \sum_i Q_{i,2}^t \frac{\frac{1}{4} (\alpha o)_i [(\partial_{\Lambda_1} \log G_{i,1}) - \frac{\sum_i G_{i,1} (\partial_{\Lambda_1} \log G_{i,1}) G_{i,2}}{\sum_i G_{i,1} G_{i,2}}]}{\alpha_2^2 G_{i,2}^2 + \frac{1}{2} (\alpha o)_i} \\
&= \sum_i \frac{2\alpha_1^2 G_{i,1}^2 (\partial_{\Lambda_1} \log G_{i,1}) + \frac{1}{2} (\alpha o)_i [(\partial_{\Lambda_1} \log G_{i,1}) - \frac{\sum_i G_{i,1} (\partial_{\Lambda_1} \log G_{i,1}) G_{i,2}}{\sum_i G_{i,1} G_{i,2}}]}{\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i} \\
&= \sum_i \frac{(\partial_{\Lambda_1} \log G_{i,1}) (\alpha_1^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i)}{\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i} \\
&- \frac{1}{2} \frac{\sum_i G_{i,1} (\partial_{\Lambda_1} \log G_{i,1}) G_{i,2}}{\sum_i G_{i,1} G_{i,2}} \sum_i \frac{(\alpha o)_i}{\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i} \tag{38}
\end{aligned}$$

Equation for $k = 1$ becomes the following

$$\begin{aligned}
0 &= \sum_i (\partial_{\Lambda_1} \log G_{i,1}) \left[\frac{(2\alpha_1^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i)}{\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,1}^2 + \frac{1}{2} (\alpha o)_i} - o_i \sum_j \frac{\frac{1}{2} (\alpha o)_j}{\alpha_1^2 G_{j,1}^2 + \alpha_2^2 G_{j,1}^2 + \frac{1}{2} (\alpha o)_j} \right] \\
&= \sum_i (\partial_{\Lambda_1} \log G_{i,1}) R_{i,1} \tag{39}
\end{aligned}$$

and analogously for $k = 2$. Recall that $G_{i,k} = \frac{1}{\sqrt{Z_k}} e^{-\frac{1}{4} (p_i - \mu_k) \Lambda_k (p_i - \mu_k)}$ where $\sqrt{Z_k} = (2\pi)^{\frac{d}{4}} |\Lambda_k|^{-\frac{1}{4}}$ and so

$$\begin{aligned}
\partial_{\Lambda_k} \log G_{i,k} &= \frac{1}{4} \partial_{\Lambda_k} \log |\Lambda_k| - \frac{1}{4} (p_i - \mu_k)^T (p_i - \mu_k) \\
&= \frac{1}{4} \frac{1}{|\Lambda_k|} \partial_{\Lambda_k} |\Lambda_k| - \frac{1}{4} (p_i - \mu_k) (p_i - \mu_k)^T \\
&= \frac{1}{4} \frac{1}{|\Lambda_k|} |\Lambda_k| \Lambda_k^{-1} - \frac{1}{4} (p_i - \mu_k) (p_i - \mu_k)^T \\
&= \frac{1}{4} \Lambda_k^{-1} - \frac{1}{4} (p_i - \mu_k) (p_i - \mu_k)^T \\
&= \frac{1}{4} [\mathbf{C}_k - (p_i - \mu_k) (p_i - \mu_k)^T] \tag{40}
\end{aligned}$$

Thus,

$$\begin{aligned}
0 &= \sum_i [\mathbf{C}_1 - (p_i - \mu_1)(p_i - \mu_1)^T] R_{i,1}^t \\
0 &= \sum_i [\mathbf{C}_2 - (p_i - \mu_2)(p_i - \mu_2)^T] R_{i,2}^t \\
&\text{or} \\
\mathbf{C}_k &= \frac{\sum_i R_{i,k}^t (p_i - \mu_k)(p_i - \mu_k)^T}{\sum_i R_{i,k}^t}
\end{aligned} \tag{41}$$

where

$$\begin{aligned}
R_{i,k}^t &= \left[\frac{(2\alpha_k^2 G_{i,k}^2 + \frac{1}{2}(\alpha o)_i)}{\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,1}^2 + \frac{1}{2}(\alpha o)_i} - o_i \sum_j \frac{\frac{1}{2}(\alpha o)_j}{\alpha_1^2 G_{j,1}^2 + \alpha_2^2 G_{j,1}^2 + \frac{1}{2}(\alpha o)_j} \right] \\
&= F_{i,k}^t + \frac{\alpha_k^2 G_{i,k}^2}{\alpha_1^2 G_{i,1}^2 + \alpha_2^2 G_{i,1}^2 + \frac{1}{2}(\alpha o)_i}
\end{aligned} \tag{42}$$

For the classical case, $\phi = \frac{\pi}{2}$, $R_{i,k}^t = 2Q_{i,k}$

$$\mathbf{C}_k = \frac{\sum_i Q_{i,k} (p_i - \mu_k)(p_i - \mu_k)^T}{\sum_i Q_{i,k}}$$

B Brief Derivation of the EM Method

Here is a brief derivation of the EM method. We denote the parameters $\alpha_k, \theta_k = \{\mathbf{C}_k, \mu_k\}$ and the set of data by $\{p_i\}$. Thus,

$$P(\{p_i\}|\{\theta_k\}) = \prod_i P(p_i|\{\theta_k\}) = \prod_i \left(\frac{P(p_i, k|\{\theta_k\})}{P(k|p_i, \{\theta_k\})} \right) \tag{43}$$

Given an estimate value $\{\theta_k\}$, the likelihood is

$$\begin{aligned}
L(\{p_i\}|\{\theta_k\}) &= \log P(\{p_i\}|\{\theta_k\}) \\
&= \sum_i \log P(p_i|\{\theta_k\}) \\
&= \sum_i \log \sum_k P(p_i, k|\{\theta_k\}) \\
&\quad \text{introducing } (Q_i(k) \geq 0 \text{ and } \sum_k Q_i(k) = 1) \\
&= \sum_i \log \sum_k Q_i(k) \frac{P(p_i, k|\{\theta_k\})}{Q_i(k)} \\
&\quad \text{using Jensen's Inequality} \\
&\geq \sum_i \sum_k Q_i(k) \log \frac{P(p_i, k|\{\theta_k\})}{Q_i(k)}
\end{aligned} \tag{44}$$

Now, for any set of distributions $Q_i(k)$, the formula (44) gives a lower-bound on $L(\{p_i\}|\{\theta_k\})$. If we have some current guess $\{\theta_k^t\}$ of the parameters, we make the lower-bound tight at that value of θ . Thus, we need for the step involving Jensen's inequality in our derivation above to hold with equality. It happens at $Q_i(k) = P(k|p_i, \{\theta_k\})$.

Then, we iterate on estimation of $[P(k|\{\theta_k\}), \{\theta_k\}]$ and $P(k|p_i, \{\theta_k\})$. We index these iterations by t . First, at $t = 0$ of the iteration procedure, a random set of values $[P^{t=0}(k|\{\theta_k\}), \theta_k^{t=0}]$ is produced.

References

- [1] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color- and texture-based image segmentation using em and its application to content-based image retrieval. In *Computer Vision, 1998. Sixth International Conference on*, pages 675–682, Jan 1998.
- [2] J. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, pages 1–281, 1998.
- [3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, Nov 2001.
- [4] M. Celenk. A color clustering technique for image segmentation. *Computer Vision, Graphics, and Image Processing*, 52(2):145 – 170, 1990.
- [5] T.-W. Chen, Y.-L. Chen, and S.-Y. Chien. Fast image segmentation based on k-means clustering with histograms in hsv color space. In *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pages 322–325, Oct 2008.
- [6] M. Cicconet, D. Geiger, and M. Werman. Complex-valued hough transforms for circles. 2015. Quebec City, Canada.
- [7] R. Feynman. *The Feynman Lectures on Physics*, volume 3. Addison Wesley, 1971.
- [8] H. Ishikawa and D. Geiger. Segmentation by grouping junctions. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 125–131, Jun 1998.
- [9] S. C. Kim and T. J. Kang. Texture classification and segmentation using wavelet packet frame and gaussian mixture model. *Pattern Recognition*, 40(4):1207 – 1221, 2007.
- [10] A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451 – 461, 2003. Biometrics.
- [11] K. McLaren. The development of the cie (l, a, b)-uniform color space, 1976.
- [12] A. Srivastava, S. H. Joshi, W. Mio, and X. Liu. Statistical shape analysis: clustering, learning, and testing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):590–602, April 2005.
- [13] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846, Jan 1998.