National Taiwan University

College of Electrical Engineering and
Computer Science

# Undergraduate Research Report

*Student:*
Chi-Ning Chou

*Advisor:*
I-Hsiang Wang


June 27, 2016

**Abstract**

This is the report for my undergraduate research with professor I-Hsiang Wang in 2015 spring. In the beginning of this semester, I studied and surveyed various topics including statistical estimation, probability metric, primal dual optimization problems[OPT], and the application of machine learning[ML0], [Lina], [Linb], [CLA], [BT], [Wau], [San], [DF], [RE0b], and [RE0a].

Later, I started to work on the minimax estimation of Gaussian Mixtures. First, I studied some techniques about constructing minimax lower bound such as Le Cam method, Fano method, and Assaud method in [Tsy09] and [Duc]. After studied some related works [JJW15a], [JJW15b], [WY14], and [WY15], I tried to construct the minimax lower bound of the problem. I divided the works into three stages and for now, I proved that the results under loose assumption are in the same order of the similar problem: support size estimation.

In this report, I'll begin with my research in minimax estimation of Gaussian mixtures, summarizing the papers that I've studied and introducing the lower bound of minimax risk in this problem. The second section will contain the summary of all the other surveys I did in this semester.

# Contents

# 1 Overview

Gaussian mixture model is widely used in signal processing, machine learning, etc. It's an intuitive model for the underlying distribution behind the data we have since the world is full of normality.

The definition of Gaussian mixture model is very simple, it's the weighted sum of finite (or in some case infinite) normal distribution, which we call *component* in the Gaussian mixture model. That is,

$$g(x; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_i \lambda_i f_i(x; \mu_i, \sigma_i) \tag{1}$$

, where $f_i(\cdot; \mu_i, \sigma_i) \sim N(\mu_i, \sigma_i)$, $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$.

Since Gaussian mixture is a common model in application, there exists various algorithms aim to find the underlying components. Most methods treat such scenario as a clustering problem and solve it with EM algorithm or K-mean method. Moreover, using these methods requires a good initial guess on the number of components in the underlying Gaussian mixture model. Although there are several adaptive methods to estimate the number of components in the beginning of the algorithms, there are no universal estimator and the corresponding error analysis.

As a result, in this research, we focus on finding the fundamental difference between counting the number of components and locating the positions of components in Gaussian mixture model. Our goal is to find out under what kind of setting counting is strictly easier than locating while some regions might not have such phenomenon. Furthermore, apart from information theoretical results, we would also like to derive some computational efficient/inefficient bounds.

## 1.1 How to compare counting and locating?

### 1.1.1 $k$ versus $k+1$

- Consider $\mathcal{H}_0 : k = k_0$ and $\mathcal{H}_1 : k = k_0 + 1$: [FL94]
- Consider $\mathcal{H}_0 : k = 2$ and $\mathcal{H}_1 : k \geq 3$:[CCK04]. It can be used in genetic problem.
- Direct construct estimator for the number of components: [Hen85],[DCG97],[Ker00]
- Consider hypothesis testing $\mathcal{H}_0 : k = k_0$ and $\mathcal{H}_1 : k = k_1$ where $k_0 < k_1$: [MR14]

### 1.1.2 Our formulation

At first glance, it seems weird to compare counting with locating since the two has totally different evaluating criteria. For counting, the loss function might simply be the difference between estimating number and the underlying number while for locating, there are many choices for the loss functions. Basically, the two estimation problem are incomparable.

However, if we slightly change the setting from estimation to testing, then we can hide the incomparable part into the choice of hypothesis and only compare the testing error. With this setting, we can adopt the suitable setting for locating problem while in the meantime making the two problem comparable.

To define the hypothesis testings for locating and counting, we first define some notations for the Gaussian mixture problem. For simplicity, in this work we only consider isotropic Gaussian mixture.

**Definition 1 (Gaussian mixture)** *We denote a Gaussian mixture with $k$ components by a length $2k$ vectors $(\boldsymbol{\mu}, \boldsymbol{\lambda})$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in (\mathbb{R}^p)^k$ records the centers of each component and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ denotes the weight of each component where $\lambda_i \geq 0$, which is the minimum weight for each component, and $\sum_{i\in[k]} \lambda_i = 1$. Thus, the pdf of Gaussian mixture $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ is $g(x; \boldsymbol{\mu}, \boldsymbol{\lambda}) = \sum_{i\in[k]} \lambda_i \frac{e^{-||x-\mu_i||_2^2/2}}{(2\pi)^{p/2}}$.*

*Denote the family containing Gaussian mixture with $k$ components with minimum weight $\tau$ and minimum distance $d$ as*

$$\mathcal{G}_{k,\tau,d} = \{(\boldsymbol{\mu}, \boldsymbol{\lambda}) | \boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in (\mathbb{R}^p)^k; \forall i \neq j \in [k], ||\mu_i - \mu_j||_2 \geq d; \\ \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k); \forall i \in [k], \lambda_i \geq \tau, \sum_{i\in[k]} \lambda_i = 1\}$$

*The family containing all Gaussian mixtures with minimum weight $\tau$ and minimum distance $d$ is $\mathcal{G}_{\tau,d} = \cup_{k\in\mathbb{N}} \mathcal{G}_{k,\tau,d}$.*

As long as we are going to locate the Gaussian mixture, we need to evaluate how close two Gaussian mixtures are. Since our goal is to compare locating with counting, if two Gaussian mixtures have different number of components, it will simultaneously be a bad instance for locating and counting, thus, it is less interesting for us. As a result, here we only define loss function for Gaussian mixtures with the same number of components.

**Definition 2** *For $(\boldsymbol{\mu}, \boldsymbol{\lambda}), (\boldsymbol{\mu}', \boldsymbol{\lambda}') \in \mathcal{G}_{k,\tau,d}$, the losses among them are*

$$\ell_\mu((\boldsymbol{\mu}, \boldsymbol{\lambda}), (\boldsymbol{\mu}', \boldsymbol{\lambda}')) = ||\boldsymbol{\mu} - \boldsymbol{\mu}'||_2^2$$
$$\ell_\lambda((\boldsymbol{\mu}, \boldsymbol{\lambda}), (\boldsymbol{\mu}', \boldsymbol{\lambda}')) = ||\boldsymbol{\lambda} - \boldsymbol{\lambda}'||_2^2$$

*Furthermore, we can define the neighborhood of Gaussian mixture* $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathcal{G}_{k,\tau,d}$ *as follow.*

$$\mathcal{B}_{\delta_\mu}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \{(\boldsymbol{\mu}', \boldsymbol{\lambda}') \in \mathcal{G}_{k,\tau,d} | \ell_\mu((\boldsymbol{\mu}, \boldsymbol{\lambda}), (\boldsymbol{\mu}', \boldsymbol{\lambda}')) \leq \delta_\mu\}$$
$$\mathcal{B}_{\delta_\lambda}(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \{(\boldsymbol{\mu}', \boldsymbol{\lambda}') \in \mathcal{G}_{k,\tau,d} | \ell_\lambda((\boldsymbol{\mu}, \boldsymbol{\lambda}), (\boldsymbol{\mu}', \boldsymbol{\lambda}')) \leq \delta_\mu\}$$

Finally, we can define the hypothesis testings for counting and locating.

**Remark 3** The reason for us to omit the locating case where the two Gaussian mixtures having different number of components is that it is also a bad instance for counting. Since here we only consider the hypothesis testing, we only care about success or not.

$$\text{(Counting } \mathcal{C}_{k,\tau,d}) \begin{cases} \mathcal{H}_0^{\mathcal{C}_{k,\tau,d}} : & \forall (\boldsymbol{\mu}_0, \boldsymbol{\lambda}_0) \in \mathcal{G}_{k,\tau,d} \\ \mathcal{H}_1^{\mathcal{C}_{k,\tau,d}} : & \forall (\boldsymbol{\mu}_1, \boldsymbol{\lambda}_1) \in \mathcal{G} - \mathcal{G}_{k,\tau,d} \end{cases} \tag{2}$$

$$\text{(Locating } \mathcal{L}_{k,\tau,d,\delta_\mu,\delta_\lambda}) \begin{cases} \mathcal{H}_0^{\mathcal{L}_{k,\tau,d,\delta_\mu,\delta_\lambda}} : & \text{For fixed } (\boldsymbol{\mu}_0, \boldsymbol{\lambda}_0) \in \mathcal{G}_{k,\tau,d} \\ \mathcal{H}_1^{\mathcal{L}_{k,\tau,d,\delta_\mu,\delta_\lambda}} : & \forall (\boldsymbol{\mu}_1, \boldsymbol{\lambda}_1) \in \mathcal{G}_{k,\tau,d} - \mathcal{B}_{\delta_\mu}(\boldsymbol{\mu}_0, \boldsymbol{\lambda}_0) - \mathcal{B}_{\delta_\lambda}(\boldsymbol{\mu}_0, \boldsymbol{\lambda}_0) \end{cases} \tag{3}$$

$\mathcal{C}_{k,\tau,d}$ is a composite testing with simple null hypothesis and composite alternative hypothesis while the null and alternative hypotheses of $\mathcal{L}_{k,\tau,d,\delta_\mu,\delta_\lambda}$ are composite.

## 1.2 Classification problem

Here, we referred to the classification defined in [AK+05] which is sufficient for locating Gaussian mixtures.

**Definition 4 (classification problem)** *Given* $k$*, the number of components,* $0 < \delta < 1$*, the success probability, and* $n$ *samples* $x_1, \ldots, x_n$ *drawn from the Gaussian mixture* $g(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\lambda})$ *with unknown* $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$*. The goal is to correctly label the samples to each component in* $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ *with probability at least* $1 - \delta$*.*

Note that as long as we can have a nice classification algorithm, then we can use its result to further locate each component after the partition.

There are several algorithms for classification problem under mild separation conditions.

## 1.3 Testing evaluation

In (2) and (3), we defined a general setting for counting ad locating problem in Gaussian mixture model. One can see that the hypotheses in both testing problems are composite, except the null hypothesis of locating problem. As a result, common hypothesis testing results such as Neyman-Pearson lemma cannot be simply applied. In this subsection, we are going to discuss how to evaluate a test in different settings.

To concentrate on mathematical reasoning, we use the following notation in the following discussion. Let the parameter space be $\Theta = \Theta_0 \cup \Theta_1$ where $\Theta_0$ forms the null hypothesis and $\Theta_1$ forms the alternative hypothesis. Distribution with parameter $\theta \in \Theta$ is denoted as $P_\theta$ with support $\mathcal{X}$. Given a test $\psi : \mathcal{X} \to \{0, 1\}$, for each $\theta \in \Theta$, we define the power function as $\beta(\theta) = \mathbb{P}_{X \sim P_\theta}[\psi(X) = 1]$, *i.e.,* the probability to reject with underlying distribution $P_\theta$. One can see that for $\theta \in \Theta_0$, we want its power function to be as small as possible. On the other hand, for $\theta \in \Theta_1$, we want its power function as close to 1 as possible.

We say a test is $\alpha$ *level* if the maximum power in null hypothesis set is at most $\alpha$, *i.e.,* $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$. And a test is *unbiased* if for any $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$, $\beta(\theta_0) \leq \beta(\theta_1)$. Finally, we say a family of tests $\{\psi_\alpha\}$ is *uniformly most powerful (UMP)* if given any $0 < \alpha < 1$, among every level $\alpha$ test, $\forall \theta_1 \in \Theta_1$, $\beta_{\psi_\alpha}(\theta_1) \geq \beta(\theta_1)$.

### 1.3.1 Neyman-Pearson setting

In Neyman-Pearson setting, our goal is to find the UMP test. When the hypothesis set is simple, *i.e.,* consisting only one distribution, by Neyman-Pearson's lemma, we know that the likelihood ratio test is optimal. Moreover, by Chernoff-Stein's lemma, we know that the type-II error exponentially converges to 0 with rate function $D_{KL}(P_{\theta_1} || P_{\theta_0})$, where $D_{KL}(\cdot||\cdot)$ is the KullbackLeibler divergence.

Now, let's extend these known results to our setting.

**Locating** Consider the hypothesis testing problem in (3), we have a simple null hypothesis and a composite alternative hypothesis. Suppose the parameter in null hypothesis is $\theta_0 = (\boldsymbol{\mu}_0, \boldsymbol{\lambda}_0, \boldsymbol{\sigma}_0) \in \mathcal{G}_k$. For arbitrary $\theta_1 = (\boldsymbol{\mu}_1, \boldsymbol{\lambda}_1, \boldsymbol{\sigma}_1) \in \mathcal{G}_k - \{(\boldsymbol{\mu}_0, \boldsymbol{\lambda}_0, \boldsymbol{\sigma}_0)\}$, by Chernoff-Stein's lemma, we have an asymptotic characterization of the optimal type-II error $\beta^*(n, \alpha)$ over all level $\alpha$ test.

**Lemma 5 (Chernoff-Stein's lemma)**
*For all $\alpha \in (0, 1)$, we have* $\lim -\frac{1}{n} \log \beta^*(n, \alpha) = D_{KL}(P_{\theta_0} || P_{\theta_1})$.

**Proof:** The proof is based on relative entropy typicality, which is worked for distributions defined on $\mathbb{R}^n$. See Chapter 11.8 of [CT12] for details. ■

From Lemma 5, we know the asymptotic optimal behavior of hypothesis testing in a 1 versus 1 Neyman-Pearson setting. Note that, since the likelihood ratio test is UMP, it can achieve this rate. Now, we are going to extend the result to a 1 versus many paradigm. In composite hypothesis setting, people like to use *generalized likelihood ratio test (GLRT)*, which is a direct extension of LRT in simple hypothesis setting. Formally, we define the generalized likelihood ratio as follow.

**Definition 6 (generalized likelihood ratio)** *For a composite testing problem with null hypothesis $\Theta_0$ and alternative hypothesis $\Theta_1$. Given observation $x^{\in}\mathcal{X}^n$, we define its generalized likelihood ratio as*

$$L(x^n; \Theta_0, \Theta_1) := \frac{\sup_{\theta_1 \in \Theta_1} P_{\theta_1}(x^n)}{\sup_{\theta_0 \in \Theta_0} P_{\theta_0}(x^n)}$$

Using generalized likelihood ratio, we define the following generalized likelihood ratio test for (3).

**Definition 7** *For a composite testing problem with null hypothesis $\Theta_0$ and alternative hypothesis $\Theta_1$. Given $0 < \alpha < 1$, we define the generalized likelihood ratio test $\psi_{n,\alpha}^{LRT} : \mathcal{X}^n \to \{0, 1\}$ in two steps. Given an observation $x^n \in \mathcal{X}^n$.*

1. *Find $\theta_1^* = \sup_{\theta_1 \in \Theta_1} \frac{P_{\theta_1}(x^n)}{P_{\theta_0}(x^n)}$ such that $L(x^n; \theta_0, \theta_1^*) = L(x^n; \theta_0, \Theta_1)$.*
2. *Let $\tau_{n,\alpha}(\theta_1^*)$ be the threshold of the optimal LRT between $\theta_0$ and $\theta_1^*$. Then output $\psi_{n,\alpha}^{LRT} = \mathbf{1}_{L(x^n;\theta_0,\theta_1^*) \geq \tau_{n,\alpha}(\theta_1^*)}$.*

Note that since the parameter that maximize the generalized likelihood ratio might not always be the same, we can not apply Neyman-Pearson lemma to assert the optimality of GLRT proposed in Definition 7. That is to say, in such complicated setting, we need to compute the error rate of GLRT case by case.

### 1.3.2 Bayesian setting

In Bayesian setting, we define error function $P_e(\psi) = \pi_0 \mathbb{P}_{\theta \in \Theta_0}[\psi(X) = 1] + \pi_1 \mathbb{P}_{\theta \in \Theta 1}[\psi(X) = 0]$ for a test with a prior $(\pi_0, \pi_1)$ among the hypotheses to weight the type-I and type-II errors. For simple hypothesis testing, the optimal asymptotic behavior of the error function in Bayesian setting is characterized by the *Chernoff information*,

**Theorem 8 (Chernoff)** *Let $P_e^{(n)}$ denote the error function of the optimal testing with $n$ sample, then we have*

$$\lim_{n\to\infty} -\frac{1}{n}\log P_e^{(n)} = D(P_{\lambda^*}||P_0) = D(P_1||P_{\lambda^*})$$

*, where $P_\lambda(x) = \frac{P_0^{\lambda(x)}P_1^{1-\lambda}(x)}{\int P_0^{\lambda}(t)P_1^{1-\lambda}(t)dt}$ and $\lambda^*$ is the value such that $D(P_{\lambda^*}||P_0) = D(P_1||P_{\lambda^*})$.*

Note that the same as in Neyman-Pearson setting, Theorem 8 only apply to simple hypothesis testing. To extend it to composite setting, one need to discuss case by case or find some useful structure.

## 1.4 Model conditions

To directly deal with general Gaussian mixture model is extremely difficult and impractical since real-life applications often have some structural assumptions which make the model easier to play with. Here, we list several common conditions such as minimum distance between each components, the shape of the components, minimum weight of the components, dimensional setting, etc. Our goal is to find out the fundamental limits in these different settings.

### 1.4.1 Minimum distance

First, we consider the minimum distance $d$ between two components in Gaussian mixture model. Clearly that the smaller the $d$ is, the more difficult to test for both counting and locating.

### 1.4.2 Shape

The shape of the components in Gaussian mixture also affect the difficulty of testing. Intuitively, if every components have the same shape, say being symmetric, then it is in some sense easier than the model which allows components to have different shapes. To model the shape of each component, we consider the variance/covariance matrix of the component. The setting could be covariance matrix being diagonal, rotation, eigenvalue is bounded, etc.

### 1.4.3 Minimum weight

If the weight of a component is small, then it is difficult to be discovered, which increase the testing error. As a result, we use $\omega$ to lower bound the minimum weight

of each component, and try to find out whether $\omega$ makes counting and locating different.

### 1.4.4 Dimension

In this report, we use $p$ to denote the dimension of our Gaussian mixture model. Basically, in the very first step, we consider the Euclidean space setting. For a more general study, we should consider general probability space.

# 2 Large deviation theory for composite testing

From Lemma 5 and Theorem 8, we can easily characterize the asymptotic optimality of simple hypothesis testing. When it comes to composite testing, there are also large deviation results in a slightly modified setting. In this section, we are going to see these asymptotic results for composite testing and apply them in our study in Gaussian mixture.

## 2.1 Setup

Here, we consider samples $x_1, \ldots, x_n \in \mathcal{X}$ and our goal is to decide whether these samples are from the null hypothesis $\mathcal{H}_0 = \{P_{\theta_0} | \theta_0 \in \Theta_0\}$ or the alternative hypothesis $\mathcal{H}_1 = \{\mathbf{P}_{\theta_1} | \theta_1 \in \Theta_1\}$. We define $\mu_n$ to be the empirical measure of the samples. Note that in the following, we use $\mu$ to denote measure over $\mathcal{X}$.

Here, we define the testing function by using a decision rule $\Omega = \{\Omega(n)\}$. Concretely, $\Omega(n) = (\Omega_0(n), \Omega_1(n))$, both $\Omega_0(n)$ and $\Omega_1(n)$ contain **measure** over $\mathcal{X}$. That is to say, we make the decision according which set the empirical measure $\mu_n$ belongs to. For simplicity, the measure we talk about here only consider countably additive measure $M_1(\mathcal{X})$. One can refer to Lemma1 of [ZG91] to see why it suffices to only consider $M_1(\mathcal{X})$. The type-I and type-II errors of $\Omega(n)$ are $\alpha(\theta_0; \Omega(n)) = P_{\theta_0}[X^n \in \Omega_1(n)]$ and $\beta(\theta_1; \Omega(n)) = P_{\theta_1}[X^n \in \Omega_0(n)]$ for any pairs of $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$. Note that we does not fix $\theta_0$ or $\theta_1$. Now, we can define the error exponent for type-II error.

**Definition 9** *Let $\Omega$ be a family of decision rules and $\theta_1 \in \Theta_1$. We define the type-II error exponent of $\Omega$ w.r.t. $P_{\theta_1}$ as*

$$J(\Theta_0, P_{\theta_1}; \Omega) := \lim \inf_{n \to \infty} -\frac{1}{n} \log \beta(\theta_1; \Omega(n))$$

*Furthermore, we an define the optimal type-II error exponent of $\Omega$ w.r.t. $P_{\theta_1}$ as*

$$J^*(\Theta_0, P_{\theta_1}) := \sup_{\Omega} \{J(\Theta_0, P_{\theta_1}; \Omega) | \forall \theta_0 \in \Theta_0, \alpha(\theta_0; \Omega(n)) \geq \lambda\}$$

*, and we call the test that achieve this error rate as worst-case optimal test.*

The goal for now is slightly different from the traditional Neyman-Pearson setting in which we let the level size $\alpha$ decrease exponentially with the number of samples. Namely, we consider

$$\begin{array}{ll} \text{maximize} & \liminf_{n\to\infty} -\frac{1}{n} \log \beta(\Omega(n)) \\ \text{subject to} & \liminf_{n\to\infty} -\frac{1}{n} \log \alpha(\Omega(n)) \geq \lambda \end{array}$$

**Remark**: This setting was formulated by Hoeffding in [Hoe65]. Note that it is somewhere between Neyman-Pearson setting and Bayesian setting.

Since the support we are going to play with is continuous instead of simply being a finite set, we need to define *$\delta$-smoothing* to help us tackle the difficulty.

**Definition 10 ($\delta$-smoothing)** *For a decision rule $\Omega$, we define its $\delta$-smoothing as follow,*

$$\Omega_1^\delta(n) = \cup_{\mu \in \Omega_1(n)} B(\mu, \delta)$$
$$\Omega_0^\delta(n) = M_1(\mathcal{X}) \backslash \Omega_1^\delta(n)$$

*, where $B(\mu, \delta)$ is a $\delta$ ball for $\mu$ w.r.t Levy's metric.*

Now, we can state the large deviation theorem for composite hypothesis testing.

## 2.2 Large deviation theorem for composite hypothesis testing

First, we consider a simple case where the null hypothesis set is simple.

**Theorem 11 (LDT for simple null hypothesis)** *Let $\Lambda^\delta$ be a decision rule such that*

$$\Lambda_1^\delta(n) = (\{\mu | \inf_{\tilde{\mu} \in B(\mu, 2\delta)} I(\tilde{\mu} || P_0) \geq \lambda\})^\delta \tag{4}$$

*, where $I(\mu||P_0) = \int_{\mathcal{X}} d\mu \log \frac{d\mu}{dP_0}$, which is the KL-divergence between $\mu$ and $P_0$. Then, for any given $\delta > 0$, we have the following.*

1. *For any* $P_0 \in M_1(\mathbb{R})$,

$$\liminf_{n\to\infty} -\frac{1}{n}\log\alpha(\Lambda^\delta(n)) \geq \lambda$$

2. *For any* $P_1 \in M_1(\mathbb{R})$,

$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta(\Lambda^\delta(n)) \geq \inf_{\mu\in\Lambda_0^\delta} I(\mu||P_1) := e(\lambda,\delta,P_1)$$

3. $\Lambda^\delta$ *is* $\delta$*-optimal in the sense that for any* $\Omega$ *such that*

$$\liminf_{n\to\infty} -\frac{1}{n}\log\alpha(\Omega^{6\delta}(n)) \geq \lambda$$

*, then*

$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta(\Omega^\delta(n)) \leq e(\lambda,\delta,P_1)$$

*Namely,* $\Lambda^\delta$ *is the worst-case optimal test when* $\delta \to 0$.

Theorem 11 gave us a test $\Lambda^\delta$ and showed that it is in some sense optimal. To us, what we are interested is more on the error component, *i.e.,* $e(\lambda,\delta) = \inf_{\mu\Lambda^\delta} I(\mu, P_1)$. Note that this error component varies among different alternative distribution. Thus, the error component we are looking for is actually the following.

$$J^*(\Theta_0, P_{\theta_1}) = \min_{P_{\theta_1}|\theta_1\in\Theta_1} e(\lambda,\delta,P_{\theta_1}) \tag{5}$$

, which is the *worst-case type-II error exponent.*

# 3 One versus one testing

As composite hypothesis testing is nontrivial, we first consider the simple hypothesis testing to get some feeling.

## 3.1 Simple counting

We formulate the simple counting hypothesis testing as follow:

$$\text{(Simple Counting } \mathcal{SC}_k) \begin{cases} \mathcal{H}_0^{\mathcal{SC}_k}: & \text{Gaussian mixture } g(x;\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\sigma}) \\ \mathcal{H}_1^{\mathcal{SC}_k}: & \text{Gaussian mixture } g(x;\boldsymbol{\lambda}',\boldsymbol{\mu}',\boldsymbol{\sigma}') \end{cases} \tag{6}$$

where $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathcal{G}_k$ and $(\boldsymbol{\lambda}', \boldsymbol{\mu}', \boldsymbol{\sigma}') \in \mathcal{G} - \mathcal{G}_k$. Note that the null hypothesis and alternative hypothesis could be any pair of distribution. However, to attain the optimal performance of the test, we need to make sure that the sample is indeed from one of the two distributions. Otherwise, it is nonsense to do hypothesis testing. As a result, we can see that the worst optimal rate in the simple hypothesis testing is actually a lower bound for the error rate of composite testing.

**Example 12** Consider $\mathcal{H}_0^{\mathcal{SC}_1} : X \sim P_0(x) = g(x; 1, 0, 1)$ and $\mathcal{H}_1^{\mathcal{SC}_1} : X \sim P_1(x) = g(x; (1-\omega, \omega), (0, d), (1, 1))$. Over all level-$\alpha$ test, by Neyman-Pearson lemma and Chernoff-Stein's lemma, we know that there's a likelihood ratio test achieve optimal type-II error rate $\lim_{n\to} -\frac{1}{n} \log \beta_{n,\alpha}^{\mathcal{SC}_1} = D_{KL}(P_0||P_1)$. See Figure 1.

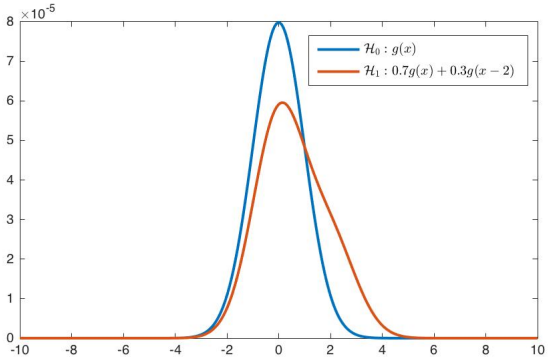


Figure 1: Simple counting test. The blue one is null hypothesis and the red one is alternative. $\omega = 0.3$ and $d = 2$.

When $d \le \frac{1}{2}$, we have a simulation result showing that $D_{KL}(P_0||P_1) \approx \frac{1}{2}\omega^2 d^2$. With this observation, we can design a simple test $\psi_{\mathcal{SC}_1}$ for $\mathcal{SC}_1$.

$$\psi_{\mathcal{SC}_1}(x^n) = \mathbf{1}_{|\frac{1}{n}\log\frac{P_0(x^n)}{P_1(x^n)} - D_{KL}(P_0||P_1)| \le \alpha} \tag{7}$$

with error rate $\frac{1}{2}\alpha^2 d^2$. We extend $\psi_{\mathcal{SC}_1}$ to a general setting in Section 4.1.

Also, there's an interesting finding that the minimum weight actually gives an asymptotic upper bound to KL-divergence for $d$ large. Concretely, the KL-divergence of the simple counting testing is $\log \frac{1}{1-\omega}$. This can be simply seem by $D_{KL}(P_0||P_1) = \int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \log \frac{e^{-x^2/2}}{(1-\omega)e^{-x/2} + \omega e^{(2xd-d^2)/2}} dx \le \int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \log \frac{1}{1-\omega} dx = \log \frac{1}{1-\omega}$.

### 3.1.1 Minimax testing region for $\mathcal{SC}_1$

Here, we are going to show a minimax testing region for $\mathcal{SC}_1$ discussed in Example 12. Recall that in this problem there are two parameters $d$ and $\omega$ decide the difficulty of testing. Intuitively, when $d$ is small it becomes more difficult to distinguish the two hypotheses, so does $\omega$. Our goal here is to find out the information-theoretical lower bound in the minimax setting and thus find out a region (w.r.t. $n, d, \omega$) where it is impossible to have a good testing.

**Theorem 13** *For any $0 < \delta < \frac{1}{2}$, $\exists \bar{c}, \underline{c}$ such that when*

- *$\omega d^2 > \frac{\bar{c}}{\sqrt{n}}$, there exists a test $\psi$ such that*

$$\sup_{\theta_0 \in \Theta_0^{\mathcal{SC}_1}, \theta_1 \in \Theta_1^{\mathcal{SC}_1}} \{\mathbb{P}_{P_{\theta_0}^n}[\psi(X^n) = 1], \mathbb{P}_{P_{\theta_1}^n}[\psi(X^n) = 0]\} \leq \delta$$

- *$\omega d^2 < \frac{\underline{c}}{\sqrt{n}}$, then for any test $\psi$,*

$$\sup_{\theta_0 \in \Theta_0^{\mathcal{SC}_1}, \theta_1 \in \Theta_1^{\mathcal{SC}_1}} \{\mathbb{P}_{P_{\theta_0}^n}[\psi(X^n) = 1], \mathbb{P}_{P_{\theta_1}^n}[\psi(X^n) = 0]\} \geq \delta$$

## 3.2 Locating

Now, we consider the simple locating testing as follow.

$$\text{(Simple Locating } \mathcal{SL}_k) \begin{cases} \mathcal{H}_0^{\mathcal{SL}_k} : & \text{Gaussian mixture } g(x; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\ \mathcal{H}_1^{\mathcal{SL}_k} : & \text{Gaussian mixture } g(x; \boldsymbol{\lambda}', \boldsymbol{\mu}', \boldsymbol{\sigma}') \end{cases} \tag{8}$$

where $(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}) \in \mathcal{G}_k$ and $(\boldsymbol{\lambda}', \boldsymbol{\mu}', \boldsymbol{\sigma}') \in \mathcal{G} - \{(\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma})\}$.

**Example 14** Consider $\mathcal{H}_0^{\mathcal{SL}_2} : X \sim P_0(x) = g(x; (\frac{1}{2} + \omega, \frac{1}{2} - \omega), (\frac{d}{2}, \frac{-d}{2}), 1)$ and $\mathcal{H}_1^{\mathcal{SL}_2} : X \sim P_1(x) = g(x; (\frac{1}{2}, \frac{1}{2}), (\frac{d}{2}, \frac{-d}{2}), 1)$. Over all level-$\alpha$ test, by Neyman-Pearson lemma and Chernoff-Stein's lemma, we know that there's a likelihood ratio test achieve optimal type-II error rate $\lim_{n\to} -\frac{1}{n} \log \beta_{n,\alpha}^{\mathcal{SL}_2} = D_{KL}(P_0 || P_1)$. See Figure 2.

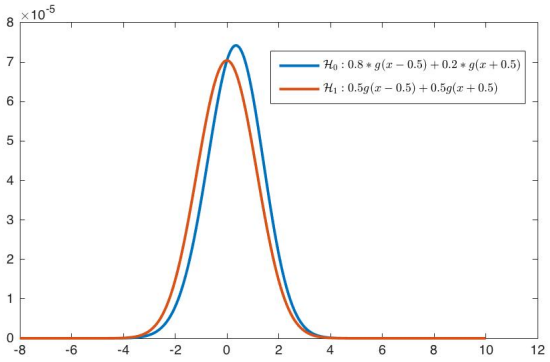


Figure 2: Simple locating test. The blue one is null hypothesis and the red one is alternative. $\omega = 0.3$ and $d = 2$.

When $d \leq \frac{1}{2}$, we have a simulation result showing that $D_{KL}(P_0||P_1) \approx \frac{1}{2}\omega^2 d^2$.

# 4 Composite test

## 4.1 Composite test for counting

Recall that in Example **??**, we construct a simple hypothesis test $\psi_{\mathcal{SC}_1}$ for simple counting test $\mathcal{SC}_1$. Now, we extend it to the composite counting test $\mathcal{C}_1$. WLOG, we consider the null hypothesis of $\mathcal{C}_1$ is $g(x; 1, 0, 1)$ and given samples $\mathbf{x} = x_1, \ldots, x_n$, we define two likelihood ratios as follow.

$$\Lambda_l(\mathbf{x}) := \frac{1}{n} \log \frac{\prod_{i=1}^n P_0(x_i)}{\prod_{i=1}^n P_l(x_i)}$$
$$\Lambda_r(\mathbf{x}) := \frac{1}{n} \log \frac{\prod_{i=1}^n P_0(x_i)}{\prod_{i=1}^n P_r(x_i)}$$
$$\Lambda_c(\mathbf{x}) := \frac{1}{n} \log \frac{\prod_{i=1}^n P_0(x_i)}{\prod_{i=1}^n P_c(x_i)}$$

, where $P_0 = g(\cdot; 1, 0, 1)$, $P_l = g(\cdot; (1-\omega, \omega), (0, -d), (1, 1))$, $P_r = g(\cdot; (1-\omega, \omega), (0, d), (1, 1))$, and $P_c = g(\cdot; (\frac{1}{2}, \frac{1}{2}), (-d, d), (1, 1))$. Then, we can define a two-sided test as follow.

$$\psi_{\mathcal{C}_1}(\mathbf{x}) = \mathbf{1}_{\Lambda_l(\mathbf{x}) \leq D_{KL}(P_0||P_l) - \alpha \text{ or } \Lambda_r \leq D_{KL}(P_0||P_r) - \alpha \text{ or } \Lambda_c \leq D_{KL}(P_0||P_c) - \alpha} \tag{9}$$

One can show that $\psi_{\mathcal{C}_1}$ achieves the same error rate as $\psi_{\mathcal{SC}_1}$.

**Proposition 15** *If $D_{KL}(P_0||P_l) \geq \alpha$, the error rate of $\psi_{\mathcal{C}_1}$ is at most $\frac{1}{2}\alpha^2 d^2$.*

**Proof:** By the AEP for relative entropy, we know that the type-I error can be controlled within $\alpha$ when $n$ large enough. As to type-II error, consider arbitrary $P \in \mathcal{H}_1^{\mathcal{C}_1}$, one can see that either $D_{KL}(P_0||P) \geq D_{KL}(P_0||P_l)$ or $D_{KL}(P_0||P) \geq D_{KL}(P_0||P_r)$ or $D_{KL}(P_0||P) \geq D_{KL}(P_0||P_c)$. (Discussed later) Assume $D_{KL}(P_0||P) \geq D_{KL}(P_0||P_l)$, by AEP for relative entropy, we have

$$\frac{1}{n}\log\frac{\prod_i P(x_i)}{\prod_i P_0(x_i)} \geq -D_{KL}(P||P_0) - \epsilon \text{ w.p. } 1-\epsilon$$
$$\frac{1}{n}\log\frac{\prod_i P(x_i)}{\prod_i P_l(x_i)} \leq D_{KL}(P||P_l) + \epsilon \text{ w.p. } 1-\epsilon$$

That is,

$$\begin{aligned}\frac{1}{n}\log\frac{\prod_i P_0(x_i)}{\prod_i P_l(x_i)} &\leq D_{KL}(P||P_l) - D_{KL}(P||P_0) + 2\epsilon \\ &= 2\epsilon \text{ w.p. } 1-2\epsilon\end{aligned}$$

■

# 5 Peace region, where locating is no harder than counting

In this section we show a region (setting of hypothesis testing) where counting is no easier than locating. Concretely, a lower bound for the testing error of locating provides a lower bound for the testing error of counting within constant multiplicative factor. Namely, a difficult instance in counting can be reduced to a difficult instance in locating.

Recall that to construct a testing lower bound, we find a pair of distribution, such that one is in the null hypothesis and the other is in the alternative. Our goal would be minimize their total variation as small as possible since total variation will be the hypothesis testing error. More generally, sometimes total variation is difficult to compute. As a result, we will turn to other f-divergences and use inequalities such as Pinsker's inequality to upper bound the total variation among the two chosen distributions.

**Theorem 16 (peace region)** *If we have $P_0 \in \mathcal{H}_0^{\mathcal{L}_k}$, $P_1 \in \mathcal{H}_1^{\mathcal{L}_k}$ and the components in $P_0$, $P_1$ do not occupy the same position. Then, we can have $Q_0 \in \mathcal{H}_0^{\mathcal{C}_k}$, $Q_1 \in \mathcal{H}_1^{\mathcal{C}_k}$ such that $D_f(Q_0||Q_1) \leq \frac{1}{2}D_f(P_0||P_1)$, where $D_f(\cdot||\cdot)$ is arbitrary f-divergences.*

**Proof:** Take $Q_0 = P_0$ and $Q_1 = \frac{1}{2}P_0 + \frac{1}{2}P_1$. As there exists component in $P_0$ not in $P_1$, $Q_1$ must have more than $k$ components and thus $Q_1 \in \mathcal{H}_1^{\mathcal{C}_k}$. By the convexity of f-divergence, we have

$$\begin{aligned} D_f(Q_0||Q_1) &= D_f(P_0||\frac{1}{2}P_0 + \frac{1}{2}P_1) \\ &\leq \frac{1}{2}D_f(P_0||P_0) + \frac{1}{2}D_f(P_0||P_1) = \frac{1}{2}D_f(P_0||P_1) \end{aligned}$$

■

Theorem 16 told us that when choosing two distributions with various components location in $\mathcal{C}_k$ can simply imply a lower bound in $\mathcal{L}_k$ with the same rate. As a result, if we want to construct a lower for $\mathcal{L}_k$, we should not consider the two distributions with different components' locations. Namely, separation instance for counting and locating only happen on two mixtures having same position but different weight.

**Remark**: As a matter of fact, here we a little abuse the setting since the construction of $Q_1$ might affect the minimum weight with factor $\frac{1}{2}$.

# 6 $\mathcal{C}_{1,\tau,d}$

In this section, we first propose a lower bound for $\mathcal{C}_{1,\tau,d}$ via multi-point Le Cam's method and then construct an optimal test for the case where $d < 1$.

## 6.1 Lower bound for $\mathcal{C}_{1,\tau,d}$

**Theorem 17 (lower bound for $\mathcal{C}_{1,\tau,d}$)** *For $0 < v < \frac{1}{4}$ and $n \leq \frac{2pC_v}{\tau^2(\cosh d^2 - 1)}$. For any test $\psi$ for $\mathcal{C}_{1,\tau,d}$, $\max_{P\in\mathcal{H}_0, Q\in\mathcal{H}_1}\{\mathbb{P}_P^n[\psi = 1], \mathbb{P}_Q^n[\psi = 0]\} \geq \frac{1}{2} - v$.*

**Proof:** We use multiple-points Le Cam's method to construct the lower bound. The strategy is straightforward, for arbitrary instance in $\mathcal{H}^{\mathcal{C}_{1,\tau,d}}$, say $P_0 = P_{(0,1)}$, we consider its neighboring 2-components mixtures, *i.e.*, $P_{((0,de_j),(1-\tau,\tau))}$, for $j = \pm 1, \ldots, \pm p$. Then, take the average over these neighboring mixtures, we have $P_1 = \frac{1}{2p}\sum_{j\in[\pm p]}$. Here, $e_j$ is the unit vector in the $|j|$-th dimension in $\mathbb{R}^p$ where the sign of $j$ decides the sign of the unit vector. For convenient, we let $[\pm p] = \pm 1, \ldots, \pm p$.

**Lemma 18** *For $0 < v < \frac{1}{4}$, let $C_v = \log[(1+8v^2) \wedge \log(\frac{e}{2-4v})]$, when $n \leq \frac{2pC_v}{\tau^2(\cosh d^2 - 1)}$, we have $\chi(P_1^n||P_0^n) \leq e^{C_v} - 1$.*

**Proof:** We defer the proof to Appendix. ■

Thus, for any test $\psi$ for $\mathcal{C}_{1,\tau,d}$, when $n \leq \frac{2pC_v}{\tau^2(\cosh d^2 - 1)}$, we have

$$\begin{aligned}
\max_{P\in\mathcal{H}_0, Q\in\mathcal{H}_1}\{\mathbb{P}_P^n[\psi=1], \mathbb{P}_Q^n[\psi=0]\} &\geq \max_{j\in[\pm p]}\{\mathbb{P}_{P_0}^n[\psi=1], \mathbb{P}_{P_{((0,de_j),(1-\tau,\tau))}}^n[\psi=0]\} \\
&= \max\{\mathbb{P}_{P_0}^n[\psi=1], \mathbb{P}_{P_1}^n[\psi=0]\} \\
(\because \text{Multiple-points Le Cam's method}) &\geq e^{-\chi(P_1^n||P_0^n)} \wedge \frac{1-\sqrt{\chi(P_1^n||P_0^n)/2}}{2} \\
(\because \text{Lemma 18}) &\geq \frac{1}{2} - v
\end{aligned}$$

■

## 6.2 Optimal test for $\mathcal{C}_{1,\tau,d}$ when $d < 1$

In this subsection, we proposed an optimal test for $\mathcal{C}_{1,\tau,d}$ when $d < 1$ based on moment method. Given samples $X^n = X_1, \ldots, X_n$, define $S(X^n) = \frac{1}{n}\sum_{i=1}^n[(X_i - \bar{X}_n)^2] - p$, define the test $\psi$ as follow:

$$\psi(X^n) = \begin{cases} 0 & , \ S(X^n) < \frac{\tau(1-\tau)d^2}{2}p \\ 1 & , \ \text{otherwise} \end{cases}$$

**Theorem 19** *For any $0 < \delta < \frac{1}{2}$ and $0 < \tau < \frac{1}{4}$, when $n \gtrsim \frac{p\log\frac{1}{\delta}}{\tau^2 d^4} \wedge \frac{\sqrt{p}\log\frac{1}{\delta}}{\tau d^2} \wedge p\log\frac{1}{\delta} \wedge \frac{\log\frac{1}{\delta}}{\tau}$, we have*

$$\max_{P\in\mathcal{H}_0, Q\in\mathcal{H}_1}\{\mathbb{P}_P[\psi=1], \mathbb{P}_Q[\psi=0]\} < \delta$$

**Proof:** The proof is divided into two parts. The first part showed that when $n > \frac{p\log\frac{1}{\delta}}{\tau^2 d^4} \wedge \frac{\sqrt{p}\log\frac{1}{\delta}}{\tau d^2}$, $\forall P \in \mathcal{H}_0$, $\mathbb{P}_P[\psi = 1] < \delta$. The second part showed that when $n \gtrsim \frac{p\log\frac{1}{\delta}}{\tau^2 d^4} \wedge \frac{\sqrt{p}\log\frac{1}{\delta}}{\tau d^2} \wedge p\log\frac{1}{\delta} \wedge \frac{\log\frac{1}{\delta}}{\tau}$, $\forall Q \in \mathcal{H}_1$, $\mathbb{P}_Q[\psi = 0] < \delta$. Before we start the proof, let's first rewrite the test statistics $S(X^n)$.

$$S(X^n) = \frac{1}{n}\sum_{i=1}^n[(X_i - \bar{X}_n)^2] - p = \frac{1}{n}\sum_{i=1}^n ||X_i - \mu||_2^2 - ||\bar{X}_n - \mu||_2^2 - p \tag{10}$$

, where $\mu = \mathbb{E}[X_i]$.

(1) For any $P \in \mathcal{H}_0$, we have $X_i - \mu \sim N(0, I_p)$ for all $i = 1, \ldots, n$ and $\bar{X}_n - \mu \sim N(0, \frac{1}{n} I_p)$.

$$\begin{aligned}\mathbb{P}_P[\psi = 1] &= \mathbb{P}_P[S(X^n) \geq \frac{\tau(1-\tau)d^2}{2}] \\ &\leq \mathbb{P}_P[\frac{1}{n}\sum_{i=1}^{n} ||X_i - \mu||_2^2 - p \geq \frac{\tau(1-\tau)d^2}{4}] + \mathbb{P}_P[||\bar{X}_n - \mu||_2^2 < \frac{p}{n} - \frac{\tau(1-\tau)d^2}{4}] \\ &\leq \delta\end{aligned}$$

Apply the following concentration inequalities for sum of chi-square random variables.

**Lemma 20** *[LM00] Let $Z_1, \ldots, Z_m$ be $m$ independent standard chi-square random variable and $Y = \sum_{i=1}^{m} Z_i$, we have*

$$\mathbb{P}[Y \geq m + 2\sqrt{mt} + 2t] \leq \exp(-t) \tag{11}$$

$$\mathbb{P}[Y \leq m - 2\sqrt{mt}] \leq \exp(-t) \tag{12}$$

*for any $t \geq 0$.*

With Lemma 20, we have

$$\mathbb{P}_P[\frac{1}{n}\sum_{i=1}^{n} ||X_i - \mu||_2^2 \geq p + 2\sqrt{\frac{pt}{n}} + 2\frac{t}{n}] \leq \exp(-t)$$

With some computation, we have

$$\mathbb{P}_P[\frac{1}{n}\sum_{i=1}^{n} ||X_i - \mu||_2^2 \geq p + \frac{\tau(1-\tau)d^2}{4}] \leq \begin{cases} \exp(-\frac{-n\tau(1-\tau)d^2}{32}) & , \text{ when } p < \frac{\tau(1-\tau)d^2}{32} \\ \exp(-\frac{n\tau^2(1-\tau)^2d^4}{256p}) & , \text{ when } p \geq \frac{\tau(1-\tau)d^2}{32} \end{cases}$$

On the other hand, by Lemma 20, we have

$$\mathbb{P}_P[||\bar{X}_n - \mu||_2^2 \leq \frac{p}{n} - 2\frac{\sqrt{pt}}{n}] \leq \exp(-t)$$

With some computation, we have

$$\mathbb{P}_P[||\bar{X}_n - \mu||_2^2 \leq \frac{p}{n} - \frac{\tau(1-\tau)d^2}{4}] \leq \exp(-\frac{n^2\tau^2(1-\tau)^2d^4}{64p})$$

That is, when $n \gtrsim \frac{p \log \frac{1}{\delta}}{\tau^2 d^4} \wedge \frac{\sqrt{p} \log \frac{1}{\delta}}{\tau d^2}$, $\mathbb{P}_P[\psi = 1] \leq \delta$.

(2) For any $Q \in \mathcal{H}_1$, let's first rewrite the error probability as follow.

$$\begin{aligned}\mathbb{P}_Q[\psi = 0] &= \mathbb{P}_Q[S(X^n) > \frac{\tau(1-\tau)d^2}{2}] \\ &\leq \mathbb{P}_Q[\frac{1}{n}\sum_{i=1}^{n} ||X_i - \mu||_2^2 - p - \tau(1-\tau)d^2 < \frac{-\tau(1-\tau)d^2}{4}] \\ &+ \mathbb{P}_Q[||\bar{X}_n - \mu||_2^2 \geq \frac{\tau(1-\tau)d^2}{4}] \\ &= A + B \end{aligned} \tag{13}$$

Rewrite random variable $X_i = N_i + L_i$, where $N_i \sim N(0, I_p)$ is the Gaussian kernel and $L_i$ is the location vector. Then,the first term can be divided into three parts.

$$\begin{aligned}A \leq \mathbb{P}_Q[\frac{1}{n}\sum_{||N_i||_2^2} -p \leq \frac{-\tau(1-\tau)d^2}{12}] \\ + \mathbb{P}_Q[\frac{1}{n}\sum_{i=1}^{n} ||L_i - \mu||_2^2 - \tau(1-\tau)d^2 \leq \frac{-\tau(1-\tau)d^2}{12}] \\ + \mathbb{P}_Q[\frac{1}{n}\sum_{i=1}^{n} 2N_i^T(L_i - \mu) \leq \frac{-\tau(1-\tau)d^2}{12}]\end{aligned}$$

Apply Lemma 20 and Hoeffding's inequality, we have

$$A \leq \exp(-\frac{n\tau^2(1-\tau)^2d^4}{576p}) + \exp(-\frac{n}{72}) + \exp(-\frac{n\tau^2(1-\tau)^2d^2}{288})$$

As to the second part of (13), we can rewrite it as follow.

$$\begin{aligned}B &\leq \mathbb{P}_Q[||\frac{1}{n}\sum_{i=1}^{n} N_i||_2^2 + ||\frac{1}{n}\sum_{i=1}^{n} L_i - \mu||_2^2 + 2(\frac{1}{n}\sum_{i=1}^{n} N_i)^T(\frac{1}{n}L_i - \mu) \geq \frac{\tau(1-\tau)d^2}{4}] \\ &\leq \mathbb{P}_Q[||\frac{1}{n}\sum_{i=1}^{n} N_i||_2^2 \geq \frac{p}{n} + \frac{\tau(1-\tau)d^2}{12}] \\ &+ \mathbb{P}_Q[||\frac{1}{n}\sum_{i=1}^{n} L_i - \mu||_2^2 \geq \frac{\tau(1-\tau)d^2}{12}] \\ &+ \mathbb{P}_Q[2(\frac{1}{n}\sum_{i=1}^{n} N_i)^T(\frac{1}{n}\sum_{i=1}^{n} L_i - \mu) \geq \frac{\tau(1-\tau)d^2}{12}]\end{aligned}$$

Apply Hoeffding's inequality, we have

$$B \leq \exp(-\frac{n^2\tau^2(1-\tau)^2d^4}{48^2p}) + 2\exp(-\frac{n\tau(1-\tau)}{(1-2\tau)}) + \exp(-\frac{n}{288p(1-\tau)^2})$$

To sum up, when $n \gtrsim \frac{p\log\frac{1}{\delta}}{\tau^2d^4} \wedge \frac{\sqrt{p}\log\frac{1}{\delta}}{\tau d^2} \wedge p\log\frac{1}{\delta} \wedge \frac{\log\frac{1}{\delta}}{\tau}$, $\mathbb{P}_Q[\psi = 0] \leq \delta$. We left the details in Appendix **??**.

■

# 7 A lower bound for locating

In this section, we construct a lower bound for $\mathcal{L}_2$ via finding difficult instance and approximate its total variation. First of all, we need the following lemma to relate total variation to testing error.

**Lemma 21** *For any hypothesis test $\psi : \mathcal{X} \to \{0,1\}$ over distribution family $\mathcal{P}$, we have*

$$\mathbb{P}_{\mathcal{H}_0}[\psi(X) = 1] + \mathbb{P}_{\mathcal{H}_1}[\psi(X) = 0] \geq 1 - \min_{P \neq Q \in \mathcal{P}} D_{TV}(P, Q)$$

As the total variation between Gaussian mixture is difficult to approximate, we approximate the KL-divergence instead. With the help of Pinsker's inequality, we can have a fairly nice upper bound for total variation.

**Lemma 22 (Pinsker's inequality)** *For any two distributions $P$, $Q$ defined on the same probability space, we have*

$$D_{KL}(P||Q) \leq D_{TV}(P, Q) \leq \sqrt{\frac{1}{2}D_{KL}(P||Q)}$$

**Corollary 23** *For any hypothesis test $\psi : \mathcal{X} \to \{0,1\}$ over distribution family $\mathcal{P}$, we have*

$$\mathbb{P}_{\mathcal{H}_0}[\psi(X) = 1] + \mathbb{P}_{\mathcal{H}_1}[\psi(X) = 0] \geq \frac{1}{2}\exp(-D_{KL}(P||Q))$$

*, where $P$ comes from $\mathcal{H}_0$ and $Q$ comes from $\mathcal{H}_1$.*

Finally, we find a difficult instance in $\mathcal{G}_2$ with minimum weight $\alpha$ and derive the following lower bound.

**Theorem 24 (lower bound for locating)** *For any test for $n$ samples $\psi_n$ of $\mathcal{L}_2$ where $\mathcal{G}_2$ has minimum weight $\alpha$ and minimum distance $d < \frac{1}{2}$, we have*

$$\mathbb{P}_{\mathcal{H}_0}[\psi_n(X) = 1] + \mathbb{P}_{\mathcal{H}_1}[\psi_n(X) = 0] \geq \exp(-cn\alpha^2 d^2)$$

*, for some constant $c$.*

**Proof:** We consider two distributions $P_0$ and $P_1$ as follow.

- $P_0$: $g(t; (\frac{1}{2} + \alpha, \frac{1}{2} - \alpha), (-\frac{d}{2}, \frac{d}{2}), (1, 1))$
- $P_1$: $g(t; (\frac{1}{2}, \frac{1}{2}), (-\frac{d}{2}), (1, 1)$

One can show that $D_{KL}(P_0||P_1) = O(\alpha^2 d^2)$. As a result, by the tensor product property of KL-divergence and Corollary 23, we have $\mathbb{P}_{\mathcal{H}_0}[\psi_n(X) = 1]+\mathbb{P}_{\mathcal{H}_1}[\psi_n(X) = 0] \geq \exp(-cn\alpha^2 d^2)$. ■

# 8 Candidate of optimal testing - Moment method

In this section, we consider the moment estimator proposed by Dacunha-Castelle, Didier and Gassiat in [DCG97]. Our goal is to find the minimax risk of this kind of moment estimator.

## 8.1 General setting

First, rewrite the definition of Gaussian mixture in (1).

$$g(x; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \int g(x; \mu, \sigma) d\lambda(\mu, \sigma) \tag{14}$$

, where $\lambda(\mu, \sigma)$ is the density function defined on the domain of mean $\mu$ and variance $\sigma$. In our Gaussian mixture setting, $\lambda$ will be a discrete measure recording the mass of each component. In the following discussion, basically we focus on the case where every component have the same unit variance. That is, $\lambda$ will be defined only on the domain of mean.

Now, given $p \in \mathbb{N}$, we define the $p$-th moment vector $\lambda(\Phi_p)$ of the density measure as follow.

$$\begin{aligned} \Phi_p &= (1, \mu, \mu^2, \ldots, \mu^{2p}) \\ \lambda(\Phi_p) &= \int \Phi_p(\mu) d\lambda(\mu) \end{aligned} \tag{15}$$

Note that $\lambda(\Phi_p)$ contains the first $2p+1$-th moments of $\lambda$. Next, define the Hankel's matrix for a length $2p+1$ vector $c^P = (c_0, c_1, \ldots, c_{2p})$ as follow.

$$H(c^p) = \begin{pmatrix} c_0 & c_1 & c_2 & \cdots & c_p \\ c_1 & c_2 & c_3 & \cdots & c_{p+1} \\ c_2 & c_3 & c_4 & \cdots & c_{p+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_p & c_{p+1} & c_{p+2} & \cdots & c_{2p} \end{pmatrix}$$

The Hankel's matrix of the $p$-th moment vector of $\lambda$ is $H(\lambda(\Phi_p))$. Finally, define $K_p := \{c^p \in \mathbb{R}^{2p+1} | c_0 = 1, \exists \text{ positive } \lambda \text{ s.t. } c^p = \lambda(\Phi_p)\}$. We have the following theorem using Hankel's matrix to test the number of components in $\lambda$.

**Theorem 25 (Main theorem of moment method)**

1. *$H(c^p)$ is non-negative iff $c^p \in K_p$.*
2. *$det(H(c^p)) = 0$ iff $\forall \lambda$ s.t. $\lambda(\Phi_p) = c^p$, $\lambda$ is discrete and supported by at most $p$ points.*

**Proof:** We have the following identity. For all $u \in \mathbb{R}^{p+1}$ and $c^p = \lambda(\Phi_p)$ for some $\lambda$,

$$\begin{aligned} u^T H(c^p) u &= \sum_{i,j=0}^{p} u_i u_j H(c^p)_{i,j} = \sum_{i,j=0}^{p} u_i u_j c_{i+j-2} \\ &= \sum_{i,j=0}^{p} u_i u_j \int \mu^{i+j-2} d\lambda(\mu) \\ &= \int (\sum_{i=0}^{p} u_i \mu^{i-1})^2 d\lambda(\mu) \end{aligned}$$

■

Using Theorem 25, we can construct the following moment testing $\psi^{moment}$ for the following modified counting problem.

$$(\text{Counting } \mathcal{C}_{k,\tau,d}) \begin{cases} \mathcal{H}_0^{\mathcal{C}_{k,\tau,d}} : & \forall (\boldsymbol{\mu}_0, \boldsymbol{\lambda}_0) \in \cup_{l \le k} \mathcal{G}_{l,\tau,d} \\ \mathcal{H}_1^{\mathcal{C}_{k,\tau,d}} : & \forall (\boldsymbol{\mu}_1, \boldsymbol{\lambda}_1) \in \cup_{l > k} \mathcal{G}_{l,\tau,d} \end{cases} \tag{16}$$

$\psi^{moment}$ is straightforward. Let $\hat{H}_{n,p}$ be the empirical Hankel's matrix of the samples $\mathbf{X} = (X_1, \ldots, X_n)$ where $(\hat{H}_{n,p})_{i,j} = \frac{1}{n}\sum_{t=1}^{n} X_t^{i+j-2}$, define the following test.

$$\psi(\mathbf{X}) := \mathbf{1}_{det(\hat{H}_{n,p}) > \frac{1}{2}\tau(1-\tau)d^2} \tag{17}$$

## 8.2 Analysis

**Definition 26 (Lipschitz function w.r.t. $\ell_p$ norm)** *$f : \mathbb{R}^n \to \mathbb{R}$ is $\lambda$-Lipschitz w.r.t. $\ell_p$ norm if for all $x, y$,*

$$|f(x) - f(y)| \leq \lambda ||x - y||_p$$

**Theorem 27 (concentration of Lipschitz function)** *Suppose $X_1, \ldots, X_n$ are independent and bounded with $a_i \leq b_i$. Then for any $\lambda$-Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ w.r.t. $\ell_1$ norm,*

$$\mathbb{P}[f \geq \mathbb{E}[f] + \epsilon] \leq e^{\frac{-2\varepsilon^2}{\lambda^2 \sum_i (b_i - a_i)^2}} \tag{18}$$

# 9 Computational issues

- It is NP-hard to minimize k-means sum-of-square error. [ADHP09]
  - It's NP-hard to compute the MLE.
- [LSW15] showed that k-means is inapproximable.
- [AK+05] presented a poly-time algorithm for mixture problem under separation constraints.
  - They also presented a poly-time approximation algorithm for mixture problem with no assumption.
  - They also mentioned a possible reduction to show the NP-hardness of exact problem.
  - [Meg90] showed that it's NP-hard to decide when a set of points can be covered by two spheres.
- [SSR] showed some empirical studies and found some gap between information limit and computational limit.

# 10 Lower bound for counting

In Section 8, we analyzed a hypothesis testing based on moment method. Intuitively, this moment estimator utilize the intrinsic characteristic of counting. As a result, we would guess that it might be the optimal estimator. In this section, we then try to derive minimax lower bound and hope to meet the solvable region of moment estimator.

## 10.1 A first trial - Le Cam's multiple-points method

The simplest way to derive hypothesis testing lower bound is to apply Le Cam's method. Let's start with introducing the Le Cam's two-points method. First, we find two distributions where one lies in the null hypothesis and the other lies in the alternative. Then compute the f-divergences of these two distributions. With some help from the inequalities among f-divergences, we can lower bound the hypothesis testing error with it. For some small dimension cases, the Le Cam's two-points method can produce optimal lower bound. However, as the dimension grows larger, most of the time the two-points construction is not optimal. As a result, we turn to the so called Le Cam's multiple-points method. The idea is actually quite simple, we consider more that one distributions in the alternative hypothesis and average them to yield a mixture. We then compute the f-divergences among a distribution from the null hypothesis and this mixture. With simple argument, we can lower bound the hypothesis testing error with this f-divergence.

Here, we consider a direct lower bound construction with Le Cam's multiple-points method. First, we pick arbitrary distribution in the null hypothesis of $\mathcal{C}_{k,\tau,d}$ with uniform weighting, say $P_0(\mathbf{x}) = \sum_{i\in[k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_i)$. Then consider the neighboring distributions of $P_0$ which have $k+1$ components, *i.e.,* $S = \{P_i | P_i = (1-\tau)P_0 + \tau g(\mathbf{x} - \boldsymbol{\mu}_i - de_1), i \in [k]\}$, where $e_1$ is the unit vector of the first dimension. Finally, we consider the mixture $P_1(\mathbf{x}) = \frac{1}{k}\sum_{i\in[k]} P_i$ and compute the $\chi^2$-divergence among $P_0$ and $P_1$.

$$\begin{aligned}\chi^2(P_1||P_0) &= \mathbb{E}_{P_0}[(\frac{P_1}{P_0})^2 - 1] \\ &= \frac{1}{k^2}\sum_{i,i'\in[k]} \mathbb{E}_{P_0}[\frac{P_i}{P_0}\frac{P_{i'}}{P_0} - 1]\end{aligned}$$

First, find a naive upper bound for the ratio term.

$$\begin{aligned}
\frac{P_i(\mathbf{x})}{P_0(\mathbf{x})} &= 1 - \tau + \tau \frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_1)}{\sum_{i' \in [k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_{i'})} \\
\frac{P_i(\mathbf{x})}{P_0(\mathbf{x})} \frac{P_{i'}(\mathbf{x})}{P_0(\mathbf{x})} &= (1-\tau)^2 + \tau(1-\tau)[\frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_1)}{\sum_{i'' \in [k]} g(\mathbf{x} - \boldsymbol{\mu}_{i''})} + \frac{g(\mathbf{x} - \boldsymbol{\mu}_{i'} - de_1)}{\sum_{i'' \in [k]} g(\mathbf{x} - \boldsymbol{\mu}_{i''})}] \\
&\quad + \tau^2 \frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_1) g(\mathbf{x} - \boldsymbol{\mu}_{i'} - de_1)}{[\sum_{i'' \in [k]} g(\mathbf{x} - \boldsymbol{\mu}_{i''})]^2}
\end{aligned}$$

Evaluate the expectation.

$$\begin{aligned}
\mathbb{E}_{P_0}[\frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_1)}{\sum_{i'' \in [k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_{i''})}] &= 1 \\
\mathbb{E}_{P_0}[\frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_1) g(\mathbf{x} - \boldsymbol{\mu}_{i'} - de_1)}{[\sum_{i'' \in [k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_{i''})]^2}] &\le k \int \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}[||\mathbf{x} - \boldsymbol{\mu}_i - de_1||_2^2 + ||\mathbf{x} - \boldsymbol{\mu}_{i'} - de_1||_2^2 - ||\mathbf{x} - \boldsymbol{\mu}_i||_2^2]} d\mathbf{x} \\
&= k \int \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}[||\mathbf{x} - \boldsymbol{\mu}_i - de_1 - de_1||_2^2 + 2de_1^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}) - 2d^2 e_1^T e_1]} \\
(\because \text{we can swap } \boldsymbol{\mu}_i \text{ and } \boldsymbol{\mu}_{i'}) &\le k e^{d^2}
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\chi^2(P_1 || P_0) &\le (1-\tau)^2 + 2\tau(1-\tau) + \tau^2 k e^{d^2} \le 1 + \tau^2 k e^{d^2} \\
\chi^2(P_1^n || P_0^n) &\le [\chi^2(P_1 || P_0) + 1]^n - 1 \le (1 + \tau^2 e^{d^2})^n - 1 \\
&\le e^{n\tau^2 k e^{d^2}} - 1
\end{aligned}$$

**Theorem 28 (Lower bound for counting (with $\tau$))** *For $\mathcal{C}_{k,\tau,d}$, when $n \le \frac{C_v}{k\tau^2 e^{d^2}}$, we have*

$$\begin{aligned}
\max_{P \in \mathcal{H}_0^{\mathcal{C}_{k,\tau,d}}, Q \in \mathcal{H}_1^{\mathcal{C}_{k,\tau,d}}} \{\mathbb{P}_P^n[\psi = 1], \mathbb{P}_Q^n[\psi = 0]\} &\ge \max_{i \in [k]} \{\mathbb{P}_{P_0}^n[\psi = 1], \mathbb{P}_{P_i}^n[\psi = 0]\} \\
&= \max\{\mathbb{P}_{P_0}^n[\psi = 1], \mathbb{P}_{P_1}^n[\psi = 0]\} \\
(\because \textit{Le Cam's multiple-points method}) &\ge e^{-\chi(P_1^n || P_0^n)} \wedge \frac{1 - \sqrt{\chi(P_1^n || P_0^n)/2}}{2} \\
&\ge \frac{1}{2} - v
\end{aligned}$$

Here, we consider a direct lower bound construction with Le Cam's multiple-points method. First, we pick arbitrary distribution in the null hypothesis of $\mathcal{C}_{k,\tau,d}$ with uniform weighting, say $P_0(\mathbf{x}) = \sum_{i\in[k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_i)$. Then consider the neighboring distributions of $P_0$ which have $k+1$ components, *i.e.*, $S = \{P_{i,j} | P_{i,j} = (1-\tau)P_0 + \tau g(\mathbf{x} - \boldsymbol{\mu}_i - de_j), i \in [k], j \in [\pm p]\}$, where $e_j$ is the unit vector of the $j$-th dimension. Finally, we consider the mixture $P_1(\mathbf{x}) = \frac{1}{kp} \sum_{i\in[k],j\in[\pm p]} P_{i,j}$ and compute the $\chi^2$-divergence among $P_0$ and $P_1$.

$$\begin{aligned}\chi^2(P_1||P_0) &= \mathbb{E}_{P_0}[(\frac{P_1}{P_0})^2 - 1] \\ &= \frac{1}{4k^2p^2} \sum_{i,i'\in[k],j,j'\in[\pm p]} \mathbb{E}_{P_0}[\frac{P_{i,j}}{P_0}\frac{P_{i',j'}}{P_0} - 1]\end{aligned}$$

First, find a naive upper bound for the ratio term.

$$\begin{aligned}\frac{P_{i,j}(\mathbf{x})}{P_0(\mathbf{x})} &= 1 - \tau + \tau \frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_j)}{\sum_{i'\in[k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_{i'})} \\ \frac{P_{i,j}(\mathbf{x})}{P_0(\mathbf{x})}\frac{P_{i',j'}(\mathbf{x})}{P_0(\mathbf{x})} &= (1-\tau)^2 + \tau(1-\tau)[\frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_j)}{\sum_{i''\in[k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_{i''})} + \frac{g(\mathbf{x} - \boldsymbol{\mu}_{i'} - de_{j'})}{\sum_{i''\in[k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_{i''})}] \\ &\quad + \tau^2 \frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_j) g(\mathbf{x} - \boldsymbol{\mu}_{i'} - de_{j'})}{[\sum_{i''\in[k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_{i''})]^2}]\end{aligned}$$

Evaluate the expectation.

$$\begin{aligned}\mathbb{E}_{P_0}[\frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_1)}{\sum_{i''\in[k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_{i''})}] &= 1 \\ \mathbb{E}_{P_0}[\frac{g(\mathbf{x} - \boldsymbol{\mu}_i - de_j) g(\mathbf{x} - \boldsymbol{\mu}_{i'} - de_{j'})}{[\sum_{i''\in[k]} \frac{1}{k} g(\mathbf{x} - \boldsymbol{\mu}_{i''})]^2}] &\le k \int \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}[||\mathbf{x} - \boldsymbol{\mu}_i - de_j||_2^2 + ||\mathbf{x} - \boldsymbol{\mu}_{i'} - de_{j'}||_2^2 - ||\mathbf{x} - \boldsymbol{\mu}_i||_2^2]} d\mathbf{x} \\ &= k \int \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}[||\mathbf{x} - \boldsymbol{\mu}_i - de_j - de_{j'}||_2^2 + 2de_{j'}^T \boldsymbol{\mu}_{i'} - 2de_{j'}^T \boldsymbol{\mu}_i - 2d^2 e_j^T e_{j'}]} \\ &\le \begin{cases} k & , \text{if } j \neq j' \\ ke^{d^2} & , \text{if } j = j' \end{cases}\end{aligned}$$

Thus, we have

$$\chi^2(P_1||P_0) \leq 1 - \tau^2 + \frac{\tau^2 k}{k^2 p^2}(k^2 p \times e^{d^2} + k^2 p \times (p-1)) - 1 = \frac{k\tau^2(e^{d^2} - 1 + p)}{p}$$

$$\chi^2(P_1^n||P_0^n) \leq [\chi^2(P_1||P_0) + 1]^n - 1 \leq (1 + \frac{k\tau^2(e^{d^2} - 1 + p)}{p})^n - 1$$

$$\leq e^{n\frac{k\tau^2(e^{d^2}-1+p)}{p}} - 1$$

**Theorem 29 (Lower bound for counting (with $p$))** *For $\mathcal{C}_{k,\tau,d}$, when $n \leq \frac{pC_v}{k\tau^2(e^{d^2}-1+p)}$, we have*

$$\max_{P \in \mathcal{H}_0^{\mathcal{C}_{k,\tau,d}}, Q \in \mathcal{H}_1^{\mathcal{C}_{k,\tau,d}}} \{\mathbb{P}_P^n[\psi = 1], \mathbb{P}_Q^n[\psi = 0]\} \geq \max_{i \in [k]} \{\mathbb{P}_{P_0}^n[\psi = 1], \mathbb{P}_{P_i}^n[\psi = 0]\}$$

$$= \max\{\mathbb{P}_{P_0}^n[\psi = 1], \mathbb{P}_{P_1}^n[\psi = 0]\}$$

$$(\because \textit{Le Cam's multiple-points method}) \geq e^{-\chi(P_1^n||P_0^n)} \wedge \frac{1 - \sqrt{\chi(P_1^n||P_0^n)/2}}{2}$$

$$\geq \frac{1}{2} - v$$

# References


[ADHP09] Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009.

[AK+05] Sanjeev Arora, Ravi Kannan, et al. Learning mixtures of separated nonspherical gaussians. *The Annals of Applied Probability*, 15(1A):69–92, 2005.

[BT] Igor Baskin and Igor Tetko. Modern machine learning techniques: Regression methods. `http://infochim.u-strasbg.fr/CS3/program/material/Baskin_Tetko.pdf`.

[CCK04] Hanfeng Chen, Jiahua Chen, and John D Kalbfleisch. Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):95–115, 2004.

[CLA] Supervised machine learning A review of classification techniques. `https://datajobs.com/data-science-repo/Supervised-Learning-%5BSB-Kotsiantis%5D.pdf`.

[CT12] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[DCG97] Didier Dacunha-Castelle and Elisabeth Gassiat. The estimation of the order of a mixture model. *Bernoulli*, pages 279–299, 1997.

[DF] Ted Dunning and Ellen Friedman. Practical machine learning: A new look at anomaly detection. `http://info.mapr.com/rs/mapr/images/Practical_Machine_Learning_Anomaly_Detection.pdf`.

[Duc] John Duchi. Statistics 311/electrical engineering 377. `http://stanford.edu/class/stats311/`.

[FL94] W David Furman and Bruce G Lindsay. Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics & Data Analysis*, 17(5):473–492, 1994.

[Hen85] Jõgi Henna. On estimating of the number of constituents of a finite mixture of continuous distributions. *Annals of the Institute of Statistical Mathematics*, 37(1):235–240, 1985.

[Hoe65] Wassily Hoeffding. Asymptotically optimal tests for multinomial distributions. *The Annals of Mathematical Statistics*, pages 369–401, 1965.

[JJW15a] Yanjun Han Jiantao Jiao, Kartik Venkat and Tsachy Weissman. Maximum likelihood estimation of functionals of discrete distributions. *arXiv:1406.6959*, 2015.

[JJW15b] Yanjun Han Jiantao Jiao, Kartik Venkat and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *arXiv:1406.6956*, 2015.

[Ker00] Christine Keribin. Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66, 2000.

[Lina] Hsuan-Tien Lin. Machine learning foundations. `https://www.coursera.org/course/ntumlone`.

[Linb] Hsuan-Tien Lin. Machine learning techniques. `https://www.coursera.org/course/ntumltwo`.

[LM00] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[LSW15] Euiwoong Lee, Melanie Schmidt, and John Wright. Improved and simplified inapproximability for k-means. *arXiv preprint arXiv:1509.00916*, 2015.

[Meg90] Nimrod Megiddo. On the complexity of some geometric problems in unbounded dimension. *Journal of Symbolic Computation*, 10(3):327–334, 1990.

[ML0] Scikit. `http://scikit-learn.org/stable/index.html`.

[MR14] Geoffrey J McLachlan and Suren Rathnayake. On the number of components in a gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5):341–355, 2014.

[OPT] Hyper textbook Optimization models and applications. `https://inst.eecs.berkeley.edu/~ee127a/book/login/index.html`.

[RE0a] Reinforcement learning: A tutorial. `http://hunch.net/~jl/projects/RL/RLTheoryTutorial.pdf`.

[RE0b] Reinforcement learning warehouse. `http://reinforcementlearning.ai-depot.com/Main.html`.

[San] Sriram Sankararaman. Practical machine learning lecture: Clustering. `https://www.cs.berkeley.edu/~jordan/courses/294-fall09/lectures/clustering/`.

[SSR] Nathan Srebro, Gregory Shakhnarovich, and Sam Roweis. When is clustering hard.

[Tsy09] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

[Wau] Fabian Wauthier. Practical machine learning lecture: Regression. `https://www.cs.berkeley.edu/~jordan/courses/294-fall09/lectures/regression/`.

[WY14] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *arXiv:1407.0381*, 2014.

[WY15] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *arXiv:1504.01227*, 2015.

[ZG91] Ofer Zeitouni and Michael Gutman. On universal hypotheses testing via large deviations. *Information Theory, IEEE Transactions on*, 37(2):285–290, 1991.