NATIONAL TAIWAN UNIVERSITY

COLLEGE OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Undergraduate Research Report

Student:
Chi-Ning Chou

Advisor:
I-Hsiang Wang

July 8, 2015

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[1], and the application of machine learning[4], [13], [14], [5], [6], [17], [15], [9], [3], and [2].

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 [16] and [8]. After studied some related works [11], [12], [18], and [19], 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 Minimax Estimation of Gaussian Mixture

1.1 Gaussian Mixture Model

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,

$$g(t) = \sum_i \lambda_i f_i(t)$$

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

Below is an example consists of three components with mean and variance as described in the figure.

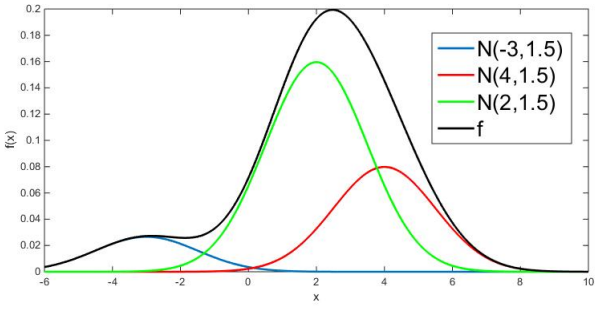


Figure 1: This Gaussian mixture consists three components. The means are -3, 4,2 respectively and the variances are all 1.5

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, I focus on the estimation of the number of components in Gaussian model. The goals are finding the tight minimax risk of this problem and constructing the rate-optimal estimator.

1.2 Estimation of The Number of Components in Gaussian Model

First, let's define the probability distribution family we consider in this problem:

$$D_{k,\delta} = \{g : g = \sum_i \lambda_i f_i, f_i \sim N(\mu_i, 1), \sum_i \lambda_i = 1, \lambda_i \geq \frac{1}{k}, |\mu_i - \mu_j| \geq \delta\}$$

, where k is the reciprocal of the minimum mass and of the component in each Gaussian mixture in this family. And δ is the least distance between the mean of each components, while what matters is the ratio between the maximum variance and the minimum distance, we can simply fix the upper bound of the variance to 1 and only control the behavior of the least distance δ.

The parameter we're going to estimate is the number of component $S : D_{k,\delta} \to \mathbf{N}$. And we want to construct estimator $\hat{S} : \mathbf{X}^n \to \mathbf{N}$, where $\mathbf{X}$ is in the domain of the Gaussian mixture model, and n is the number of samples. Note that, the domain in the problem setting can be any multi-dimension continuous space as long as multi-normal distribution can be defined on it. For simplicity, here we assume the domain is $\mathbf{R}^m$.

Thus we can write down the risk for a single estimator and the minimax risk for this problem according to the problem setting above:

$$r_{n,k,\delta}(\hat{S}) := \sup_{\mathbf{X}^n \leftarrow f \in D_{k,\delta}} E[|\hat{S}(\mathbf{X}^n) - S(f)|^2]$$
$$R_{n,k,\delta} := \inf_{\hat{S}} r_{n,k,\delta}(\hat{S})$$

Our goal is to find a tight lower bound for the minimax risk and the rate-optimal estimator. Namely,

$$R_{n,k,\delta} \gtrsim l(n,k,\delta)$$
$$r_{n,k,\delta}(S^*) \approx l(n,k,\delta)$$

where l is a function of n, k, δ which is the optimal rate. For instance, l could be $l(n,k,\delta) = \frac{k^2 n}{\delta}$.

1.3 Support Size Estimation

The estimation about the number of components in the Gaussian mixture can be reduced to a discrete analogy problem: *support size estimation*. The goal of the support size estimation problem is to estimate the number of non-zero mass points. Thus, we can simply view it as the limit problem of Gaussian mixture model as we let the variance of each component shrinks to zero or set the least distance between components to infinity.

The minimax risk analysis of this problem and the rate-optimal estimator was solved by Yihong Wu and Pengkun Yang [19]. The lower bound of the minimax risk is:

$$R_{n,k,\delta} \gtrsim k^2 \exp(-\frac{n}{k}), if\ n \gtrsim k \log k$$

$$R_{n,k,\delta} \gtrsim k^2 \exp(-\sqrt{\frac{n \log k}{k}})), if\ n \lesssim k \log k$$

The lower bound is constructed with techniques involves Poisson sampling, generalized Le Cam's two point method, moment matching, and best polynomial approximation. And the rate-optimal estimator is constructed with the Chebyshev polynomial.

With this result and the reduction from Gaussian mixture model to support size estimation, we can yield a direct lower bound to the minimax risk in the number of components estimation. Thus, our goal is to find whether there also exists a rate-optimal estimator with the same order as the lower bound for the Gaussian mixture model. Or, we need to find a tighter lower bound and optimal estimator.

1.4 Assumption

However, there are two difficulties for us to seek a tight lower bound in the estimation of component number: locating mean and adjusting distance.

Locating mean

The first one is how to locate the position of each components and extract useful statistics from the samples? This can be solve by quantization on the domain space. However, things might become more difficult as the distribution of components are very random. Namely, we cannot directly divide the domain space evenly and find where the means are in order to calculating the histogram.

Adjusting distance

The second difficulty is about the distance between two components. Once there are two components getting too close, the region where both distributions have non-negligible mass will grow larger and larger. As a result, the chance of wrong estimation will be higher.

With respect to the two difficulties discussed above, I divide the problem into three stages:

1. Fixed central position and large δ
2. Any central position and large δ
3. Any central position and small δ

Fixed central position requires the center of each component to appear only on certain position. For example, the mean of each component can only appear on the coordinate where δ multiplied by an integer.

In the following analysis, the research will be based the above three stages. And in this semester, I only finish part of the first stage.

1.5 Lower Bound and Rate-Optimal Estimator

In this semester, I've only finished the stage one of the research. The following is the derivation of the minimax risk lower bound and rate-optimal estimator with the fixed central position and large δ assumption.

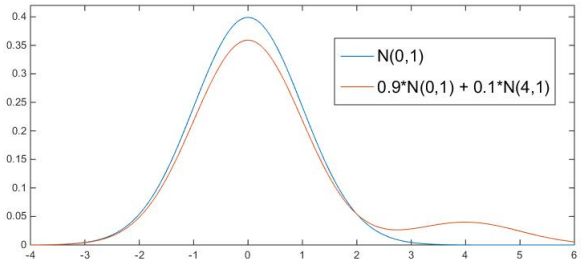
In the following, I only consider the case where $n \geq \alpha k \log k$.

1.5.1 Lower Bound

To obtain the lower bound, let's apply the Le Cam's two points method just as the support size problem did. The following considers the $n \gtrsim k \log k$ case, the $n \lesssim k \log k$ can be reduced using similar approach in [19].

First, consider two distribution: $P \sim N(0,1), Q \sim (1 - \frac{1}{k})N(0,1) + kN(0,\delta)$. The following is an example with $\delta = 4, k = 10$.

Next, consider the upper bound of the distortion between these two distributions and use it to construct the desired lower bound. We can utilize the convexity of



KL-divergence to give an upper bound:

$$D_{kl}(P||Q) \leq \frac{\delta^2}{2k}$$

Apply the Le Cam's two points method [16]

$$\begin{aligned} R_{n,k,\delta} &\geq \frac{1}{4}(S(P) - S(Q))^2 \exp(-nD_{kl}(P||Q)) \\ &= \frac{1}{4}\exp(-\frac{n\delta^2}{2k}) = \frac{1}{4}\exp)(-\frac{n\delta^2}{2k} - \frac{4n\delta^2}{2k\alpha} + \frac{4n\delta^2}{2k\alpha}) \\ &\geq \frac{1}{4}\exp(-(\frac{1}{2} + \frac{2}{\alpha})\frac{n\delta^2}{k} + \log(k^{2\delta^2})) \\ &= \frac{1}{4}k^{2\delta^2}\exp(-C\frac{n}{k}) \end{aligned}$$

,where $C = (\frac{1}{2} + \frac{2}{\alpha})$.

The intuition of all the lower bound techniques are about utilizing the indistinguishable level between two or many distributions with respect to the number os samples n and the δ. The harder the distributions are distinguished, the more difficult for the

1.5.2 Rate-Optimal Estimator

After finding an upper bound, now we are going to construct an estimator and argue that it's minimax risk is of the same order as the lower bound $O(k^{2\delta^2}\exp(-C\frac{n}{k}))$.

To construct an estimator requires a good statistics from the sample, and here I choose histogram as the sufficient statistics. The reason why I choose histogram is that histogram reveals local information in the samples. As a result, we can conveniently make inference on the number of components.

In addition to histogram, using empirical distribution to estimate is also a good approach. However, the techniques involved are much more difficult. As a result, I choose histogram instead of empirical distribution for now.

The construction of histogram requires the quantization of domain space. And in the first stage it's trivial to do so. We just simply evenly divide the domain space according to the position of mean.

In this report, I only consider the stage 1 case: fixed central position and large δ. With the first assumption, suppose that the center of each component only appears on the multiplicity of ρ. That is, $\mu_i = n_i\rho,\ n_i \in \mathbf{N}\ \forall i$. Thus, the histogram can be easily defined as

$$\{N_s : \sum_j \mathbf{1}_{\{(s-\frac{1}{2})\rho \le X_j < (s+\frac{1}{2})\rho\}}\}$$

For instance, consider a Gaussian model with $\rho = 1$ and consists of three components with their center on -4, 5, 6 respectively.

With the second assumption, large δ, the ρ is simultaneously being larger. For convenience, let's consider the case that $\rho \ge 4$

Once we define our histogram, we an simply construct an intuitive estimator:

$$\hat{S} := \sum_s \mathbf{1}_{\{N_s \ge \frac{n}{k}\}}$$

Before we calculate the minimax risk, let's have a little observation in advance. For each bar N_s in the histogram, we can regard it as a collector which counts the number of samples in its interval. Actually, we can think of this as a Poisson distribution, such that for a single random sample from the Gaussian mixture model, the probability for it to go into bar N_s is the total probability mass of the corresponding interval. Formally, the total probability mass in bar N_s is

$$\begin{aligned} P_s &= \mathbf{P}(X \in [(s-\frac{1}{2})\rho, (s+\frac{1}{2})\rho)) \\ &= \sum_i \lambda_i \int_{(s-\frac{1}{2})\rho}^{(s+\frac{1}{2})\rho} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x-\mu_i)^2}{2}) dx \end{aligned}$$

Note that with the large δ and fixed central means assumptions, there's at most one component in the Gaussian mixture model lies in the interval $[(s-\frac{1}{2})\rho, (s+\frac{1}{2})\rho)$, and thus we can approximate the total probability in bar N_s as:

$$P_s \approx \sum_i 0.95\lambda_i \mathbf{1}_{\{\mu_i = s\rho\}}$$

, while taking $\rho = 4$ forms a 95% confident interval. Rigorous argument about this approximation will be explained in future research.

To calculate the minimax risk of this estimator, let's fix a single distribution P and find it's corresponding error.

First we consider the bias:

$$\begin{aligned}|E[\hat{S}(P) - S(P)]| &= |\sum_s \mathbf{P}(N_s \geq \frac{n}{k}) - \mathbf{1}_{\{\exists i \text{ s.t. } \mu_i = s\rho\}}| \\ &= \sum_s (1 - \mathbf{P}(N_s \geq \frac{n}{k}))\mathbf{1}_{\{\exists i,\ \mu_i = s\rho\}} \\ &+ \sum_s \mathbf{P}(N_s \geq \frac{n}{k})\mathbf{1}_{\{\forall i,\ \mu_i \neq s\rho\}}\end{aligned}$$

The bias is composed of false negative error and false positive error.

$$\begin{aligned}\sum_s (1 - \mathbf{P}(N_s \geq \frac{n}{k}))\mathbf{1}_{\{\exists i,\ \mu_i = s\rho\}} &= \sum_s \exp(-nP_s)\mathbf{1}_{\{\exists i,\ \mu_i = s\rho\}} \\ &\leq \sum_i \exp(-0.95\lambda_i n) \\ &\leq k\exp(-0.95\frac{n}{k})\end{aligned}$$

$$\begin{aligned}\sum_s \mathbf{P}(N_s \geq \frac{n}{k})\mathbf{1}_{\{\forall i,\ \mu_i \neq s\rho\}} &= \sum_s \exp(-n(1 - P_s))\mathbf{1}_{\{\forall i,\ \mu_i \neq s\rho\}} \\ &\leq \sum_i \exp(-0.05\lambda_i n) \\ &\leq k\exp(-0.05\frac{n}{k})\end{aligned}$$

Thus,

$$|E[\hat{S}(P) - S(P)]| \leq k\exp(-C'\frac{n}{k})$$

And the variance can be bounded in a similar fashion

While we take $\rho \geq 4$, $k \leq k^{\delta^2}$. Thus, the upper bound for the minimax risk of $\hat{S}$ is obtained, which is in the same order as the lower bound we found in the previous section.

To sum up, in this semester I finish the rate-optimal estimation of the components number in Gaussian mixture model with the assumption that the mean of each component is on the multiplicity of ρ, $\rho \geq 4$, and $n \geq \alpha k \log k$. The optimal minimax risk rate is $O(k^{2\delta^2}\exp(-C\frac{n}{k}))$ and the rate-optimal estimator is $\hat{S} := \sum_s \mathbf{1}_{\{N_s \geq \frac{n}{k}\}}$.

1.6 Future

The results so far is reasonable since we can view the support size problem as letting the variance of Gaussian mixtures converge to zero. As a result, the error order of Gaussian mixtures estimation must be greater than that of support size estimation. For more details, it's welcome to see my slide for this research: Number of components in Gaussian mixture model: Lower bound and optimal estimator for minimax risk.

However, for now I only complete the very beginning of the estimation of components number in Gaussian mixture problem. There are still lots of issues that are waited to be solved such as the location of the means and distinguishable problem of two close normal distribution.

I plan to keep working on this interesting problem. There are still some techniques I'm not very familiar with such as moment matching, best polynomial approximation, Poisson sampling, etc. So, the next step of this research is to study these related works while in the meantime keep finding the optimal minimax risk rate for general cases.

2 Surveys

In this semester, I've done some surveys on various topic such as statistical estimation, problems in machine learning, minimax risk lower bound techniques, and inequalities between probability distribution metrics.

In this section, I will briefly introduce my works. And all the slides and handouts I made are available online.

2.1 Lower Bound For Minimax Risk Estimation

I studied the lower bound techniques for minimax risks in [16] and [8]. And I made a handout about it. The following is the url to the handout: Minimax Lower Bound Techniques

This handout is divided into 4 sections:

1. Minimax risk
2. From estimation to testing

3. Le Cam method
4. Others methods

Please take a look at it if you are interested in.

2.2 Introduction to Statistical Estimation

This survey consists two parts:

1. How to **find** an estimator?
2. How to **evaluate** an estimator?

2.2.1 How to find an estimator?

In the first part, I introduce maximum likelihood estimation (MLE), bayes approach, methods of moments, and EM algorithm.

2.2.2 How to evaluate an estimator?

In the second part, I introduced two ways to evaluate an estimator.

The first way is through **quantified** properties, which consider the error for single realization. For example, bias and variance.

The second way is through **behavioral** properties. Namely, the asymptotic behavior of the estimator. There are two central properties: consistency and asymptotic normality, which both describe the limiting behavior of the estimator.

To learn more details, please refer to the slides I made for this survey: Introduction to Statistical Estimation

2.3 Problems for Machine Learning

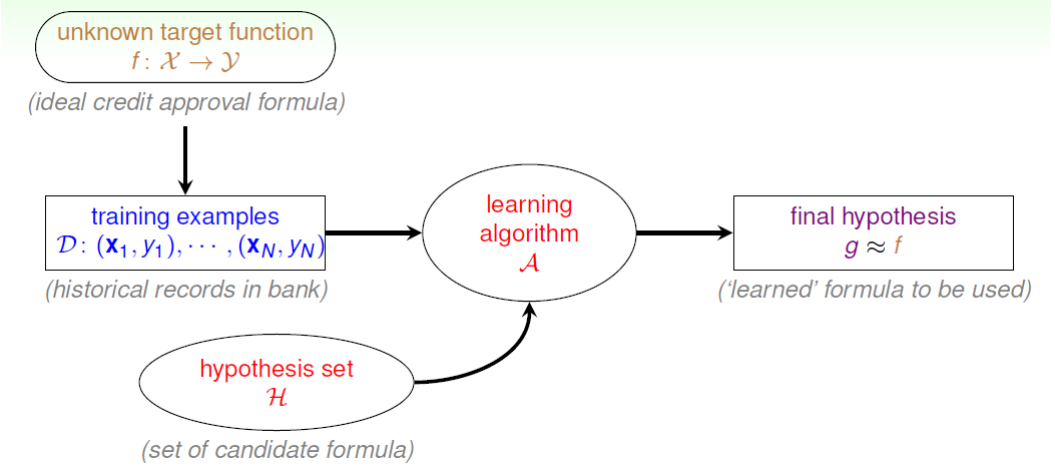
This survey aimed to give an introduction to the problems in machine learning. The presentation is divided into three parts:

1. Machine learning flow
2. Types of learning

3. Categories of machine learning problems

2.3.1 Machine learning flow

The following is a figure showing the machine learning framework from professor Hsuan-Tien Lin's online course: *Machine Learning Foundations* [13]. For more



details, please refer to [13] and [14].

2.3.2 Types of learning

According to the given date types, we can categorize the problems into three types:

- Discrete-valued
- Continuous-valued
- Label-valued

Moreover, we can categorize the problems according to whether there's label attached on the input date.

- Supervised learning
- Semi-supervised learning
- Unsupervised learning

2.3.3 Categories of learning problems

In this survey, I introduces five categories of learning problems:

- Classification
- Regression
- Clustering
- Anomaly detection
- Reinforcement learning

Classification Classification problem wants to classify the data according to some property. The famous classification techniques are decision trees, neural network, bayes approach, kNN, SVM, and rule learners etc. The following is a comparison of the above techniques:

	Decision Trees	Neural Network	Bayes	kNN	SVM	Rule Learners
Accuracy in General	**	***	*	**	****	**
Attributes Type: Discrete, Binary, Continuous	****	*** (no discrete)	*** (no continuous)	*** (no directly discrete)	** (no discrete)	*** (no directly continuous)
Speed of classification	****	****	****	*	****	****
Speed of Learning	***	*	****	****	*	**
Dealing with Overfitting	**	*	***	***	**	**
Tolerance to Noise	**	**	***	*	**	*
Explanation Ability of Classifications	****	*	****	**	*	****

Figure 2: Comparison between various classification techniques

You can find more details here [5].

Regression You can find details in [6], [17].

Clustering You can find details in [15].

Anomaly detection You can find details in [9].

Reinforcement learning You can find details in [3], [2].

The slide is available at: Problems for Machine Learning

2.4 Inequalities Between Probability Metric

The slide can be found here: Probability metrics and their inequalities.

[10] introduces 10 probability metric and their inequalities. It's a really complete and detailed resource.

[7] pays more attention on KL-divergence, total variation and Hellinger distance. It gives more aspect on measure theory and is very rigorous.

References

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

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

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

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

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

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

[7] Pollard David. *Asymptopia*. 2013.

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

[9] 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.

[10] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

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

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

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

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

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

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

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

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

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