CONVEX FORMULATION FOR NONPARAMETRIC ESTIMATION OF MIXING DISTRIBUTION

Kazuho Watanabe[1] *and Shiro Ikeda*[2]

[1]Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0192, JAPAN, wkazuho@is.naist.jp
[2] The Institute of Statistical Mathematics,
10-3 Midori-cho, Tachikawa-shi, Tokyo, 190-8562 JAPAN, shiro@ism.ac.jp

ABSTRACT

We discuss a nonparametric estimation method of the mixing distribution in mixture models. We propose an objective function with one parameter, where its minimization becomes the maximum likelihood estimation or the kernel vector quantization in special cases. Generalizing Lindsay's theorem for the nonparametric maximum likelihood estimation, we prove the existence and discreteness of the optimal mixing distribution and devise an algorithm to calculate it. Furthermore, we show the connection between the unifying estimation framework and the rate-distortion problem. It is demonstrated that with an appropriate choice of the parameter, the proposed method is less prone to overfitting than the maximum likelihood method.

1. INTRODUCTION

Mixture models are widely used for clustering and density estimation. We discuss a nonparametric estimation method of mixture models where an arbitrary distribution, including a continuous one, is assumed over the component parameter. It was proved by Lindsay [1] that the maximum likelihood estimate of the mixing distribution is given by a discrete distribution whose support consists of distinct points, the number of which is no more than the sample size. This provides a framework for determining the number of mixture components from data. The mixture estimation algorithm developed in [2] can be considered as a procedure for estimating such discrete distributions. However, it is vulnerable to overfitting because of the flexibility of the nonparametric estimation.

In this study, we propose a nonparametric mixture estimation method defined by minimization of an objective function with one parameter β. With specific choices of β, the proposed method reduces to the maximum likelihood estimation (MLE) and the kernel vector quantization (KVQ) [3]. Generalizing Lindsay's theorem for the nonparametric MLE, we prove the existence and discreteness of the optimal mixing distribution. Then, we provide an algorithm to calculate the optimal discrete distribution, that is specifically tailored to the proposed objective function from the procedure in [2]. Numerical experiments demonstrate that there exists an appropriate choice of β in terms of the average generalization error. Furthermore, we relate the proposed mixture estimation method to the rate-distortion problem [4] to build insight into the selection of the width of the component density.

2. MIXTURE MODELLING

Given n training samples, $\{x_1, \cdots, x_n\}$, $x_i \in R^d$, consider nonparametric estimation of the mixing distribution $q(\theta)$ of the following mixture density of the model $p(x|\theta)$ with parameter $\theta \in \Omega$,

$$r(x) = r(x;q) = \int p(x|\theta)q(\theta)d\theta. \tag{1}$$

Let $r_i = r(x_i;q) = \int p(x_i|\theta)q(\theta)d\theta$. We choose $q(\theta)$ as the optimal function of the following problem,

$$\hat{q}(\theta) = \operatorname*{argmin}_q F_\beta(q),$$

where

$$F_\beta(q) = \begin{cases} \frac{1}{\beta}\log\left(\frac{1}{n}\sum_{i=1}^n r_i^{-\beta}\right), & (\beta \neq 0) \\ -\frac{1}{n}\sum_{i=1}^n \log r_i & (\beta = 0). \end{cases} \tag{2}$$

The objective function $F_\beta(q)$ is continuous with respect to $\beta \in R$. This estimation boils down to the MLE when $\beta = 0$ [1]. As $\beta \to \infty$, it becomes the minimization of $\max_i(-\log r_i)$, that is, KVQ with the kernel function, $K(x,\theta) = p(x|\theta)$ [3][1].

For $\beta \neq 0$, it is also expressed as

$$F_\beta(q) = -\frac{1}{\beta}\min_{\boldsymbol{p}\in\Delta}\left\{\beta\sum_{i=1}^n p_i \log r_i + \sum_{i=1}^n p_i \log\frac{p_i}{1/n}\right\}, \tag{3}$$

where $\Delta = \{\boldsymbol{p} = (p_1, p_2, \cdots, p_n)|p_i \geq 0, \sum_{i=1}^n p_i = 1\}$. This expression is verified through the fact that the minimum is attained by

$$p_i = \frac{r_i^{-\beta}}{\sum_{j=1}^n r_j^{-\beta}}, \tag{4}$$

and will be used for deriving a simple learning procedure in the next section.

[1] The original KVQ restricts the possible support points of $q(\theta)$ to the training data set $\{x_1, \cdots, x_n\}$. That is $q(\theta) = \sum_{i=1}^n q_i\delta(\theta - x_i)$, $q_i \geq 0$, $\sum_{i=1}^n q_i = 1$.

3. OPTIMAL MIXING DISTRIBUTION

3.1. Discreteness of the Optimal Mixing Distribution

We can show the convexity of F_β with respect to $\boldsymbol{r} = (r_1, \cdots, r_n)$ for $\beta \geq -1$.

Therefore, for $\beta \geq -1$, there exists a unique $\boldsymbol{r}$ that minimizes F_β at the boundary of the convex hull of the set $\{\boldsymbol{p}_\theta = (p(x_1|\theta), \cdots, p(x_n|\theta)) | \theta \in \Omega\}$ where Ω is the parameter space. From Caratheodory's theorem, this means that the optimal $\boldsymbol{r}$ is expressed by a convex combination, $\sum_{l=1}^k q_l \boldsymbol{p}_{\theta_l}$, with $q_l \geq 0$, $\sum_{l=1}^k q_k = 1$ and $k \leq n$, indicating that the optimal mixing distribution is $q(\theta) = \sum_{l=1}^k q_l \delta(\theta - \theta_l)$, the discrete distribution whose support size is no more than n.

3.2. Learning Algorithm

The KKT condition for the optimal $q(\theta)$ is given by $\mu(\theta) \leq 1$ for all θ where

$$\mu(\theta) = \sum_{i=1}^n \alpha_i p(x_i|\theta), \tag{5}$$

and

$$\alpha_i = \frac{r_i^{-\beta-1}}{\sum_{j=1}^n r_j^{-\beta}}. \tag{6}$$

Hence the mixing distribution $q(\theta)$ can be optimized by Algorithm 1 which sequentially augments the set of the support points until the maximum of $\mu(\theta)$ approach 1 [2].

Algorithm 1 Decoupled Approach to Mixture Estimation

```
1: Initialize $k = 0$ and $\alpha_i = 1/n$ and prepare a small positive constant $\epsilon$.
2: repeat
3:     Let $\hat{\theta}_k = \underset{\theta}{\mathrm{argmax}}\, \mu(\theta)$ and $k = k + 1$, where $\mu(\theta)$ is given by eq.(5).
4:     Define the discrete distribution, $q_k(\theta) = \sum_{l=1}^k \pi_l \delta(\theta - \hat{\theta}_l)$. Optimize $\{\pi_l, \hat{\theta}_l\}_{l=1}^k$ by minimizing $F_\beta(q_k)$.
5:     Compute $\{\alpha_i\}_{i=1}^n$ by eq.(6) with $r_i = \sum_{l=1}^k \pi_l p(x_i|\hat{\theta}_l)$.
6: until $\max_\theta \mu(\theta) < 1 + \epsilon$ holds.
```

3.3. EM Updates for Finite Mixtures

Eq.(3) is equivalent to a weighted sum of negative log-likelihood and an EM-like algorithm is available for the optimization of $\{\pi_l, \hat{\theta}_l\}_{l=1}^k$ in Step 4. Its updating rule is obtained as follows,

$$\pi_j^{(t+1)} = \sum_{i=1}^n p_i^{(t)} \nu_{ij}, \quad \text{and} \quad \hat{\theta}_j^{(t+1)} = \frac{\sum_{i=1}^n p_i^{(t)} \nu_{ij} x_i}{\sum_{i=1}^n p_i^{(t)} \nu_{ij}},$$

where $p_i^{(t)} = \frac{r_i^{(t)-\beta}}{\sum_{j=1}^n r_j^{(t)-\beta}}$, $r_i^{(t)} = \sum_{l=1}^k \pi_l^{(t)} p(x_i|\hat{\theta}_l^{(t)})$ and

$$\nu_{ij} = \frac{\pi_j^{(t)} p(x_i|\hat{\theta}_j^{(t)})}{\sum_{m=1}^k \pi_m^{(t)} p(x_i|\hat{\theta}_m^{(t)})} \tag{7}$$

is the posterior probability that the data point x_i is assigned to the cluster center $\hat{\theta}_l$.

We can prove for $\beta \leq 0$ that the above update monotonically decreases the objective F_β since this minimization is expressed by the double minimization over $\{\pi_l, \hat{\theta}_l\}_{l=1}^k$ and $\{p_i\}_{i=1}^n$ from eq.(3). However, the similar proof does not apply for $\beta > 0$. Hence, we switch to another update rule for $\beta > 0$, which is omitted in this paper.

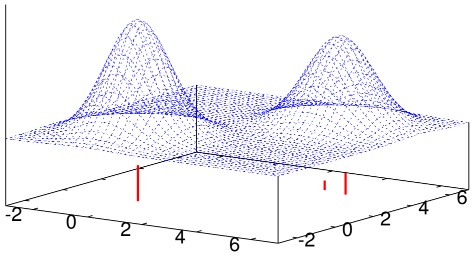


Figure 1. Example of the estimated mixture for $\beta = -0.2$ and $\sigma^2 = 1$. Corresponding mixing distributions are illustrated in the x-y planes where the location and the height of the red lines are respectively the mean parameter $\hat{\theta}_l$ and the weight $\hat{\pi}_l$ of each component.

4. EXPERIMENTS

In this section, we demonstrate the properties of the estimation method by a numerical simulation focusing on the case of 2-dimensional Gaussian mixtures where

$$p(x|\theta) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{||x - \theta||^2}{2\sigma^2}\right). \tag{8}$$

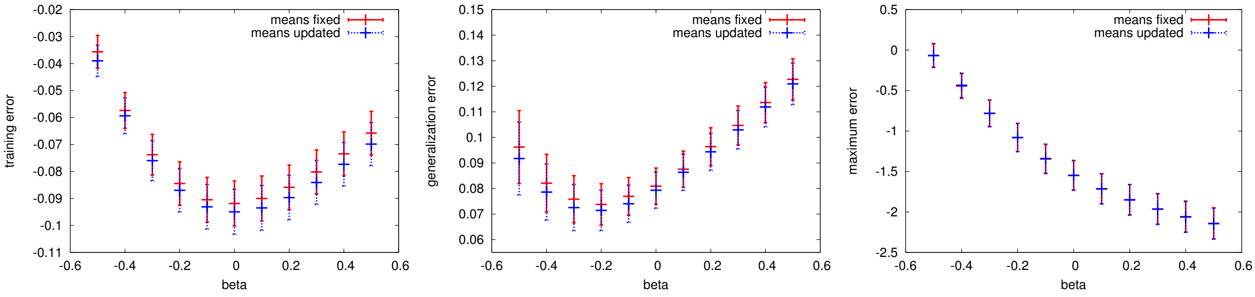
We generated synthetic data by the true distribution,

$$p^*(x) = \frac{1}{2} N(x|\theta_1^*, I_2) + \frac{1}{2} N(x|\theta_2^*, I_2), \tag{9}$$

where $\theta_1^* = (0, 0)^T$, $\theta_2^* = (4, 4)^T$ and $N(x|\theta, \sigma^2 I_2) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{||x-\theta||^2}{2\sigma^2}\right)$ is the Gaussian density function.

We assumed that the kernel width σ^2 in eq.(8) was known and $p(x|\theta)$ was set to $N(x|\theta, I_2)$. Let $\hat{q}(\theta)$ be an estimated mixing distribution. The optimal mixing distribution $q(\theta)$ is given by $\frac{1}{2}\delta(\theta - \theta_1^*) + \frac{1}{2}\delta(\theta - \theta_2^*)$ in this case. An example of the estimated mixture model for $\beta = -0.2$ and $\sigma^2 = 1$ is demonstrated in Figure 1.

Figure 2(a) and Figure 2(b) respectively show the training error, $\frac{1}{n} \sum_{i=1}^n \log \frac{p^*(x_i)}{\int p(x_i|\theta)\hat{q}(\theta)d\theta}$, and the generalization error, $\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \log \frac{p^*(\tilde{x}_i)}{\int p(\tilde{x}_i|\theta)\hat{q}(\theta)d\theta}$, for test data $\{\tilde{x}_i\}_{i=1}^{\tilde{n}}$ generated from the true distribution (9). All results were averaged over 100 trials for different data sets generated by (9). The number of training data is $n = 50$ and that of test data is $\tilde{n} = 200000$. We also applied the original version of the algorithm in [2], where only $\{\pi_l\}$ are updated by the EM algorithm with the weight p_i in eq.(4) for each sample in Step 4. These results are indicated as "means fixed". We see that the average training error takes the minimum at $\beta = 0$ as expected while the average generalization error is minimized around $\beta = -0.2$.



(a) Training error

(b) Generalization error

(c) Maximum error

Figure 2. Training error (a), generalization error (b) and maximum error (c) against β. The error bars show 95% confidence intervals.

Figure 2(c) shows the average of the maximum error, $\max_i \left(-\log \int p(x_i|\theta)\hat{q}(\theta)d\theta\right) - \max_i \left(-\log p^*(x_i)\right)$, which corresponds to the objective function of the KVQ. As expected, the monotone decrease of it with respect to β implies the estimation approaches the KVQ as $\beta \to \infty$.

In Figure 3, we show the number of estimated components remaining after the elimination of components with sufficiently small mixing proportions (less than $\frac{1}{n^2}$). Since it strongly depends on ϵ, we also applied hard assignments to cluster centers for each data point and counted the number of hard clusters, which is also plotted in Figure 3. Here, each point x_i is assigned to the cluster center $\hat{\theta}_l$ that maximizes the posterior probability (7). The number of components $\hat{k}$ as well as that of hard clusters increase as β becomes larger. This reduces the average generalization error when β takes slightly negative value as we just observed in Figure 2(b).

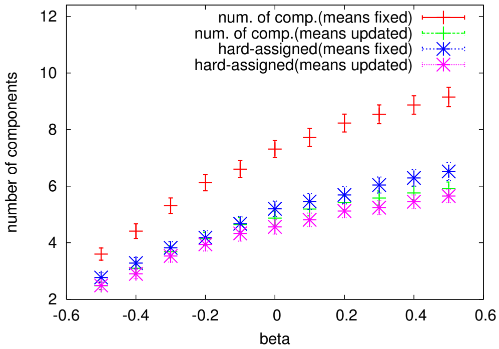


Figure 3. Number of components (cross) and number of hard clusters (asterisk) against β.

5. CONNECTION TO RATE-DISTORTION PROBLEM

The rate-distortion (RD) problem encoding the source random variable X with density $p^*(x)$ to the output Θ is reformulated to solving the following optimization problem [4, 5],

$$\inf_q - \int p^*(x) \log \int q(\theta) \exp(sd(x,\theta)) d\theta dx. \tag{10}$$

Here $d(x,\theta)$ is the distortion measure and s is a Lagrange multiplier. It provides the slope of a tangent to the RD curve and hence has one-to-one correspondence with a point on the RD curve. This problem reduces to the MLE ($F_\beta(q)$ when $\beta = 0$) with $p(x|\theta) \propto \exp(sd(x,\theta))$ if the source $p^*(x)$ is replaced with the empirical distribution. In the case of the Gaussian mixture with $d(x,\theta) = ||x-\theta||^2$, s specifies the kernel width by $\sigma^2 = -\frac{1}{2s}$.

For general β, the expression (3) and the optimal output distribution $\hat{q}(\theta) = \sum_{l=1}^{\hat{k}} \hat{\pi}_l \delta(\theta - \hat{\theta}_l)$ imply the RD function of the source, $\sum_{i=1}^n p_i \delta(x - x_i)$, with the rate

$$\sum_{i=1}^n \sum_{l=1}^{\hat{k}} p_i \nu_{il} \log \frac{\nu_{il}}{\sum_{j=1}^n p_j \nu_{jl}},$$

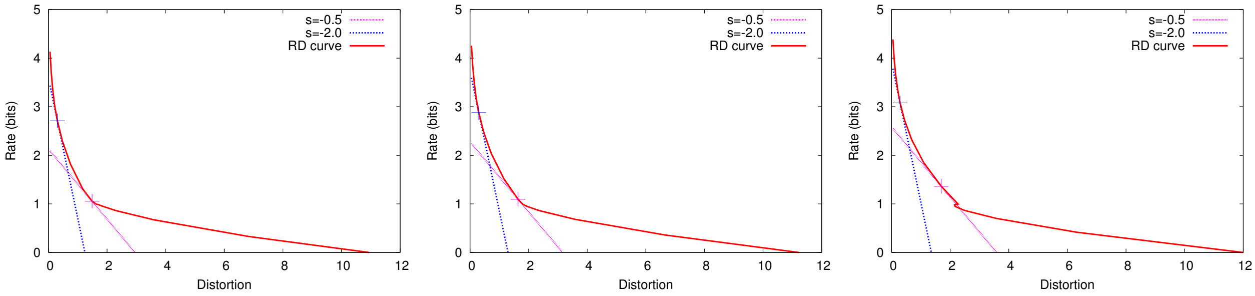
and the average distortion

$$\sum_{i=1}^n \sum_{l=1}^{\hat{k}} p_i \nu_{il} d(x_i, \hat{\theta}_l),$$

where ν_{il} is the posterior probability defined by eq.(7). Since the rate is the mutual information between X and Θ, it is bounded from above by the entropy, $-\sum_{l=1}^{\hat{k}} \hat{\pi}_l \log \hat{\pi}_l$ and further by $\log \hat{k}$. However, the source depends on p_i, which depends on $q(\theta)$ as in eq.(4) and hence the above pair of rate and distortion does not necessarily inherit properties of the usual RD function such as convexity.

Figure 4 demonstrates examples of RD functions obtained by the minimization of $F_\beta(q)$ for $\beta = -0.2$, $\beta = 0$ and $\beta = 0.5$ in the case of the Gaussian mixture used in Section 4.

The three curves show similar behavior such as a monotone decreasing trend although only that for $\beta = 0.5$ loses convexity. This suggests the usage of the RD curve for determining the kernel width σ^2, e.g., by prespecifying a desired rate or average distortion. If we keep the desired rate or distortion to determine σ^2 for different choices of β, then β can be chosen among them for example by CV.



(a) Rate-distortion curve for $\beta = -0.2$.

(b) Rate-distortion curve for $\beta = 0.0$.

(c) Rate-distortion curve for $\beta = 0.5$.

Figure 4. Examples of rate-distortion curves. The lines with slope s passing through the point corresponding to s (cross) are also illustrated for $s = -0.5$ (magenta) and $s = -2.0$ (blue). The rate is scaled by $\log 2$ to yield bits.

6. EXTENSION TO OTHER CONVEX OBJECTIVE FUNCTIONS

The proposed algorithm in Section 3.2 is based on the decoupled approach developed in [2]. The general objective function considered in [2] includes the MLE and the KVQ to estimate $q(\theta)$. We proved in Section 3.1 by extending Lindsay's theorem that the estimated $q(\theta)$ is a discrete distribution consisting of distinct support points no more than n, the number of training data. This statement can be generalized to other objective functions as long as they are convex with respect to $\boldsymbol{r} = (r_1, \cdots, r_n)$ and hence to $q(\theta)$. More specifically, the following four objective functions are demonstrated as examples in [2]. Here, $\rho = \min_i r_i$ and C is a constant.

1. MLE: $-\sum_{i=1}^{n} \log r_i$
2. KVQ: $-\rho$
3. Margin-minus-variance:
   $-\rho + \frac{C}{n} \sum_{i=1}^{n} (r_i - \rho)^2$
4. Mean-minus-variance:
   $-\frac{1}{n} \sum_{i=1}^{n} r_i + \frac{C}{n} \sum_{i=1}^{n} \left(r_i - \frac{1}{n} \sum_{j=1}^{n} r_j \right)^2$

The objective function F_β in eq.(2) combines the first two objectives by the parameter β. The other two objectives above are convex with respect to $\boldsymbol{r}$ as well and hence can be proven to have optimal discrete distributions $q(\theta)$ with support size no more than n. Note that since $\boldsymbol{r}$ is a linear transformation of $q(\theta)$, the convexity on $\boldsymbol{r}$ is equivalent to that on $q(\theta)$ as long as $q(\theta)$ appears in the objective function only with the form of $r_i = \int p(x_i|\theta) q(\theta) d\theta$. Furthermore, we have developed a simple algorithm for finite mixture models to minimize F_β in Section 3.3. Note that, to apply the general framework of Section 3.2 to specific objective functions, we need learning algorithms for optimizing them for finite mixture models.

7. CONCLUSION

We proposed an objective function for learning of mixture models, which unifies the MLE and the KVQ with the parameter β. We proved that the optimal mixing distribution is a discrete distribution with distinct support points no more than the sample size and provided a simple algorithm to calculate it. We discussed the nature of the objective function in relation to the rate-distortion theory and demonstrated its less proneness to overfitting with an appropriate choice of the parameter.

8. REFERENCES

[1] B. G. Lindsay, *Mixture Models: Theory, Geometry and Applications*, Institute of Mathematical Statistics, 1995.

[2] S. Nowozin and G. Bakir, "A decoupled approach to exemplar-based unsupervised learning," in *Proceedings of the 24th International Conference on Machine Learning (ICML 2008)*, 2008.

[3] M. Tipping and B. Scholkopf, "A kernel approach for vector quantization with guaranteed distortion bounds," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2001.

[4] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[5] D. Lashkari and P. Golland, "Convex clustering with exemplar-based models," in *Advances in Neural Information Processing Systems 19*, 2007.