Entropic risk minimization for nonparametric estimation of mixing distributions

Kazuho Watanabe · Shiro Ikeda

Abstract We discuss a nonparametric estimation method for the mixing distributions in mixture models. The problem is formalized as a minimization of a one-parameter objective functional, which becomes the maximum likelihood estimation or the kernel vector quantization in special cases. Generalizing the theorem for the nonparametric maximum likelihood estimation, we prove the existence and discreteness of the optimal mixing distribution and provide an algorithm to calculate it. It is demonstrated that with an appropriate choice of the parameter, the proposed method is less prone to overfitting than the maximum likelihood method. We further discuss the connection between the unifying estimation framework and the rate-distortion problem.

1 Introduction

When we utilize a statistical model $p(x|\theta)$ in order to analyze i.i.d. samples from an unknown distribution, a commonly used approach is to compute a single point $\hat{\theta}$ from the parameter space. Another natural approach is to make an inference as a distribution of θ, which is natural in Bayesian statistics or ensemble learning. In this paper, we focus on the following mixture distribution:

K. Watanabe (✉)
Department of Computer Science and Engineering, Toyohashi University of Technology,
1-1, Hibarigaoka, Tempaku-cho, Toyohashi 441-8580, Japan
e-mail: wkazuho@cs.tut.ac.jp

S. Ikeda
The Institute of Statistical Mathematics, 10-3, Midori-cho, Tachikawa-shi, Tokyo 190-8562, Japan
e-mail: shiro@ism.ac.jp

$$r(x; q) = \int p(x|\theta)q(\theta)d\theta$$

and discuss estimating the mixing distribution $q(\theta)$ nonparametrically, where $q(\theta)$ is arbitrary, including continuous distributions. It was proved by Lindsay (1983) that when samples $\{x_1, \ldots, x_n\}$ are provided, the maximum likelihood estimator of the mixing distribution is a discrete distribution and the number of its support points, in other words the probability mass points, is no more than the sample size. This provides a guideline for determining the number of mixture components from data. The mixture estimation algorithm developed in Nowozin and Bakir (2008) can be utilized for estimating such discrete distributions. However, it is vulnerable to overfitting because of the high flexibility of the nonparametric estimation.

In this paper, we define an objective functional with a single parameter β, called entropic risk measure (Rudloff et al. 2008) and propose a nonparametric mixture estimation method as a minimization problem of it. With specific choices of β, the method reduces to the maximum likelihood estimation (MLE) (Lindsay 1983, 1995) and the kernel vector quantization (KVQ) (Tipping and Schölkopf 2001). We generalize Lindsay's theorem for the proposed method and prove the discreteness of the optimal mixing distribution for general β. Then, we provide an algorithm which is an extension of the procedure in Nowozin and Bakir (2008) to calculate the optimal mixing distribution for the entropic risk measure. Numerical experiments indicate that an appropriate choice of β will reduce the generalization error. We discuss the estimation bias and variance to show that the range of optimal β depends on the sample size. We also discuss the relation between the proposed mixture estimation method and the rate-distortion problem (Berger 1971).

The paper is organized as follows. Section 2 introduces the entropic risk measure as the objective functional for estimating the mixture model. Section 3 proves the discreteness of the optimal mixing distribution with an overview of Lindsay's proof, and a concrete estimation algorithm for the mixing distribution is shown. Section 4 examines its properties through numerical experiments for the Gaussian mixture model. In Sect. 5, we consider the range of β that will improve the generalization ability and describe the relation to the rate-distortion theory. Section 6 discusses the extension to other objective functionals than the entropic risk measure, and Sect. 7 concludes this paper.

2 Mixture model and objective functional

We consider a problem of estimating an unknown underlying distribution $p^*(x)$ behind n training samples, $\{x_1, \ldots, x_n\}, x_i \in \mathbf{R}^d$. As is common for density estimation and clustering, we use the following mixture density of the model $p(x|\theta)$ with parameter $\theta \in \Omega$ and the mixing distribution $q(\theta)$:

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

where $q(\theta) \geq 0$ for $\forall \theta \in \Omega$ and $\int q(\theta)d\theta = 1$. For further discussion, we assume $p(x|\theta)$ is bounded for $\forall x \in \mathbf{R}^d$ and $\forall \theta \in \Omega$.

If $q(\theta)$ is a single point distribution, computing $q(\theta)$ is a point estimation; and if $q(\theta)$ is a parametric distribution, the inference of $q(\theta)$ is known as the empirical Bayesian approach. Instead, we consider the problem nonparametrically.

Let us start by showing two approaches which are closely related to our framework. One is the Maximum Likelihood Estimation (MLE) (Lindsay 1983, 1995); the problem is denoted as follows:

$$\hat{q}(\theta) = \underset{q}{\mathrm{argmin}} \left[-\frac{1}{n} \sum_{i=1}^{n} \log r_i \right]$$

where $r_i = r(x_i; q) = \int p(x_i|\theta)q(\theta)d\theta$. The other is the Kernel Vector Quantization (KVQ) (Tipping and Schölkopf 2001),[1]

$$\hat{q}(\theta) = \underset{q}{\mathrm{argmin}} \max_i \left[-\log r_i \right].$$

In this paper, we extend these ideas and consider the following optimization problem:

$$\hat{q}(\theta) = \underset{q}{\mathrm{argmin}} \, F_\beta(q),$$

where

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

This objective functional corresponds to the entropic risk measure (of $\log r(x)$) in the literature of mathematical finance (Rudloff et al. 2008).[2] Note that $F_\beta(q)$ is continuous with respect to $\beta \in \mathbf{R}$, and convex for $\beta \geq -1$. We will discuss the convexity of it in Sect. 3.1. The above optimization problem becomes the MLE for $\beta = 0$ and the KVQ for $\beta \to \infty$. We will discuss other choices of the convex objective functional in Sect. 6.

In the rest of this section, we investigate the properties of F_β. Let $p^*(x)$ be the true distribution that is generating the data $\{x_1, \ldots, x_n\}$. The law of large numbers ensures that F_β converges to

$$\frac{1}{\beta} \log \int p^*(x) r(x)^{-\beta} dx$$

as $n \to \infty$. Let us recall the definition of the Renyi divergence (Renyi 1961),

$$D_\alpha(p_1, p_2) = \frac{1}{\alpha - 1} \log \int p_1(x)^\alpha p_2(x)^{1-\alpha} dx \geq 0, \quad D_\alpha(p, p) = 0 \quad (\alpha > 0).$$

From the non-negativity of the Renyi divergence, the following is shown easily by setting $p_1 = \widetilde{p}^*$, $p_2 = r$ and $\alpha = \beta + 1$:

$$\frac{1}{\beta} \log \int p^*(x) r(x)^{-\beta} dx \geq \frac{1}{\beta} \log \int p^*(x) \widetilde{p}^*(x)^{-\beta} dx \quad (3)$$

where

$$\widetilde{p}^*(x) = \frac{p^*(x)^{\frac{1}{1+\beta}}}{\int p^*(x)^{\frac{1}{1+\beta}} dx}. \quad (4)$$

The inequality in Eq. (3) implies that the estimated mixture distribution $\hat{r}(x)$ approaches $\widetilde{p}^*(x)$, which is equivalent to the escort distribution of p^*. The escort distribution is a distribution derived from the properties of the nonextensive form of entropy proposed by Tsallis

---

[1] The original KVQ assumes the kernel function which is connected to the probability model as $p(x|\theta) \propto K(x, \theta)$, and the possible support points of $q(\theta)$ are fixed to the sample set $\{x_1, \ldots, x_n\}$. That is $\hat{q}(\theta) = \sum_{i=1}^{n} q_i \delta(\theta - x_i)$, $q_i \geq 0$, $\sum_{i=1}^{n} q_i = 1$, where $\delta(\cdot)$ is Dirac's delta function.

[2] The entropic risk measure was originally defined only for $\beta > 0$.

([2009](#)). To put it another way, the optimal mixture distribution $r(x)$ satisfies the following relation:

$$p^*(x) = \frac{r(x)^{1+\beta}}{\int r(x)^{1+\beta} dx}.$$

The differential entropy, $\widetilde{H} = -\int \widetilde{p}^*(x) \log \widetilde{p}^*(x) dx$, of the escort distribution defined in Eq. ([4](#)) increases as β increases since $\frac{d\widetilde{H}}{d\beta} \geq 0$ for $\beta \geq -1$, and the escort distribution converges to the uniform distribution as $\beta \to \infty$.

3 Optimal mixing distribution

3.1 Discreteness of the optimal mixing distribution

In this section, we generalize Lindsay's theorem ([Lindsay 1983](#), [1995](#)) to prove that the optimal mixing distribution $q(\theta)$ which minimizes F_β in Eq. ([2](#)) is discrete. Furthermore, this enables us to rely on the decoupled approach in [Nowozin and Bakir (2008)](#). We will see this in Sect. [3.2](#).

First, we show the convexity of F_β with respect to $r = (r_1, \ldots, r_n)$ for $\beta \geq -1$. We denote F_β as a function of r by $F_\beta(r)$. The function $F_\beta(r)$ is convex if the following inequality holds for any two points, $r_0 = (r_{01}, \ldots, r_{0n})$ and $r_1 = (r_{11}, \ldots, r_{1n})$, and for $0 \leq \eta \leq 1$:

$$F_\beta(r_0 + \eta(r_1 - r_0)) \leq (1 - \eta)F_\beta(r_0) + \eta F_\beta(r_1).$$

This is equivalent to $\frac{d^2}{d\eta^2} F_\beta(r_0 + \eta(r_1 - r_0)) \geq 0, 0 \leq \eta \leq 1$ since F_β is twice differentiable. Note here that the convexity of F_β with respect to r is equivalent to the convexity with respect to q because F_β depends linearly on q through $r_i = r(x_i; q)$. Let $r_i(\eta) = (1 - \eta)r_{0i} + \eta r_{1i}$, and we have

$$\frac{d^2}{d\eta^2} F_\beta(r_0 + \eta(r_1 - r_0)) = \left[(\beta + 1) \sum_{i=1}^{n} r_i(\eta)^{-\beta-2}(r_{1i} - r_{0i})^2 \sum_{l=1}^{n} r_l(\eta)^{-\beta} \right.$$

$$\left. - \beta \left\{ \sum_{i=1}^{n} r_i(\eta)^{-\beta-1}(r_{0i} - r_{1i}) \right\}^2 \right] \frac{1}{\left(\sum_{j=1}^{n} r_j(\eta)^{-\beta} \right)^2}$$

$$= \left[(\beta + 1) \left\{ \left(\sum_{i=1}^{n} a_i^2 \right) \left(\sum_{i=1}^{n} b_i^2 \right) - \left(\sum_{i=1}^{n} a_i b_i \right)^2 \right\} \right.$$

$$\left. + \left(\sum_{i=1}^{n} a_i b_i \right)^2 \right] \frac{1}{\left(\sum_{j=1}^{n} r_j(\eta)^{-\beta} \right)^2} \geq 0$$

where $a_i = r_i(\eta)^{-\frac{\beta}{2}-1}(r_{1i} - r_{0i})$, $b_i = r_i(\eta)^{-\frac{\beta}{2}}$ and the fact $\beta \geq -1$ and the Cauchy-Schwarz inequality are used. This shows that $F_\beta(r)$ is convex.

Next, we show the directional derivative of $F_\beta(r)$. The directional derivative from r_0 to r_1 is defined as follows:

$$F'_{\beta, r_0}(r_1) = \lim_{\eta \to 0} \frac{F_\beta(r_0 + \eta(r_1 - r_0)) - F_\beta(r_0)}{\eta} = 1 - \frac{\sum_{i=1}^{n} r_{0i}^{-\beta-1} r_{1i}}{\sum_{j=1}^{n} r_{0j}^{-\beta}}. \tag{5}$$

Note that Eq. (5) is valid for $\beta = 0$ as well.

If r^* minimizes $F_\beta(r)$, $F'_{\beta,r^*}(r_1) \geq 0$ must hold for any r_1. It has been proved for $\beta = 0$ that there exists a unique r that minimizes F_β at the boundary of the convex hull of the set $\{p_\theta = (p(x_1|\theta), \ldots, p(x_n|\theta))|\theta \in \Omega\}$, where Ω is the parameter space (Lindsay 1983, 1995). This result can be generalized for the case $\beta \geq -1$ because of the convexity of F_β. From Caratheodory's theorem, this means that the optimal r is expressed by a convex combination, $\sum_{l=1}^k \pi_l p_{\theta_l}$, with $\pi_l \geq 0$, $\sum_{l=1}^k \pi_l = 1$ and $k \leq n$, indicating that the optimal mixing distribution is $q(\theta) = \sum_{l=1}^k \pi_l \delta(\theta - \theta_l)$, $\theta_l \in \Omega$, which is a discrete distribution where the number of the support points is no more than n.

3.2 Optimization of mixing distribution

In this section, we derive an estimation algorithm for $q(\theta)$ following Nowozin and Bakir (2008). This algorithm iterates the subproblem that augments a new point to the support of $q(\theta)$ and the learning of the finite mixture model. The minimization of F_β over finite mixture models is implemented with a simple updating rule by the expectation-maximization (EM) algorithm (Dempster et al. 1977; Barber 2012).

3.2.1 Learning procedure

If r^* minimizes $F_\beta(r)$, $F'_{\beta,r^*}(p_\theta) \geq 0$ for any p_θ. Furthermore, $F'_{\beta,r^*}(p_\theta) = 0$ holds for $\theta = \theta_l$ where $\{\theta_l\}_{l=1}^k$ is the set of support points of the optimal mixing distribution. This follows from the fact that if $F'_{\beta,r^*}(p_{\theta_l}) > 0$ holds for some θ_l, there must exist a $\theta_{l'}$ such that $F'_{\beta,r^*}(p_{\theta_{l'}}) < 0$,[3] and $F_\beta(r)$ can be decreased by adding more weight $\pi_{l'}$ on $\theta_{l'}$. Thus, the optimal condition for the mixing distribution $q(\theta)$ is summarized as

$$\begin{cases} \mu(\theta) = 1 \ (\theta = \theta_l, l = 1, 2, \ldots, k) \\ \mu(\theta) \leq 1 \ (otherwise) \end{cases}$$

where

$$\mu(\theta) = \sum_{i=1}^n \alpha_i \, p(x_i|\theta), \tag{6}$$

$$\alpha_i = \frac{r_i^{-\beta-1}}{\sum_{j=1}^n r_j^{-\beta}}, \tag{7}$$

and $r_i = \int p(x_i|\theta)q(\theta)d\theta$. This yields Algorithm 1 for the optimization of the mixing distribution $q(\theta)$, which sequentially augments points θ, satisfying $\mu(\theta) > 1$ until the maximum of $\mu(\theta)$ approaches 1 (Nowozin and Bakir 2008). In each iteration, the maximum of $\mu(\theta)$ is calculated (Step 3). If the maximum is larger than 1, one point is added to the set of support points. Then locations and weights of the support points are optimized (Step 4), that is, parameters of a finite mixture model are estimated. We derive an EM-like algorithm for this step in Sect. 3.2.2. In the case of the KVQ ($\beta \to \infty$), this step was originally formulated and solved by linear programming (Tipping and Schölkopf 2001; Nowozin and Bakir 2008).

The above algorithm updates $\{\theta_l\}$ as well as $\{\pi_l\}$ in Step 4. This is an extension of Nowozin and Bakir (2008), where only $\{\pi_l\}$ is updated in the algorithm. Algorithm 1 requires a constant ϵ and strongly depends on it especially when only $\{\pi_l\}$ is updated. In the numerical

---

[3] This is proved simply from the fact that $F'_{\beta,r^*}(r^*) = \sum_{l=1}^k \pi_l F'_{\beta,r^*}(p_{\theta_l}) = 0$.

Algorithm 1 Decoupled Approach to Mixture Estimation

1: Initialize $k = 0$ and $\alpha_i = 1/n$ and prepare a small positive constant ϵ.
2: **repeat**
3:　Let

$$\theta_k = \underset{\theta}{\text{argmax}}\, \mu(\theta)$$

　　and $k = k + 1$, where $\mu(\theta)$ is given by Eq. (6).
4:　Define the discrete distribution

$$q_k(\theta) = \sum_{l=1}^{k} \pi_l \delta(\theta - \theta_l).$$

　　Optimize $\{\pi_l, \theta_l\}_{l=1}^{k}$ by minimizing $F_\beta(q_k)$.
5:　Compute $\{\alpha_i\}_{i=1}^{n}$ by Eq. (7) with $r_i = \sum_{l=1}^{k} \pi_l p(x_i | \theta_l)$.
6: **until** $\max_\theta \mu(\theta) < 1 + \epsilon$ holds.

experiments in the next section, we set $\epsilon = 0.01$. From the assertion in Sect. 3.1, this learning procedure is guaranteed to stop before the support size of $q(\theta)$ exceeds n if the learning procedure is started with an empty support set and updating both $\{\theta_l\}$ and $\{\pi_l\}$.

3.2.2 EM updates for finite mixtures

In this subsection, we discuss the learning algorithm of a finite mixture model that is required in Step 4 of Algorithm 1. We separate the learning algorithm for the cases $-1 \leq \beta < 0$ and $\beta \geq 0$, and derive the algorithm for each case. Let us start with $-1 \leq \beta < 0$. First we define the following function $F_\beta(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{w})$ for $\beta \neq 0$:[4]

$$F_\beta(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{w}) = -\sum_{i=1}^{n} w_i \log r(x_i; q) - \frac{1}{\beta} \sum_{i=1}^{n} w_i \log w_i - \frac{1}{\beta} \log n \qquad (8)$$

where $\boldsymbol{w} \in \Delta = \{\boldsymbol{w} = (w_1, w_2, \ldots, w_n) | w_i \geq 0, \sum_{i=1}^{n} w_i = 1\}$, $\boldsymbol{\theta} = \{\theta_l\}_{l=1}^{k}$ and $\boldsymbol{\pi} = \{\pi_l\}_{l=1}^{k}$. Note that $F_\beta(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{w})$ is convex for $\boldsymbol{w}$ when $-1 \leq \beta < 0$ and

$$\min_{\boldsymbol{w} \in \Delta} F_\beta(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{w}) = F_\beta(\boldsymbol{r})$$

where the minimum is attained when

$$w_i = \frac{r_i^{-\beta}}{\sum_{j=1}^{n} r_j^{-\beta}}. \qquad (9)$$

Since $F_\beta(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{w})$ is convex with respect to $\boldsymbol{w}$, when $-1 \leq \beta < 0$, we can derive a double minimization algorithm as follows:

$$w_i^{(t)} = \frac{r_i^{(t)-\beta}}{\sum_{j=1}^{n} r_j^{(t)-\beta}}, \quad r_i^{(t)} = \sum_{l=1}^{k} \pi_l^{(t)} p(x_i | \theta_l^{(t)}),$$

---

[4] $F_\beta(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{w})$ is related to the fact that the conjugate function of the log-sum-exp function is the entropy function, $-\sum_{i=1}^{n} w_i \log w_i$ (Boyd and Vandenberghe 2004).

$$\pi_l^{(t+1)} = \sum_{i=1}^{n} w_i^{(t)} v_{il}^{(t)}, \tag{10}$$

$$\theta_l^{(t+1)} = \underset{\theta}{\mathrm{argmax}} \sum_{i=1}^{n} w_i^{(t)} v_{il}^{(t)} \log p(x_i|\theta) \tag{11}$$

where

$$v_{il}^{(t)} = \frac{\pi_l^{(t)} p(x_i|\theta_l^{(t)})}{\sum_{m=1}^{k} \pi_m^{(t)} p(x_i|\theta_m^{(t)})}$$

is the posterior probability that the data point x_i is assigned to the cluster center θ_l. In fact, at the stationary point, either $\mu(\theta_l) < 1$ and $\pi_l = 0$ or $\mu(\theta_l) = 1$ and $\pi_l > 0$ hold. This is an EM-like algorithm which can be seen from the fact that Eq. (8) is equivalent to a weighted sum of negative log-likelihood.

When $p(x|\theta)$ is a member of the exponential family with the sufficient statistic $T(x)$, that is, $p(x|\theta) = h(x) \exp\{\theta^T T(x) - G(\theta)\}$, Eq. (11) is simplified to

$$\theta_l^{(t+1)} = (\nabla G)^{-1} \left(\sum_{i=1}^{n} w_i^{(t)} v_{il}^{(t)} T(x_i) \right)$$

where $(\nabla G)^{-1}$ is the link function to the natural parameter space (Banerjee et al. 2005).

We can prove that the above update monotonically decreases the objective F_β for $\beta \leq 0$.[5]

Let us move to the case $\beta > 0$. When $\beta > 0$, $F_\beta(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{w})$ is concave with respect to $\boldsymbol{w}$ and the previous EM-like algorithm does not work.[6] To directly minimize $F_\beta(q)$ with respect to $\{\boldsymbol{\theta}, \boldsymbol{\pi}\}$, we switch to the following updating rules:

$$\pi_l^{(t+1)} \propto \left\{ \sum_{i=1}^{n} (v_{il}^{(t)})^{1+\beta} p(x_i|\theta_l)^{-\beta} \right\}^{\frac{1}{1+\beta}},$$

$$\theta_l^{(t+1)} = \theta_l^{(t)} + H(\theta_l^{(t)})^{-1} \sum_{i=1}^{n} (v_{il}^{(t)})^{1+\beta} p(x_i|\theta_l^{(t)})^{-\beta} \frac{\partial \log p(x_i|\theta_l^{(t)})}{\partial \theta}$$

where

$$H(\theta_l^{(t)}) = \sum_{i=1}^{n} (v_{il}^{(t)})^{1+\beta} p(x_i|\theta_l^{(t)})^{-\beta} \left\{ -\frac{\partial^2 \log p(x_i|\theta_l^{(t)})}{\partial \theta \partial \theta^T} \right.$$

$$\left. + \beta \left(\frac{\partial \log p(x_i|\theta_l^{(t)})}{\partial \theta} \right) \left(\frac{\partial \log p(x_i|\theta_l^{(t)})}{\partial \theta} \right)^T \right\}$$

is the Hessian matrix. These updates monotonically decrease $F_\beta(q_k)$ for $\beta > 0$ and are derived as the Newton-Raphson step (Boyd and Vandenberghe 2004, Sect. 9.5) to decrease the right hand side of the inequality

---

[5] The algorithm was defined for $-1 \leq \beta < 0$, but it can be easily checked that the algorithm works for $\beta = 0$.

[6] For $\beta > 0$, it holds that

$$\max_{\boldsymbol{w} \in \Delta} F_\beta(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{w}) = F_\beta(\boldsymbol{r}),$$

and the maximum is attained by $\boldsymbol{w}$ satisfying Eq. (9).

$$\sum_{i=1}^{n} \left(\sum_{l=1}^{k} \pi_l \, p(x_i|\theta_l) \right)^{-\beta} \le \sum_{i=1}^{n} \sum_{l=1}^{k} v_{il}^{1+\beta} \, (\pi_l \, p(x_i|\theta_l))^{-\beta},$$

which is Jensen's inequality for the convex function $x^{-\beta}$ with $\beta > 0$.

3.3 Pre-imaging for generation of support points

We discuss the relationship between the proposed algorithm and kernel-based learning algorithms. In this subsection, we focus on the case in which the component $p(x|\theta)$ is a location family and is represented as $p(x|\theta) \propto f(x - \theta)$ for some function f, such as the Gaussian density in Eq. (12).

As mentioned in Nowozin and Bakir (2008), the maximization of $\mu(\theta)$ in Eq. (6) reduces to the pre-image problem (Schölkopf et al. 1999) if the likelihood function $p(x|\theta) \propto f(x-\theta)$ is, up to multiplication of a constant, given by the kernel function $K(x, \theta)$ associated with a reproducing kernel Hilbert space. This is because, for the feature map $\phi(x)$ satisfying $f(x - \theta) = K(x, \theta) = \phi(x)^T \phi(\theta)$, the maximization of $\mu(\theta)$ in Eq. (6) is equivalent to the minimization of the norm (squared) in the Hilbert space

$$\left\| \phi(\theta) - \sum_{i=1}^{n} \alpha_i \phi(x_i) \right\|^2 = K(\theta, \theta) - 2 \sum_{i=1}^{n} \alpha_i K(x_i, \theta) + \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$$

when $K(\theta, \theta)$ is constant. Note here that the coefficients $\{\alpha_i\}$ depend on β as in Eq. (7). More specifically, the coefficient α_i is identified by w_i in Eq. (9) and $r_i = \int K(x_i, \theta) q(\theta) d\theta$ from Eq. (7) as

$$\alpha_i = \frac{w_i}{r_i}.$$

The reciprocal dependence on r_i means that maximizing Eq. (6) yields the new support point around which the finite mixture constructed so far has low density. Further, if $\beta \ne 0$, w_i weighs each sample point according to Eq. (9).

4 Numerical experiments

In this section, we investigate the properties of the estimation method by a numerical simulation with 2-dimensional Gaussian mixtures where

$$p(x|\theta) = \frac{\gamma}{\pi} \exp\left(-\gamma \|x - \theta\|^2 \right). \tag{12}$$

We generated synthetic data by the following distribution:

$$p^*(x) = \frac{1}{2} N(x|\theta_1^*, I_2) + \frac{1}{2} N(x|\theta_2^*, I_2) \tag{13}$$

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. Figure 1 shows $p^*(x)$ and an example of a data set ($n = 50$). We applied Algorithm 1 to the synthetic data drawn from $p^*(x)$ in Eq. (13) and estimated $q(\theta)$. We also applied the original version of the algorithm in Nowozin and Bakir (2008), where each θ_l is not updated in Step 4 of Algorithm 1 but fixed once generated in Step 3, and only $\{\pi_l\}$ are updated. Results for this case will be indicated as "means fixed."
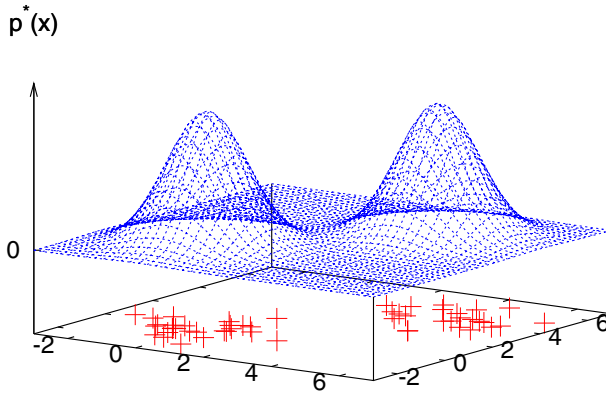
$$p^*(x)$$

Fig. 1 The density function of $p^*(x)$ (3D plot) and an example of a data set (*crosses*) ($n = 50$)

Let $\hat{q}(\theta)$ be an estimated mixing distribution. We calculated the likelihood as the training error

$$-\frac{1}{n}\sum_{i=1}^{n}\log r(x_i;\hat{q}) = -\frac{1}{n}\sum_{i=1}^{n}\log\int p(x_i|\theta)\hat{q}(\theta)d\theta, \tag{14}$$

the prediction error for the test data $\{\tilde{x}_i\}_{i=1}^{\tilde{n}}$ drawn from the true distribution in Eq. (13)

$$-\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\log r(\tilde{x}_i;\hat{q}) = -\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\log\int p(\tilde{x}_i|\theta)\hat{q}(\theta)d\theta, \tag{15}$$

and the maximum error for the training data $\{x_i\}_{i=1}^{n}$

$$\max_{1\le i\le n}\left[-\log\int p(x_i|\theta)\hat{q}(\theta)d\theta\right] \tag{16}$$

which corresponds to the objective functional of the KVQ. The number of training data is $n = 50$ and that of test data is $\tilde{n} = 200{,}000$.

Furthermore, we investigated the number of estimated components. Since this number strongly depends on ϵ, we also applied hard assignments to cluster centers for each data point and counted the number of hard clusters. Here, each point x_i is assigned to the cluster center $\hat{\theta}_l$ that maximizes the posterior probability

$$p(\hat{\theta}_l|x_i) = v_{il} = \frac{\hat{\pi}_l p(x_i|\hat{\theta}_l)}{\sum_{m=1}^{\hat{k}}\hat{\pi}_m p(x_i|\hat{\theta}_m)} \tag{17}$$

where we have assumed $\hat{q}(\theta) = \sum_{l=1}^{\hat{k}}\hat{\pi}_l\delta(\theta-\hat{\theta}_l)$. The number of the hard clusters is usually smaller than the number of the mixture components, that is, there are some components which will never be selected by the hard assignment.

This posterior probability and the number of components will be used in connection with rate-distortion function in Sect. 5.2. All results were averaged over 100 trials for different data sets generated by (13).
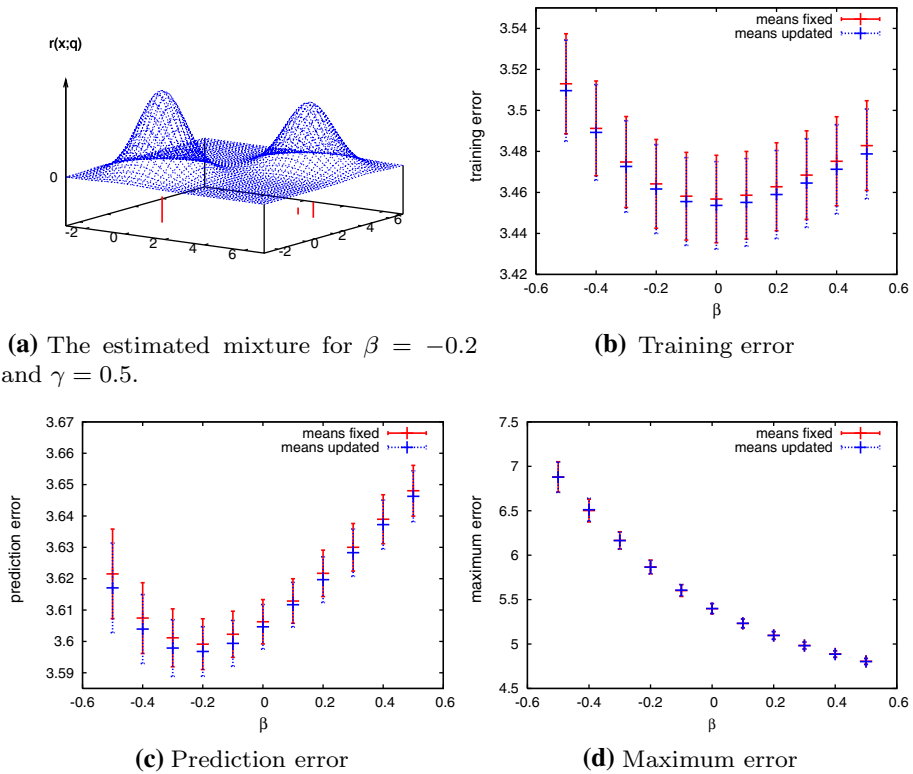
(a) The estimated mixture for $\beta = -0.2$ and $\gamma = 0.5$.

(b) Training error

(c) Prediction error

(d) Maximum error

Fig. 2 **a** Example of the estimated mixture for $\beta = -0.2$ and $\gamma = 0.5$. The corresponding mixing distribution is illustrated in the x–y plane where the location and the height of the *solid lines* are, respectively, the mean parameter $\hat{\theta}_l$ and the weight $\hat{\pi}_l$ of each component. **b–d** Training error, prediction error and maximum error against β, respectively. The *error bars* show 95 % confidence intervals

4.1 Prediction with known Kernel width

First, we assumed that the kernel width γ in Eq. (12) was known and $p(x|\theta)$ was set to $p(x|\theta) = \frac{1}{2\pi} \exp\left(-\frac{||x-\theta||^2}{2}\right)$. The distribution $p^*(x)$ is realized in this case by the mixing distribution $q(\theta) = \frac{1}{2}\delta(\theta - \theta_1^*) + \frac{1}{2}\delta(\theta - \theta_2^*)$. An example of the estimated mixture model for $\beta = -0.2$ and $\gamma = 0.5$ is demonstrated in Fig. 2a.

Figure 2b, c, respectively, show the training error (14) and the prediction error (15).

We see that the average training error is minimized at $\beta = 0$ as expected, while the minimum of average prediction error is attained around $\beta = -0.2$. Figure 2d shows the average of the maximum errors of Eq. (16). As expected, it monotonically decreases with respect to β, which is consistent with the fact that the estimation approaches the KVQ as $\beta \to \infty$.

In Fig. 3, we show the average number of estimated components remaining after the elimination of components with sufficiently small mixing proportions (less than $\frac{1}{n^2}$) and the average number of hard clusters.

The number of components $\hat{k}$ as well as the number of hard clusters increase as β becomes larger. The discussion in Sect. 2 suggests that as β grows, more components are estimated to
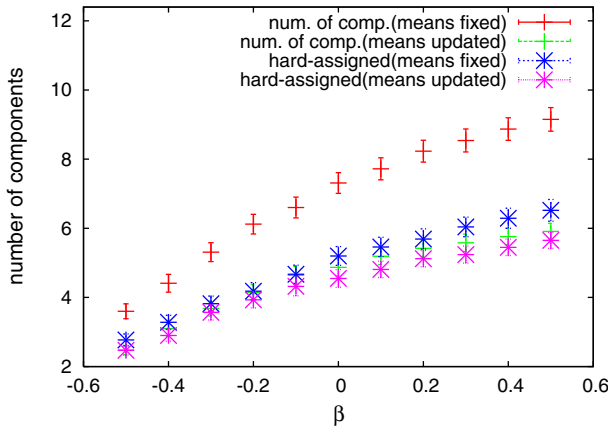
Fig. 3 Number of components (*crosses*) and number of hard clusters (*asterisks*) against β

increase the entropy of the mixture $r(x; q)$. This regularization reduces the average prediction error when β takes a slightly negative value as we have just observed in Fig. 2c. In Sect. 5.1, we will discuss the effective range of β which reduces the generalization error.

4.2 Mismatched Kernel width

Next we assumed $\gamma \neq 0.5$, that is, the variance of a component has a mismatch.

When $\gamma < 0.5$, the true distribution in Eq. (13) cannot be realized with the model in Eq. (1). This will induce a larger objective F_β and a larger training error by the order of $O(1)$. The top panels of Fig. 4a–c show examples of estimated mixtures and the average prediction error as a function of β for $\gamma = 0.05$, $\gamma = 0.2$ and $\gamma = 0.4$, respectively. We see that the prediction error is much larger for $\gamma = 0.05$ and $\gamma = 0.2$ than for $\gamma = 0.4$.

When $\gamma \geq 0.5$, the distribution $p^*(x)$ is realizable with the mixture distribution in Eq. (1) using the mixing distribution, $q(\theta) = \frac{1}{2}N(\theta|\theta_1^*, (1-1/(2\gamma))I_2) + \frac{1}{2}N(\theta|\theta_2^*, (1-1/(2\gamma))I_2)$. The top panels of Fig. 4d–f show examples of estimated mixtures and the average prediction error as a function of β for $\gamma = 0.6$, $\gamma = 1.0$ and $\gamma = 2.0$, respectively. For $\gamma = 1.0$ and $\gamma = 2.0$, the minimum is achieved when $\beta > 0$. This is expected from the fact that infinitely many components are required for $r(x; q)$ to be identical to the true distribution.

The results in Fig. 4c ($\gamma = 0.4$) and Fig. 4d ($\gamma = 0.6$) are similar to those presented in Sect. 4.1 because γ is close to the true value, 0.5. The prediction error increases when the mismatch of γ is large. The above results imply that it is possible to use cross-validation to select β and the kernel width γ. But a practical procedure needs to be explored further. The next section is devoted to discussing the selection of these parameters.

5 Selection of parameters

This section first discusses the optimal parameter β that minimizes the average generalization error. The relationship between the mixture estimation and the rate-distortion problem is then described.
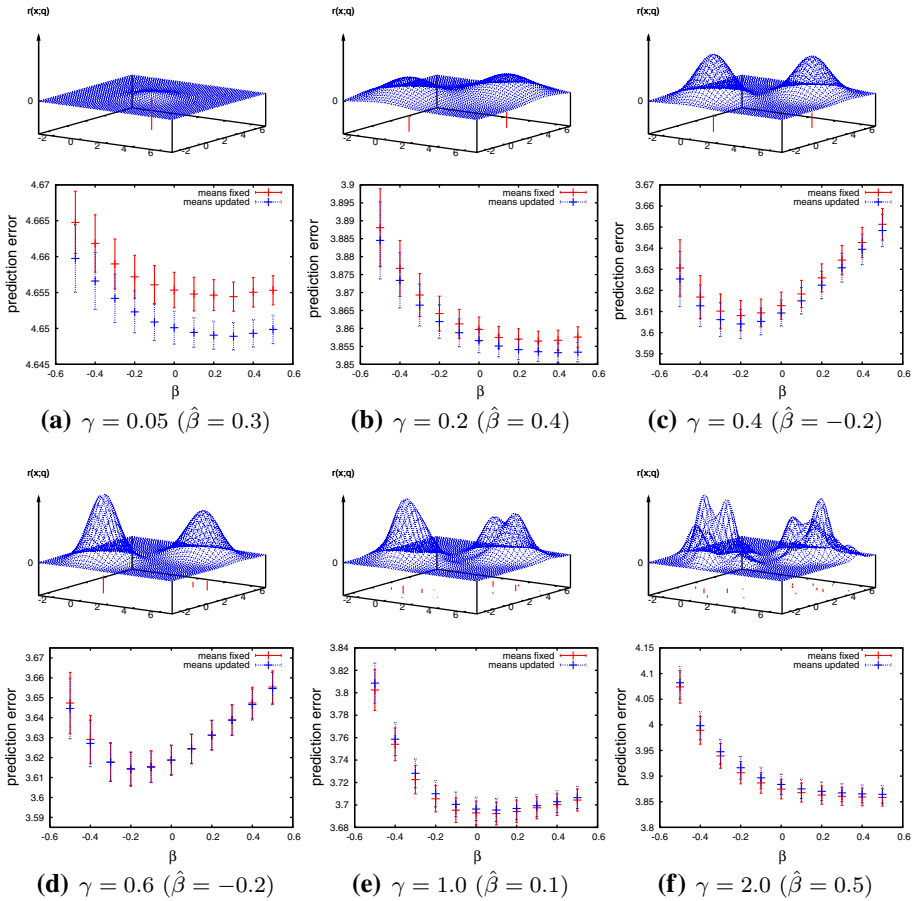
Fig. 4 Examples of estimated mixtures (*top row*) and prediction errors as a function of β (*bottom row*) for **a** $\gamma = 0.05$ ($\hat{\beta} = 0.3$), **b** $\gamma = 0.2$ ($\hat{\beta} = 0.4$), **c** $\gamma = 0.4$ ($\hat{\beta} = -0.2$), **d** $\gamma = 0.6$ ($\hat{\beta} = -0.2$), **e** $\gamma = 1.0$ ($\hat{\beta} = 0.1$) and **f** $\gamma = 2.0$ ($\hat{\beta} = 0.5$). In *each column*, the estimated mixture and the mixing distribution are displayed in the same way as in Fig. 2a, and the prediction error is displayed as in Fig. 2c. The value of β, indicated in parentheses as $\hat{\beta}$, was selected so as to minimize the average prediction error when the means are updated

5.1 Effective range of β

Let $\hat{r}_\beta(x)$ be the estimated mixture model for β. We discuss the β which minimizes the average generalization error

$$E\left[\int p^*(x) \log \frac{p^*(x)}{\hat{r}_\beta(x)} dx\right]$$

where E denotes the expectation with respect to the distribution of the data sets, $\prod_{i=1}^{n} p^*(x_i)$. The optimal β is related to the number of data n. As described in Sect. 2, $\hat{r}_\beta(x)$ approaches the escort distribution $\widetilde{p}^*(x) \propto p^*(x)^{\frac{1}{1+\beta}}$. This brings a bias to the estimation. We use the Kullback-Leibler divergence $KL(p^*, \widetilde{p}^*) = \int p^*(x) \log \frac{p^*(x)}{\widetilde{p}^*(x)} dx$ as the indicator of the bias.
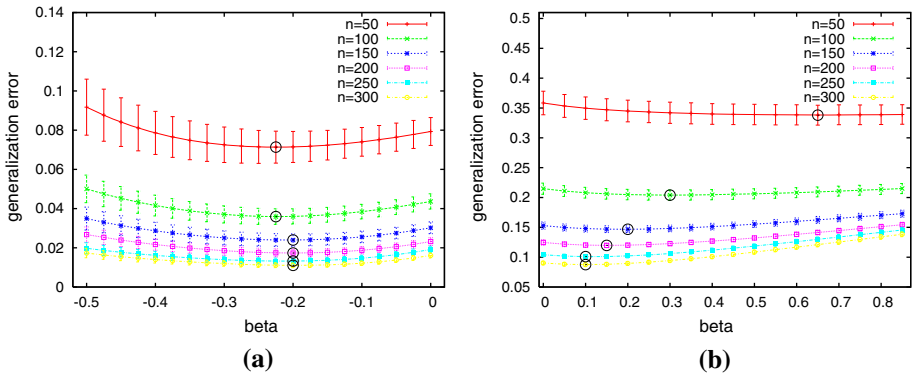
Fig. 5 Generalization errors against β for different n when **a** $\gamma = 0.5$ and **b** $\gamma = 2.0$. Minimums are *marked with circles*

It follows from the Taylor expansion of the divergence around $\beta = 0$ that

$$KL(p^*, \widetilde{p}^*) = \frac{\mathrm{Var}_{p^*}[\log p^*(x)]}{2}\beta^2 + O(\beta^3) \tag{18}$$

where $\mathrm{Var}_{p^*}[\log p^*(x)] = \int p^*(x)(\log p^*(x))^2 dx - \left(\int p^*(x)\log p^*(x)\right)^2$.

We focus on the Gaussian mixture used in Sect. 4 and consider the condition that the influence of the bias does not exceed the reduction in variance of estimation. It is conjectured that the log-likelihood ratio of the finite mixture models is in the order of $\log\log n$ when the model has redundant components (Hartigan 1985). This implies that $\sum_{i=1}^{n}\log\frac{\hat{r}_0(x_i)}{p^*(x_i)} = O_p(\log\log n)$. We further assume symmetry between the training error, $E\left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{p^*(x_i)}{\hat{r}_0(x_i)}\right]$, and the generalization error, $E\left[KL(p^*, \hat{r}_0)\right]$. The symmetry means that the training and generalization errors converge to zero symmetrically from below and above zero as $n \to \infty$ (Amari et al. 1992; Watanabe 2005). Here, we assume the symmetry holds in the slightly weaker sense that both errors have the same order with respect to n. More specifically,

$$E\left[KL(p^*, \hat{r}_0)\right] = O\left(\frac{\log\log n}{n}\right). \tag{19}$$

From Eq. (18) and the above result, the order of the optimal β that minimizes the generalization error is

$$\hat{\beta} = O\left(\sqrt{\frac{\log\log n}{n}}\right)$$

since otherwise the bias in Eq. (18) has a larger order than that of the variance around $\beta = 0$.

Figure 5a, b show the average generalization error, $\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\log\frac{p^*(\tilde{x}_i)}{r(\tilde{x}_i;\hat{q})}$ against β for different sample sizes, n, when $\gamma = 0.5$ and $\gamma = 2.0$ respectively.[7] As can be seen from the results for $\beta = 0$ in both figures, the assumption in Eq. (19) is reasonable. We also see that the generalization error is minimized by β with the smaller absolute value as n increases. This tendency becomes more apparent for $\gamma = 2.0$, that is, when the variance is underestimated.

---

[7] We did not show but a similar trend was observed for the case where $\{\theta_l\}$ are fixed (see Sect. 4).

5.2 Connection to the rate-distortion problem

The convex clustering in Lashkari and Golland (2007) corresponds to a special case of our proposal, that is, $\beta = 0$ (MLE), and the support points of $q(\theta)$ are fixed to the training data set $\{x_1, \ldots, x_n\}$. For this restricted version of the problem, it is pointed out that the kernel width, γ in Eq. (12), has a relationship to the rate-distortion (RD) function of source $\hat{p}(x)$ (the empirical distribution) and distortion associated with $p(x|\theta)$ (for example, the squared distortion for Gaussian) (Lashkari and Golland 2007). In this section, we investigate the relationship of our proposal with the RD theory in a general case where the support points of $q(\theta)$ are not restricted to the sample points and $\beta \neq 0$. The relationship of mixture modeling with the RD theory is also partly discussed in Banerjee et al. (2005) for the finite mixture of exponential family distributions under the constraint that the cardinality of the support of $q(\theta)$ is fixed.

Let us start with a short summary of the RD theory. The source random variable X with density $p^*(x)$ is reproduced to Θ with the conditional distribution $p(\theta|x)$, where the distribution is chosen to minimize the rate, *i.e.*, the mutual information $I(X; \Theta)$ under the constraint of the average distortion measure. This is formulated by the Lagrange multiplier method as follows and is reformulated as the optimization problem in Eq. (21) (Berger 1971):

$$
\min_{q(\theta|x)} \left[I(X; \Theta) - s E[d(X, \Theta)] \right]
$$

$$
= \min_{q(\theta|x)} \left[\int p^*(x) q(\theta|x) \log \frac{q(\theta|x)}{\int p^*(x) q(\theta|x) dx \exp(sd(x, \theta))} dx d\theta \right]
$$

$$
= \min_{q(\theta|x), q(\theta)} \left[\int p^*(x) q(\theta|x) \log \frac{q(\theta|x)}{q(\theta) \exp(sd(x, \theta))} dx d\theta \right] \tag{20}
$$

$$
= \min_{q(\theta)} \left[- \int p^*(x) \{ \log \int q(\theta) \exp(sd(x, \theta)) d\theta \} dx \right] \tag{21}
$$

where Eqs. (20) and (21) follow from the facts that the minimization on $q(\theta)$ and $q(\theta|x)$ are respectively attained by

$$
q(\theta) = \int p^*(x) q(\theta|x) dx
$$

and

$$
q(\theta|x) = \frac{q(\theta) \exp(sd(x, \theta))}{\int q(\theta) \exp(sd(x, \theta)) d\theta}.
$$

Here $d(x, \theta)$ is the distortion measure and the negative real variable s is a Lagrange multiplier. s provides the slope of a tangent to the RD curve and hence has one-to-one correspondence with a point on the RD curve. The problem in Eq. (21) 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 $\gamma = -s$.

For a general β, the expression (8) and the optimal reconstruction distribution $\hat{q}(\theta) = \sum_{l=1}^{\hat{k}} \hat{\pi}_l \delta(\theta - \hat{\theta}_l)$ imply the RD function of the source, $\sum_{i=1}^{n} w_i \delta(x - x_i)$, with the rate

$$
R = \sum_{i=1}^{n} \sum_{l=1}^{\hat{k}} w_i v_{il} \log \frac{v_{il}}{\sum_{j=1}^{n} w_j v_{jl}}
$$

and the average distortion

(a) RD curve for the true Gaussian mixture defined in Eq. (13) and its Shannon lower bound.

(b) RD curve for $\beta = 0.0$.

(c) RD curve for $\beta = -0.2$.
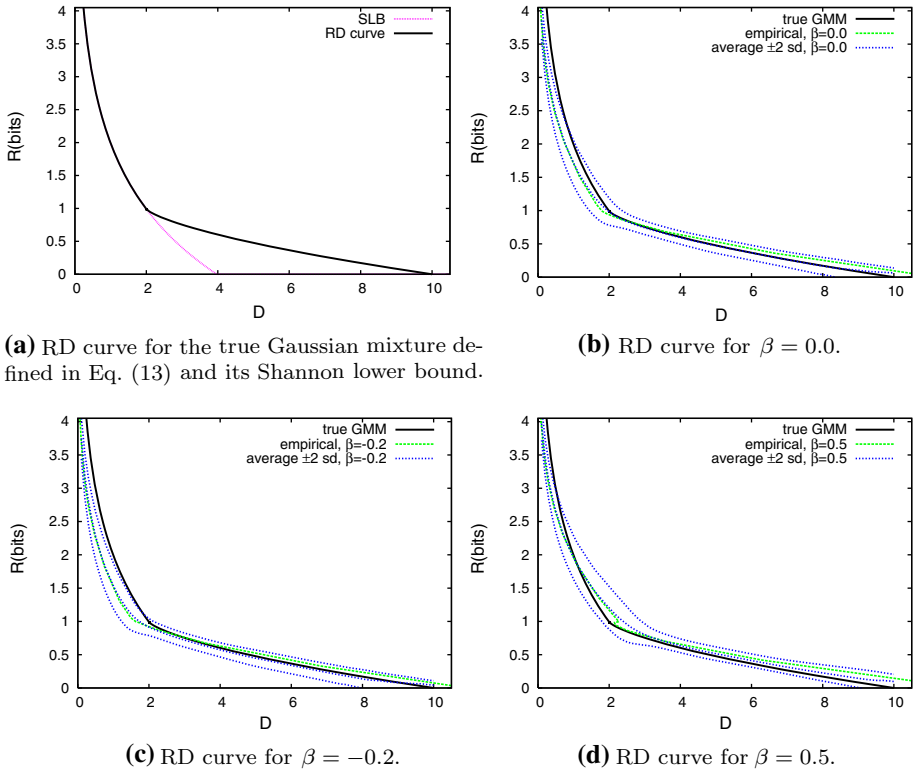
(d) RD curve for $\beta = 0.5$.

Fig. 6 Examples of RD curves. **a** RD curve for the true Gaussian mixture in Eq. (13) (*solid line*) and its Shannon lower bound (*dotted line*). **b–d**: RD curve for an empirical distribution (*dashed line*) and the average of linearly interpolated RD curves over 100 empirical data sets with the upper and lower bands indicating 2 standard deviations (*dotted lines*) when **b** $\beta = 0.0$, **c** $\beta = -0.2$ and **d** $\beta = 0.5$. The rate is scaled by log 2 to yield bits

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

where v_{il} is the posterior probability defined by Eq. (17). Since the rate is equivalent to 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 $\log \hat{k}$.

Figure 6a shows the RD curve for $p^*(x)$ given by the Gaussian mixture in Eq. (13) and its Shannon lower bound (Berger 1971). To draw the RD curve, we used the following facts: The optimal reconstruction distribution $q(\theta)$ is the Dirac delta distribution centered at $\bar{\theta} = \frac{\theta_1^* + \theta_2^*}{2} = (2, 2)^T$ for a distortion larger than $D_{max} = \min_\theta \int p^*(x)\|\theta - x\|^2 dx = 10$. The optimal $q(\theta)$ is the Gaussian mixture $\frac{1}{2} N\left(\theta | \theta_1^*, (1 + \frac{1}{2s}) I_2\right) + \frac{1}{2} N\left(\theta | \theta_2^*, (1 + \frac{1}{2s}) I_2\right)$ for a distortion less than 1, the variance of the component. The RD curve is equal to the Shannon lower bound for this range of the distortion. For a distortion between 1 and D_{max}, the optimal $q(\theta)$ is a two-component discrete distribution, $\frac{1}{2}\delta(\theta - \bar{\theta} - a\mathbf{1}) + \frac{1}{2}\delta(\theta - \bar{\theta} + a\mathbf{1})$, where $\mathbf{1} = (1, 1)^T$ and a is a real number between 0 and 2, which we identified for each s ($s \geq -0.5$) by the optimality condition of the minimization problem in Eq. (21). These

$\textcircled{2}$ Springer

facts agree with the result of Rose (1994), which proves that for the squared distortion, if the Shannon lower bound is not tight, then the optimal reconstruction distribution is discrete.

Figure 6b overwrites Fig. 6a (without the Shannon lower bound) with the RD curve for the empirical distribution given by a data set ($n = 50$) generated by $p^*(x)$. We used Algorithm 1 with $\beta = 0$ for estimating $q(\theta)$ for each s and interpolated linearly to draw the RD curve. Figure 6c, d show RD curves for $\beta = -0.2$ and $\beta = 0.5$, respectively, drawn by using Algorithm 1 for each s. Note here that, in the above interpretation of the proposed optimization as an RD problem, the source depends on w_i, which depends on $q(\theta)$ as in Eq. (9) whereas in the original RD problem, the source does not depend on the reconstruction distribution. Hence the above pair of rate and distortion does not necessarily inherit properties of the usual RD function such as convexity, except for $\beta = 0$. In fact, the RD curve for $\beta = 0.5$ loses convexity as in Fig. 6d.

Figure 6b–d also show the average and twice the standard deviation of the linearly interpolated RD curves for 100 empirical data sets. We can see that when compared to the MLE ($\beta = 0$), the RD curves for $\beta = -0.2$ have small variation around the point $(D, R) = (2, 1)$, and those for $\beta = 0.5$ are, on average, close to the RD curve for the true Gaussian mixture in the small distortion region, that is, for a small variance of the Gaussian component. These imply the observation in Sects. 4.1 and 4.2 that $\beta \neq 0$ can reduce the generalization error of the MLE. For $\beta > 0$, the learning algorithm developed in Sect. 3 can be considered as an algorithm for computing Renyi's analog of the rate-distortion function previously appearing in Arikan and Merhav (1998) in the context of guessing.

6 Further topics

By extending Lindsay's theorem, we proved in Sect. 3.1 that the estimated $q(\theta)$ is a discrete distribution consisting of distinct support points no greater in number than the number of training data. If $p(x|\theta)$ is bounded for all x and θ, this statement can be generalized to other objective functionals as long as they are convex with respect to $q(\theta)$ and hence to $r = (r_1, \ldots, r_n)$. The proposed algorithm in Sect. 3.2 is based on the decoupled approach developed in Nowozin and Bakir (2008). The general objective functional considered in Nowozin and Bakir (2008) includes the MLE and the KVQ. More specifically, the following four objective functionals are demonstrated as examples in Nowozin and Bakir (2008). 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 functional F_β in Eq. (2) combines the first two objectives by the parameter β. The other two objectives above are convex with respect to r as well and hence can be proven to have optimal discrete distributions $q(\theta)$ with no more than n support points. Note that since r is a linear function of $q(\theta)$, the convexity on r is equivalent to that on $q(\theta)$. Furthermore, we have developed in Sect. 3.2.2 a simple algorithm for finite mixture models to minimize F_β. In fact, this optimization algorithm for large β is used for approximate computation of the prior distribution achieving the normalized maximum likelihood in the context of universal coding (Barron et al. 2014). Note that, to apply the general framework

of Sect. 3.2.1 to specific objective functionals, we need learning algorithms for optimizing them for finite mixture models.

Another aspect of the choice of the objective functional is the robustness of the estimation. In Sect. 2, we demonstrated that the minimization of F_β is related to that of the Renyi divergence. We further discuss its relationship to the divergence minimization that was proposed for the purpose of robust estimation. The gamma divergence (Fujisawa and Eguchi 2008; Eguchi et al. 2011) is defined for non-negative densities g and h with a real parameter $\gamma \geq -1$ as

$$D_\gamma(g, h) = d_\gamma(g, h) - d_\gamma(g, g) \tag{22}$$

where d_γ is the gamma cross entropy

$$d_\gamma(g, h) = -\frac{1}{\gamma} \log \frac{\int g(x)h(x)^\gamma dx}{\left\{\int h(x)^{1+\gamma} dx\right\}^{\frac{\gamma}{1+\gamma}}}.$$

The following relation holds:

$$F_\beta(q) = (1 + \gamma)d_\gamma(\hat{p}, r^{\frac{1}{1+\gamma}})$$

where $\beta = -\frac{\gamma}{1+\gamma}$ and $\hat{p}(x)$ is the empirical distribution.

The beta divergence in Murata et al. (2004) and Eguchi and Kato (2010) is a generalization of the Kullback-Leibler divergence, which consists of a cross entropy term as above d_γ, and is identical with the power divergence in Basu et al. (1998).

The expression (8) of F_β can be viewed as a weighted version of the log-likelihood function. When $\beta < 0$, Eq. (9) provides a downweighting for outlying observations. This downweighting is equivalent to what is referred to in Basu et al. (1998) as a relative-to-the-model downweighting. This implies that the robustness of the estimation, the main feature of minimization of these divergences, carries over to F_β minimization for $\beta < 0$. We observed that this can alleviate overfitting in Sect. 4.1 where the generalization error is minimized with a slightly negative value of β. It is an interesting direction to explore a class of robustness-inducing objective functionals of $q(\theta)$.

7 Conclusion

In this article, a nonparametric estimation method of mixing distributions is discussed. We have proposed an objective functional for the learning of mixing distributions of mixture models which unifies the MLE and the KVQ with the parameter β. By extending Lindsay's result, we proved that the optimal mixing distribution is a discrete distribution with distinct support points no greater in number than the sample size, and we provided a simple algorithm to calculate it. It has been demonstrated through numerical experiments and analyzed theoretically that the estimated distribution is less prone to overfitting for some range of β. We have further discussed the nature of the objective functional in relation to the RD theory. Finally, we have shown certain open problems. We believe these results open a new direction for further research.

References

Amari, S., Fujita, N., & Shinomoto, S. (1992). Four types of learning curves. *Neural Computation*, *4*(4), 605–618.

Arikan, E., & Merhav, N. (1998). Guessing subject to distortion. *IEEE Transactions on Information Theory*, *44*(3), 1041–1056.

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of Machine Learning Research*, *6*, 1705–1749.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge: Cambridge University Press.

Barron, A. R., Roos, T., Watanabe, K. (2014). Bayesian properties of normalized maximum likelihood and its fast computation. In *Proceedings of the 2014 IEEE International Symposium on Information Theory* (pp. 1667–1671).

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*(3), 549–559.

Berger, T. (1971). *Rate distortion theory: A mathematical basis for data compression*. Englewood Cliffs, NJ: Prentice-Hall.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39–B*, 1–38.

Eguchi, S., & Kato, S. (2010). Entropy and divergence associated with power function and the statistical application. *Entropy*, *12*, 262–274.

Eguchi, S., Komori, O., & Kato, S. (2011). Projective power entropy and maximum Tsallis entropy distributions. *Entropy*, *13*, 1746–1764.

Fujisawa, H., & Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, *99*(9), 2053–2081.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer* (Vol. 2, pp. 807–810).

Lashkari, D., Golland, P. (2007). Convex clustering with exemplar-based models. In Advances in neural information processing systems 19.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, *11*(1), 86–94.

Lindsay, B. G. (1995). *Mixture models: Theory geometry and applications*. Hayward, CA: Institute of Mathematical Statistics.

Murata, N., Takenouchi, T., Kanamori, T., & Eguchi, S. (2004). Information geometry of U-boost and Bregman divergence. *Neural Computation*, *16*(7), 1437–1481.

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

Renyi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 547–561). University of California Press, Berkeley.

Rose, K. (1994). A mapping approach to rate-distortion computation and analysis. *IEEE Transactions on Information Theory*, *40*(6), 1939–1952.

Rudloff, B., Sass, J., & Wunderlich, R. (2008). Entropic risk constraints for utility maximization. In C. Tammer & F. Heyde (Eds.), *Festschrift in celebration of Prof. Dr. Wilfried Grecksch's 60th Birthday* (pp. 149–180). Aachen: Shaker.

Schölkopf, B., Mika, S., Burges, C. J. C., Knirsch, P., Müller, K. R., Ratsch, G., et al. (1999). Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, *10*, 1000–1017.

Tipping, M. & Schölkopf, B. (2001). A kernel approach for vector quantization with guaranteed distortion bounds. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* (AISTATS).

Tsallis, C. (2009). *Introduction to nonextensive statistical mechanics*. New York: Springer.

Watanabe, S. (2005). Algebraic geometry of singular learning machines and symmetry of generalization and training errors. *Neurocomputing*, *67*, 198–213.