Notes on Mixture Models

Parametric mixture models are partway between parametric models and nonparametric models. They are useful for classification, image segmentation, clustering and many other tasks. Despite being parametric, they exhibit strange behavior and do not behave like typical parametric models.[1]

1. Introduction

Flip a coin with success probability η. If heads, draw X from density $f(x)$. If tails, draw X from density $g(x)$. Then the density of X is

$$p(x) = \eta f(x) + (1-\eta)g(x) \tag{1}$$

which is called a mixture of f and g. Figure 1 shows a mixture of two Gaussian distributions. Define $Z \sim \text{Bernoulli}(\eta)$ to be the unobservable coin flip. We can also write $p(x)$ as

$$p(x) = \sum_{z=0,1} p(x,z) = \sum_{z=0,1} p(x \mid z)p(z) \tag{2}$$

where $p(x \mid z=0) \equiv f(x)$, $p(x \mid z=1) \equiv g(x)$ and $p(z) \equiv \eta^z(1-\eta)^{1-z}$. Equation (2) is called the hidden variable representation.

More generally, starting with a parametric family $\{p(x;\theta) : \theta \in \Theta\}$, a parametric mixture model takes the form

$$p(x;\psi) = \sum_{j=1}^{k} \eta_j p(x;\theta_j) \tag{3}$$

where $\eta_j \geq 0$, $\sum_{j=1}^{k} \eta_j = 1$ and $\psi = (\eta_1, \ldots, \eta_k, \theta_1, \ldots, \theta_k)$ are the unknown parameters. Generally, even if $\{p(x;\theta) : \theta \in \Theta\}$ is an exponential family model, the mixture need no longer be an exponential family.

Let $\phi(x;\mu_j,\sigma_j^2)$ be the probability density function of Gaussian distribution with mean μ_j and variance σ_j^2. A typical example is the mixture of Gaussians. In one dimension we have

$$p(x;\psi) = \sum_{j=1}^{k} \eta_j \phi(x;\mu_j,\sigma_j^2) \tag{4}$$

which has $3k-1$ unknown parameters, due to the restriction $\sum_{j=1}^{k} \eta_j = 1$. A mixture of d-dimensional multivariate Gaussians is

$$p(x) = \sum_{j=1}^{k} \frac{\eta_j}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right\}. \tag{5}$$

[1] These notes were written Larry Wasserman and John Lafferty.

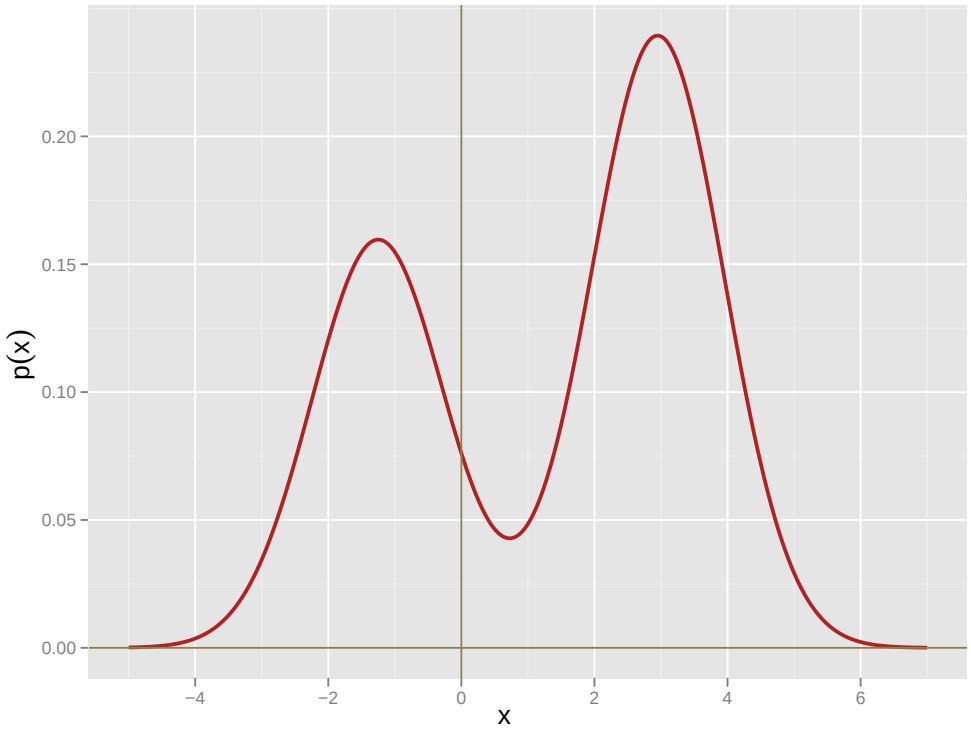


FIG 1. *A mixture of two Gaussians,* $p(x) = \frac{2}{5}\phi(x; -1.25, 1) + \frac{3}{5}\phi(x; 2.95, 1)$.

.

We can write a mixture more generally as

$$p(x; \psi) = \int p(x; \theta) dQ(\theta) \tag{6}$$

where Q is the distribution that puts mass η_j on θ_j. But Q can also be continuous.

Sometimes we can formally write a complicated distribution as a mixture of simpler distributions. For example, let $p(x)$ be the density function for a random variable with a t distribution with ν degrees of freedom,

$$p(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)(1+x^2)^{(\nu+1)/2}}. \tag{7}$$

Some calculations show that

$$p(x) = \int \phi(x; 0, 1/\sqrt{u}) dQ(u) \tag{8}$$

where Q is the distribution for a Gamma ($\nu/2$,$\nu/2$) random variable. Thus, a t distribution is a continuous mixture of Gaussians.

2. Convexity for Mixtures

Certain forms of mixture models result in convex estimation problems. It can be useful to recognize and exploit this convexity. First, suppose that we fix the density parameters θ_j in the parametric

mixture, we have

$$p(x;\eta) = \sum_{j=1}^{k} \eta_j \, p(x;\theta_j) \tag{9}$$

where $\eta = (\eta_1, \ldots, \eta_k)^T$ are the unknown mixing coefficients. The log-likelihood function

$$\ell(\eta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \eta_j \, p(x;\theta_j) \right) \tag{10}$$

is a concave function of η. This can be seen since $\log \left(\sum_{j=1}^{k} \eta_j \, p(x_i;\theta_j) \right)$ is the logarithm of an affine function of η, and $\log(x)$ is concave. Thus, optimizing the mixing weights, holding the component parameters fixed, is a convex optimization problem. For the same reason, if we hold the mixing parameters η fixed, and the model $p(x;\theta_j)$ is a multinomial, then the log-likelihood of θ

$$\ell(\theta) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \eta_j \prod_{\ell} \theta_{j\ell}^{x_{i\ell}} \right) \tag{11}$$

is concave in θ. However, it is not jointly concave in η and θ.

Example 2.1. [A Simple Model for Statistical Machine Translation] Imagine the following generative process for transforming an English sentence $x = (x_1, x_2, \ldots, x_n)$ into a French sentence $y = (y_1, y_2, \ldots, y_m)$:

For each position j in the French sentence.

(a) Select a position $a(j) \in \{1, \ldots, n\}$ in the English.

(b) Translate word $x_{a(j)}$ into y_j, with probability $\theta_{x_{a(j)},y_j} = \mathbb{P}(y_j \,|\, x_{a(j)})$.

This is therefore a mixture of multinomials. Conditioned on an English sentence of length n, the French words are generated independently from a mixture of multinomials, one for each English word, where the mixing weights are uniform. The conditional distribution can be written as

$$\mathbb{P}_\theta \left(y \,|\, x \right) = \prod_{j=1}^{m} \mathbb{P}_\theta \left(y_j \,|\, x \right) = \prod_{j=1}^{m} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_\theta \left(y_j \,|\, x_i \right) \right) = \prod_{j=1}^{m} \left(\frac{1}{n} \sum_{i=1}^{n} \theta_{x_i,y_j} \right). \tag{12}$$

The conditional log-likelihood is then

$$\log \mathbb{P}_\theta \left(y \,|\, x \right) = \sum_{j=1}^{m} \log \left(\frac{1}{n} \sum_{i=1}^{n} \theta_{x_i,y_j} \right). \tag{13}$$

When estimated on a large corpus of aligned English-French sentences, the maximum likelihood estimate for θ gives a reasonable translation dictionary for word-to-word translation between the

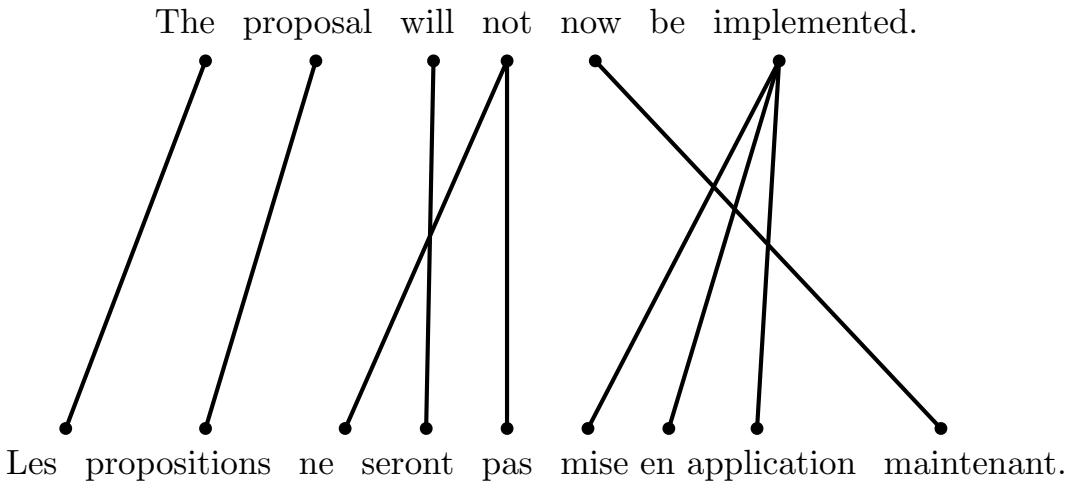


FIG 2. *A sample word alignment from a mixture model defined in terms of word translation probabilities, estimated on a large corpus of French-English sentence pairs.*

languages. Similar mixture models, as simple as they are, have proven to be useful for various problems in image processing.

After fitting the model, the most probable alignment of words in a given sentence pair can be computed. An example of such an alignment is shown in Figure 2.

In a modified version of the model, the mixing weights are not taken to be uniform, but rather are estimated, in a parameterization that depends on the relative word positions in the English and French. In this case the resulting mixture has a log-likelihood function that fails to be concave.

3. Estimation

Let $\psi = (\eta_1, \ldots, \eta_k, \theta_1, \ldots, \theta_k)$. The likelihood of ψ based on the observations $x_1, \ldots, x_n$ sampled from $p(x; \psi)$ is

$$\mathcal{L}(\psi) = \prod_{i=1}^{n} p(x_i; \psi) = \prod_{i=1}^{n} \left(\sum_{j=1}^{k} \eta_j p(x_i; \theta_j) \right) \tag{14}$$

and, as usual, the maximum likelihood estimator is the value $\widehat{\psi}$ that maximizes $\mathcal{L}(\psi)$. (But not quite; see below.) For fixed θ, the log-likelihood is typically a concave function of the mixing parameters η_j, as noted above. However, for fixed $\eta_1, \ldots, \eta_k$, it is not generally concave with respect to $\theta_1, \ldots, \theta_k$.

One way to find $\widehat{\psi}$ is to apply your favorite optimizer directly to the log-likelihood. A convenient algorithm for finding the maximum likelihood estimates of a mixture is the expectation-maximization algorithm or EM algorithm. We will discuss the EM algorithm in detail in a later chapter. Here we only provide a brief summary of the main points of this algorithm for fitting mixture models.

Let $p(x) = \sum_{j=1}^{k} \eta_j p_j(x)$ be a mixture of k densities. When we generate $X \sim p(x)$, it comes from one of the k densities. Let $Z = (Z_1, \ldots, Z_k)$ be a random vector of length k where $Z_j = 1$ if X came from p_j and $Z_j = 0$ otherwise. For example, $Z = (0, 0, 1, 0, 0, 0, 0)$ means that X came from the third component. Then $\mathbb{P}(Z_j = 1) = \eta_j$. We can write

$$p(x) = \sum_{z} p(x|Z = z)\mathbb{P}(Z = z). \tag{15}$$

where $z = (z_1, \ldots, z_k)$. Suppose now that we observed $Z = z$, Then the joint distribution of (X, Z) is

$$p(x, z) = \sum_{j=1}^{k} \eta_j p_j(x)^{z_j}. \tag{16}$$

For a sample of size n, $\{(x_i, z_i)\}_{i=1,\ldots,n}$, where $z_i = (z_{i1}, \ldots, z_{ik})$, the likelihood is

$$\mathcal{L}_c(\psi) = \prod_{i=1}^{n} \prod_{j=1}^{k} (\eta_j p_j(x_i))^{z_{ij}} \tag{17}$$

which is called the complete likelihood.

A Sketch of the EM Algorithm The EM algorithm iterates the following two steps:

1. *E-step*. Let ψ^{old} be the current parameter value. We compute

$$Q(\psi) = \sum_{z_1,\ldots,z_n} p(z_1, \ldots, z_n \,|\, x_1, \ldots, x_n; \psi^{\text{old}}) \ \log p(z_1, \ldots, z_n, x_1, \ldots, x_n; \psi).$$

2. *M-step*. Set ψ^{new} to be the value that maximizes $Q(\psi)$.

The algorithm should be repeated from several different starting values.

Example 3.2. [EM algorithm for a mixture of Gaussians] Let's consider a d-dimensional mixture of Gaussian distribution

$$p(x) = \sum_{j=1}^{k} \eta_j N(x; \mu_j, \Sigma_j),$$

where $N(x; \mu_j, \Sigma_j)$ is a d-dimensional Gaussian density with mean parameter μ_j and covariance parameter Σ_j. The EM steps turn out to be (The detailed derivation is left as an exercise, see Exercise 1):

1. *E-Step*: set

$$\gamma_{ij} = \mathbb{P}(z_{ij} = 1 \,|\, x_1, \ldots, x_n) = \frac{\eta_j N(x_i; \mu_j, \Sigma_j)}{\sum_\ell \eta_\ell N(x_i; \mu_\ell, \Sigma_\ell)}. \tag{18}$$

2. *M-Step*: Update

$$\eta_j \longleftarrow \frac{1}{n}\sum_{i=1}^{n}\gamma_{ij} \tag{19}$$

$$\mu_j \longleftarrow \frac{\sum_{i=1}^{n}\gamma_{ij}x_i}{\sum_{i=1}^{n}\gamma_{ij}} \tag{20}$$

$$\Sigma_j \longleftarrow \frac{\sum_{i=1}^{n}\gamma_{ij}\,(x_i-\mu_j)(x_i-\mu_j)^T}{\sum_{i=1}^{n}\gamma_{ij}}. \tag{21}$$

These are just the weighted sample mean and covariance, where the weight γ_{ij} is the posterior probability that example i was generated from mixture component j. These weights are sometimes called "responsibilities," suggesting that it is the probability that the latent variable j was responsible for example i.

Example 3.3. [Old Faithful Geyser Data.]

We study a version of the eruptions data "Old Faithful" geyser in Yellowstone national park. The geyser was named by the Washburn expedition of 1870. They were impressed by its size and frequency. It is not the biggest or most regular geyser in Yellowstone but it is the biggest regular geyser. Furthermore, it has been erupting in nearly the same fashion throughout the recorded history of Yellowstone. Through the years, it has become one of the most studied geysers in the park.

The data are provided by Azzalini and Bowman (1990), which contain continuous measurements from August 1 to August 15, 1985. There are two variables with 299 observations. The first variable ,"Duration", represents the numeric eruption time in minutes. The second variable, "waiting", represents the waiting time to next eruption. This data is believed to have two modes. We fit a mixture of two Gaussians using EM algorithm. To illustrate the EM step, we purposely choose a bad starting point. The EM algorithm quickly converges in six steps. Figure 3 illustrates the fitted densities for all the six steps. We see that even though the starting density is unimodal, it quickly becomes bimodal. This is consistent with the previous analysis in Azzalini and Bowman (1990).

4. Strangeness

No one doubts that mixture models are useful. But mixture models do not behave like the usual parametric models. Here are some of the bizarre features of mixtures that you should be aware of.

Infinite Likelihood. Let $p(x;\psi)=\sum_{j=1}^{k}\eta_j\phi(x;\mu_j,\sigma_j^2)$ be a mixture of Gaussians. Given a sample of size n, let $\mathcal{L}(\psi)=\prod_{i=1}^{n}p(x_i;\psi)$ be the likelihood function. Then $\sup_\psi \mathcal{L}(\psi)=\infty$. To see this, set $\mu_j=x_1$ for some j. Then $\phi(x_1;\mu_j,\sigma_j^2)=(\sqrt{2\pi}\sigma_j)^{-1}$. Now let $\sigma_j\to 0$. This also shows that $\widehat{\sigma}_1=\cdots=\widehat{\sigma}_k=0$. Fortunately all is not lost. If we define the maximum likelihood estimate to be the mode of $\mathcal{L}(\psi)$ in the interior of the parameter space, we usually get a well-defined estimator.

Multimodal Likelihood. Typically, $\mathcal{L}(\psi)$ is multimodal so there are many local maxima. This makes finding $\widehat{\psi}$ difficult. Usually, we choose many random starting points and apply our optimizer

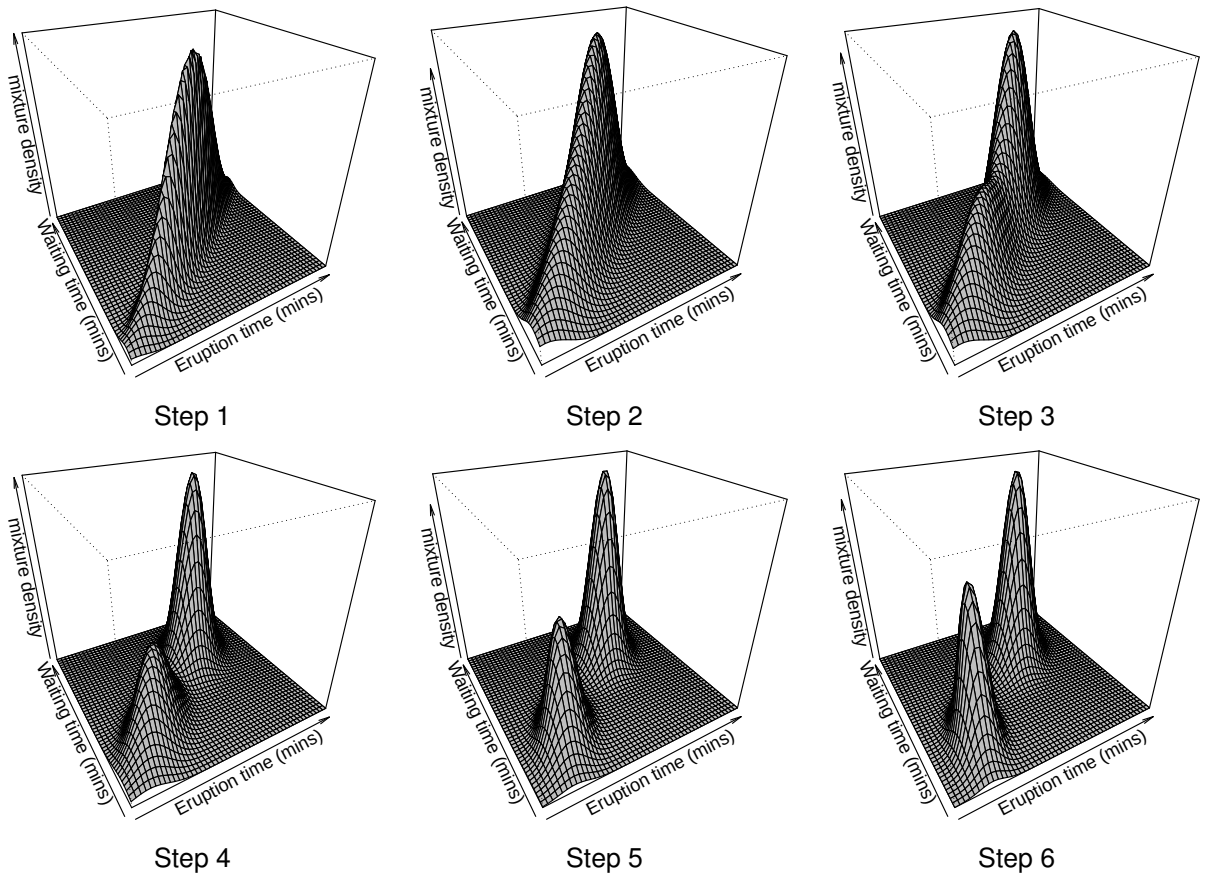


FIG 3. *Fitting a mixture of two Gaussians on the Mexican Stamp Data. The starting values are* $\eta_1 = \eta_2 = 0.5$. $\mu_1 = (4, 70)^T$, $\mu_2 = (3, 60)^T$, $\Sigma_1 = \Sigma_2 = \left(\begin{smallmatrix} 0.8 & 7 \\ 7 & 70 \end{smallmatrix}\right)$. *We see that even though the starting density is bimodal, the EM algorithm quickly converges to a bimodal density.*

many times.

Multimodality of the Density. Consider the mixture of two Gaussians

$$p(x) = (1 - \eta)\phi(x; \mu_1, \sigma^2) + \eta\phi(x; \mu_2, \sigma^2). \tag{22}$$

You would expect $p(x)$ to be multimodal but this is not necessarily true. The density $p(x)$ is unimodal when $|\mu_1 - \mu_2| \leq 2\sigma$ and bimodal when $|\mu_1 - \mu_2| > 2\sigma$. One might expect that the maximum number of modes of a mixture of k Gaussians would k. However, there are examples where a mixture of k Gaussians has more than k modes (Carreira-Perpiñán and Williams, 2003). Figure 4 provides a simple example from David Mackay and Chris Williams.

Nonidentifability. A model $\{p(x; \theta) : \theta \in \Theta\}$ is identifiable if

$$\theta_1 \neq \theta_2 \quad \text{implies} \quad p(x; \theta_1) \neq p(x; \theta_2). \tag{23}$$

Mixture models are nonidentifiable in two different ways. First, there is nonidentifiability due to permutation of labels. For example

$$p_1(x) = 0.3\phi(x; 0, 1) + 0.7\phi(x; 2, 1) \tag{24}$$

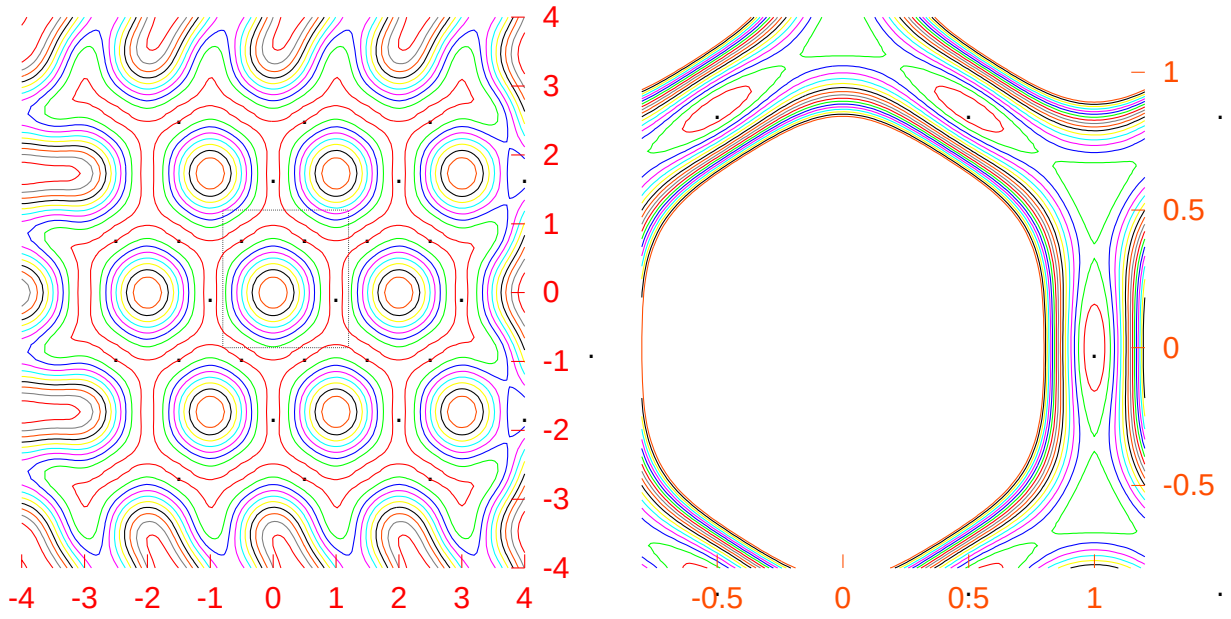


FIG 4. *An illustration example that shows mixture of k Gaussians models may asymptotically have $\frac{5}{3}k$ modes. Left plot: each black dot corresponds to the center of one Gaussian. Right plot: for adjacent Gaussians, new maxima appear in the centers of the triangles, so there are roughly $\frac{5}{3}k$ maxima (neglecting boundary effects). These two pictures are provided by David Mackay and Chris Williams.*

and

$$p_2(x) = 0.7\phi(x; 2, 1) + 0.3\phi(x; 0, 1) \tag{25}$$

then $p_1(x) = p_2(x)$ even though $\psi_1 = (0.3, 0.7, 0, 2, 1) \neq (0.7, 0.3, 2, 0, 1) = \psi_2$. This is not a serious problem although it does contribute to the multimodality of the likelihood.

A more serious problem is local nonidentifiability. Suppose that

$$p(x; \eta, \mu_1, \mu_2) = (1 - \eta)\phi(x; \mu_1, 1) + \eta\phi(x; \mu_2, 1). \tag{26}$$

When $\mu_1 = \mu_2 = \mu$, we see that $p(x; \eta, \mu_1, \mu_2) = \phi(x; \mu)$. The parameter η has disappeared. Similarly, when $\eta = 1$, the parameter μ_2 disappears. This means that there are subspaces of the parameter space where the family is not identifiable. This local nonidentifiability causes headaches (and generates papers) for theoreticians.

For the model (26), to this day, there is no simple theory to describe the distribution of the likelihood ratio test for $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$. The best available theory is very complicated. We will return to this point later.

Irregularity. Mixture models are irregular, that is, they do not satisfy the usual regularity conditions that make parametric models so easy to deal with. Consider the following simple example; see Chen (1995) for more details.

Consider a one-dimensional mixture of two Gaussians distribution:

$$p(x; \theta) = \frac{2}{3}\phi(x; -\theta, 1) + \frac{1}{3}\phi(x; 2\theta, 1). \tag{27}$$

Then $I(0) = 0$ where $I(\theta)$ is the Fisher information. Moreover, no estimator of θ can converge faster than $n^{-1/4}$ if the number of components is not known in advance. Compare this to a Normal family $\phi(x; \theta, 1)$ where the Fisher information is $I(\theta) = n$ and the maximum likelihood estimator converges at rate $n^{-1/2}$.

Nonintinuitive Group Membership. Mixtures are often used as a parametric method for finding clusters. For two clusters in one dimension a common model is

$$p(x) = (1 - \eta)\phi(x; \mu_1, \sigma_1^2) + \eta\phi(x; \mu_2, \sigma_2^2). \tag{28}$$

Suppose that $\mu_1 < \mu_2$. We can classify an observation as being from cluster 1 or cluster 2 by computing the probability of being from the first or second component, denoted $Z = 0$ and $Z = 1$. We get

$$\mathbb{P}(Z = 0|X = x) = \frac{(1 - \eta)\phi(x; \mu_1, \sigma_1^2)}{(1 - \eta)\phi(x; \mu_1, \sigma_1^2) + \eta\phi(x; \mu_2, \sigma_2^2)}. \tag{29}$$

Define $Z(x) = 0$ if $\mathbb{P}(Z = 0|X = x) > 1/2$ and $Z(x) = 1$ otherwise. When σ_1 is much larger than σ_2, Figure 5 shows $Z(x)$. We end up classifying all the observations with large x_i to the leftmost component. Technically this is correct, yet it seems to be an unintended consequence of the model and does not capture what we mean by a cluster.

Improper Posteriors. Suppose that $X \sim N(\mu, 1)$ and $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ be a sample of size n. In a Bayesian analysis, it is common to use an improper prior $\pi(\mu) = 1$. This is improper because

$$\int \pi(\mu) d\mu = \infty.$$

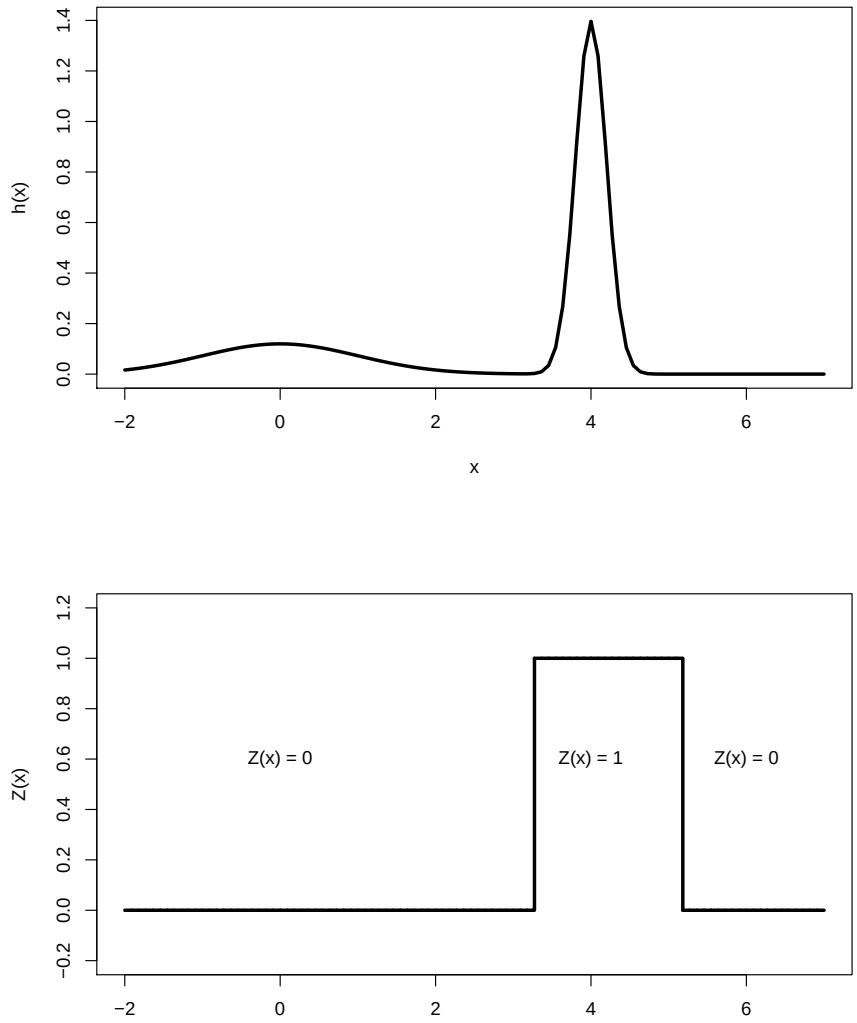


FIG 5. *Mixtures are used as a parametric method for finding clusters. Observations with x near 4 are classified into the second component. Otherwise observations are classified into the first (leftmost) component.*

Nevertheless, the posterior $p(\mu \,|\, \mathcal{D}_n) \propto \mathcal{L}(\mu)\pi(\mu)$ is a proper distribution, where $\mathcal{L}(\mu)$ is the data likelihood of μ. In this case, the posterior for μ is $N(\overline{x}, 1/\sqrt{n})$ where $\overline{x}$ is the sample mean. The posterior inferences coincide exactly with the frequentist inferences. In most parametric models, the posterior inferences are well defined even if the prior is improper and usually they approximate the frequentist inferences. Not so with mixtures. Let

$$p(x; \mu) = \frac{1}{2}\phi(x; 0, 1) + \frac{1}{2}\phi(x; \mu, 1). \tag{30}$$

If $\pi(\mu)$ is improper then so is the posterior. Moreover, Wasserman (2000) shows that the only priors that yield posteriors in close agreement to frequentist methods are data-dependent priors.

5. Estimating the Number of Components

Having been sufficiently warned about the dangers of using mixtures, let's throw caution to the wind and use them. Assuming the maximum likelihood estimates are well-defined and we can compute them, the major practical hurdle is choosing the number of components k. Sometimes, background knowledge suggests k; without such background information, we must estimate k from the data. Here we enter the slippery domain of parametric versus nonparametric. As long as we have some fixed upper bound K on k, we shall regard this as a parametric problem.

We can't use maximum likelihood to estimate k; this will lead to choosing $\widehat{k} = K$ (see Exercise 2 for more details). Here are five different approaches. To be concrete, we shall assume the model is

$$p(x) = \sum_{j=1}^{k} \eta_j \phi(x; \mu_j, \sigma_j^2) \tag{31}$$

where ϕ denotes, as usual, the Gaussian density.

Testing. Let H_ℓ be the hypothesis that $k = \ell$. In the testing approach we proceed as follows.

1. Set $\ell = 1$.
2. Test the hypothesis that $k = \ell$. If we accept, then stop and report $\widehat{k} = \ell$. Otherwise, set $\ell = \ell + 1$ and repeat (but stop when $\ell = K$.)

To implement this idea, we need to be able to test:

$$H : \text{the distribution is a mixture of } \ell \text{ components.} \tag{32}$$

Consider testing a mixture of two Gaussians versus a single Gaussian. The usual parametric test is the likelihood ratio test:

$$T = \frac{\text{maximum of likelihood over big model}}{\text{maximum of likelihood over small model}}. \tag{33}$$

For many parametric models, under the null hypothesis that the smaller model is true, $W = 2 \log T$ has, asymptotically, a χ^2_v distribution where v is the dimension of the larger model minus dimension

of the smaller model. We then reject the smaller model if $W > \chi^2_{p,\alpha}$ to obtain a level $(1-\alpha)$ test. By now you will not be surprised that for mixtures, W is not asymptotically χ^2_v. The distribution of W is, for all practical purposes, unknown.

Usually, one estimates the distribution of W using the parametric bootstrap. (We will discuss the bootstrap in greater generality later.) The steps are:

Hypothesis Test based on Parametric Bootstrap Let $x_1, \ldots, x_n$ be n data points.

1. Fit the simpler model $p_{\text{small}}(x; \widehat{\psi})$ and the bigger model $p_{\text{big}}(x; \widehat{\theta})$ and compute

$$W = 2\left(\sum_{i=1}^{n} \log(p_{\text{big}}(x_i; \widehat{\theta})/p_{\text{small}}(x_i; \widehat{\psi}))\right). \tag{34}$$

2. Sample n data points $x_1^*, \ldots, x_n^*$ from the parametric density $p_{\text{small}}(x; \widehat{\psi})$.
3. Using (34), find $\widehat{\psi}^*$ and $\widehat{\theta}^*$ from the simulated data and compute W^*.
4. Repeat the last step N times to get $W_1^*, \ldots, W_N^*$.
5. To obtain a level $1-\alpha$ test, we reject the null hypothesis if

$$\frac{1}{N}\sum_{j=1}^{N} I(W_j^* > W) < \alpha. \tag{35}$$

What makes the bootstrap work is that the empirical distribution

$$\widehat{F}(t) = \frac{1}{N}\sum_{j=1}^{N} I(W_j^* \leq t)$$

approximates the true null distribution of W. However, it appears this has not been proved for mixtures.

AIC. The AIC (Akaike Information Criterion) method chooses $\widehat{k}$ to maximize

$$\text{AIC}(k) = \log \mathcal{L}(\widehat{\psi}_k) - d(k), \tag{36}$$

where $\mathcal{L}(\widehat{\psi}_k)$ is the log-likelihood of the maximum likelihood estimates for the mixture model with k components, and $d(k)$ is the number of parameters in the k-component mixture model. Some texts instead define $\text{AIC}(k) = -2\log \mathcal{L}(\widehat{\psi}_k) + 2d(k)$ in which case we minimize instead of maximize.

Here is the motivation for AIC. Recall that the Kullback-Leibler distance between two densities $f(x)$ and $g(x)$ is

$$D(f\|g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx. \tag{37}$$

Let $\widehat{p}_1, \ldots, \widehat{p}_K$ be estimates of p from K different models. Thus $\widehat{p}_k(x) = p(x; \widehat{\psi}_k)$. We want to choose $\widehat{k}$ to minimize $D(p\|\widehat{p}_k)$ where $p(x)$ is the true density. But

$$D(p\|\widehat{p}_k) = \int p(x) \log p(x) dx - \int p(x) \log \widehat{p}_k(x) dx. \tag{38}$$

So minimizing $D(p\|\widehat{p}_k)$ is the same as maximizing

$$J(k) = \int p(x) \log \widehat{p}_k(x) dx. \tag{39}$$

We don't know $p(x)$ and it is tempting to estimate $J(k)$ with $n^{-1}\sum_{i=1}^n \log \widehat{p}_k(x_i)$. But this is a biased estimate of $J(k)$ for the same reason that the training error is a biased estimate of prediction error. In regular parametric models the bias is approximately $d(k)/n$, which leads to

$$\widehat{J}(k) = \frac{1}{n}\sum_{i=1}^n \log \widehat{p}_k(x_i) - \frac{d(k)}{n} = \frac{1}{n}\left(\log \mathcal{L}(\widehat{\psi}_k) - d(k)\right), \tag{40}$$

which is $\mathrm{AIC}(k)/n$. It is not known if the bias adjustment is correct for mixtures.

BIC. The BIC (Bayesian Information Criterion) method chooses $\widehat{k}$ to maximize

$$\mathrm{BIC}(k) = \log \mathcal{L}(\widehat{\psi}_k) - \frac{d(k)}{2}\log n \tag{41}$$

where $d(k)$ is the number of parameters in the model. Outside of mixture models, the justification for BIC is that it approximates the log-posterior of k from a Bayesian analysis, it consistently chooses the true k if one exists and it has a coding interpretation. Unfortunately, similar theory in the mixture case is lacking.

Data Splitting. Let $p(x)$ be the true density and $\widehat{p}_k(x)$ be the estimated density mixture of a mixture model with k components. We try to choose k to minimize

$$\int (p(x) - \widehat{p}_k(x))^2 dx \tag{42}$$

which is equivalent to minimizing

$$J(k) = \int \widehat{p}_k(x) dx - 2\int \widehat{p}_k(x) p(x) dx. \tag{43}$$

Randomly split the data $\mathcal{D} = \{x_1, \ldots, x_n\}$ in half. Let $\mathcal{D}_1 = \{x_1, \ldots, x_{\lfloor n/2 \rfloor}\}$ and $\mathcal{D}_2 = \{x_{\lfloor n/2 \rfloor + 1}, \ldots, x_n\}$ denote the two halves of the data. Construct $\widehat{p}_1, \ldots, \widehat{p}_K$ using the first half of the data. Estimate $J(k)$ with

$$\widehat{J}(k) = \int \widehat{p}_k(x) dx - \frac{2}{n}\sum_{i=\lfloor n/2 \rfloor + 1}^{n} \widehat{p}_k(x_i) \tag{44}$$

where the sum is over the second half of the data. Choose $\widehat{k}$ to minimize $\widehat{J}(k)$.

Clearly, $\mathbb{E}(\widehat{J}(k)|\mathcal{D}_1) = J(k)$. If each $\widehat{p}_k$ is bounded, it follows from Hoeffding's inequality that, for any $\epsilon > 0$,

$$\mathbb{P}(|\widehat{J}(k) - J(k)| > \epsilon) \leq c_1 \exp\left(-n c_2 \epsilon^2\right) \tag{45}$$

for some $c_1, c_2 > 0$. Set $\epsilon = \sqrt{\log n / n}$ and conclude that

$$\mathbb{P}(\max_k |\widehat{J}(k) - J(k)| > \sqrt{\log n / n}) \leq c_1 K n^{-c_2} \to 0. \tag{46}$$

We then have the following oracle inequality. Let $\widehat{k}$ minimize $\widehat{J}(k)$ and let k_* minimize $J(k)$. Then, except on a set of probability tending to 0,

$$J(\widehat{k}) \;\leq\; \widehat{J}(\widehat{k}) + \sqrt{\frac{\log n}{n}} \leq \widehat{J}(k_*) + \sqrt{\frac{\log n}{n}} \leq J(k_*) + 2\sqrt{\frac{\log n}{n}}.$$

That is,

$$\mathbb{P}\left(J(\widehat{k}) > J(k_*) + 2\sqrt{\frac{\log n}{n}} \right) \to 0. \tag{47}$$

Bayes. From a Bayesian perspective, we use a random variable K to represent the number of components in a mixture model. Put a prior $\gamma_k = \mathbb{P}(K = k)$ on k and a prior π_k on ψ_k. Let $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ be a sample of size n. By Bayes' theorem, the posterior on K is

$$\mathbb{P}(K = k \,|\, \mathcal{D}_n) = \frac{\gamma_k m_k}{\sum_j \gamma_j m_j} \tag{48}$$

where

$$m_j = \int \mathcal{L}(\psi_k)\pi_k(\psi_k)d\psi_k. \tag{49}$$

A crude approximation to the log-posterior is the BIC. To compute $\mathbb{P}(K = k \,|\, \mathcal{D}_n)$ we face two challenges. First, we must specify each $\pi_k(\psi_k)$. Second, we have to compute the integral in the definition of m_k. Later, we discuss Markov chain Monte Carlo methods for approximating the integral.

6. Examples

Example 6.4. [Mexican Stamps]

This is a famous example involving the thickness of 485 stamps. (source xxxx) Based on historical records, it is believed that there may be seven printings of this stamp. Figure 6 shows the data and two estimates of the density. The estimate in the top plot uses a nonparametric estimator, the kernel estimator, defined later. Note that this estimator has 7 peaks, in remarkable agreement with expectations. The estimate in the bottom plot uses a mixture of three normals. It has been reported that three normals is a reasonable fit although there is room for debate. AIC, BIC and cross-validation all seem to favor $k = 2$.

Example 6.5. [Latent Dirichlet Allocation]

The Latent Dirichlet Allocation (LDA) of Blei et al. (2003) is an example of a complex mixture model that is useful in modeling text. It is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

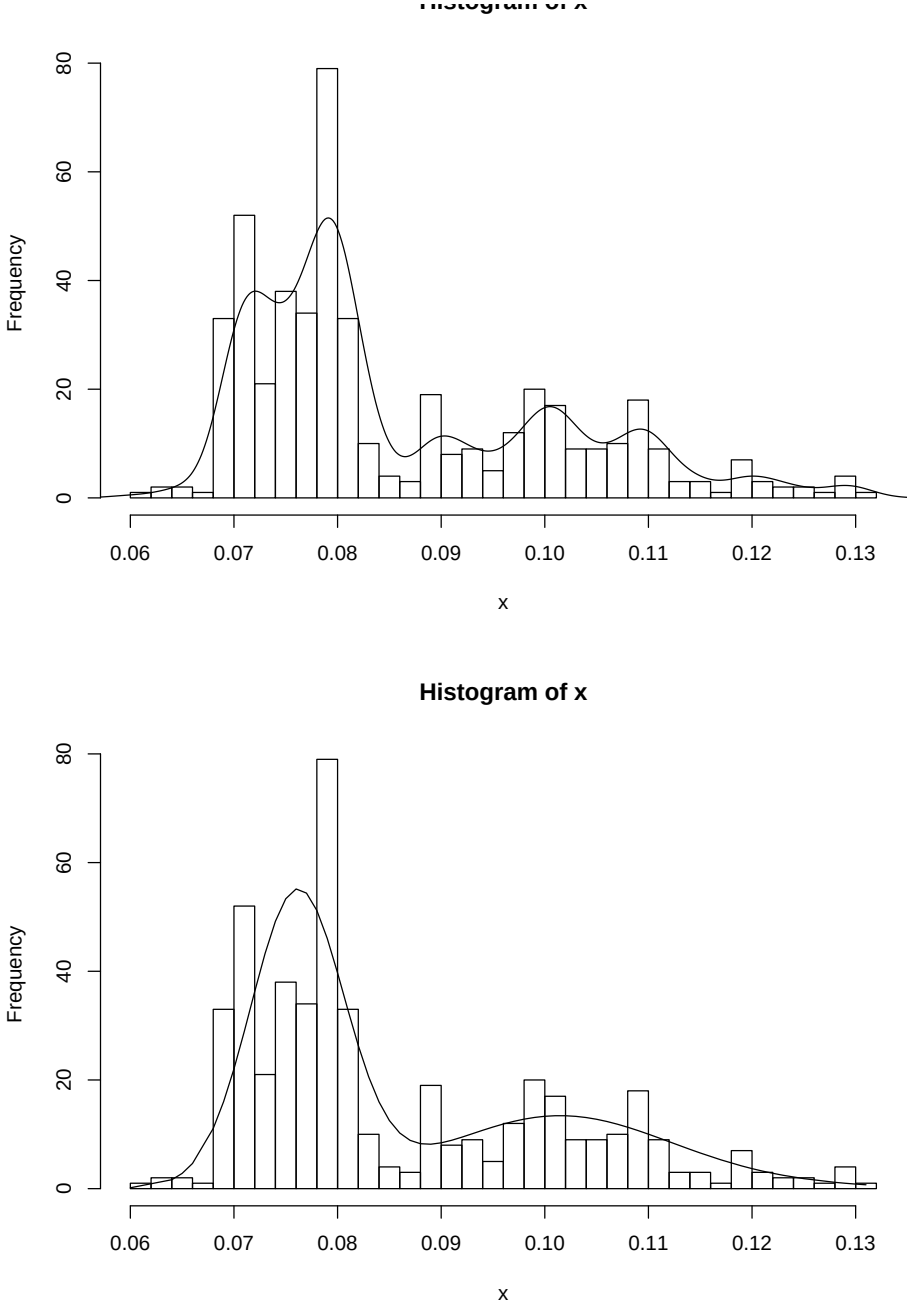


FIG 6. *Estimating the density of the Mexican Stamp Data. The estimate in the top plot uses a nonparametric kernel density estimator. Note that this estimator has 7 peaks, in remarkable agreement with expectations. The estimate in the bottom plot uses a mixture of three Gaussians.*

We have a corpus of documents $D_1, \ldots, D_n$. Each document has the form $D_i = \{w_{i1}, \ldots, w_{im_i}\}$ where each w_{ij} is a word taking values in a vocabulary $\{1, \ldots, V\}$. The w_{ij} are the only observable variables, and the other variables are latent variables.

In LDA, each document is viewed as a mixture of various topics. Let W be a random variable representing a word with observed value w. The distribution of words depends on a latent variable Z representing the topic. Thus,

$$p(w) = \sum_{z=1}^{k} p(w \mid z)p(z) \tag{50}$$

is a mixture over topics. Given $Z = z$, $W \mid Z = z \sim \text{Multinomial}(\beta_z)$. The topics are also multinomial, $Z \sim \text{Multinomial}(\theta)$. The parameters θ of the topic distribution are given by a Dirichlet distribution

$$p(\theta) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}. \tag{51}$$

The distribution of the observable W can be written

$$p(w) = \int \sum_{z=1}^{k} p(w \mid z)p(z \mid \theta)p(\theta)d\theta \tag{52}$$

which is a mixture of mixtures.

This model can be used to classify documents by using a different LDA model for each class. It can also be used to find clusters of words representing topics. We will show examples later when we explain how to estimate the parameters in this mixture model.

Example 6.6. [Classification] Figure 7 shows 200 data points $(x_1, y_1), \ldots, (x_n, y_n)$ from two classes. The class label $y_i 1\{0, 1\}$. The class densities $p_0(x)$ and $p_1(x)$ were estimated using mixtures of bivariate Gaussians. The estimated densities are shown in the top right and bottom left. The resulting classifier is $h(x) = 1$ if $\widehat{p}_1(x)/\widehat{p}_0(x) > (1-\widehat{\pi})/\widehat{\pi}$ and $h(x) = 0$ otherwise where $\widehat{\pi} = n^{-1} \sum_{i=1}^{n} y_i$. The set $\{h(x) = 1\}$ is shown in the bottom right plot.

7. Greedy Mixtures

Greedy methods can be attractive for fitting mixtures for high dimensional problems. Given $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, the following algorithm is due to Li and Barron (1999). Let $\{p(x; \theta) : \theta \in \Theta\}$ be a parametric model.

A similar approach is proposed by Meek et al. (2002).

We have the following theoretical result due to Rakhlin et al. (2005). Let

$$D(f \| g) = \int f(x) \log \left(\frac{f(x)}{g(x)} \right) dx \tag{54}$$

denote the Kullback-Leibler distance between two densities $f(x)$ and $g(x)$.

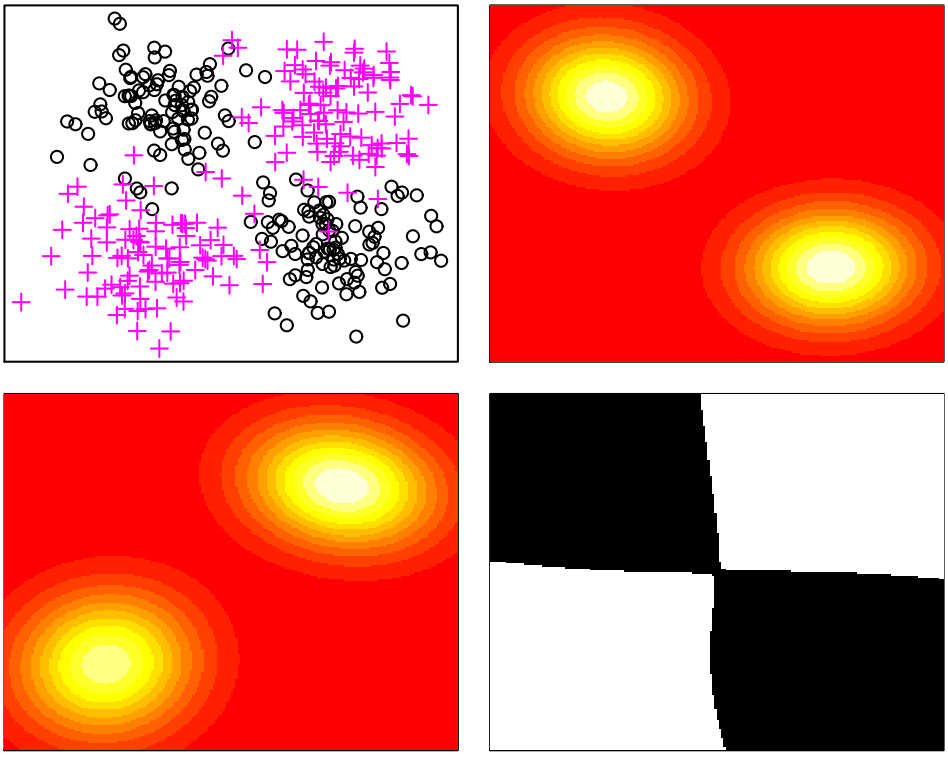


FIG 7. *Top left: scatter plot of the data from two classes. Top right: estimate of $p_0(x)$ using a mixture of two Gaussians. Bottom left: estimate of p_1 using a mixture of two Gaussians. Bottom right: classification rule $\widehat{p}_1(x)/\widehat{p}_0(x) > (1-\widehat{\pi})/\widehat{\pi}$.*

Greedy Methods for Fitting High Dimensional Mixture Models

1. Set $k = 1$. Find the maximum likelihood estimator $\widehat{\theta}$. Set $\widehat{p}_1(x) = p(x;\widehat{\theta})$.
2. Find $(\widehat{\theta}, \widehat{\eta})$ to maximize
$$\sum_{i=1}^{n} \log\Big[(1-\eta)\widehat{p}_k(x_i) + \eta p(x_i;\theta)\Big]. \tag{53}$$
3. Set $\widehat{p}_{k+1}(x) = (1-\widehat{\eta})\widehat{p}_k(x) + \widehat{\eta}p(x;\widehat{\theta})$ and $k \leftarrow k+1$.
4. Repeat.

Theorem 7.7. *Let $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ be a sample of size n and each $x_i \in \mathcal{R}^d$. We denote*

$$\mathcal{M} = \left\{ p : \ p(x) = \int_\Theta p(x;\theta) dQ(\theta), \ \text{ for some } Q \right\}. \tag{55}$$

Suppose that there exist positive constants a, b, A, B such that

$$a \leq \inf_x p(x) \leq \sup_x p(x) \leq b, \tag{56}$$

$$\sup_x |\log p(x;\theta) - \log p(x;\theta') \leq B \sum_{j=1}^d |\theta_j - \theta'_j| \tag{57}$$

and $\Theta \subset [-A, A]^d$.

For any target density $p(x)$, the greedy estimator $\widehat{p}_k(x)$ after k-steps satisfies

$$\mathbb{E}(D(p\|\widehat{p}_k)) - \inf_{g\in\mathcal{M}} D(f\|g) \leq \frac{c_1}{k} + \frac{c_2}{\sqrt{n}} \tag{58}$$

where c_1 and c_2 depend on a, b, A, B

By letting $k \asymp \sqrt{n}$, this result implies that

$$\mathbb{E}(D(p\|\widehat{p}_k)) - \inf_{g\in\mathcal{M}} D(f\|g) = O_P(1/\sqrt{n}). \tag{59}$$

This might be surprising since $\mathcal{M}$ looks nonparametric. It suggests that we can use as many components in the mixture as we want and still achieve a parametric rate.

The result might be misleading. The conditions of the theorem imply that $\mathcal{M}$ is essentially parametric. For example, the conditions rule out mixtures of Gaussians unless we bound the variances to be larger than some constant $h > 0$. The constant k then acts like a nonparametric smoothing parameter. In a sense, these results hide all the difficulties.

8. Bayesian Mixture Models

Assume that the total number of components is a prefixed constant k, we consider the following mixture of k Gaussians from a Bayesian perspective

$$p(x) = \sum_{j=1}^k \eta_j \phi(x; \mu_j, \sigma_j^2). \tag{60}$$

As discussed in the previous chapter, a fully Bayesian procedure will treat all the unknown parameters as random variables and specify a prior for them. Let $x_1, \ldots, x_n$ be the observed data. Let $z_i = (z_{i1}, \ldots, z_{ik})$ be a random vector of length k where $z_{ij} = 1$ if x_i came from the jth component $\phi(x; \mu_j, \sigma_j^2)$.

A popular choice of the prior distributions is

$$\begin{aligned}
\xi &\sim \text{Dirichlet}(\beta) \text{ where } \beta \in \mathbb{R}_+^k && (61)\\
z_1, \ldots, z_n &\sim \text{Multinomial}(\xi) && (62)\\
\sigma_1^2, \ldots, \sigma_k^2 &\sim \text{Inverse-Gamma}(v_0, \sigma_0^2) && (63)\\
p\left(\mu_1, \ldots, \mu_k \,|\, \sigma_1^2, \ldots, \sigma_k^2\right) &= \prod_{j=1}^{k} \phi(\mu_j; \mu_0, \sigma_j^2) && (64)\\
p(x_1, \ldots, x_n \,|\, z_1, \ldots, z_n, \mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_k^2) &= \prod_{i=1}^{n} \Big(\sum_{j=1}^{k} z_{ij} \phi(x_i; \mu_j, \sigma_j^2) \Big) && (65)
\end{aligned}$$

where $\beta, v_0, \mu_0, \sigma_0^2$ are pre-specified hyperparameters.

With this setup, standard Bayesian inference can be conducted to infer the posterior distribution of the parameters $\mu_1, \ldots, \mu_k$, $\sigma_1^2, \ldots, \sigma_k^2$, and $\eta_1, \ldots, \eta_k$ given the data $x_1, \ldots, x_n$.

9. Summary

Mixtures models are very flexible and arise naturally in many domains. However, they are both numerically and theoretically difficult to deal with.

10. Bibliographic Remarks

A thorough text on parametric mixture models is *Finite Mixture Models* by McLachlan and Peel (2000). Detailed derivations of the EM algorithms for mixture of Gaussians and mixture of Bernoulli's are provided in Bishop (2007).

Exercises

1. Derive the E-step and M-step updating rules in Example 3.

2. Let $x_1, \ldots, x_n$ be n data points sampled from a mixture of k Gaussians,

$$p(x; k, \theta) = \sum_{j=1}^{k} \eta_j \phi(x; \mu_j, \sigma_j^2) \text{ where } \sum_{j=1}^{k} \eta_j = 1. \tag{66}$$

   Let $\mathcal{A} = \arg\min_{1 \le k \le K} \prod_{i=1}^{n} p(x_i; k, \theta)$. Show that $K \in \mathcal{A}$.

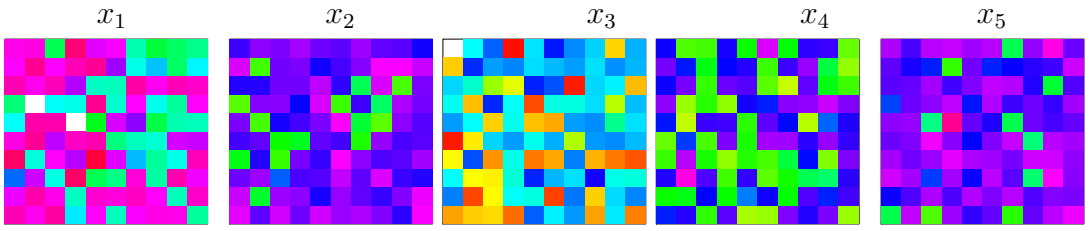
3. Show that a t distribution with ν degrees of freedom can be written as a continuous mixture of Gaussians.

4. Let
$$p(x;\theta) = \frac{2}{3}\phi(x;-\theta,1) + \frac{1}{3}\phi(x;2\theta,1).$$
   (a) Find the Fisher information function $I(\theta)$, and plot it as a function of θ. Comment.
   (b) What is the complete data (X, Z), in the language of the EM algorithm?
   (b) Find the complete data Fisher information $I_{(X,Z)}(\theta)$ and the missing data Fisher information $I_{Z\,|\,X}(\theta)$.
   (c) Let $n = 100$ and simulate n observations $x_1, \ldots, x_n$ from the distribution $p(x; \frac{1}{2})$. Plot the log-likelihood function for θ and find the maximum likelihood estimate $\widehat{\theta}$ by direct optimization (for example, using Newton's method).
   (d) Now estimate the maximum likelihood estimate θ using the EM algorithm with different initial values. Compare the convergence rate to the ratio of Fisher informations $\frac{1}{n}\sum_{i=1}^{n} I_{Z\,|\,x_i}(\theta^*)/I_{(X,Z)}(\theta^*)$ and discuss.

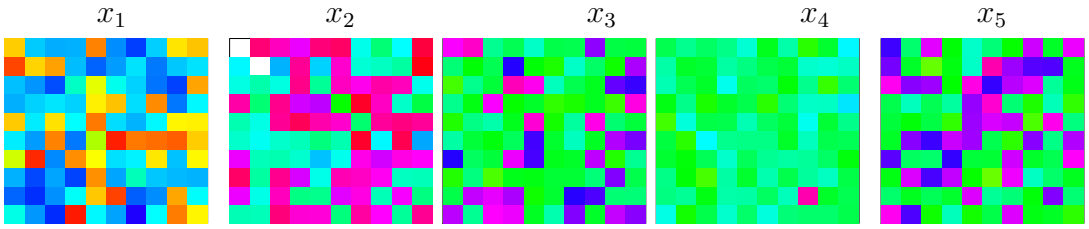
5. Analyze the Mexican stamp data. Use the bootstrap-hypothesis testing approach to choose k.

6. (a) Write program to fit a mixture of multivariate Gaussians using the EM algorithm (with the number of components k known).
   (b) Simulate 1000 observations from the model
$$\frac{1}{5}N(0, I) + \frac{4}{5}N(\mu, I)$$
   where $\mu = (3, 3, 3, 3, 3)$ and I is the 5 by 5 identity matrix. Use your program to find the MLE.
   (c) Now repeat but take k as unknown. Use AIC and BIC to choose k.

7. In this problem you will use mixture models for a classification task. The training data consist of 100 examples (x_i, y_i), where $x_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ and each $x_{ij} \in \mathbb{R}^{100}$ is a 100-dimensional vector, and $y_i = \{c_{ij}\}$ is an unordered set of up to five labels, with each $c_{ij} \in \{1, 2, \ldots, 10\}$. Two examples are shown below:



classes = {1, 5, 6, 9}



classes = {1, 7, 9}

In the first example, there are four classes. Each of the five x_i was generated from exactly one of these four classes; thus, one of the four classes must have generated two of the x_i. Considering the alignment of each of the five images with one of the classes as a latent variable $z = (z_1, z_2, z_3, z_4, z_5)$, a possible alignment is $z = (9, 5, 9, 1, 6)$ where class 9 generates x_1, class 5 generates x_2, class 9 generates x_3, and so on.

The test data consist of 100 unlabeled examples x_i. The task is to label each of the test examples with an *ordered* set of labels $y_i = (c_{i1}, c_{i2}, c_{i3}, c_{i4}, c_{i5})$ where c_{ij} is the predicted class of x_{ij}.

You may use any method and model you choose, but it should make use of mixture models in some way.

8. Let $(z_1, x_1, y_1), \ldots, (z_n, x_n, y_n)$ be observations from the random variables (Z, X, Y) that are generated as follows:

$$
\begin{aligned}
Z &\sim \text{Bernoulli}(\eta) \\
X &\sim \text{Uniform}(0, 1) \\
\epsilon &\sim N(0, \sigma^2) \\
Y &\sim \begin{cases} 5X + \epsilon & \text{if } Z = 0 \\ -5X + \epsilon_i & \text{if } Z = 1. \end{cases}
\end{aligned}
$$

   (a) Assume we do not observe the z_i's or ϵ_i's. Write the distribution $p(x, y)$ of X and Y as a mixture.

   (b) Write down the likelihood function for η.

   (c) Write down the steps for the EM algorithm.

   (d) Find a consistent estimator of η that avoids using EM. Hint: find $\mathbb{E}(Y \mid X = x)$.

9. (a) Write an `R` program to fit a mixture of multivariate Gaussians using the EM algorithm,

estimating both the mean and the covariance of each Gaussian component. (with the number of components k known).

(b) Simulate 1000 observations from the model

$$p(x;\psi) = \frac{1}{5}N(x;0,I) + \frac{4}{5}N(x;\mu,I)$$

where $\mu = (3,3,3,3,3)$ and I is the 5 by 5 identity matrix. Use your program to find the maximum likelihood estimate.

(c) Now repeat but take k as unknown. Use BIC to choose k.

References

Azzalini, A. and Bowman, A. (1990). A look at some data on the old faithful geyser. *Applied Statistics*, 39:357–365.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 2006. corr. 2nd printing edition.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Carreira-Perpiñán, M. and Williams, C. (2003). On the Number of Modes of a Gaussian Mixture. In *Scale Space Methods in Computer Vision*, volume 2695 of *Lecture Notes in Computer Science*, chapter 44, pages 625–640. Springer Berlin / Heidelberg.

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23(1):221 – 233.

Li, J. Q. and Barron, A. R. (1999). Mixture density estimation. In *In Advances in Neural Information Processing Systems 12*, pages 279–285. MIT Press.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics.

Meek, C., Thiesson, B., and Heckerman, D. (2002). Staged mixture modeling and boosting. In *In Proceedings of Eighteenth Conference on Uncertainty in Artificial Intelligence, Edmonton, Alberta*. Morgan Kaufmann.

Rakhlin, A., Panchenko, D., and Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, 9:220–229.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107.