Chapter 2

Finite Mixture Models

A finite mixture model is a convex combination of two or more probability density functions. By combining the properties of the individual probability density functions, mixture models are capable of approximating any arbitrary distribution [145]. Consequently, finite mixture models are a powerful and flexible tool for modeling complex data. Mixture models have been used in many applications in statistical analysis and machine learning such as modeling, clustering, classification and latent class and survival analysis. In this chapter, we will introduce the basics about mixture models. Thereby, we define the statistical and computational framework that will be further explored for specific bioinformatics applications in the subsequent chapters. All the content covered in this chapter is a review of established research in the area and can be found, for example, in the textbooks [93, 142, 145].

First, we describe the basic concepts and notations used through this thesis (Section 2.1). Then, we introduce mixture models formally (Section 2.2), show how a mixture model can be efficiently estimated with the expectation-maximization (EM) algorithm (Section 2.3), give an example of mixture models with multivariate Gaussians (Section 2.3.3) and discuss some aspects of model selection and determination of the number of components (Section 2.3.5).

2.1 Basics

A continuous L-dimensional random variable will be denoted as $X = (X_1, ..., X_l, ..., X_L)$, where X_l corresponds to the lth variable. Lower case letters will be used for a particular observation (or realization) $x = (x_1, ..., x_l, ..., x_L)$ of a variable X. Bold face letters, such as $\mathbf{X}$, will denote a data of N observations of variable X or, equivalently, a $N \times L$ matrix, where x_{il} is the value of the ith observation for the lth variable in $\mathbf{X}$. This notation is based on the one introduced in the textbook [93].

A probability density function (pdf) $p(x)$ is any function defining the probability density of a variable X such that $p(x) \geq 0$ and $\int_{-\infty}^{\infty} p(x) = 1$. By integrating $p(x)$ over an interval,

we obtain the probability that variable X assumes values in the interval $[a, b]$, that is

$$\mathbf{P}[a \leq X_i \leq b] = \int_a^b p(x)\, dx.$$

For a given pdf $p(x)$, the expectation of X is defined as,

$$E[X] = \int_{-\infty}^{\infty} xp(x)\, dx. \tag{2.1}$$

In relation to the model parameters, we use the "hat" symbol to indicate an estimator. For example $\hat{\theta}$ is the estimator of parameter θ.

2.2 Mixture Models

Let $X = (X_1, ..., X_j, ..., X_L)$ be a L-dimensional continuous random variable and $x = (x_1, ..., x_L)$ be an observation of X. A probability density function (pdf) of a mixture model is defined by a convex combination of K component pdfs [145],

$$p(x|\Theta) = \sum_{k=1}^{K} \alpha_k p_k(x|\theta_k), \tag{2.2}$$

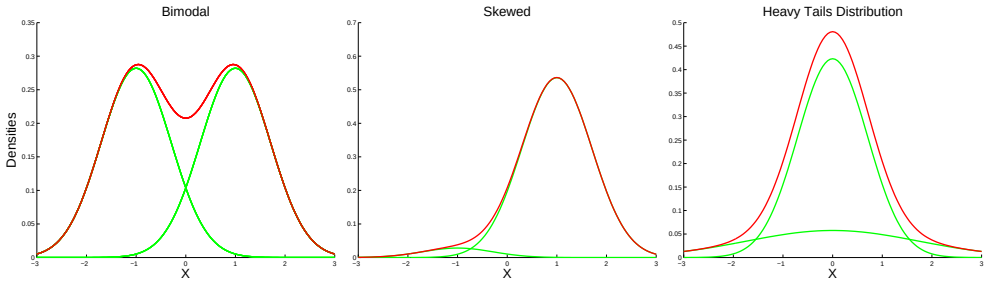
where $p_k(x|\theta_k)$ is the pdf of the kth component, α_k are the mixing proportions (or component priors) and $\Theta = (\alpha_1, ..., \alpha_K, \theta_1, ..., \theta_K)$ is the set of parameters. We assume that

$$\alpha_k \geq 0, \text{ for } k \in \{1, ..., K\}, \text{ and} \tag{2.3}$$

$$\sum_{k=1}^{K} \alpha_k = 1. \tag{2.4}$$

By the property of convexity, given that each $p_k(x|\theta_k)$ defines a probability density function, $p(x|\Theta)$ will also be a probability density function.

The most straightforward interpretation of mixture models is that the random variable X is generated from K distinct random processes. Each of these processes is modeled by the density $p_k(x|\theta_k)$, and α_k represents the proportion of observations from this particular process. For example, the mixture in Figure 2.1 (a) models a bimodal density generated by two independent processes. A mixture can also, by combining simpler densities, model pdfs of arbitrary shapes. For example, with two Gaussian densities as components, we can model a skewed density Figure 2.1 (b), or a heavy tail density Figure 2.1 (c).



(a) (b) (c)

Figure 2.1: *Examples of densities modeled by mixtures of two Gaussians pdfs. Green lines indicate the individual component densities and red lines the mixture densities. In Figure (a), we have a highly overlapping bimodal density, while in Figure (b), we depict an unimodal density skewed to the left, while in Figure (c) a density with heavy tails. These are only a few examples representing the power of mixture models in modeling densities of arbitrary shapes.*

2.3 Mixture Model Estimation

For a given data $\mathbf{X}$ with N observations, the likelihood of the data assuming that x_i are independently distributed is given by

$$p(\mathbf{X}|\Theta) = \mathcal{L}(\Theta|\mathbf{X}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \alpha_k \cdot p_k(x_i|\theta_k). \tag{2.5}$$

The problem of mixture estimation from data $\mathbf{X}$ can be formulated as to find the set of parameters Θ that gives the maximum likelihood estimate (MLE) solution

$$\Theta^* = \arg\max_{\Theta} \mathcal{L}(\Theta|\mathbf{X}). \tag{2.6}$$

The summation inside the product in Eq. 2.5 prevents the possibility of analytical solutions. One alternative is to maximize the complete likelihood in an expectation-maximization (EM) approach [61].

2.3.1 Expectation-maximization Algorithm

The expectation-maximization (EM) algorithm is a general method for finding maximum likelihood estimates when there are missing values or latent variables [61]. In the mixture model context, the missing data is represented by a set of observations $\mathbf{Y}$ of a discrete random variable Y, where $y_i \in \{1, ..., K\}$ indicates which mixture component generated the observation x_i. For now, we will assume that the number K is fixed and known a priori.

The likelihood of the complete data $(\mathbf{X}, \mathbf{Y})$ takes the following multinomial form

$$\begin{aligned} p(\mathbf{X}, \mathbf{Y}|\Theta) = \mathcal{L}(\Theta|\mathbf{X}, \mathbf{Y}) &= p(\mathbf{X}|\mathbf{Y}, \Theta)p(\mathbf{Y}|\Theta) \\ &= \prod_{k=1}^{K}\prod_{i=1}^{N}(\alpha_k \cdot p_k(x_i|\theta_k))^{\mathbf{1}(y_i=k)} \end{aligned} \tag{2.7}$$

where $\mathbf{1}$ is the indicator function, i.e. $\mathbf{1}(y_i = k) = 1$ if $y_i = k$ holds, and $\mathbf{1}(y_i = k) = 0$ otherwise.

The EM algorithm is derived as follows. Let Q be an auxiliary function, the conditional expectation of the complete data $(\mathbf{X}, \mathbf{Y})$, given the observed data $\mathbf{X}$ and a parameterization Θ^{p-1},

$$\begin{aligned} Q(\Theta, \Theta^{p-1}) &= E[\log(p(\mathbf{X}, \mathbf{Y}|\Theta))|\mathbf{X}, \Theta^{p-1}) \\ &= \sum_{\mathbf{Y}\in\mathcal{Y}} p(\mathbf{Y}|\mathbf{X}, \Theta^{p-1}) \log(p(\mathbf{X}, \mathbf{Y}|\Theta)), \end{aligned} \tag{2.8}$$

where $\mathcal{Y}$ is the space of all possible values of $\mathbf{Y}$ and $p(\mathbf{Y}|\mathbf{X}, \Theta^{p-1}) = \prod_{i=1}^{N} p(y_i|x_i, \Theta^{p-1})$. As $\mathcal{Y}$ is the space of all possible values of $\mathbf{Y}$, it follows that

$$\sum_{\mathbf{Y}\in\mathcal{Y}} p(\mathbf{Y}|\mathbf{X}, \Theta^{p-1}) = 1. \tag{2.9}$$

By Bayes rule we can re-write the likelihood function (Eq. 2.5) as

$$p(\mathbf{X}|\Theta) = \frac{p(\mathbf{X}, \mathbf{Y}|\Theta)}{p(\mathbf{Y}|\mathbf{X}, \Theta)}. \tag{2.10}$$

Then, applying the logarithm function to Eq. 2.10 and by Eq.2.9, it follows that

$$\log p(\mathbf{X}|\Theta) = \sum_{\mathbf{Y}\in\mathcal{Y}} p(\mathbf{Y}|\mathbf{X}, \Theta^{p-1}) \log p(\mathbf{X}, \mathbf{Y}|\Theta) - \sum_{\mathbf{Y}\in\mathcal{Y}} p(\mathbf{Y}|\mathbf{X}, \Theta^{p-1}) \log p(\mathbf{Y}|\mathbf{X}, \Theta). \tag{2.11}$$

Next, by replacing the definition of Q (Eq. 2.8) in Eq. 2.11, we can represent the ratio $\log(p(\mathbf{X}|\Theta)/p(\mathbf{X}|\Theta^{p-1}))$ by

$$\begin{aligned} \log p(\mathbf{X}|\Theta) - \log p(\mathbf{X}|\Theta^{p-1}) &= Q(\Theta, \Theta^{p-1}) - Q(\Theta^{p-1}, \Theta^{p-1}) \\ &\quad + \sum_{\mathbf{Y}\in\mathcal{Y}} p(\mathbf{Y}|\mathbf{X}, \Theta^{p-1}) \log \frac{p(\mathbf{Y}|\mathbf{X}, \Theta^{p-1})}{p(\mathbf{Y}|\mathbf{X}, \Theta)} \end{aligned} \tag{2.12}$$

The last term of this equation is equal to the relative entropy between the two densities, and by definition have always positive value [54]. Thus, it follows that

$$\log p(\mathbf{X}|\Theta) - \log p(\mathbf{X}|\Theta^{p-1}) \geq Q(\Theta, \Theta^{p-1}) - Q(\Theta^{p-1}, \Theta^{p-1}). \tag{2.13}$$

Given a parameterization Θ^p such that

$$\Theta^p = \arg\max_{\Theta} Q(\Theta, \Theta^{p-1}), \tag{2.14}$$

and substituting Θ^p in Eq 2.13, we obtain

$$\begin{aligned} \log p(\mathbf{X}|\Theta^p) - \log p(\mathbf{X}|\Theta^{p-1}) &\geq Q(\Theta^p, \Theta^{p-1}) - Q(\Theta^{p-1}, \Theta^{p-1}) \\ &\geq Q(\Theta, \Theta^{p-1}) - Q(\Theta^{p-1}, \Theta^{p-1}) \\ &\geq 0 \end{aligned}$$

and consequently

$$\log p(\mathbf{X}|\Theta^p) \geq \log p(\mathbf{X}|\Theta^{p-1}). \tag{2.15}$$

Intuitively, this means that by maximizing Q (Eq. 2.8) in regard to a parameterization Θ^{p-1}, we obtain a parameterization Θ^p that maximizes the log likelihood (Eq. 2.5). Based on this result, the EM algorithm works by iterating between two steps. In the first (E-step), it finds the expected value of the complete likelihood given the current parameterization Θ^{p-1}. In the second step (M-step), it looks for the set of parameters Θ^p that maximize the expectation from the E-step. At each iteration, the EM increases the log-likelihood converging to a local maximum [61]. These steps are repeated P times or until a convergence criterion is fulfilled.

Before proceeding with the deduction, we need to define the posterior probability of $y_i = k$, given x_i. By Bayes rule this can be defined as follows [145],

$$\begin{aligned} p(y_i = k|x_i, \Theta) &= \frac{p(y_i = k)p(x_i|y_i = k, \theta_k)}{p(x_i|\Theta)} \\ &= \frac{\alpha_k p_k(x_i|\theta_k)}{\sum_{k'=1}^{K} \alpha_{k'} p_{k'}(x_i|\theta_{k'})} \end{aligned} \tag{2.16}$$

For simplicity of notation we denote $p(y_i = k|x_i, \Theta)$ by r_{ik}.

In the case of mixture models, Eq. 2.8 can be re-written, after some mathematical manipulations [27], as follows

$$Q(\Theta, \Theta^{p-1}) = \sum_{k=1}^{K} \sum_{i=1}^{N} r_{ik} \log(\alpha_k \cdot p_k(x_i|\theta_k^{p-1})). \tag{2.17}$$

For the E-Step, we need to find the expected value of $\mathcal{L}(\Theta|\mathbf{X}, \mathbf{Y})$ given x_i and the current parameterization. As $\log(\mathcal{L}(\Theta|\mathbf{X}, \mathbf{Y}))$ is linear in x_i, this step reduces to calculating the

expected value $y_i = k$ given x_i and the previous parameterization Θ^{p-1}, that is

$$\begin{aligned} E[y_i = k|x_i, \Theta^{p-1}] &= p(y_i = k|x_i, \Theta^{p-1}) \\ &= r_{ik}. \end{aligned} \tag{2.18}$$

The M-Step can be formally described as

$$\Theta^p = \arg\max_{\Theta} Q(\Theta, \Theta^{p-1}). \tag{2.19}$$

To find the parameter estimates, we need to integrate Eq. 2.8 in relation to its parameters Θ in a maximum likelihood fashion.

For the α_k, the MLE estimate can be obtained as

$$0 = \left[\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik}\log(\alpha_k \cdot p_k(x_i|\theta_k)) + \lambda(\sum_{k=1}^{K}\alpha_k - 1)\right]\frac{\partial}{\partial \alpha_k} \tag{2.20}$$

$$0 = \sum_{i=1}^{N}\frac{1}{\alpha_k}r_{ik} + \lambda \tag{2.21}$$

where λ is a Lagrange multiplier that guarantees stochasticity (Eq. 2.4). Setting $\lambda = -N$, we have

$$\alpha_k = \frac{\sum_{i=1}^{N} r_{ik}}{N}. \tag{2.22}$$

The estimates of θ_k will be specific to the choice of the component densities. For many families of densities, such as exponential type densities, there are analytical solutions (see Section 2.3.3). Even for cases where the maximum likelihood estimate cannot be found, it is sufficient to find a parameterization Θ^{p-1}, such that

$$Q(\Theta^p, \Theta^{p-1}) > Q(\Theta, \Theta^{p-1}). \tag{2.23}$$

This is the case, for example, when Hidden Markov Models (HMM) are used as the component densities. In this scenario, we can apply the Baum-Welch algorithm [21] for each component of the mixture at the M-Step of the EM algorithm. This procedure estimates a local maximum likelihood estimate of a HMM, and meets Eq. 2.23. This estimation method is known as the generalized expectation-maximization algorithm [27].

2.3.2 Method Initialization

An important point of the EM algorithm is the selection of the initial parameterization Θ^0 of the model. A standard way to obtain Θ^0 is to choose random r_{ik} values uniformly from $[0, 1]$ and estimating the individual models with the M-Step. In order to deal with the effects of the random initialization, all estimations are repeated a number of times (usually 15), and the solution with highest likelihood is selected [143].

2.3.3 Mixture of Multivariate Gaussians

As an example, we show how the estimates of a mixture with multivariate Gaussians can be computed. The probability density function of X is defined as

$$p(x|\theta) = \frac{1}{\sqrt{2\pi|\Sigma_x^{-1}|}} \exp\left(-\frac{1}{2}(x-\mu_x)\Sigma_x^{-1}(x-\mu_x)^T\right) \tag{2.24}$$

where μ_x is a vector of means $(\mu_{x_1}, ..., \mu_{x_L})$, Σ_x is the $L \times L$ covariance matrix, and $\theta = (\mu_x, \Sigma_x)$. By replacing 2.24 in 2.17, we obtain,

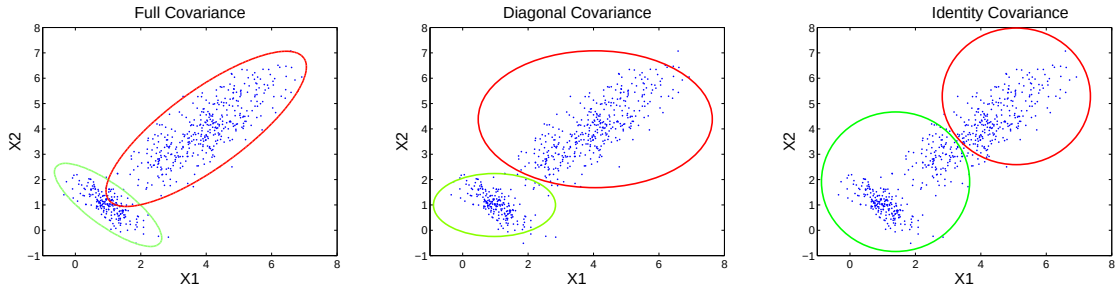
$$\begin{aligned} Q(\Theta, \Theta^{p-1}) &= \sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik}\log(\alpha_k) - \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik}\log(2\pi|\Sigma_{x|k}^{-1}|) \\ &\quad -\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{N} r_{ik}(x-\mu_{x|k})\Sigma_{x|k}^{-1}(x-\mu_{x|k})^T, \end{aligned} \tag{2.25}$$

where $\theta_k = (\mu_{x|k}, \Sigma_{x|k})$ are the parameters of the pdf p_k. Subscripts on parameter the $\mu_{x|k}$ indicate that the parameter μ is an estimate of the variable X and it is conditioned on the mixture model component k. By taking the derivative of Eq. 2.25 in respect to $\theta_k = (\mu_{x|k}, \Sigma_{x|k})$, we obtain the following estimates,

$$\hat{\mu}_{x|k} = \frac{\sum_{i=1}^{N} r_{ik}x_i}{\sum_{i=1}^{N} r_{ik}}, \text{and,} \tag{2.26}$$

$$\hat{\Sigma}_{x|k} = \frac{\sum_{i=1}^{N} r_{ik}(x_i-\mu_{x|k})(x_i-\mu_{x|k})^T}{\sum_{i=1}^{N} r_{ik}}. \tag{2.27}$$

Mixture of multivariate Gaussians are able to model groups of observations in ellipsoidal regions of the Euclidean space with any orientation and size. See Figure 2.2 for an example. In many situations, it may be desirable to use models with simpler assumptions, and consequently fewer parameters. One alternative is to restrict the covariance matrix



(a) (b) (c)

Figure 2.2: *Example of solutions in a two dimensional data found by a mixture of Gaussians with full covariance matrices (a), mixture of Gaussians with diagonal covariance matrices (b), and mixture of Gaussians with identity covariance matrices (c). The ellipsoids correspond to the region with 95% of the component density. With the full covariance matrix, the mixture fits the two groups shapes well. With the diagonal covariance matrix, the components also model similar groups of observations compared to the full covariance matrix. However, for the former, the density cover regions of the space without observations. Gaussians with identity covariance matrices, which can only find spherical and equal size components, cannot model the two groups of observations well.*

(Eq. 2.27) to the diagonal entries

$$\Sigma_{x|k}^{d} = \text{diag}(\Sigma_{x|k}), \tag{2.28}$$

where $\text{diag}(\Sigma)$ denotes a matrix, which has same values as the diagonal of the matrix Σ and zero for all off diagonal entries. In this case, we obtain pdfs with ellipsoidal shape, but with orientation parallel to the coordinate (see Figure 2.2). Another possibility is to restrict all covariance matrices of the components to be the identical, which leads to all components having the same shape and orientation. The most simplistic assumption is the use of identity covariance matrices such that $\Sigma_k^* = \sigma^2 I$ and $\alpha_k = 1/K$. In this case, all components cover spherical and equal size regions of the space (see Figure 2.2 for a comparison of distinct parameterization in a toy data). See [10] and [37] for a complete listing of possible parameterizations of the covariance matrix of a multivariate Gaussian.

2.3.4 EM and Local Maxima

Ideally, one would like to use the full covariance matrix parameterization, as it model all covariance between variables. However, with such covariance matrices, the EM usually returns local maximizers, characterized by having a component with few observations assigned to it [143]. In other words, the mixture fits perfectly a small part of the data, obtaining a high likelihood, but does not achieve a good fit for other regions of the space. This follows from the fact that the likelihood function is unbounded on boundaries of the

parameter space (very low values of α or diagonal entries of Σ.) In particular, when the number of observations (N) in the data is low, or in the presence of outliers, such solutions will be often found by the EM algorithm [143].

To prevent this, there are several techniques available. One simple method [95] is to constrain the diagonal values of the covariance matrices to never be below a given threshold value. Another alternative, which minimizes the effects of outliers, is to use alternative density functions, such as the student [144], or the use of noise components [10].

A more principled approach is to define prior density functions on the mixture parameters and perform a maximum-a-posteriori (MAP) estimation with Monte Carlo Markov Chains (MCMC) [65, 84, 180]. This requires the specification of a proper conjugate prior on the parameters. For example, [65] considers a Wishart density function as a prior on Σ_k and Dirichlet distributions for the component responsibilities α. However, MCMC has a higher computational cost than the EM algorithm. Recently, [80] showed that for multivariate Gaussians, where the posterior mode solution is given with the use of conjugate priors, EM estimation with point MAP estimates achieves comparable results to those obtained with the computationally costly MCMC.

2.3.5 Determining the Number of Components

We cannot rely on maximal likelihood to estimate the number of components, since overfitted solutions, such as one component per observation would arise (see Figure 2.3). We need to balance between fit versus generality. This is commonly done with a penalized likelihood approach, as the Bayesian information criterion (or BIC for short) [191], and further extensions [26, 38, 227]. The problem of finding the number of components can also be tackled in a Full Bayesian setting using Dirichlet Process priors [75]. However, this approach requires the use of the computationally expensive MCMC. Despite its simplicity, BIC performs well in simulation studies [145]. Thus, it will be the methodology used throughout this thesis for selecting the number of components.

We can tackle the selection of the number of components in a Bayesian framework by comparing two mixture models Θ_K and Θ_{K+1} with Bayes Factors. We calculate the ratio of posterior,

$$B_{K,K+1} = p(X, Y|\Theta_K)/p(X, Y|\Theta_{K+1}), \tag{2.29}$$

where Θ_K and Θ_{K+1} are the parameters of two mixture models with K respectively $K+1$ components. It is possible to compare several models at once, rather than two by two as in frequentist statistical test. When we use the EM-algorithm to estimate maximum likelihood mixture models, approximate Bayes factors can be easily deduced from the Bayesian information criterion (BIC) [191],

$$-2\log p(X, Y|\Theta_K) \approx -2\log \mathcal{L}(\Theta_K|\mathbf{X}, \mathbf{Y}) + \psi_K \log N, \tag{2.30}$$

where K is the number of components, $\mathcal{L}(\Theta_K|\mathbf{X}, \mathbf{Y})$ is the maximized mixture log-likeli-

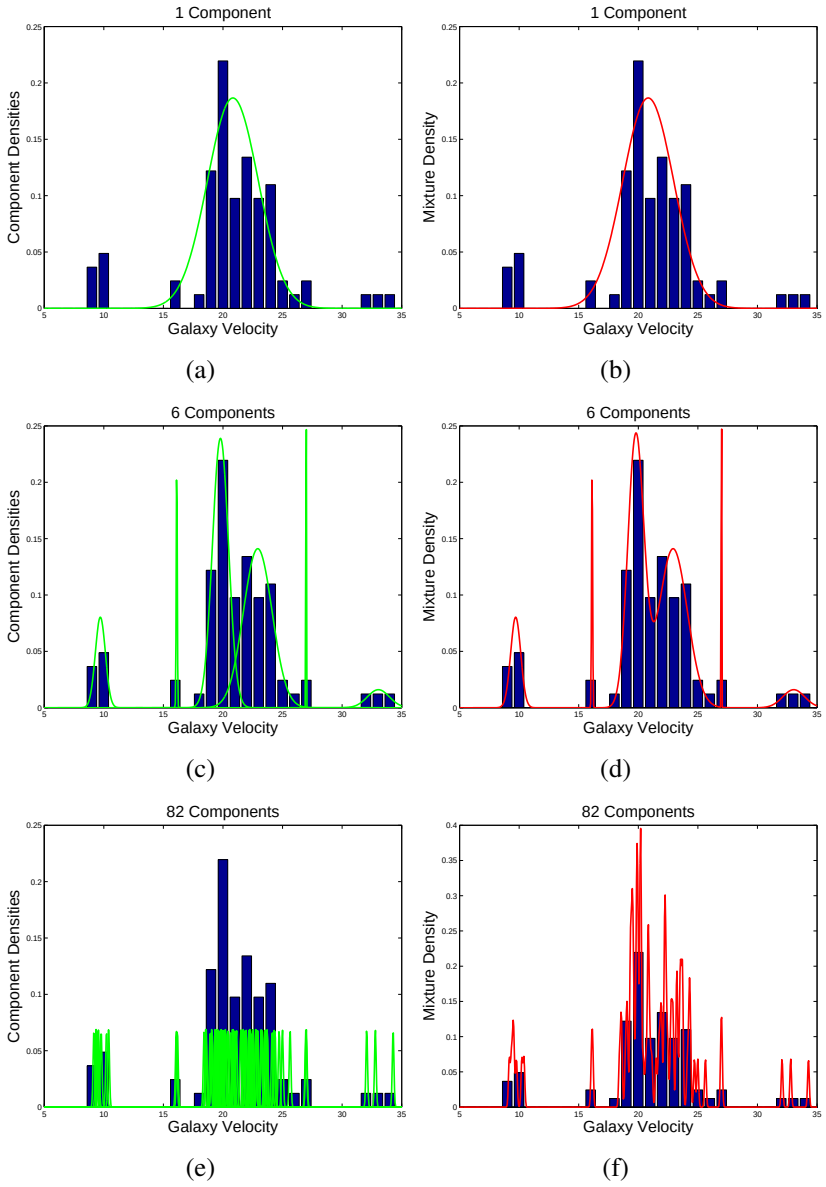


Figure 2.3: *Examples of mixture models with 1, 6 and 82 components fitting the Galaxy Velocity data [176]. On the left plots, we have the density of individual components and the histogram of the data, while in the right we have the mixture density and the data histogram. The mixture with one component models roughly the density in the range* $[15, 25]$*, and imposes zero density to other ranges of the density, plots (a) and (b). The mixture with 82 components, the maximum likelihood solution for number of components equal to the number of observations, simply over-fits the data, plots (e) and (f). The solution with 6 components offers a trade off between these two solutions, providing a good fit of the data, modeling well all ranges of the density, plots (c) and (d) . This mixture was presented in [145] as the optimal solution for the Galaxy data.*

hood with K components (Eq. 2.7), ψ_K is the number of free parameters in Θ_K and N is the number of observations in $\mathbf{X}$.

The term $\psi_K \log N$ penalizes more complex models, since the fit of a model tends to improve as the number of parameters increases. The smaller the value of BIC, the better the model. It has been shown that BIC does not underestimate the number of true components asymptotically and performs well in simulation studies [145]. In the case of a multivariate Gaussians, parameterized by (μ_x, Σ_x), the number of free parameters in a model θ_k is equal to $L + L(L-1)/2$. Hence,

$$\psi_K = K * (L + L(L-1)/2). \tag{2.31}$$

This chapter covered the basics aspects on mixture models and their estimation. In the next chapter, we show how mixture models can be used in the context of clustering. Furthermore, for specific applications, as the ones described in Chapter 4 or in Chapter 5, we take advantage of the characteristics of the data at hand, and choose the component models accordingly.