CHAPTER 1

The Wide Scope

The mixture model has long been a challenge to the statistician, whether beginner, practitioner or theoretician. Recent times have seen great advances in our understanding of the some basic mathematical features of this model, and these notes are meant to be a unification of the work I have carried out, jointly with many wonderful collaborators, in this area. Based on lectures given in 1993 at a regional conference of the Conference Board of the Mathematical Sciences, the notes are directed toward a mixed audience of advanced graduate students and research workers in this and related areas. For the sake of newcomers to the mixture model, I will attempt to be complete enough for the text to make sense in itself, but must at some points refer the reader to other more extensive treatments.

Unfortunately, the goal of timeliness in the end also forced some truncation of the subject matter in the original lecture notes. On the other hand, some subjects have been given enhanced development because they are truly new, and the audience I have in mind will appreciate a deeper presentation of background and of the beautiful geometric structures of the model.

The first chapter of these notes, corresponding to Lecture 1, lays out the mixture landscape as I see it, and the practical side of my motivation for interest in the area. There are two major points:

- There are *many* statistical topics, some quite extensive by themselves, that can rightly be called mixture model subtopics. They all share the mixture model structure, and have similar inferential goals. One of the themes of this chapter is, therefore, that the many names for the mixture model hide its universality. In these notes I am aiming for the universal aspects that lie beneath.
- There is a nonparametric approach to maximum likelihood in the mixture model that gives us an extremely powerful set of tools to use, both in a nonparametric approach and in diagnosing the ailments of parametric approaches. I might add that, to my mind, it is mathematically elegant and fun.

The more sophisticated reader knows well that the sphere of mixture modeling includes random and mixed effects models, empirical Bayes, latent class and trait models, clustering, deconvolution and many other key words and phrases, and for such a person this chapter could perhaps be shorter. Just the same, it was an adventure and an amazement to me just to construct an itemized list.

My own understanding of this subject has been greatly enhanced by studying the geometric structures of the model, and as such the emphasis in later chapters will many times be on fundamental geometric ideas whose truth can sometimes be transmitted more easily pictorially than through detailed mathematical arguments. In addition to giving those pictures, I will sometimes sacrifice the generality of an argument in order to make it more transparent, with the understanding that the reader pursuing the matter further should go to the cited sources. The reason for this is that the *goal* in these notes is not to repeat what has been done, but to try to provide a reader of modest background a clear understanding of a sometimes difficult topic.

Although this is not a textbook, I am presuming that some readers will wish to confirm and enhance their understanding by the active process of doing calculations and drawing pictures rather than through the more passive activity of reading. For this purpose, I have marked various features of the text as exercises (denoted by *Italic* type and frequently, but not always, within brackets). None should take long to perform, given that the desired insight has been obtained.

1.1. The finite mixture problem. The simplest and most natural derivation of the mixture model arises when one samples from a population that consists of several homogeneous subpopulations, which we will call the *components* of the population. The number of components will generally be denoted m, but if we wish to emphasize that it is not known, we will use ν. The components will be indexed by $j = 1, \ldots, m$. Suppose we sample from such a population, recorded as data (X_i, J_i), for $i = 1, \ldots, n$, where $X_i = x_i$ is a measurement on the ith sampled unit and $J_i = j_i$ indicates the index number of the component to which the unit belongs.

Further, suppose that if we were sampling just from the jth component, it is known that there would be an appropriate probability model for the sampling distribution, say

$$\Pr[X = x | J = j] = f(x; \theta, \xi_j). \tag{1.1}$$

Please note that although the left side of the above expression formally refers to a discrete variable X, we will use this same symbol—and its like—in these notes to mean a density function, whether discrete or absolutely continuous. In (1.1), f represents a known density function, most naturally called the *component density*.

The variable ξ_j in (1.1) is an unknown parameter, called the *component parameter*, that describes the specific attributes of the jth component population, and θ is a vector of parameters that describes unknown characteristics

common to the entire population. For the moment, the parameter θ is of secondary interest and will be dropped from the notation, although we will return to it later.

The proportion of the total population that is in the jth component will be denoted by π_j and called the *component weight*. We therefore have $\sum_j \pi_j = 1$. The component weights π_j are usually unknown parameters. Since we suppose that the population has been sampled at random, the probability that an observation comes from the jth component is $\Pr(J = j) = \pi_j$. We can conclude that the variables (X_i, J_i), for $i = 1, \ldots, n$, are a random sample from the joint density described by

$$\begin{aligned}\Pr(X = x, J = j) &= \Pr(X = x | J = j) \Pr(J = j) \\ &= f(x; \xi_j) \cdot \pi_j.\end{aligned}$$

The mixture model arises if the component label data $J_1, \ldots, J_n$ is missing, so that we observe only the sample $X_1, \ldots, X_n$ from the marginal density of X. Thus the observed data are a sample from the *mixture* density

$$g(x; \boldsymbol{\pi}, \boldsymbol{\phi}) = \sum P(X = x | J = j) P(J = j) = \sum \pi_j f(x; \xi_j). \tag{1.2}$$

If there are m components in the mixture, it will be called an m-component *finite mixture model*. In a general finite mixture model, there is a set of $2m$ parameters

$$\begin{pmatrix} \pi_1 & \cdots & \pi_m \\ \xi_1 & \cdots & \xi_m \end{pmatrix},$$

each column corresponding to a component. The weights satisfy the constraints that

$$\pi_j \geq 0 \quad \text{and} \quad \sum_j \pi_j = 1.$$

A very important special case occurs when there is just one component; in this case the density $f(x; \xi)$ will be the called the *unicomponent* mixture density.

1.1.1. *A simple example.* To illustrate some of the fundamental characteristics of the problem, it is useful to consider a simple example and examine the mixture densities that arise. Suppose that we have a population of animals consisting of two component types, say component 1 = male and component 2 = female, and the characteristic, say X = length, is normally distributed in both components when considered alone. Suppose for simplicity that the two component groups have the same standard deviation σ for X, but that they have different means, ξ_1 and ξ_2, respectively. Further, assume that the males have the smaller mean ξ_1, so $\xi_1 < \xi_2$ (e.g., black widow spiders). Let π be the population proportion of component 1 so that $\bar{\pi} := 1 - \pi$ is the proportion from component 2. If we sample from the two components without gender label, then the resulting distribution for heights is a mixture of two normals,

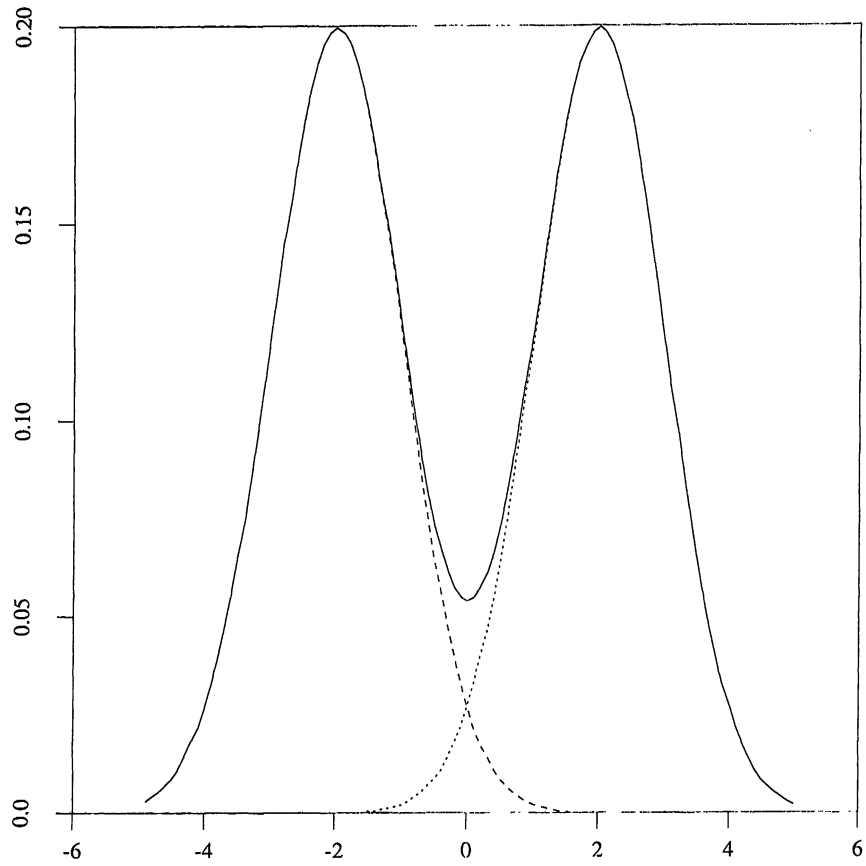

FIG. 1.1. *Mixed normal densities with means four standard deviations apart.*

which we write as

$$\pi N(\xi_1, \sigma^2) + \bar{\pi} N(\xi_2, \sigma^2).$$

Here we use the symbol $N(\xi, \sigma^2)$ to represent the normal probability measure with mean ξ and variance σ^2.

If the components have equal component weights of $\pi = (1 - \pi) = 1/2$ and the means are four standard deviations apart, then heights from the population are described by the solid curve in Figure 1.1.

In this case the mixed nature of the density is revealed through its bimodality. Moreover, it is clear that we can obtain from the value of the variable X = height a great deal of information about whether the measurement was on a male or female.

We might contrast this favorable situation with one in which the means are just one-half as far apart, two standard deviations. One might not expect the dramatic change that going from four to two standard deviations makes. In Figure 1.2 we see that the population density is now unimodal.

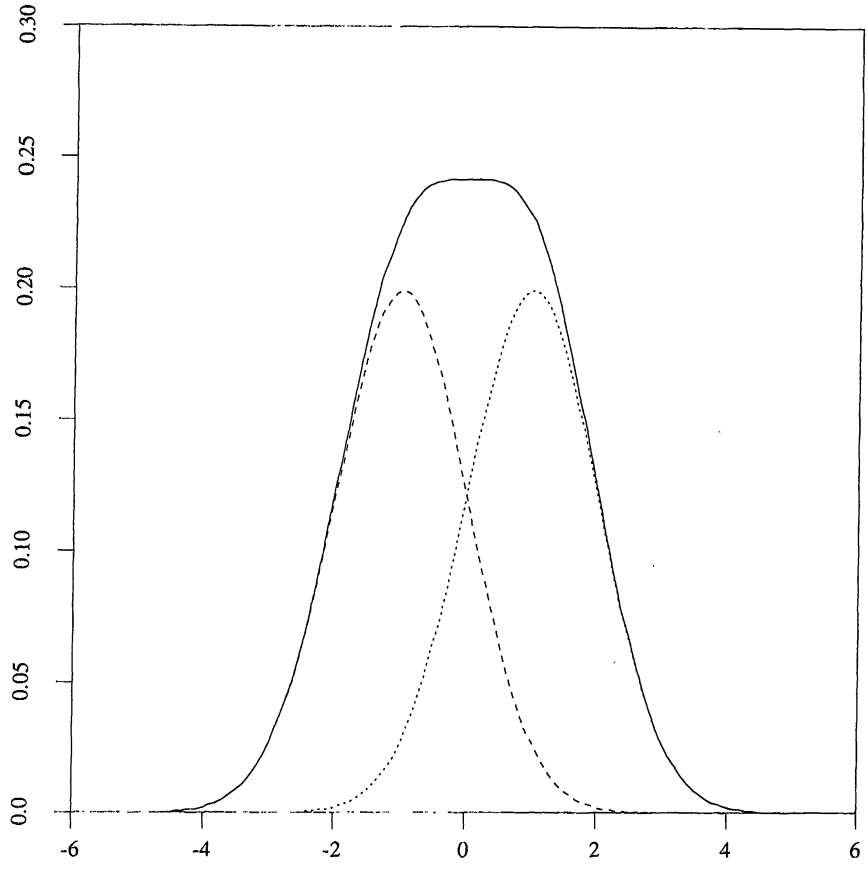

FIG. 1.2. *Mixed normal densities with means two standard deviations apart.*

Suppose that we wished to construct a classification rule that took the variable X and made a determination of the sex. In this case, the best rule (fewest misclassifications) would be to assign those spiders smaller than the population mean to be males and those larger than this to be females. As an indication of the loss of information about the true number of males and females, we note that using this rule would cause many males and females to be misclassified. (We will learn more about the difficulties associated with the loss of information about the mixture due to the closeness of the components in a later section of this chapter.)

1.1.2. *More complicated applications.* The most compelling examples of the mixture model occur when there are physically identifiable components in a true population. We would like to point out here just two of the many interesting such applications in the literature.

MacDonald and Pitcher (1979) considered a situation in which the population consists of the fish of a single species in a lake, and each component

consists of the fish of a single yearly spawning of that species. The components are thus relatively homogeneous and fish in any one component j might, therefore, have X = length that would be adequately modeled with a normal distribution with unknown mean and variance. The parameters ξ_j are then the parameters in this normal distribution and the weights π_j represent the relative abundance of the different age groups. The weights would often be the parameters of interest. Note that determining the values of these parameters over several years would be useful for determining the relative mortality of the yearly cohorts.

We note that this example illustrates one of the useful features of mixture analysis. It enables one to use a *surrogate* measure, in this case X = length, in place of an ideal measurement, here J = age group, in experimental situations in which the ideal measurement is expensive or impossible to obtain. However, in MacDonald and Pitcher's case, this ideal measurement is actually available, so they can compare their mixture model analysis with that attained by knowing the ages of the fish.

Another example of this type is given by Do and McLachlan (1984). In this case the scientific interest was in the population of rats being eaten by a particular species of owl. The components of the population were a number of distinct rat species. Although it would be very difficult to directly survey the owls for dietary preference, it was easy to collect owl feces, from which the rat skulls could be extracted. Various measurements X were taken on the skulls. A mixture model can be constructed as follows: Let π_j be the proportion of the jth species in the owls' diet and let $f(x;\xi_j)$ be the intraspecies density of these characteristics within the jth species, where f is the multivariate normal density and ξ_j are the parameters for the jth species.

For the reader who wishes to read more about the many direct applications of the finite mixture model, there are three good books available on the topic: Everitt and Hand (1981), Titterington, Smith and Makov (1985) and McLachlan and Basford (1988). We return to the basic mathematical structures of the model.

1.2. The latent (or mixing) distribution. An extremely important aspect of the finite mixture problem that we have just described is that we can identify the unknown parameters with a distribution. We next show how this is done.

1.2.1. *The discrete latent distribution.* Define the *latent* random variables $\Phi_1, \ldots, \Phi_n$ to be the values of the parameter ξ corresponding to the sampled components $J_1, \ldots, J_n$; that is, if the ith observation came from the jth component, then define $\Phi_i = \xi_j$. In symbols,

$$\{J_i = j\} \iff \{\Phi_i = \xi_j\}.$$

It is conventional to let the realized value of a random variable, such as Φ_i, be denoted by the lowercase version of the same letter, such as $\Phi_i = \phi_i$. Thus we will have two symbols, ϕ and ξ, representing elements

of the component parameter space. It is useful to do so in order to avoid ambiguity about whether a symbol, such as ϕ_j, refers to the component parameter of the *jth component* (here denoted ξ_j) or the latent variable that was sampled in the *jth observation* (here denoted ϕ_j).

The Φ_i are a random sample from the discrete probability measure Q that puts mass π_j at the support point ξ_j; that is,

$$\Pr[\Phi = \xi_j] = Q(\{\xi_j\}) = \pi_j.$$

We can in this way equate the set of unknown parameters π_j and ξ_j uniquely with a discrete probability measure Q on the parameter space for ξ, with m points of support $\{\xi_1, \ldots, \xi_m\}$ and corresponding masses $\{\pi_1, \ldots, \pi_m\}$. Thus we have the extremely important concept that estimating the unknown parameters in the mixture model (1.2) is the same as determining an unknown distribution Q with support on the parameter space.

The distribution Q is usually called the *mixing distribution*, but to avoid confusion with the expression "mixture distribution" and because it is associated with the latent variable Φ, here we will call it the *latent distribution*, in line with terminology found in the social sciences literature. We will also replace the component and weight parameters in the notation with Q, so that when it is known that we are in a finite mixture situation,

$$Q \equiv \begin{pmatrix} \pi_1 & \cdots & \pi_m \\ \xi_1 & \cdots & \xi_m \end{pmatrix}.$$

With this change in perspective, there is also a natural change in terminology, in which the component parameters ξ_j are now the *support points* of the latent distribution and the component weights π_j are the *probability masses* of the latent distribution.

We note that we can also view the hypothetical complete data set as being (X_i, Φ_i), $i = 1, \ldots, n$, rather than (X_i, J_i). Since we arrive at the same marginal distribution for the observable variables X_i, we can choose the representation that is most useful. The latent variable representation (X_i, Φ_i) links the mixture problem to the standard Bayesian paradigm: The mathematical model for the pair (X, Φ) is identical to a Bayesian model in which the conditional distribution of X given the realized parameter $\Phi = \phi$ is $f(x|\phi) = f(x; \phi)$, and Q is the prior distribution on Φ.

The distinction here is that we have more than one observation of the random variable X, corresponding to multiple realizations of the latent variable Φ, so that we are able to do frequentist inference about the prior Q itself. We return to this point later in discussing empirical Bayes. We note that a Bayesian approach to the mixture problem we are considering here can be found in the literature under the key phrase *hierarchical Bayes*.

We next note that if the component density f does not depend on the component other than through the parameter ξ, then we can write the mixture density of an observation X as an expectation, or average, over the latent

distribution Q of Φ :

$$f(x; Q) := E[f(x; \Phi)] = \int f(x; \phi)\, dQ(\phi).$$

[*Exercise.*] This last representation of the mixture density as an integral of a known component density function f with respect to an unknown probability measure Q will be the basis of many of the results of this monograph.

1.2.2. *The continuous latent distribution.* Although we have derived the mixture model in the context of a population model with finitely many distinct components because it provides a fundamental level of understanding, it is important that there is a natural extension of the model in which the latent variable Φ has a continuous density, say $dQ(\phi) = q(\phi)\, d\phi$, so that the mixture density becomes

$$f(x; Q) := \int f(x; \phi)\, dQ(\phi) = \int f(x; \phi)\, q(\phi)\, d\phi.$$

Sometimes the continuous latent variable has a direct physical interpretation, in the sense that one could, at least in theory, measure it exactly, but that it was not measured on the ith unit. An example could be a variable such as Φ = age in a population with year-round births. The density $f(x; \phi)$ then represents the conditional distribution $f(x|\phi)$ of X given the missing variable $\Phi = \phi$. In this situation, it is possible to have a subsample of data in which both Φ and X are measured, with joint density $f(x; \phi)q(\phi)$.

Many other times Φ represents a more abstract quantity presumed to have a strong influence on the measurement X, and about which inference may be desired, but which cannot itself be directly determined by physical measurement. For example, we might suppose that there is a latent *mathematical ability* that largely determines the test score in mathematics on a certain exam, such as the SAT, but we realize that there is randomness in the outcome of an exam for any one subject arising from a variety of other factors.

We can allow for this by constructing a model in which the latent variable Φ_i is a subject-specific parameter that determines the probability that the ith subject will get a question right. Then each subject has a random number of correct answers, but his overall distribution is determined by that subject's latent *ability* parameter ϕ_i. Since the true value of this parameter can only be exactly determined through the taking of infinitely many test items of the same type, it has only abstract meaning.

Much of our interest in these notes will be in the case where the distribution Q is treated as completely unspecified as to whether it is discrete, continuous or in any particular family of distributions; this will be called the *nonparametric mixture model.* If the parameter θ is present, it will be called the *semiparametric mixture model.* In addition, it should be noted that we will later consider further complications to these models, such as the presence of covariates.

We note that these models have many applications, but using the nonparametric model is not appropriate when, for physical or other reasons, there is a discrete mixing distribution with a known number of components. However, the insights gained into the models by the nonparametric approach will be useful in this case as well, as we intend to demonstrate.

1.3. Many more variations and names. In this section we wish to make a brief compendium of the many statistical problems that have mixture structure, in the sense that there is an unknown probability distribution Q that enters naturally into the model construction. The mixture structure goes by many names and the main objective here is simply to alert the reader to the large number of application areas for the methodology discussed here. Many of the subjects mentioned here have vast literatures of their own, which we will not try to summarize. However, they all share the following feature: The likelihood on a observation can be written in the form

$$L_i(\theta, Q) = \int L_i(\theta, \phi)\, dQ(\phi),$$

where some features of the distribution Q are unknown and to be inferred from the data. The term $L_i(\theta, \phi)$ will be called the *likelihood kernel* and it represents the form of the density for the ith observation, conditional on $\Phi = \phi$.

1.3.1. *Known component densities.* There are many interesting and important examples in which there are a finite set of component densities that are completely known and so there are no unknown component parameters ξ_j. In this case we can retain the formal structure defined above by equating the latent variable Φ with the component variable J. The known density function for X conditional on $J = j$ can then be written as

$$\Pr[X = x | J = j] = f(x; j)$$

and the only unknown parameters are the component weights π_j. In this case, the latent distribution Q is the distribution of the variable J, and so knowing the component distributions is equivalent to specifying that the latent distribution Q has a known set of support points $\{1, 2, \ldots, m\}$. This case, called hereafter the *known component density model*, is simpler than the general problem, but is worth studying in its own right both for its many applications and for the insight it gives into the general problem. We therefore offer several examples of its use.

This author had his earliest exposure to the mixture problem in a consulting problem involving fish stock analysis. The components involved were salmon subpopulations, each identifiable as spawning in a single river system and possessing, therefore, a great amount of genetic homogeneity relative to the entire population. The measured variable X on each fish was a genetic typing determined by electrophoresis, a variable that had only a finite number of

outcomes. Correspondingly, the jth river system would have a discrete density

$$f(t; j) = P(X = t|J = j)$$

describing the distribution in that river of the finite set of genetic types t. The data of interest arose when salmon were caught in a particular region of open ocean by fishermen and it was desired to know the fractions π_j of the region's salmon population that came from each river system. The research had political implications in that the river systems were in two countries, the United States and Canada. [For more on applications of this type, see Millar (1987).]

A second example of the known component type comes from Roeder, Devlin and Lindsay (1989). The data consisted of the genetic types of seeds that had been collected from a mother plant and it was desired to know what fraction π_j of the seed generation could be ascribed to each of several competing father plants. In this case, knowing the jth father's genetic type and the mother's genetic type, it was possible to construction a probability density $f(t; j)$ for the genetic types of the seed generation.

A third example arises in the use of positron emission tomography. In this problem, a subject ingests a radioactive substance which is designed to congregate in some bodily part of interest, say a tumor in the brain. An array of sensors is set up around the region of interest and the radioactive emissions are observed. The emissions have the following characterization: at the time of disintegration, two rays shoot off in opposite directions (180° apart). When they arrive at two opposing detectors, it can be deduced that the emission occurred at some unknown point on the line between the two detectors, but nothing more. Since the array has a finite number of detectors, there are a finite number of such opposing detector pairs. The index t will refer to the possible detector pairs and each emission generates an observed variable X taking on values in this set.

We desire to find the hot spots in the brain that are generating the most emissions. To do so, we create a grid of possible emission regions. The variable j will index this grid of sites in the brain, and we define a latent variable by letting $J_i = j$ mean that the ith emission came from site j. Thus we are interested in inference on $\pi_j = \Pr(J = j)$. The geometry of the array and the nature of the emissions determine exactly the component density $P(X = t|J = j) = f(t; j)$, the probability that an emission from j will be observed in the detector pair t.

Thus we have a simple mixture model for the observed process of emissions: $P(X = t) = \sum \pi_j f(t; j)$. For more details, see Shepp and Vardi (1982) and Vardi, Shepp and Kaufman (1985).

1.3.2. *Linear inverse problems.* We offer yet another perspective on the mixture problem. Our basic unknown is the latent distribution Q that satisfies the relationship

$$g(x) = \int f(x; \phi)\, dQ(\phi). \tag{1.3}$$

Here f is known, but g is observed with some error— in our case because we see a sample from it. In the mathematics literature, solving for Q in (1.3) is called a *Fredholm integral equation of the first kind* [Wing and Zahrt (1991)].

If the density for X is discrete, with finite range $1, 2, \ldots, T$, and Q is discrete on a known support set $\{1, \ldots, m\}$, as in the examples of the preceding section, then we can write (1.3) as $g(t) = \sum_s f(t; s)\pi(s)$. This can be written as a matrix equation of the form

$$\mathbf{g} = \mathbf{F}\boldsymbol{\pi},$$

in which we wish to solve for the vector $\boldsymbol{\pi}$, where $\sum \pi_j = 1$ and the π_j are nonnegative. This, in the terminology of Vardi and Lee (1993), is a *linear inverse problem* with *positivity constraints*.

A recurring theme of the mathematical literature is that the linear inverse problem is difficult and unstable to solve. In the matrix case, this is clearly related to the numerical instability of solving such linear equations when the column vectors are highly correlated. One of the key points of Vardi and Lee is that the Expectation-Maximization (EM) algorithm (Chapter 3) is a stable and reliable way to handle the numerical difficulties involved in solving linear inverse problems.

1.3.3. *Random effects models.* Consider the classic one-way ANOVA model. We can visualize the data as being in a two-way array:

$$\begin{array}{cccc} X_{11} & X_{12} & \ldots & X_{1n_1} \\ X_{21} & X_{22} & \ldots & X_{2n_2} \\ \vdots & \vdots & \vdots & \vdots \\ X_{p1} & X_{p2} & \ldots & X_{pn_p} \end{array}.$$

In the *fixed effects* version of this model, row i consists of a sequence of i.i.d. observations from a $N(\phi_i, \sigma^2)$ density, where the parameters $\phi_1, \ldots, \phi_p$ are the unknown means of the p rows. In the *random effects* version of this model, the parameters are viewed as having been sampled from a larger population and so have a latent distribution Q in that larger population. The resulting marginal density of the observations in a single row is the mixture model

$$f(\mathbf{x}; Q, \sigma^2) = \int f(x_1; \phi, \sigma^2) \cdots f(x_n; \phi, \sigma^2)\, dQ(\phi).$$

The usual assumption in the normal theory random effects model is that Q is a normal distribution. If we make this assumption, then the observations $\mathbf{X}$ in a row have a multivariate normal distribution, with a positive and equal correlation between observations that is induced by the fact that the entire row has the latent variable ϕ in common. In fact, the covariance matrix for the row vector $\mathbf{X}$ has the form $\sigma^2 I + \sigma_\Phi^2 E$, where I is the identity matrix and E is a matrix of ones. [*Exercise*: This is easiest to derive by writing $\mathbf{X} = \mathbf{Z} + \Phi \cdot \mathbf{1}$.]

Although the assumption of the normality of the latent row parameters may be tenable in many examples and is the simplest path, it is possible to take a nonparametric approach to the problem. This latter approach seems especially

natural when the data are discrete so that normality of the observations is no longer a tenable assumption. We will consider some aspects of an example of this type in Chapter 8.

1.3.4. *Repeated measures models.* Another statistical problem with nearly the same structure as the random effects model is the *repeated measures model*. If we think of the row vector $\mathbf{X}$ as being a sequence of repeated measurements on a single individual and assume that $\mathbf{X}_i$ has a multivariate normal distribution with

$$\text{Var}(X_{ia}) = \text{Var}(X_{ib}) = \sigma^2$$

and $\text{Corr}(X_{ia}, X_{ib}) = \rho$, then one has the exact same multivariate normal distribution for the observable vectors of $\mathbf{X}$ as in the normal random effects model above, with one exception: In the random effects model, ρ must be nonnegative.

We note that these models start from different modeling points of view, but arrive at essentially the same model. The random effects model arises from assuming the existence of subject-specific latent means, conditional upon which the individual observations within a subject are independent. The marginal correlation of observations within a subject is a consequence of that structure. In the repeated measures model, we have directly modeled the whole vector of observations, including its correlation structure. As a bonus, modeling the covariance structure directly allows us to consider dependencies other than the equicorrelation of all pairs, which, for example, would not be most natural if the repeated observations had occurred in a time series.

These different points of view are largely reconciled in the normal model, at least for equicorrelation, but when the normality assumption is not tenable, they have led to two different schools of modeling data. If we say that the observations in a row form a *cluster* because of their related nature, having all come from the same subject or school or other unit, then we expect them to have some correlation, most likely positive, when viewed from the perspective of the entire population. That is, the population covariance matrix of the $\mathbf{X}$ vector should show nonnegligible correlations. We can choose to model this directly, and do so in the so-called *population average* approach. The other model building approach is *cluster specific*. That is, we can model the effect of being in a cluster (row) by a cluster-specific latent parameter ϕ and assume that this parameter has been sampled from a population Q. This then induces a correlation structure.

When the observations have a binary nature, these two approaches to modeling result in rather different models. See, for example, Neuhaus, Kalbfleisch and Hauck (1991).

1.3.5. *Latent class and latent trait models.* As noted earlier, we have borrowed the terminology "latent variable" from the social sciences literature, where the expression is often used in the mixture model analysis of categorical data. In this literature, we have multinomial observations $\mathbf{X}_i$ whose

probabilistic behavior is determined by some unobserved variable Φ_i. If the distribution of Φ is discrete, then the possible values of Φ (the support points) correspond to the *latent classes* of the population. If the variable is continuous, then its values correspond to some *latent trait* of the population. For a review of this literature, see Clogg (1995) and Heinen (1993).

1.3.6. *Missing covariates and data.* Whenever variates are missing at random from a data set, whether covariates or response variables, then the distribution of the remaining values comes from integrating the missing values out of the joint density function. If we have a regression model, say $f(y|x,z) = f(y|x'\beta + z'\gamma)$, where the z's are entirely missing, then we can write the conditional model for the observable data as $f(y|x) = \int f(y|x'b + \phi)\, dQ(\phi|x)$, where Q is the conditional distribution of $z'\gamma$. If the measured covariates are independent of the missing ones, so that $dQ(\phi|x) = dQ(\phi)$, then the resulting model is of mixture form, with a random intercept in the regression. That is, using a random intercept in a regression model is a method that allows for the additional uncertainty in the inference due to missing covariates.

1.3.7. *Random coefficient regression models.* Another area which has seen much work of a nonparametric kind is the random coefficient regression model. The previous examples in this section have largely been of a kind where a regression model for a particular cluster might include a random intercept term, as in $\beta' z + \phi$. There exist a number of studies in the pharmacokinetics literature of models in which one or more of the β's are random as well. See, for example, Mallet (1986), Mallet, Mentre, Steimer and Lokiec (1988) and Davidian and Gallant (1992). We also note the extensive study of this model by Beran and Hall (1992).

1.3.8. *Empirical and hierarchical Bayes.* Yet another application of the mixture methodology arises in the subject known as *empirical Bayes*. The basic conceptual framework is as we have described, but now the emphasis is on making inferences on the set of realized values of the latent variable $\Phi_1 = \phi_1, \ldots, \Phi_n = \phi_n$. If the distribution Q had been known in advance, then we are in a Bayesian setting and so the best mean-squared error estimator of an individual ϕ is the Bayes estimator

$$E[\Phi|X = x] = \frac{\int \phi f(x;\phi)\, dQ(\phi)}{\int f(x;\phi)\, dQ(\phi)}.$$

Since Q is unknown in our case, one natural tactic is to replace the Q in the above formula with an estimate, either parametric or nonparametric. However, there are many other strategies; see the book by Maritz and Lwin (1989).

One important aspect of empirical Bayes methods is that it has been learned that these methods have advantages in estimating a set of parameters $\phi_1, \ldots, \phi_n$ even if they did not arise as a sample from a distribution Q. To be more precise, the comparison we wish to make is between the fixed

effects model, in which the ϕ parameters are treated as unknown parameters and so the likelihood has the form

$$L_{\mathrm{FE}}(\phi_1, \ldots, \phi_n) = \prod_i f(x_i; \phi_i),$$

and the random effects model, which has a likelihood of the form

$$\prod_i \int f(x_i; \phi)\, dQ(\phi) = \int \cdots \int L_{\mathrm{FE}}(\phi_1, \ldots, \phi_n)\, dQ(\phi_1) \cdots dQ(\phi_n).$$

The claim is that there are advantages to using the random effects model, even if conceptually the fixed effects model is more applicable, when one is estimating the values of the realized sequence of parameter values. For a readable introduction to this subject, the *Scientific American* article of Efron and Morris (1977) is highly recommended.

The rough idea is that if one estimates the parameters individually, say by $\hat{\phi}_i$, then the set of estimates $\{\hat{\phi}_1, \ldots, \hat{\phi}_n\}$ tends to be more dispersed than the set of parameters being estimated. To take an extreme case, if the true ϕ's are all equal, then the actual parameters have *no* dispersion, but the estimates have a dispersion corresponding to the variance of their distribution.

Bayesian formulations of this problem fall under the keyword *hierarchical Bayes*. In its simplest form, we say Q has some simple parametric form, say normal, and put priors on the parameters of this density.

1.3.9. *Nuisance parameter models.* The "empirical Bayes effect" just described also arises when estimating a nonlatent parameter of interest θ in the presence of many nuisance parameters $\phi_1, \ldots, \phi_n$, the famous Neyman–Scott problem (1948). This is a topic which has ended up being truncated from the notes, but not due to lack of importance.

The contrast is between the fixed effects approach, which treats the ϕ values as unknown parameters, and random effects approach, which treats them as an aggregate by assuming they come from an unknown distribution. We have already noted that there are efficiency advantages to the empirical Bayes approach when making inferences about the parameters ϕ_i. It turns out that there also are advantages in estimating the structural parameters θ.

1.3.10. *Measurement error models.* Suppose that our goal is to make scientific inferences about the relationship between two population variables, say $Y :=$ blood cholesterol level and $\Phi :=$ dietary cholesterol level. For example, we may wish to fit some sort of regression model, with parameter β, to the conditional density $f(y|\Phi; \beta)$. However, we have available only a *surrogate* measure X that is highly correlated with Φ, such as

$$X := \text{estimate of } \Phi \text{ from dietary questionnaire.}$$

It is well known that if we use X in place of Φ, then the resulting regression will show a diminished effect of the covariate X as compared with the use of latent variable Φ. Indeed, in a worst case scenario, where one covariate is measured with error and a second one is not, the relative importance of the two variables can be reversed.

One easy way to illustrate this point is to assume that (Y, Φ) have zero means and covariance matrix

$$\begin{bmatrix} \sigma_Y^2 & \rho\sigma_Y\sigma_\Phi \\ \rho\sigma_Y\sigma_\Phi & \sigma_\Phi^2 \end{bmatrix}.$$

In such a setting, the correct (latent) regression slope is $\beta = \rho\sigma_Y/\sigma_\Phi$.

Assume the surrogate satisfies $X = \Phi + \varepsilon$, where the measurement error ε is independent of the other variables. Since $\mathrm{Cov}(X, Y) = \mathrm{Cov}(\Phi, Y)$, the covariance matrix for (Y, X) differs from the above only in the lower right corner, where we obtain $\mathrm{Var}(X) = \sigma_\Phi^2 + \sigma_\varepsilon^2$, and so the surrogate (Y, X) regression coefficient is

$$b = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \beta\frac{\sigma_\Phi^2}{\sigma_\Phi^2 + \sigma_\varepsilon^2},$$

so that $0 \le b/\beta \le 1$.

There are two approaches to the measurement error problem, corresponding closely to whether we treat the unobserved $\Phi_1, \ldots, \Phi_n$ as unknown parameters (by conditioning on their realized values $\phi_1, \ldots, \phi_n$) or as latent random variables from some unknown population distribution Q with its own set of unknown latent parameters. The former approach runs into the problem of infinitely many nuisance parameters mentioned earlier. Taking the latent variable or mixture approach, we write the density of the observed variables as

$$f(y, x) = \int f(y, x|\phi)\, dQ(\phi) = \int f(y|\phi, x) f(x|\phi)\, dQ(\phi).$$

Assuming that in the knowledge of ϕ, the variable X contributes no further information about Y, we replace $f(y|\phi, x)$ with $f(y|\phi)$, the regression distribution of interest. The density

$$f(x|\phi) = \Pr(X = x|\Phi = \phi)$$

is determined by the measurement error process and, typically, as in the above example, X would be modeled as having a normal distribution with mean $\Phi = \phi$. The important feature to us is that the mixture model structure holds in this case, with likelihood kernel

$$L_i(\phi) = f(y_i|\phi, x_i) f(x_i|\phi).$$

In Chapter 8 we will return to this model to analyze some properties of the nonparametric approach to estimating the distribution Q.

1.3.11. *Deconvolution problems.* Deconvolution problems are mixture problems with a special additional piece of structure. If the data X can be written as $\Phi + Z$, where Φ is a latent variable and Z has a known density f, then the resulting density for X has the mixture form

$$g(x; Q) = \int f(x - \phi)\, dQ(\phi).$$

This structure, as we have just seen, arises as a natural model for measurement error processes. Because of this structure, one can employ special techniques to attempt the inversion problem. In particular, because of the identity of characteristic functions, $\Psi_x(t) = \Psi_\Phi(t)\Psi_z(t)$, one can develop methods to solve for the characteristic function of latent variable Φ from the known characteristic function of Z and an estimated characteristic function of X.

Since the cumulant generating functions in a convolution are additive, one also has the cumulant identity $\kappa_r(X) = \kappa_r(\Phi) + \kappa_r(Z)$. If Z is normal, then this implies that the cumulants of X and Φ are identical for $r \geq 3$, since $\kappa_r(Z) = 0$ on this range. This simple structure has led to substantial use of moment methods in the signal processing literature, yet another statistics topic closely related to the mixture topic.

1.3.12. *Robustness and contamination models.* It is a common statistical practice to study the robustness of a statistical procedure by constructing a simple class of alternative mixture models. One can construct a simple symmetric model alternative to the normal model via the normal scale mixture,

$$(1 - \alpha)N(\theta, \sigma^2) + \alpha N(\theta, k\sigma^2),$$

where k is large and α is small, representing the proportion of observations being measured with larger errors. In fact, many families of symmetric distributions that are commonly used as heavy tailed alternatives to the normal have *normal scale mixture* structure. For example, the family of t distributions are scale mixtures of normals. (*Hint:* Write $t = \Phi Z$, where Z is normal and Φ is independently distributed as the inverse of the square root of a chi-square divided by its degrees of freedom.)

Another approach to constructing robust procedures through mixtures are *contamination models*, such as $(1 - \alpha)N(\theta, \sigma^2) + \alpha N(\phi, \tau^2)$. If α is small, ϕ quite different from θ and τ small, then this model generates *outliers* near the value ϕ.

Aitkin and Wilson (1980) explicitly used maximum likelihood with the above mixture models in order to obtain robust procedures, with apparent success. There appears to be little theoretical work on this approach to robustness.

1.3.13. *Overdispersion and heterogeneity.* The last section suggests that we might find additional advantages for the mixture type model in cases where there is no physical meaning to the latent variable Φ. It is just a simple means of extending our class of models to allow for any lack of fit of a basic model. As we will study in Chapter 2, the construction of a mixture model from a basic model will always give us not only flexibility, but a class of models that allows for more dispersion than the original model. This approach has long been used to construct models that allow for extra dispersion. For example, the beta binomial model, used as an extension of the binomial model, or the negative binomial model, used as an extension of the Poisson, both have natural derivations as mixtures of the basic models. These points will be discussed briefly in Section 3.1.1, which deals with conjugate families of distributions.

1.3.14. *Hidden mixture structures.* There are a variety of nonparametric statistical models with hidden mixture structure. By this we mean, first of all, that the class of distribution functions in the model is *convex*. That is, if F and G are in the model, then so is $(1-\alpha)F + \alpha G$. Second, there is a representation of all elements of the model as convex combinations of some basic class of distributions, say $\{F_\phi\colon\ \phi \in \Omega\}$. If so, we can write an arbitrary element of the model as $\int F_\phi\, dQ(\phi)$. (This basic class should be in the set of extreme points of the convex class if we seek identifiability for Q.)

The simplest example is the class of all distribution functions. If we let F_ϕ be the degenerate distribution at ϕ and let Q be the latent distribution, then Q is also the mixture distribution. We here have the equality of the latent variable Φ with the observed variable X.

A more sophisticated example concerns the class of distribution functions with nonincreasing density functions on $[0,\infty)$. This class is clearly convex and we seek a class of basic distributions F_ϕ. The solution is to let F_ϕ be the uniform distribution on $[0,\phi]$. [*Exercise.*] Here the latent distribution appears to have no intrinsic statistical meaning.

1.3.15. *Clustering: A second kind.* We have already pointed out one of the mixture-related uses of the word cluster. However, mixture methods are also prominently used in another area that is, technically, nonstatistical. Suppose we have a set of vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$, not obtained by sampling. They might represent a vector of numerical characteristics for a set of n species. We desire to find those species which are most similar to each other in these characteristics, forming thereby *clusters* of species. One of the approaches to such problems is to treat the data as if it were a sample from a mixture distribution. For example, if we desire two clusters, we might construct a reasonable two component mixture model, such as a mixture of two multivariate normal densities with mean vectors ϕ_1 and ϕ_2 and common covariance matrix Σ. After fitting the mixture, say by maximum likelihood, one can assign a data point to cluster 1 or cluster 2 depending on its posterior probability of being from that component. That is, we calculate

$$\Pr(J = 1 | X = x) = \frac{\pi_1 f_1(x)}{[\pi_1 f_1(x) + \pi_2 f_2(x)]}$$

and assign the observation to cluster 1 if and only if this conditional probability is greater than 0.5. This approach to *cluster analysis* is extensively discussed in the book by McLachlan and Basford (1988).

1.4. Be aware of limitations. We hope that the reader is now convinced that there are a multitude of interesting statistical applications that involve the estimation of an unknown distribution function Q and whose likelihoods have the formal structure of mixture problem. However, we must also confess that inference is a very difficult task, as we take observations from Q only indirectly. This section contains some warnings about the limitations of our procedures and our knowledge.

1.4.1. *Robustness characteristics.* It is natural and desirable to ask the question: What are the consequences of slight errors in the specification of the model? When a mixture model is being specified, and I specify a mixture of two normals with different means, what are the consequences if the mixture is actually of two t distributions or is actually a mixture of three normals, two of which are very close together? How stable are my parameter estimates under contamination? At present we would seem to know very little about this side of the subject. Most of the robustness literature is not relevant, as it deals with location-scale regression type modeling. See McLachlan and Basford (1988) for an attempt to adapt these methods to the multivariate normal mixture problem.

There seem to be two other feasible approaches. One is to exploit special structures of the densities involved to create diagnostic procedures and goodness-of-fit tests, thereby constructing, in stages, a model suitable to the data at hand. Another approach is to use, in conjunction with maximum likelihood, a more robust procedure based on minimum distance ideas. Both areas have seen relatively little development in the mixture model.

1.4.2. *Extracting signal from noise.* Another important warning relates to understanding that we cannot possibly discern very much of the fine detail about the distribution Q. In particular, estimating its density $dQ(\phi)$ with a realistic sample size is virtually impossible. Moreover, it is not uncommon that the goodness-of-fit of a mixture model to a data set does not change very much if we switch from a continuous latent distribution to a discrete one, or whether the discrete distribution has two components or four components.

To make this point more clearly, we consider the very simplest of mixture scenarios. Suppose we have a mixture of two normals, say

$$g(x) = \pi N(-a, 1) + \bar{\pi} N(+a, 1).$$

Assume for the moment that a is fixed and known, so that the only unknown parameter is π. If the variance of the normals were very small, the resulting data would appear much like a Bernoulli distribution, in which nearly all the observations would be quite close to $\pm a$. If the data were exactly Bernoulli, the Fisher information about π in a single observation would be $1/\pi\bar{\pi}$. Thus we know that for reasonable accuracy in estimating π when it is near 0.5, one would need a sample of roughly Gallup poll size, say 1000, yielding a standard error of $1/\sqrt{4000} \approx 0.02$.

How much information is in the mixture distribution? As an *exercise*, check the following calculations. The information in π at $\pi = 0.5$ has the form, writing the normal density as n,

$$i_a = E\left[\frac{n(X; a, 1) - n(X; -a, 1)}{0.5n(X; a, 1) + 0.5n(X; -a, 1)}\right]^2.$$

The information relative to the ideal information $1/\pi\bar{\pi} = 4$ is then $i_a/4$. This relative information has been plotted as a function of $2a$, the separation of the means, in Figure 1.3.

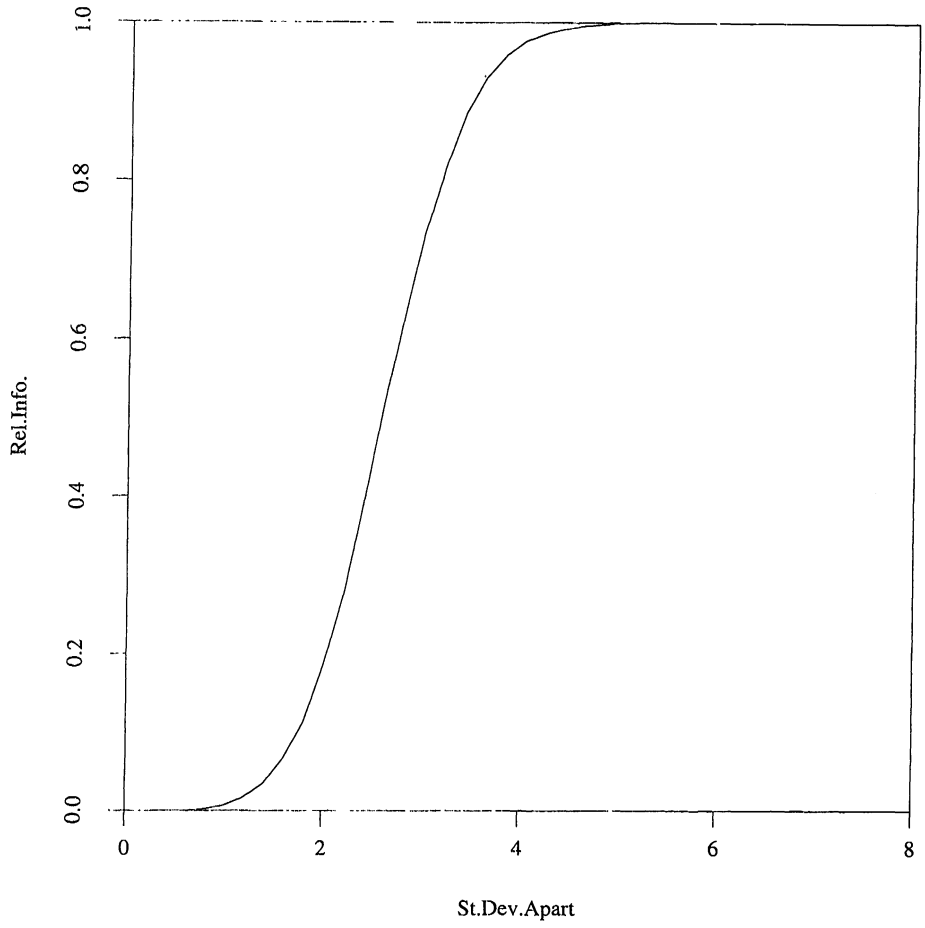

FIG. 1.3. *The information about π as a function of the separation of the means.*

We can see that the relative information is very nearly 1 when the means are four or more standard deviations apart. We can interpret this as showing that we can infer the group label J from the observation X quite accurately and that the mixture maximum likelihood estimator of π is very nearly equal to $n^{-1}\sum J_i$. However, as we move the difference in means from four to two standard deviations, the information falls off dramatically, until at two standard deviations separation we see that in order to attain the same accuracy of estimation as in the Bernoulli trials situation, it would take a sample size *10 times* as large. Recall that, as shown in Figure 1.2, this is exactly where the density is on the unimodal–bimodal boundary. When the separation of means is between 0 and 2, there is virtually no information about π.

This analysis is clearly relevant to the problem of estimating a latent density function $dQ(\phi)$, as indicated by our difficulty in separating out the relative contributions π of two nearby values of ϕ. Indeed, by setting $a = h$ and

letting h go to zero, we have

$$i_h = 4h^4 E\left[\frac{n''(x;0)}{n(x;0)}\right]^2 + o(h^4).$$

The absence of terms of lower order than h^4 is another indicator of the small amount of information in this model when the components are close together. The term inside the expectation will show up again when we consider the $C(\alpha)$ test in Chapter 4.

(If we modified this problem so that the two normal location parameters were unknown, then clearly the information about π is smaller. However, the order of magnitude is not changed from the above calculation.)

This low level of information is also an indication that algorithmic methods will have difficulty. In particular, the EM algorithm will be very slow in this situation.

1.5. The likelihoods. In the examples above and in many typical examples, we will be concerned with independent samples for which the likelihoods have the form

$$L(Q) = \prod_{i=1}^{n} L_i(Q),$$

where $L_i(Q)$ has the integral form $\int L_i(\phi)\, dQ(\phi)$. The corresponding log likelihood will be denoted

$$l(Q) = \ln(L(Q)) = \sum \ln(L_i(Q)). \tag{1.4}$$

The standard analyses of a model with mixture structure involves making one of two assumptions:

- There are a fixed and known number of components, so that the unknown parameters are the π's and ϕ's.
- The distribution Q is from some known parametric family of distributions, say $q(\phi;\gamma)$, typically chosen to be either the normal density with unknown mean and variance or the conjugate family of distributions to the density f.

As a basis for statistical likelihood procedures, both standard modeling methods have their drawbacks. For the fixed component models, there is the problem that the likelihood is high dimensional and known to be multimodal. (A further awkwardness, which we will study in depth in Chapter 4, is the aberrant behavior of the limiting distribution of the likelihood ratio statistic when we test for the number of components.) In the continuous case, numerical integrations are usually needed to carry out the procedure and, regardless of this, one can end up again with a multimodal likelihood.

If, on the other hand, we make no restrictions on the number of components, but view the above likelihood as a function of a completely unknown distribution function Q, then the solution has some extremely nice properties, as will be discussed shortly. Before proceeding on this point, we introduce some special likelihood structures.

1.5.1. *The multinomial likelihood.* First, a number of simplifications in the study of mixture models occur when the component densities are discrete, so that if the sampling is i.i.d., we have a *multinomial likelihood.* Suppose that we have an i.i.d. sample from a discrete mixture density function $f(t;Q) = \Pr(X = t;Q)$ with range $t = 0,\ldots,T$. Then the likelihood of a sample $X_1,\ldots,X_n$ can be written as

$$L(Q) = \prod_{i=1}^{n} f(x_i;Q) = \prod_{t=0}^{T} f(t;Q)^{n(t)}.$$

Here $n(t)$ is the count of the number of the X's that took on value t and it is clear from the representation of the likelihood that the values $n(0),\ldots,n(T)$ are sufficient statistics for the parameter Q. The log likelihood (1.4) can therefore also be written as

$$l(Q) = \sum_{t=0}^{T} n(t)\ln(f(t;Q)). \tag{1.5}$$

1.5.2. *Partly classified data.* Another likelihood structure that occurs in many examples of interest arises when, in a sampling situation with physically identifiable components, the data are *partly classified.* That is, a portion of the sample, the *unclassified* part, has the variable J_i missing and so has the mixture model structure. In the remainder of the sample the J_i are *not* missing and so this group could be identified as being *fully classified.*

In this case the likelihood will be a product of two terms: the first, say $L_1(Q)$, being the usual mixture likelihood for the first sample, whereas the second will have the form

$$L_2(Q) = \prod P(X_i = x_i | J_i = j_i)P(J_i = j_i) = \prod f_i(x_i;\xi_j)\pi_{j_i}.$$

We next show that one can write this likelihood in the mixture form (1.4) and so this complication of the data structure does not require any additional theory.

We start by defining the likelihood kernel

$$L_i(\phi) := f_i(x_i;\phi)\mathcal{I}[\phi = \xi_{j_i}]. \tag{1.6}$$

Here the symbol $\mathcal{I}[\cdot]$ represents the indicator function, a function of all the arguments included in its brackets, which takes on the value 1 when the bracketed statement is correct and 0 when it is false. Note that although L_i is a function of the data, we can and do express this only through the subscript i. (We can do so because in a likelihood, the data are fixed.) Moreover, the corresponding term in the likelihood can be written as an integral over the mixing distribution Q: As an *exercise*, check that for Q discrete we can write

$$L_2(Q) = \prod f_i(x_i;\xi_{j_i})\pi_{j_i} = \prod \int L_i(\phi)\,dQ(\phi).$$

For the theory of maximum likelihood we will present, it is not mandatory that L be a density function—only that it is nonnegative, a prescription satisfied by L_i in this case.

A similar likelihood situation arises when the latent variable $\Phi_i = \phi_i$, possibly continuous, has been directly measured in a subsample. In this case the appropriate likelihood kernel analogue of (1.6) is

$$L_i(\phi) := f_i(x_i; \phi)\mathcal{I}[\phi = \phi_i].$$

The distinction between (1.6) and this case is that in the former the exact value of the latent variable Φ_i is not observed; all that is known is that it is the value of the parameter ξ associated with the j_i component. Thus (1.6) has less information about the latent distribution than when the latent variable is itself observed.

1.6. The mixture NPMLE theorem. The idea of finding a nonparametric maximum likelihood estimator (NPMLE) of a latent distribution is an old one. The idea was suggested in an abstract by Robbins (1950), and later received substantial theoretical development by Kiefer and Wolfowitz (1956). Although the latter showed that the method had great theoretical promise as a method of providing consistent estimators in problems with many nuisance parameters, there was no development of numerical methods for the computation of such an estimator. The development of such methods, together with further properties of the estimator $\hat{Q}$, arose in papers that came 20 years and more later, particularly Simar (1976), Laird (1978), Jewell (1982) and Lindsay (1981, 1983a, b). In this chapter we will summarize the most important developments by informally describing the "fundamental theorem of nonparametric mixture maximum likelihood estimation." We will return to the details and proofs in Chapter 5.

1.6.1. *The fundamental theorem.* We consider the problem of maximizing the objective function

$$l(Q) = \sum_{s=1}^{D} n(s) \ln(L_s(Q))$$

over all distribution functions Q, where

$$L_s(Q) := \int L_s(\phi)\, dQ(\phi)$$

and we have written the likelihood allowing for multiple observations $n(s)$ of a single L_s. We assume that the L_s are all distinct, in that no two arise from identical likelihood kernels. We assume all $n(s) > 0$, but they need not be integers.

These assumptions give D a precise meaning as the *number of distinct summands* in the log likelihood. The only substantial further assumption is that the individual kernels of the likelihood, namely, $L_i(\phi) := f_i(x_i; \phi)$, are both nonnegative and bounded as functions of ϕ. There do exist models with unbounded likelihoods for which the theory is therefore inapplicable without some modification.

Technically, assuring existence does require another two assumptions that will be discussed in Chapter 5, but we know of no cases where they present a genuine difficulty.

PART 1. *Existence and discreteness.* The first part of the theorem states that, under the assumptions above, there *exists* a maximum likelihood estimator (MLE) $\hat{Q}$ that is a *discrete* distribution with no more than D distinct points of support ξ_r. This bound guarantees, of course, that the number of support points is no more than n, the sample size.

The implication of this part of the result is that we can carry out the calculation of this estimator using known techniques from the theory of finite mixture models and that there is an upper bound on the complexity of this distribution.

PART 2. *Gradient characterization.* The second part of the theorem gives us a way of testing whether a given latent distribution, say Q_0, is the MLE. The idea is relatively simple. Suppose we have a distribution Q_0 that is a candidate to be the MLE. We cannot determine if it is the MLE by a direct search process because the space of distributions is infinite dimensional.

However, there is a simple function we can calculate to determine if the solution has been found. We first form a *path* in the space of distribution functions from Q_0 to any other distribution, say Q_1, by letting $Q_\alpha = (1-\alpha)Q_0 + \alpha Q_1$. For every α, this generates an intermediate distribution, with $\alpha = 0$ and 1 corresponding to the original two distributions of interest.

Next, we compute the likelihood along this path, obtaining a one parameter likelihood function $L^*(\alpha) = L(Q_\alpha)$. The derivative of $\ln(L^*(\alpha))$ at $\alpha = 0$ is the *directional derivative* corresponding to this path from Q_0 to Q_1 and [*exercise*] it has the simple form

$$D_{Q_0}(Q_1) = \sum_{i=1}^{D} n(i)\left(\frac{L_i(Q_1)}{L_i(Q_0)} - 1\right).$$

A special case of this derivative occurs when we look along paths where Q_1 is degenerate; that is, a point mass at a single point ϕ, which we will denote Δ_ϕ. We define the *gradient function* to be

$$D_{Q_0}(\phi) := D_{Q_0}(\Delta_\phi) = \sum_{i=1}^{D} n(i)\left(\frac{L_i(\phi)}{L_i(Q_0)} - 1\right). \tag{1.7}$$

Note that

$$D_{Q_0}(Q_1) = \int D_{Q_0}(\phi)\, dQ_1(\phi).$$

[*Exercise.*] This has the important implication that once the gradient function has been determined, we can also determine the value of the directional derivative toward *any* distribution Q_1 by integrating the gradient function.

Next, it is clear that if the gradient function $D_Q(\phi)$ takes on positive values at any ϕ, then the likelihood along the path from Q in the direction of Δ_ϕ

is increasing at Q, so that Q cannot be the maximum likelihood estimator. However, we will show later that a much stronger result holds: that Q is a maximum likelihood estimator *if and only if*

$$D_Q(\phi) \overset{\text{set}}{\leq} 0 \qquad \forall \phi. \tag{1.8}$$

That is to say, all the information we need to find a maximum likelihood estimator is contained in the gradient function.

Moreover, if the *gradient inequality* (1.8) fails for candidate Q at some ϕ_0, not only do we learn that we are not at the maximum, but we also know one way to increase the likelihood—simply move some mass to ϕ_0. Algorithms based on this idea will be discussed in Chapter 6.

PART 3. *Support point properties.* The third part of the theorem regards the location of the support points $\hat{\xi}_j$ of $\hat{Q}$. The result is that if ξ is a support point for any $\hat{Q}$ that maximizes the likelihood, then

$$D_{\hat{Q}}(\xi) = 0.$$

Together with the gradient inequality (1.8) this implies that the support points will be local maxima of the gradient function $D_{\hat{Q}}(\phi)$. One of the consequences of this result is that a gradient-based algorithm need not keep track of the support points in $\hat{Q}$ because they can be recovered from the gradient function at the end of the algorithm. This result is also very useful in proofs of the uniqueness of the MLE $\hat{Q}$.

PART 4. *Uniqueness.* The final main result is that the *fitted values* of the likelihood, namely,

$$\mathbf{L}(\hat{Q}) := (L_1(\hat{Q}), \ldots, L_D(\hat{Q})),$$

are uniquely determined. That is, even if there were two distributions maximizing the likelihood, they would generate the same vector of likelihood fitted values. This elementary geometric result, together with Part 3, can be extended by a much more sophisticated analysis to prove the uniqueness of $\hat{Q}$ within various families of component distributions. See Chapter 5.

AN IMPORTANT REMARK. One the most striking features of the above theory is the complete lack of regularity conditions on the models and the complete generality with regard to the parameter space of ϕ.

1.7. Related nonparametric problems. We earlier introduced some classical nonparametric models with hidden mixture structure. We will use them here to illustrate the workings of the NPMLE of a latent distribution.

1.7.1. *The MLE of an unknown distribution.* We start with the simplest, but still profound, result. Suppose we have a sample $x_1, \ldots, x_n$ from a completely unknown distribution F. Suppose further that we can restrict attention to distributions that are discrete. That we can do so is not obvious, but a number of authors have developed approaches to nonparametric maximum likelihood that lead to this conclusion in this problem. See, for example, Scholz (1980).

With this step taken, the likelihood can be meaningfully written as $L(F) = \prod F(\{x_i\})$, where we use the convention of using the symbol F for both distribution function and measure. If we let $\pi_i = F(\{x_i\})$, then the problem is to maximize $\prod \pi_i$ while maintaining the obvious constraints on the π_i. This can be carried out by using a Lagrange multiplier for the inequality constraint $\sum \pi_j \leq 1$, which results in $\hat{\pi}_j = 1/n$. [*Exercise.*] The resulting distribution is called the *empirical distribution function* and is denoted $\hat{F}$.

The mixture version arises as follows. First, replace F by Q in the notation and let the likelihood kernel function be defined by

$$L_i(\phi) = \mathcal{I}[\phi = x_i].$$

If this is done, then $L_i(Q) = Q(\{x_i\})$ and the mixture likelihood corresponds exactly to the above likelihood. [*Exercise.*]

The latent distribution NPMLE theorem can then be used to prove that the empirical distribution function gives the maximum likelihood estimator. We need only to check the gradient inequality and it is easily verified [*exercise*] that

$$D_{\hat{F}}(\phi) = \begin{cases} 0, & \text{if } \phi = x_i \text{ for some } i, \\ -n, & \text{else.} \end{cases}$$

Note that, in accordance with Part 3, the estimated support points are local maxima of the gradient.

1.7.2. *Accurate and error-prone measurements.* The preceding example leads naturally to the appropriate technique to use when some of the ϕ's are seen directly, say $\phi_1, \ldots, \phi_a$, and others are observed indirectly, through $X_i | \Phi = \phi_i$. This occurs, for example, in some measurement error problems, where in order to ascertain the level of the measurement error, on a small subsample one takes much more accurate measurements (presumably also more expensive and/or time-consuming). The observed ϕ's are then assumed to be the gold standard whose relationship to response variable Y is desired. In such a case, as we have earlier indicated, we can write the likelihood in two parts, one part being from the gold standard measurements and so it has indicator functions for the likelihood kernel, as in the preceding example, and the second part has the mixture form corresponding to the density of X given Φ. See Roeder, Carroll and Lindsay (1993).

1.7.3. *Monotone density problems.* We have already indicated that the class of nonincreasing density functions on $[0, \infty)$ is a model with hidden mixture structure.

It is well known that the nonparametric MLE of such a nondecreasing density function can be characterized as having a compound distribution function (CDF) $\hat{G}$ that is the "least concave majorant" of the empirical distribution function [e.g., Groeneboom and Wellner (1992)]. That is, $\hat{G}$ is that concave function whose graph lies above that of $\hat{F}$, but is closest to it. The graph is piecewise linear; its points of contact with $\hat{F}$ are the points where it bends and these points are some subset of the observations.

It can be checked that this description implies that the nonparametric MLE of the latent distribution Q has its mass at some subset of the observations and so the mixture solution has the representation $\sum \pi_i U(0, x_i)$, where U indicates a uniform distribution on the specified range.

The gradient characterization of this problem can be used to obtain this solution. The reader may find it a helpful *exercise* to do so, following this line of argument. First, by examining the gradient it can be determined that it must have all its local maxima at the observations x_i, so, by Part 3 of the theorem, the support points must be among this set. Next, it can be shown that if ϕ_1 is the smallest support point of the latent distribution, then $f(x; \hat{Q})$ must be constant on the interval $(0, \phi_1)$, and the gradient inequality on the interval $[0, \phi_1]$ implies that $\hat{F}(t) \leq F(t; \hat{Q})$, for $t < \phi_1$, with equality at $t = \phi_1$ (since it is a support point). Since $F(t; \hat{Q})$ is linear on this interval and must be concave overall, it is clear that this defines the first support point as corresponding to that point $(x_i, \hat{F}(x_i))$ first intersected by a ray from the origin that is rotated from the y axis toward the x axis. Thus we have shown that the solution "majorizes" the empirical distribution function on this first interval. We can then continue to the next support point ϕ_2 and slightly modify this argument to show majorization over the interval $[\phi_1, \phi_2]$, with equality at the endpoints.

1.7.4. *Censoring problems.* Another important class of nonparametric problems that have hidden mixture structure arise in censoring problems. For example, consider the problem of finding the distribution function F that maximizes the likelihood $L(F) = \prod F(\{x_i\}) \prod F([c_j, \infty))$. This is the likelihood that arises under so-called noninformative right censoring, in which the x_is correspond to observed lifetimes, but all that is known about the observations in the second set is that they fell to the right of the censoring values c_j.

We can turn this into a mixture problem, as before by using indicator functions, where we now use $\mathcal{I}[\phi = x_i]$ for the observed data and $\mathcal{I}[\phi \geq c_j]$ for the censored data. Once again we can show that the gradient function has all its local maxima for ϕ in the observed data set, so the support points can be restricted to this set. The solution is the product limit estimator, also known as the Kaplan–Meier estimator. [*Exercise*: Use the gradient characterization to derive this result.]

Jewell, Malani and Vittinghoff (1994) have shown that the mixture NPMLE theorem can be used in much more complicated interval censoring problems that arises in various AIDS studies.

1.8. Similar statistical problems. As a final note to this discussion of the vast range of the mixture problem, it should also be pointed out that there are still more statistical areas that are closely related mathematically and so they carry techniques and theory that are relevant to the study of mixture models.

The mixture NPMLE theorem is, in a mathematical sense, simply a restatement, with statistical interpretation, of a basic result in the maximization of a concave objective function over a convex set. The theory of optimal design [Silvey (1980)] hinges on exactly such an optimization and it has a theorem of exactly the same form, but with other interpretations. The algorithmic literature from optimal design theory can be carried directly over to the mixture problem with slight modification.

The theory of order-restricted inference [Robertson, Wright and Dykstra (1986)] also has large areas of overlap. For example, the monotone density problem of the previous section is an example of an estimation problem carried out under a order restriction. Those restrictions often can be expressed in a way involving convexity, and the estimation problem again relates to finding the minimum or maximum of a functional over a convex set. Some of the relationships between the two will be made clearer when we deal with the likelihood ratio problem in depth in Chapter 4.