CHAPTER 2

# Structural Features

This chapter is devoted to developing a mathematical understanding of the structures that are inherent to the mixture model, ranging from the simple properties of moments up to rather complicated features of exponential family mixtures. Sections 2.1 and 2.2 contain the material of greatest practical importance because they address features of the mixture model useful for diagnostic purposes. The material thereafter is very important for understanding the issues of identifiability of the latent distribution $Q$, but can be skimmed and returned to as needed for the later chapters.

## 2.1. Descriptive features.

2.1.1. *Some simple moment results.* One of the nicest mathematical features of the mixture model is the simple way in which the latent distribution $Q$ enters into the calculation of expectation. Simply by reordering the order of integration (or summation), we obtain the fact that if $t(x)$ has expectation $\tau(\phi)$ under the unicomponent model $f(x;\phi)$, then it has expectation $\int \tau(\phi)\, dQ(\phi)$ under the mixture model $f(x;Q)$. This is easily shown using the latent variable $\Phi$:

$$E[t(X);Q] = E[E[t(X)|\Phi]] = E[\tau(\Phi)].$$

Using the latent variable also simplifies the calculation of variances under the mixture model:

$$\text{Var}[t(X);Q] = \text{Var}(E[t(X)|\Phi]) + E(\text{Var}[t(X)|\Phi]). \tag{2.1}$$

To illustrate these formulas, suppose that $X$ comes from a mixture of Poisson densities with mean parameter $\phi$. Then the following simple relationships between the marginal mean and variance of $X$ and the latent variable $\Phi$ hold:

$$\begin{aligned} E(X;Q) &= E[\Phi], \\ \text{Var}(X;Q) &= \text{Var}[\Phi] + E[\Phi]. \end{aligned}$$

[*Exercise.*] Manipulation of these equations then shows that the variance of $X$ in a Poisson mixture model is inflated, compared to a unicomponent Poisson

model, in the sense that the *variance-to-mean ratio* is inflated to a value larger than the ratio 1 of the unicomponent model.

Examination of the variance formula (2.1) shows that there is a general sense in which the mixture model will create extra variation over the unicomponent models that generate it. This section of this chapter is devoted to two important diagnostic features of the mixture model related to the way the extravariability will show up in the observable distribution of $X$:

- There is a tendency for the presence of the mixture model to be evident in the form of multimodality.
- A comparison of a multicomponent mixture distribution with the unicomponent model yields a very strong form of stochastic ordering between the two, related to the heavier tails the mixture distribution will have.

2.1.2. *Shape and modality.* We have already seen that in the normal mixture model, having two components is not synonymous with having two modes. For more details on the exact conditions under which the two-component normal mixture is bimodal, see Robertson and Fryar (1969).

Thus if we were to examine a histogram of data that were unimodal, we could not discard the possible presence of two or more normal components mixed together. However, by considering ratios of densities, we can greatly increase the sensitivity of plots. Suppose that

$$g_2(x) = \bar{\pi} n(x; \phi_1, \sigma^2) + \pi n(x; \phi_2, \sigma^2)$$

is a two-component normal mixture density, with

$$\text{mean } E[X] = E[\Phi] \text{ and variance } \operatorname{Var}(X) = \sigma^2 + \operatorname{Var}[\Phi].$$

Let $g_1(x) = n(x; E[X], \operatorname{Var}(X))$ be the unicomponent normal density with the same mean and variance as the two-component mixture. Lindsay and Roeder (1992b) show that the ratio $g_2(x)/g_1(x)$ is *always* bimodal. Moreover, as $x$ goes from $-\infty$ to $+\infty$, the centered ratio

$$\frac{g_2(x) - g_1(x)}{g_1(x)}$$

will have the sign sequence $(-, +, -, +, -)$, reflecting four crossing points of the density functions $g_1$ and $g_2$.

The remarkable feature of this result is that the bimodality and sign changes occur even if the component densities are arbitrarily close together, and no matter the magnitude of $\pi$.

We will not give a complete proof of this result here nor a detailed look at how one can use this for a diagnostic in the normal problem. For the latter, see Roeder (1994). The complete proof relies on results in totally positive kernels and some properties of moment-matched distributions. The interested reader will find that the technical material later in this chapter provides much of the background needed for understanding the proof.

2.1.3. *Overdispersion and sign changes.* It is very common to find that the component densities used in a mixture model come from an exponential family. This is fortunate because there is much that can be said about the structural features of such exponential family mixtures. We have already noted that in the normal family we can be very precise about the relationship between the densities of the two-component and one-component models that have the same mean and variance. We will now develop similar results for the *one parameter exponential family*.

We suppose that the component densities have the form

$$f(x;\phi) = \exp(\phi x - \kappa(\phi)) \tag{2.2}$$

with respect to some supporting measure $dF_0(x)$ on $\mathbf{R}^1$. The supporting measure contains all parts of the density not depending on $\phi$ and can be chosen to be one of the members of the family.

For the reader unfamiliar with these representations, we give an example of a statistical model that can be put into this canonical form. For example, in the binomial distribution with sample size parameter $n$ and success parameter $p$, which we denote $\mathrm{Bin}(n, p)$, we can let $dF_0(x)$ be the binomial distribution with $p = 0.5$. To write the density in the canonical form, $\phi$ is the log odds parameter $\ln[p/(1-p)]$ and

$$\kappa(\phi) = n\ln(1-p) - n\ln(2).$$

Shaked (1980) established a number of important properties of exponential family mixtures. For a given latent distribution $Q$, suppose that $f(x;\phi_0)$ is the unicomponent model that has the same mean for $X$ as does the mixture distribution; that is, $E[X;\phi_0] = E[X;Q]$. Define the *ratio function* by

$$R(x) = \frac{f(x;Q)}{f(x;\phi_0)}. \tag{2.3}$$

The key results are:

- $R(x)$ is a convex function of $x$.
- $R(x) - 1$ has the sign sequence $(+,-,+)$ as $x$ traverses the real axis.

We will prove these results in the next subsection.

These properties show that the mixture density has heavier tails than the mean-matched unicomponent model. Shaked used the convexity result to show a type of stochastic ordering between the multicomponent distribution and the corresponding unicomponent distribution:

$$\text{the distribution } F_Q \text{ is a dilation of } F_{\phi_0}.$$

Here we say that a distribution $G$ is a *dilation* of distribution $F$ if $\int x\,dF(x) = \int x\,dG(x)$ and if, for every convex function $c(x)$,

$$\int c(x)\,dG(x) \geq \int c(x)\,dF(x).$$

Notice that Shaked's dilation result implies directly the overdispersion result

$$\mathrm{Var}(X; Q) \geq \mathrm{Var}(X; \phi_0).$$

[*Exercise.*]

We note further that $G$ is a dilation of $F$ if and only if they have the same mean and there exists a family of distributions $K(y|x)$, with $\int y\, dK(y|x) = x$, such that $G(y) = \int K(y|x)\, dF(x)$. In our case, this means that there exists $K(y|x)$ such that

$$F_Q(y) = \int K(y|x)\, dF_{\phi_0}(x).$$

Thus in the mixture case we can think of the mixture variable $Y$ as being generated in two steps: first generate $X = x$ from the corresponding unicomponent model, then dilate it by generating $Y$ from a kernel distribution with mean $x$. This, curiously, reverses the original representation, putting the unicomponent distribution in the role of the latent distribution and $K$ in the role of the component density family.

As an illustration of how these dispersion properties show up in data, we consider the data set in Table 2.1, which identifies the number of male children in 6115 sibships of size 12, collected in Saxony, Germany [Geissler (1889)].

Given a parenting couple, we might model their children as being born with independent sex determination, like a sequence of Bernoulli trials with some constant probability $p$ of having a male child. If so, the number of male children $X$ in a family with 12 children would be distributed as a Bin(12, $p$) random variable. In this context it is natural to ask if the probability of a male birth $p$ is a latent variable, varying from family to family, or is constant across families. If $p$ does vary, then we could associate with each couple a

TABLE 2.1
*Number of male children in sibships of size* 12

| # Males | Obs. Count | Obs. vs. Fit | Bin. Fit |
|---|---|---|---|
| 0 | 3 | > | 0.9 |
| 1 | 24 | > | 12.1 |
| 2 | 104 | > | 71.8 |
| 3 | 286 | > | 258.5 |
| 4 | 670 | > | 628.1 |
| 5 | 1033 | < | 1085.2 |
| 6 | 1343 | < | 1367.3 |
| 7 | 1112 | < | 1265.6 |
| 8 | 829 | < | 854.3 |
| 9 | 478 | > | 410.0 |
| 10 | 181 | > | 132.8 |
| 11 | 45 | > | 26.1 |
| 12 | 7 | > | 2.3 |
| Total | 6115 | | 6115.0 |

latent parameter $p_i$ representing their propensity to have male children. The result is that the data would be from a mixture of binomials.

Table 2.1 also shows the expected values of those counts, assuming that they arose from sampling from a unicomponent binomial distribution Bin(12, $p$), where $p$ was estimated from the sample to be 0.51. It is clear that the observed distribution has heavier tails than would be expected from the unicomponent binomial model and, moreover, that the difference *observed – expected* has the sign change behavior (+, −, +) predicted under the above results for a mixture model.

2.1.4. *Log convexity of ratios.* Establishing the first part of Shaked's results regarding the convexity of $R(x)$ is simple and instructive. Indeed, we can just as easily show a stronger result, that

$$\ln(R(x)) \text{ is a convex function of } x. \tag{2.4}$$

(Why is this stronger?) We learn this by examining the structure more closely. We can write

$$\begin{aligned} R(x) &= \frac{\int \exp(\phi x - \kappa(\phi))\, dQ(\phi)}{\exp(\phi_0 x - \kappa(\phi))} \\ &= \int \exp((\Phi - \phi_0)x)\, d\Gamma(\Phi). \end{aligned} \tag{2.5}$$

Here the positive measure $\Gamma$ is defined by

$$d\Gamma(\phi) = \exp(\kappa(\phi) - \kappa(\phi_0)) \cdot dQ(\phi).$$

This demonstrates that $R(x)$ has the mathematical structure of a *Laplace transform*, that is, it is a scalar multiple of a moment generating function. Since this means its logarithm is a scalar translation of a cumulant generating function, this implies that $\ln(R(x))$ is a convex function of $x$.

(*Technical note*: In an exponential structure model, such as the binomial, it is possible that there exist distributions that correspond to infinite values of the natural parameter $\phi$. Thus, when the binomial success parameter $p$ is 0, the natural parameter $\phi$ is $-\infty$. Including these points in the analysis would involve technical difficulties that we would rather avoid here, so we will assume $Q$ assumes no mass at infinity, although it is not usually necessary to do so.)

Indeed, we gain further insight by differentiating twice, to find that

$$\begin{aligned} \frac{d}{dt}\ln(R(t)) &= E[(\Phi - \phi_0)|X = t], \\ \frac{d^2}{dt^2}\ln(R(t)) &= \mathrm{Var}(\Phi|X = t). \end{aligned}$$

[*Exercise.*] That is, differentiation of the log ratio function generates the cumulants of the posterior distribution of the latent variable, given the observation.

Now the convexity of $\ln(R(x))$ implies the convexity of the centered ratio function

$$R(x) - 1 = \frac{f(x;Q) - f(x;\phi_0)}{f(x;\phi_0)},$$

so we have only to prove the sign change result, which can be done by showing that the difference $f(x;Q) - f(x;\phi_0)$ has the sign pattern $(+,-,+)$. That is, the convexity of the function $R$ shows that there can be at most two crossings of zero, so what remains is to show that there are exactly two. Although this can be done directly, we instead take a diversion into much stronger results relating moments and sign changes. The reader may wish to skim the following section on first passage.

2.1.5. *Moments and sign changes.* The following striking result relates moments and the sign crossing properties of density functions.

PROPOSITION 1. *Suppose $f$ and $g$ are two density functions on $\mathbf{R}$ with supporting measure $d\mu(x)$ and possessing the same first $M$ moments $E(X^k)$, $k = 1,\ldots,M$. Then the difference between the densities,*

$$\Delta(x) := f(x) - g(x),$$

*has at least $M+1$ sign changes, unless the distributions are identical.*

PROOF. Suppose not and that there are $K$, with $K \le M$, nodes $t_1,\ldots,t_K$ such that $\Delta(x)$ has a constant sign between nodes, with signs alternating on adjacent internodal intervals. We can construct a polynomial

$$p(x) = \pm(x - t_1)\cdots(x - t_K)$$

of degree $K$ that has the same sign between nodes as does $\Delta(x)$. Hence $p(x)\cdot\Delta(x) \ge 0$. However, if we integrate this function, the equality of the first $K$ moments implies we get zero. The conclusion is that $p(x)\Delta(x) = 0$, almost everywhere $d\mu$. This shows that $\Delta(x)$ is zero ($\mu$ a.e.) between the nodes. This indicates that $f$ and $g$ can yield different probability measures only if they have a discrete component that differs only on the nodes. However, we can then apply the following lemma, which shows that two discrete densities having a common support set of $K$ points, with their first $K$ moments matching, must be identical. □

In addition to its use in the above proof, the following result is very important in the theory of the method of moments in mixture distributions [Lindsay (1989a)]. It will be useful in this chapter to have a special notation for the construction of a vector consisting of the powers of a basic variable, so we define

$$\mathbb{x}^\dagger := \begin{pmatrix} 1 \\ x \\ \vdots \\ x^K \end{pmatrix}.$$

LEMMA 2. *Suppose that $F$ and $G$ are two distributions with support on some fixed set of $K+1$ (or fewer) points $\{t_0, t_1, \ldots, t_K\}$. Further, suppose that they match in their first $K$ moments,*

$$\int \mathbf{x}^\dagger \, dF(x) = \int \mathbf{x}^\dagger \, dG(x).$$

*Then $F = G$ and the masses at the support points are given by the matrix equation* (2.6) *below.*

PROOF. For any given set of moments $\int \mathbf{x}^\dagger \, dF(x)$ on a known support set $\{t_0, \ldots, t_K\}$, the corresponding masses $\pi$ are determined by the matrix equation

$$\int \begin{pmatrix} 1 \\ x \\ \vdots \\ x^K \end{pmatrix} dF(x) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_0 & t_1 & \cdots & t_K \\ \vdots & \vdots & \vdots & \vdots \\ t_0^K & t_1^K & \cdots & t_K^K \end{bmatrix} \begin{pmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \pi_K \end{pmatrix}. \tag{2.6}$$

Thus if the square matrix in the equation is invertible, we have a unique solution. However, this is the well known *Vandermonde matrix*, with determinant known to be $\prod_{i>j}(t_i - t_j)$. □

Finally, we can turn the result concerning density functions into a result regarding distribution functions by using integration by parts.

PROPOSITION 3. *If two distribution functions $F$ and $G$ have their first $K$ moments in common, then either they are equal or the difference $F - G$ has at least $K$ sign changes.*

The proof is an *exercise*. See Lindsay and Roeder (1992b) for more details.

If we now return to the problem of the sign change behavior of $R(x)$, we see that since we have two densities with the same first moment, they must cross two times and the sign pattern must be $(+, -, +)$. We can additionally say that the distribution functions cross exactly once, with sign sequence $(+, -)$.

2.1.6. *Dispersion models.* Lindsay (1986) investigated in some detail the construction of parametric mixture models that had exponential family structure. We briefly survey these and related results.

One of the important uses of a mixture model is as a means of allowing for *overdispersion*. As a strategy for dealing with overdispersion, one might consider constructing an exponential family model that contains the model of interest, but contains an additional parameter to account for overdispersion. As such, consider the two parameter exponential families with densities, with respect to $dF_0(x)$, of the form

$$f(x; \alpha, \beta) = \exp(\alpha x + \beta t(x) - K(\alpha, \beta)),$$

for some function $t(x)$. Note that $\beta = 0$ generates the original model. We then ask, which functions $t(x)$ will cause this model to be overdispersed relative to the original model?

The preceding results indicate that if $f(x; \alpha, \beta)$ is to be a mixture model in the sense that there is some $Q$ depending on $\alpha$ and $\beta$ with $f(x; Q) = f(x; \alpha, \beta)$, then the log ratio function must be convex, from which it follows that $t(x)$ must be convex. [Lindsay (1986) gives necessary and sufficient conditions on $t(x)$.]

Gelfand and Dalal (1990) took this idea one step further and showed that if $t(x)$ is convex, even if the resulting density is not a mixture density, it still is an overdispersed density relative to the unicomponent model, in the strong sense of dilation. One example of an overdispersion model that is not necessarily a mixture model is Efron's double exponential family [Efron (1986)].

However, the more usual method for the construction of a two parameter overdispersed family is to use the *conjugate distribution*, a subject we will introduce in the next chapter.

**2.2. Diagnostics for exponential families.** We now turn to using the above insights for diagnostic purposes. We ask the question, for a given set of data, does the mixture model fit and, if so, do we need to use more than one component? We use the information that the ratio function is convex and also log convex. For more details on this section, see Lindsay and Roeder (1992a).

2.2.1. *Empirical ratio plots.* In a discrete sample space, with observed proportions $\hat{p}(t)$, it is natural to attempt to estimate the ratio function $R(t)$ [see (2.3)] by its empirical counterpart:

$$\hat{R}(t) = \frac{\hat{p}(t)}{f(t; \hat{\phi})}.$$

Here $\hat{\phi}$ is the maximum likelihood estimator of $\phi$ in the unicomponent model. Such an estimation clearly relies on the sample size being sufficiently large that $\hat{p}(t)$ is a good estimator of the true density. If the unicomponent model is correct, then the empirical ratio converges to 1 for every $t$. If the alternative of a mixture model is correct, then the empirical ratio will converge to the *convex* ratio function $R(t)$.

This suggests plotting $(t, \hat{R}(t) - 1)$ or $(t, \ln(\hat{R}(t)))$ and examining the plot for convexity. If the plot is clearly nonconvex, then the mixture model cannot possibly fit well. If the plot is nearly linear, then the unicomponent model is likely to fit well and strict convexity is diagnostic for mixture structure. We note that convexity is a very particular prediction for the shape of the plot, and it is something easy to identify visually. Such an empirical ratio plot is given for the data of Table 2.1 in Figure 2.1.

2.2.2. *Gradient function plots.* A second diagnostic method for looking for mixture structure is to consider the gradient function $D_Q(\phi)$ introduced in Chapter 1. One advantage to this approach is that it enables us to create plots similar to the empirical ratio plot even when the data are not discrete.

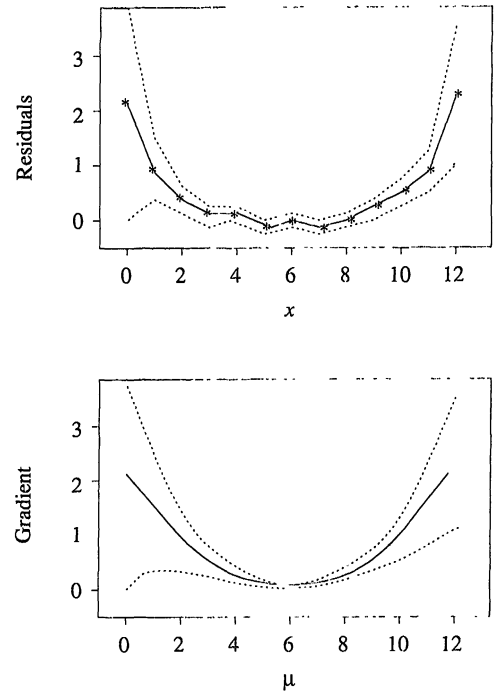


FIG. 2.1. *The ratio and gradient plots for the sibship data.*

We evaluate the gradient at $Q = \Delta_{\hat{\phi}}$, the best fitting one-component model, calling

$$D_1(\phi) := D_{\Delta_{\hat{\phi}}}(\phi)$$

the *unicomponent gradient function*. We know from the NPMLE theorem that this gradient function is diagnostic for whether or not the degenerate distribution $\Delta_{\hat{\phi}}$ is the maximum likelihood estimator of the latent distribution $\hat{Q}$. In fact, if

$$D_1(\phi) \leq 0 \qquad \forall \phi,$$

then the one-component model fits better, in the sense of higher likelihood, than any mixture model with *any* other number of components. On the other hand, if the gradient inequality is violated, we know we can increase the likelihood by adding components. This suggests that examination of $D_1(\phi)$ may be a useful diagnostic for the presence of a mixture alternative to a unicomponent model.

It is somewhat surprising that we can be precise about the predicted shape of the graph of the unicomponent gradient under the mixture model. The reason for this is most clear in the discrete case, where we can draw a close relationship between the unicomponent gradient $D_1$ and the empirical ratio

function $\hat{R}(t)$:

$$
\begin{aligned}
D_1(\phi) &= \sum_t n(t)\left(\frac{f(t;\phi)}{f(t;\hat{\phi})} - 1\right) \\
&= n\sum f(t;\phi)\left(\frac{\hat{p}(t)}{f(t;\hat{\phi})} - 1\right) \\
&= n\sum f(t;\phi)\cdot[\hat{R}(t) - 1].
\end{aligned} \tag{2.7}
$$

That is, the unicomponent gradient function can be represented as a kernel-smoothed version of $[\hat{R}(t) - 1]$, where the smoothing kernel is $f(t;\phi)$. This suggests that since the empirical ratio function is asymptotically convex under the mixture model, then perhaps the gradient function could be as well.

To make this statement completely correct, we have to find the correct reparameterization of $\phi$ in which to plot the gradient. In order to make the result more general, we drop the multinomial assumption and consider the asymptotic limit of the unicomponent gradient function when the mixture model is correct. Check that if $Q$ is the latent distribution and $\hat{\phi} \to$ some $\phi_0$, depending on $Q$, that

$$
\begin{aligned}
n^{-1}D_1(\phi) \to D^*(\phi) &:= \int \frac{f(x;\phi)}{f(x;\phi_0)} f(x;Q)\,dF_0(x) - 1 \\
&= \int R(x)f(x;\phi)\,dF_0(x) - 1.
\end{aligned}
$$

Recalling that the ratio function $R(x)$ is a convex function of $x$, we then ask: When will a convex function $R(x)$, smoothed by a kernel $f(x;\phi)$, yield a convex function $D^*(\phi)$? When $f(x;\phi)$ is an exponential family, it will happen when the parameter $\phi$ is the mean value parameter of the exponential family. That is, if we replace $\phi$ with $\mu(\phi) = E[X;\phi]$, then

$$\text{a plot of } (\mu, D^*(\mu)) \text{ is convex.}$$

The proof of this is relatively simple, but requires high powered results from total positivity: see Lindsay and Roeder (1992a) for details, including the calculation of statistical error bounds for the plots.

The gradient plot for the sibship data can be found in Figure 2.1, together with error bounds (calculated pointwise). Notice that it appears very much to be a smoothed version of the residual plot.

Further justification for examining the residual plot can be given by considering the *normalized gradient function*, where we divide the gradient function by its asymptotic standard error under the unicomponent model. In Chapter 4 we will show that the likelihood ratio test statistic for one component versus two components is asymptotically equivalent to the square of the maximum of the normalized gradient function, so that, in terms of the gradient plot, the likelihood ratio test is equivalent to rejecting the unicomponent model if the gradient crosses the upper confidence line. Methods for adjusting the critical value for simultaneous inference are given in Chapter 4.

TABLE 2.2
*Number of male children in sibships of size 8*

| # Males | Obs. Count | Obs. vs. Fit | Bin. Fit |
|---|---|---|---|
| 0 | 215 | > | 165.22 |
| 1 | 1,485 | > | 1,401.69 |
| 2 | 5,331 | > | 5,202.65 |
| 3 | 10,649 | < | 11,034.65 |
| 4 | 14,959 | > | 14,627.60 |
| 5 | 11,929 | < | 12,409.87 |
| 6 | 6,678 | > | 6,580.24 |
| 7 | 2,092 | > | 1,993.78 |
| 8 | 342 | > | 264.30 |
| Total | $n = 53,680$ | | 53,680.00 |

2.2.3. *Comparing gradient and ratio plots.* The residual plot requires a larger data set because it has no smoothing feature, but as a consequence it reveals more structure. One striking example of this occurs when we turn to the sibship data for families of size 8, presented in Table 2.2.

This table is based on a much larger sample than Table 2.1, and it shows a striking lack of fit of the mixture model in that there is too much deviation from convexity, relative to the standard error bounds, as can be seen from the ratio residual plot in Figure 2.2.

These data were examined by Fisher (1925), who said:

> The observed series differs from expectation markedly in two respects: one is the excess of unequally divided families; the other is the irregularity of the central values, showing an apparent bias in favor of even values. No biological reason is suggested for the latter discrepancy, which therefore detracts from the value of the data.

We note that the gradient plot (Figure 2.2) captures the overdispersion, but smooths out the fine structure. Such a large data set with interesting fine detail is probably fairly unusual, so it might be anticipated that the gradient, with its smoothness, is more generally the appropriate tool. In addition, it is a natural by-product of the nonparametric approach.

Extensions of these diagnostic ideas into the domain of generalized linear models has been carried out by Lambert and Roeder (1995).

**2.3. Geometry of multinomial mixtures.** We have already introduced, in Chapter 1, an important subclass of the mixture model where the component densities are treated as known. We now start giving a more detailed picture of the mixture model by examining the geometric structure in these, the simplest of mixture models. This will lead to a better understanding of several statistical issues, such as the identifiability of the latent distribution and how mixture models with different numbers of components are related.

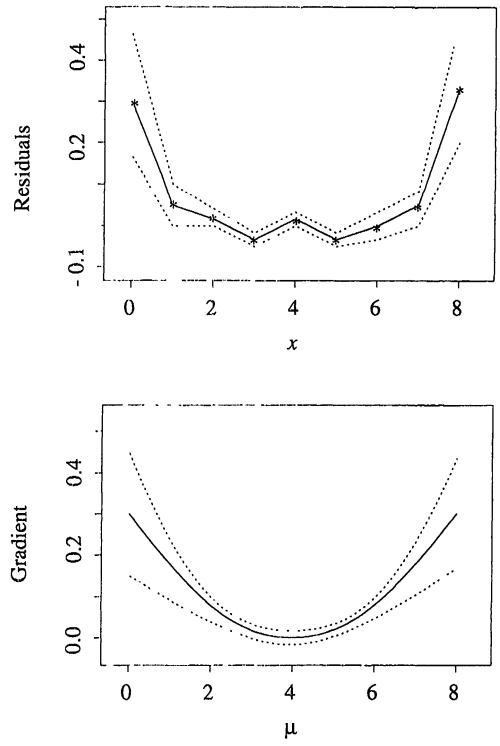


FIG. 2.2. *The ratio and gradient plots for the second sibship data.*

2.3.1. *Known component densities.* The setting is the multinomial density, with sample space $\{0, 1, \ldots, T\}$. We start with a fixed set of known component densities, say $f_1(t), \ldots, f_m(t)$, and consider mixtures of the form $f(t; Q) = \sum \pi_j f_j(t)$. We start by turning each component density function into a vector, by setting

$$\mathbf{f}_j := \begin{pmatrix} f_j(0) \\ \vdots \\ f_j(T) \end{pmatrix},$$

calling this the *density vector* for component $j$. Note that the entries of the vector are nonnegative and sum to 1.

Next, we need a definition. Let $\mathbf{v}_1, \ldots, \mathbf{v}_m$ be vectors in $(T+1)$-dimensional Euclidean space $\mathbf{R}^{T+1}$. If $\pi_1, \ldots, \pi_m$ is a set of nonnegative weights summing to 1, then the linear combination

$$\pi_1 \mathbf{v}_1 + \cdots + \pi_m \mathbf{v}_m$$

is called a *convex combination* of $\mathbf{v}_1, \ldots, \mathbf{v}_m$.

The fundamental result that gives power to a geometric analysis of the mixture model is extremely simple: *mixture density vectors are convex combinations of the component density vectors.*

For example, for three components with latent masses $\pi_1, \pi_2, \pi_3$, the mixture density

$$f(t; Q) = \pi_1 f_1(t) + \pi_2 f_2(t) + \pi_3 f_3(t)$$

has the vector representation

$$\mathbf{f}_Q = \pi_1 \mathbf{f}_1 + \pi_2 \mathbf{f}_2 + \pi_3 \mathbf{f}_3.$$

Moreover, if we define the matrix $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3]$, then the above relationship can be expressed as a matrix equation:

$$\mathbf{f}_Q = \mathbf{F}\pi.$$

Thus we have a linear model, with constraints, in the parameters $\pi_j$ for the mixture multinomial probabilities $\mathbf{f}_Q$.

In such a model, with known components, the mixture is *identifiable* if we can determine the values of $\pi$ given the values of $\mathbf{F}$ and $\mathbf{f}_Q$. Although such a question can be addressed directly through the theory of matrices, we believe it is more insightful to use convex geometry.

2.3.2. *Basic convex geometry.* A *convex set* $\mathbf{C}$ is a set of vectors that contain every finite convex combination of its elements. Pictorially, a convex set contains all the lines connecting any two points of the set. Given a set of vectors $\mathbf{V}$, the *convex hull* of $\mathbf{V}$, denoted conv($\mathbf{V}$), is the smallest convex set containing $\mathbf{V}$.

Of particular interest to us is the case when the vectors are multinomial density vectors. That is, the vectors $\mathbf{v}$ have nonnegative entries, with entries summing to 1. Such vectors live in the *probability simplex*

$$\mathbf{P}^T = \{\mathbf{p}\colon\ \mathbf{p}'\mathbf{1} = 1,\ \mathbf{p} \geq \mathbf{0}\},$$

a convex set with dimension $T$ in $\mathbf{R}^{T+1}$. (The set's dimension is reduced by 1 due to the linear constraint on the coordinates, $\sum p_j = 1$.) The probability simplex can be represented as the convex hull of $(\mathbf{e}_0, \ldots, \mathbf{e}_T)$, where $\mathbf{e}_j$ is defined to be that $(T+1)$-vector with a one in position $j$ and zeros elsewhere. That is, the $\mathbf{e}_j$ are the usual basis vectors for Euclidean space. [*Exercise*: Sketch $\mathbf{P}^1$ and $\mathbf{P}^2$.]

We examine $\mathbf{P}^2$, a two-dimensional surface in $\mathbf{R}^3$. In Figure 2.3 we show this surface, rotated about so that it lies in the plane of the page. In addition, we show the location of the basis vectors $\mathbf{e}_j$ and a set of three multinomial component density vectors $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$. From the above discussion, it is clear that the set of mixture density vectors

$$\mathcal{M} := \{\mathbf{f}_Q\}$$

is the convex hull of the set $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$. The hatched set in the figure represents the possible mixture density vectors allowed under this model.

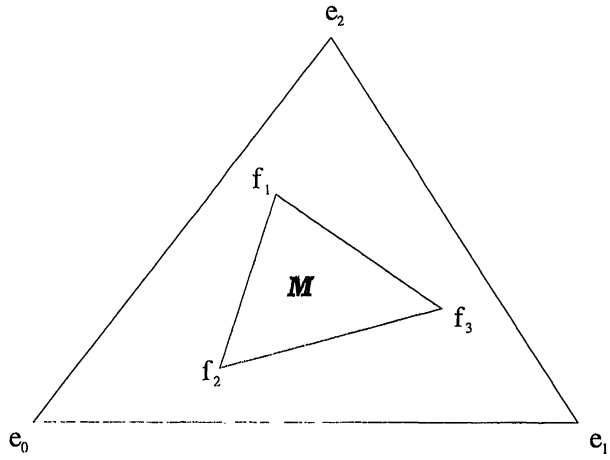


FIG. 2.3. *The two-dimensional probability simplex.*

2.3.3. *Identifiability of weight parameters.* We return to the question of identifiability. In essence, the weight parameters are identifiable if we can solve uniquely for them from the distribution of the observables, in our case from the density $f(t; Q)$ or the vector $\mathbf{f}_Q$. In a general linear problem, we can solve uniquely for the parameters $\pi$ in $\sum \pi_j \mathbf{f}_j = \mathbf{f}_Q$ if and only if the vectors $\mathbf{f}_1, \ldots, \mathbf{f}_m$ are linearly independent. If we add the mixture model requirement that $\sum \pi_j = 1$, then the uniqueness of the solution is guaranteed under the weaker condition of *affine independence*. However, this concept need not concern us here, because when all the vectors $\mathbf{f}_j$ involved are in the probability simplex, the affine independence of the vectors is equivalent to their linear independence.

The reader should consider the geometric consequences of this. For example, in Figure 2.3 the vectors are linearly independent. A set of three density vectors that were not linearly independent would lie on a line, as in Figure 2.4. (Remember that one dimension is missing from the plot and the origin is not pictured.)

One simple conclusion from such an identifiability analysis is that the weight parameters $\pi$ cannot be identifiable for all mixtures $\mathbf{f}_Q$ if the number of components $m$ is greater than the number of multinomial categories $T + 1$. That is,

$$m > T + 1 \quad \Longrightarrow \quad \text{weights not identifiable.} \tag{2.8}$$

In Figure 2.5 the reader should visualize why mixtures of four density vectors give nonidentifiable weights.

We note, however, that even if the weights $\pi$ cannot all be identified, there may be identifiable linear combinations of scientific interest. See Roeder, Devlin and Lindsay (1989). Also, as we shall see, it is possible to have unique estimates of these parameters even when identifiability fails.

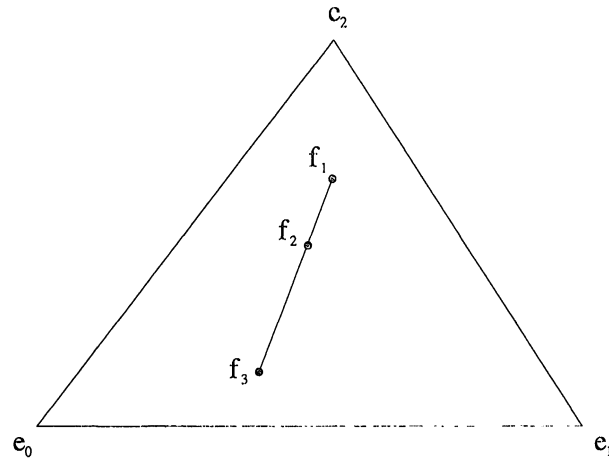


FIG. 2.4. *A mixture set with nonidentifiable weights.*

2.3.4. *Carathéodory's theorem.* One result from convex geometry that will be extremely useful to us is a classic theorem about the representation of elements of the convex hull of $\mathbf{V} \subset \mathbf{R}^K$ in terms of convex combinations of the elements of the generating set $\mathbf{V}$. *Carathéodory's theorem* says that if $\mathbf{u} \in \text{conv}(\mathbf{V})$, then there exists at least one representation of $\mathbf{u}$ as a convex combination of $K + 1$ or fewer elements of $\mathbf{V}$, say

$$\mathbf{u} = \pi_1 \mathbf{v}_1 + \cdots + \pi_{K+1} \mathbf{v}_{K+1},$$

for some $\mathbf{v}_1, \ldots, \mathbf{v}_{K+1}$ in $\mathbf{V}$.

Since the density vectors involved lie in a $T$-dimensional subspace, it follows that there exist representations of a multinomial mixture model vector $\mathbf{f}_Q$ in terms of the convex combination of some set of $T + 1$ or fewer components $\mathbf{f}_i$,

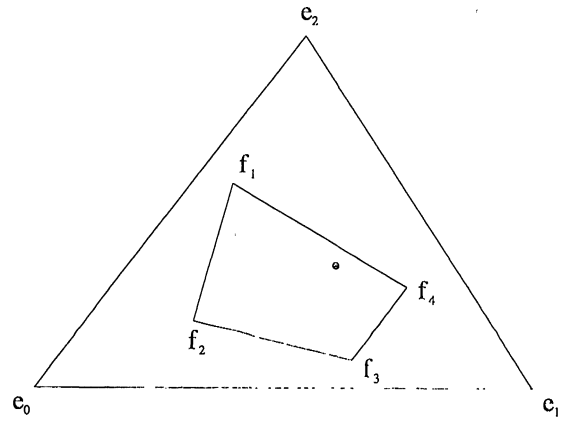


FIG. 2.5. *A second mixture set with nonidentifiable weights.*

whether or not the mixture is identifiable. As an exercise in visualization, the reader should consider the geometric truth of this theorem for models in $\mathbf{P}^2$, where it says that every point in a convex hull can be represented by three or fewer elements of its generating set.

**2.4. Exponential family geometry.** We now consider the question of identifiability in the one parameter exponential family model (2.2). For a multinomial exponential family, the range of the random variable $X$ will typically be a subset of the integers of the form $\{0, 1, 2, \ldots, T\}$, with $T$ possibly infinite. If this occurs, we will say we are in the *lattice case* and we will have special results when such structures exist.

The previous section gave us some preliminary insights into this problem, and we already know that if the range of $X$ is finite, then there is no hope of solving uniquely for $Q$ if it has too many support points. However, there is certainly information about the latent distribution, and the goal of this section is to identify just what structures are estimable.

2.4.1. *Identifiable functions.* A functional $h(Q)$ will be called *identifiable in the nonparametric sense* if whenever $Q_1$ and $Q_2$ are latent distributions that generate the same mixture distribution for $X$, then $h(Q_1) = h(Q_2)$. Thus the value of $h(Q)$ can be determined uniquely from the observable distribution of $X$.

In the case when $X$ has a lattice distribution, we return to the representation of the ratio function as a moment generating function (2.5). We can conclude that, for $t = 0, \ldots, T$, the ratio function $R(t)$ is the $t$th moment of $\exp(\Phi)$ under the measure $\Gamma$. It follows that the first $T$ moments of $\exp(\Phi)$ under $\Gamma$ can be determined from the distribution of the observable $X$ and so are identifiable functions of the latent distribution $Q$. Moreover, they are a full set in the sense that any other identifiable functional must be a function of them. [*Exercise.*]

This result is perhaps not satisfying in itself because the identifiable functions do not have a natural statistical interpretation. In some special cases, the identifiable functions have a direct interpretation in terms of the measure $Q$ as well. For example, we have the following proposition:

PROPOSITION 4. *If $X$ is a mixture of* $\mathrm{Bin}(T, \phi)$ *distributions, with latent distribution $Q$ on $\phi$, then the first $T$ moments of $\Phi$ under the distribution $Q$ are identifiable from the mixed density and all other identifiable functions are functions of these moments.*

PROOF. We prove this for the case $T = 2$; the extension to arbitrary $T$ is an exercise. We start with the matrix identity

$$\begin{pmatrix} (1-\phi)^2 \\ 2\phi(1-\phi) \\ \phi^2 \end{pmatrix} = \begin{bmatrix} 1 & -2 & 1 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} 1 \\ \phi \\ \phi^2 \end{pmatrix}.$$

We write this in the symbolic form $\mathbf{f}_\phi = \mathbf{A}\boldsymbol{\phi}^\dagger$, where $\mathbf{f}_\phi$ is the density vector for a binomial model with parameter $\phi$ and $\boldsymbol{\phi}^\dagger$ is the power vector introduced earlier. If we integrate both sides of this equation with respect to $dQ(\phi)$, we find that the left side becomes the mixture density vector $\mathbf{f}_Q$ and the right side is a linear function of the moments of $\Phi$, which we write as $\mathbf{A} \cdot E(\boldsymbol{\Phi}^\dagger)$. Since the matrix $\mathbf{A}$ is clearly invertible, we can solve for the moments of $Q$, given the mixed density, via the equation $E(\boldsymbol{\Phi}^\dagger) = \mathbf{A}^{-1}\mathbf{f}_Q$. □

See Lindsay, Clogg and Grego (1991) for another example, the Rasch model, where there is a natural set of identifiable mixture parameters in the form of posterior expectations $E[g_i(\Phi)|X = x]$ that therefore allow fully identified nonparametric empirical Bayes estimation.

However, one cannot consistently estimate the latent distribution function $Q(t)$ at any value of $t$ without further external information, such as knowing that the distribution function lies in some parametric class. In particular, the nonparametric approach of Chapter 1 cannot consistently estimate the distribution function.

Just the same, features of $Q$ that are not identifiable are usually *partially identified*. For example, in the binomial model, knowing the first $T$ moments of $Q$ does limit the set of allowable distributions, the more so the larger $T$ is. In fact, Tchebysheff developed an optimal system of upper and lower bounds for the distribution function evaluated at a point, given a set of its moments [Uspensky (1937)]. For any given functional being estimated, it is at least theoretically possible to construct upper and lower bounds that would give the degree of determination of that function.

This point is relevant statistically because it is therefore possible to construct informative confidence intervals for nonidentifiable functions of $Q$. Although such bounds will not shrink to zero in width asymptotically, but rather to the limits of knowledge of that function, for any fixed sample size the width due to randomness could greatly exceed that due to indeterminacy. Lindsay, Clogg and Grego (1991) consider nonparametric bounds in the Rasch model for some nonidentifiable empirical Bayes functionals of interest.

2.4.2. *Identifiability of weights, m fixed.* We leave the nonparametric setting and consider a situation where the number of components is assumed to be known, say $m$. The relevant question here is: If we restrict attention to latent distributions $Q$ that have $m$ or fewer points, will the latent distribution be identifiable, in that there will be exactly one possible latent distribution in this class that generates any one $X$ distribution? Again, we restrict attention to discrete one parameter exponential families.

This subsection deals with the simplest case in which the support points $\xi_1, \ldots, \xi_m$ of $Q$ are known and fixed, so that our concern is with the identifiability of the weight parameters. We have already developed the appropriate basic theory for this case because this is exactly the situation of Section 2.3.1. We consider first the binomial model, for which we have the following proposition:

PROPOSITION 5. *The parameters* $\pi_1, \ldots, \pi_m$ *in the mixture*

$$\pi_1 \operatorname{Bin}(T, \theta_1) + \cdots + \pi_m \operatorname{Bin}(T, \theta_m)$$

*are identifiable provided that* $m \leq T + 1$. [*Note from* (2.8) *that this is the maximal number of identifiable components for this binomial family.*]

PROOF. We need to establish the linear independence of the vectors $\mathbf{f}_j := \mathbf{f}_{\theta_j}$. It therefore suffices to consider $m = T + 1$. In the notation of the proof of the last proposition, we have for an appropriate nonsingular matrix $\mathbf{A}$,

$$\begin{aligned}\det[\mathbf{f}_1, \ldots, \mathbf{f}_{T+1}] &= \det\{\mathbf{A}[\boldsymbol{\theta}_1^\dagger, \ldots, \boldsymbol{\theta}_{T+1}^\dagger]\} \\ &= \det \mathbf{A} \cdot \det[\boldsymbol{\theta}_1^\dagger, \ldots, \boldsymbol{\theta}_{T+1}^\dagger].\end{aligned}$$

It follows that it suffices to show that $\det[\boldsymbol{\theta}_1^\dagger, \ldots, \boldsymbol{\theta}_{T+1}^\dagger] \neq 0$. However, this is again the well known Vandermonde determinant, equaling $\prod_{i>j}(\theta_i - \theta_j)$. □

We next consider how this result might be extended to other one parameter exponential families. The key to the identifiability in the binomial family is the nonsingularity of the matrix of probability vectors $[\mathbf{f}_1, \ldots, \mathbf{f}_{T+1}]$, which we might believe quite difficult to deal with for an arbitrary exponential family. However, there is a quite amazing theory that relates the nonsingularity of such matrices to the maximal number of zeros of certain polynomial equations. The interested reader should dig into the difficult but impressive works of Karlin and Studden (1966) and Karlin (1968). We give a brief outline of the fundamental ideas here.

A system of functions $\tau_0(\phi), \ldots, \tau_T(\phi)$ of the real variable $\phi$ is called a *Chebyshev system* if every polynomial $\sum_{j=0}^{T} w_j \tau_j(\phi)$, whose coefficients $w_j$ are not all zero, has at most $T$ zeros in $\phi$. The most familiar example of such a system is the $1, \phi, \ldots, \phi^T$, where we can apply the fundamental theorem of algebra to bound the number of real zeros by $T$.

Suppose we have a family of multinomial densities $f(t; \phi)$ and we define $\tau_j(\phi) := f(j; \phi)$. Further, suppose these $\tau_j$ constitute a Chebyshev system. We may conclude that

$$m \leq T + 1 \quad \Longrightarrow \quad \text{weights identifiable,}$$

any fixed set of $m$ component densities, using the following argument:

Suppose not. Then there exists a vector $\mathbf{w}$ such that

$$[\mathbf{f}_{\phi_1}, \ldots, \mathbf{f}_{\phi_{T+1}}]\mathbf{w} = \mathbf{0}.$$

However, these equations can be written out row-by-row to show that the $\phi_j$ are $T+1$ solutions to $\sum w_j \tau_j(\phi) = 0$, a contradiction to the Chebyshev system property.

Although it is not insightful to our present task to prove this result, it is known that every finite discrete exponential family generates a Chebyshev system. There are other useful models that form Chebyshev systems. One

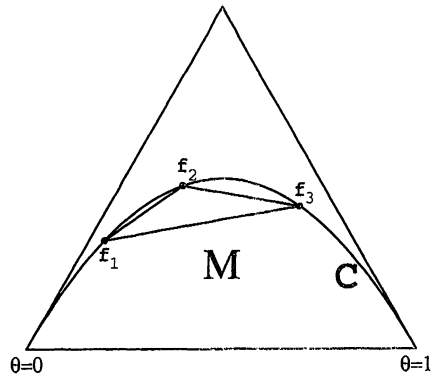


FIG. 2.6. *The binomial unicomponent density curve.*

such extension is a continuous exponential family that has been discretized into bins, as done when performing a chi-squared goodness-of-fit test. See Lindsay and Roeder (1993) for more examples and details on these results.

In Figure 2.6 we show the implications of the above proposition for the binomial model when $T = 2$. We construct such a plot as follows. If we plot for each value of $\theta$ the vector for the binomial density function $\mathbf{f}_\theta = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)'$, then we trace out a curve in $\mathbf{P}^2$ that we call the *unicomponent density curve* and denote by

$$\mathbf{C} = \{\mathbf{f}_\theta\colon\ \theta \in [0,1]\}.$$

[*Exercise*: Find the points in the figure corresponding to $\theta = 0$, 1 and 0.5.]

The curve $\mathbf{C}$ corresponds to all the possible density vectors obtainable under the unicomponent binomial model. As we have already learned, the set of all probability densities obtainable as mixtures of binomials is the convex hull of $\mathbf{C}$, denoted $\mathbf{M} = \text{conv}(\mathbf{C})$. If we consider any fixed set of three binomial densities, we can see pictorially that they cannot fall on a single line, so that the mixture set determined by any three distinct $\theta$ gives identifiable parameters $\pi$.

2.4.3. *Full identifiability of m components.* We now raise the level of difficulty by considering the more statistically important question: If we allow the location of the support points $\xi$ of $Q$ to be unknown, but restrict their number to be no more than $m$, when will both sets of parameters $\pi$ and $\xi$ be identifiable? This difficult question has a very simple and elegant solution in the case of the one parameter exponential family. To motivate the answer, we return to Figure 2.6. Geometrically, we can consider three distinct cases, depending on where the true mixture density vector $\mathbf{f}_Q$ is located.

- Suppose that $m = 1$. Notice that in this case the density vector $\mathbf{f}_Q$ must be an element of $\mathbf{C}$. This class of densities must be identifiable since they are just the (identifiable) binomial densities. Hence $m = 1$ implies identifiability.

- Suppose $\mathbf{f}_Q \in \text{int}(\mathbf{M})$, the *interior* of $\mathbf{M}$ (relative to the simplex). Note that one can draw infinitely many lines that pass through $\mathbf{f}_Q$ and connect two points on $\mathbf{C}$. The intersection points correspond to two parameter values, say $\phi_1$ and $\phi_2$, and we can correspondingly write

$$\mathbf{f}_Q = \pi \mathbf{f}_{\phi_1} + (1 - \pi)\mathbf{f}_{\phi_2},$$

  for some $\pi$ for every such pair. We can draw two conclusions: first, that every interior density has a representation with $m = 2$, but second, that no such two-component mixture can be identifiable.
- If $m = 2$ and $\mathbf{f}_Q \in \text{bdry}(\mathbf{M})$, the boundary of $\mathbf{M}$, then $\mathbf{f}_Q$ must lie on the bottom line of the triangle, where it is clear that it can be represented by a convex combination of the component density vectors corresponding to the two extreme values of $\theta$, namely, 0 and 1. Moreover, this is a unique representation with two points.

Thus we have a situation in which $m = 1$ implies identifiability, but only some two-component mixtures are identified. If we change our method of counting support points somewhat, however, we can report a very simple rule for identifiability of the latent distribution.

Thus we define index($Q$) to be the number of support points in the latent distribution $Q$, with the *special rule* that the support points at the left and right extremes of the parameter space be counted as 1/2 a support point. Thus, for example, in the above binomial example, a mixture of Bin(2, 0) and Bin(2, 0.5) corresponds to a latent distribution with index 1.5, and the identifiable mixtures on the bottom edge of the triangle have index 1.

With this said, we can summarize the binomial example by saying that the distribution $Q$ can be determined from $\mathbf{f}_Q$ if and only if index($Q$) $\leq 1$. This result can be generalized, but we first develop some important tools in the following subsection.

2.4.4. *Hyperplanes and convex sets.* A useful tool for working with convex sets is the idea of the support hyperplane. For every vector of length 1, say $\mathbf{w}$, in $\mathbf{R}^d$, and every constant $c$, there exists a *hyperplane* $\mathbf{H} = \mathbf{H}(\mathbf{w}, c)$, defined to be the set

$$\mathbf{H} = \{\mathbf{v} \in \mathbf{R}^d \colon \mathbf{v}'\mathbf{w} = c\}.$$

It is a $(d - 1)$-dimensional linear surface. We can think of it as a translation of the linear subspace consisting of all vectors orthogonal to $\mathbf{w}$, namely,

$$\mathbf{H}(\mathbf{w}, c) = c\mathbf{w} + \langle \mathbf{w} \rangle^{\perp}.$$

[*Exercise.*]

In $\mathbf{R}^2$, a hyperplane is just a line in the plane. In $\mathbf{R}^3$ it is a two-dimensional planar surface. Each hyperplane can be associated with a *lower half space* $\{\mathbf{v} \in \mathbf{R}^d \colon \mathbf{v}'\mathbf{w} \leq c\}$ and an *upper half space* $\{\mathbf{v} \in \mathbf{R}^d \colon \mathbf{v}'\mathbf{w} \geq c\}$.

A *support hyperplane* to a convex set $\mathbf{B}$ is a hyperplane that bounds the set on some side. More formally, for each direction vector $\mathbf{w}$, let

$$c^*(\mathbf{w}) = \sup\{\mathbf{w}'\mathbb{b}: \mathbb{b} \in \mathbf{B}\}.$$

Verify that for every $c < c^*$ the convex set $\mathbf{B}$ must have a *nonempty* intersection with the upper half space of $\mathbf{H}(\mathbf{w}, c)$. For every $c > c^*$, the convex set $\mathbf{B}$ must have an *empty* intersection with the upper half space, and so it is contained in the lower half space. In this case, $\mathbf{H}(\mathbf{w}, c^*(\mathbf{w}))$ is a support hyperplane.

If the convex set $\mathbf{B}$ is closed, then it is clear that the support hyperplane $\mathbf{H}$ intersects $\mathbf{B}$ along a boundary and that the intersection consists of those points $\mathbf{b} \in \mathbf{B}$ satisfying $\mathbf{w}'\mathbf{b} = c^*$, whereas $\mathbf{w}'\mathbf{b} < c^*$ for all other $\mathbf{b} \in \mathbf{B}$. Thus the closed convex set lies completely in one of the half spaces generated by the hyperplane, with some of its boundary points in the hyperplane. (Indeed, a closed convex set can be represented as the intersection of the lower half spaces of its support hyperplanes.)

As an aside, we note that if $\mathbf{B}$ is a convex set, then $\mathbf{B}^* = \{\mathbf{v}: \mathbf{v} \cdot \mathbf{b} \leq 1$ for all $\mathbf{b} \in \mathbf{B}\}$ is a *dual* convex set to $\mathbf{B}$. Note that if $\mathbf{b}^* \in \mathbf{B}^*$, then the set $\mathbf{B}$ is in the lower half space $\{\mathbf{u}: \mathbf{u} \cdot \mathbf{b}^* \leq 1\}$. It follows that if $\mathbf{B}$ is closed, then $\mathbf{B}^*$ is closed and its boundary points correspond to the support hyperplanes of the set $\mathbf{B}$.

When we are working with a convex set $\mathbf{B}$ in the probability simplex, the entire set lies within the hyperplane $\mathbf{H}(\mathbf{1}, 1)$. Thus this hyperplane is a support hyperplane, but not a very interesting one as far as describing the set $\mathbf{B}$. If $\mathbf{w} \neq \mathbf{1}$, then a hyperplane $\mathbf{H}(\mathbf{w}, c)$ will intersect $\mathbf{H}(\mathbf{1}, 1)$ in a linear manifold of dimension $d-2$, and there will be multiple hyperplanes $\mathbf{H}(\mathbf{w}, c)$ that generate the same manifold within the simplex.

In particular, if $\mathbf{p}$ is in $\mathbf{P}^T$ and lies in the hyperplane $\sum w_j p_j = c$ determined by $\mathbf{H}(\mathbf{w}, c)$, then, since $\sum p_i = 1$, it also lies in the hyperplane $\sum(w_j - c)p_j = 0$, which is created by the hyperplane $\mathbf{H}(\mathbf{w} - c\mathbf{1}, 0)$.

The fact that in the probability simplex, and therefore in the mixture problem, one can reduce attention to hyperplanes with $c = 0$, and therefore containing the origin, turns out to be quite important in reducing the dimensionality of the mixture problem.

2.4.5. *Identifiability of weights and supports.* We now use the tools of support hyperplanes to turn questions about the structure of the boundary of a mixture set into questions about polynomials and so solve identifiability questions.

PROPOSITION 6. *If $f(x; \phi)$ is a discrete exponential family density (more generally a Chebyshev system density) with $T + 1$ points of support, then the class of identifiable mixtures is those in the boundary of $\mathbf{M}$, which is exactly those elements satisfying*

$$\text{index}(Q) \leq T/2.$$

PROOF. For the proof here we simply use the binomial model, because the results can then be derived using well known results regarding polynomials. We start by showing that boundary points have the specified bound on their index. Suppose point $\mathbf{f}$ is in the boundary of the mixture density set $\mathbf{M}$ and so lies in a support hyperplane $\mathbf{H}$ defined by

$$\mathbf{w}'\mathbf{m} \leq 0 \quad \text{for all } \mathbf{m} \in \mathbf{M},$$

with $\mathbf{w}'\mathbf{f} = 0$. It follows that $\mathbf{f}$ can be represented as a mixture of the binomial vectors $\mathbf{f}_\theta$ that lie in that same hyperplane, hence satisfying $\mathbf{w}'\mathbf{f}_\theta = 0$; otherwise, one could use the mixture representation to show $\mathbf{w}'\mathbf{f} < 0$. However, the function

$$g(\theta) := \mathbf{w}'\mathbf{f}_\theta$$

is a polynomial in $\theta$ of degree $T$, so it has at most $T$ roots. Moreover, since we are in a support hyperplane, with $\mathbf{w}'\mathbf{f}_\theta \leq 0$ for all $\theta$, these roots must correspond to local maxima of $g$. Hence any root in the open interval $(0, 1)$ must be a root of even multiplicity to the polynomial. Since each root corresponds exactly to a potential support point, this means that if we count each support point in $(0, 1)$ with weight 2 and mass points at the extremal values of 0 and 1 with weight 1, then the total cannot exceed $T$. We can now apply the definition of index to argue that the maximal index of the latent distribution corresponding to $\mathbf{f}$ is $T/2$. Moreover, since the roots correspond to all of the possible support points, the preceding proposition shows that this latent distribution is unique and so is identifiable. To show that every mixture of index no more than $T/2$ is in the boundary, we can construct a polynomial with the roots corresponding to the support points and show that this polynomial implies the existence of a bounding hyperplane. On the other hand, if $\mathbf{f}$ is in the interior of $\mathbf{M}$, then we can again argue that we can draw many lines through it connecting two boundary points, each of which corresponds to a different mixture representation, and so nonidentifiability holds. □

As an exercise, the reader should consider how the proof might extend from the binomial to the case of the Chebyshev system.

More precise descriptions of elements of the interior of $\mathbf{M}$ are available, including the existence of exactly two representations of $\mathbf{f}$ in terms of mixtures of index $(T+1)/2$. If $T$ is even, then there is one representation involving each of the two extreme parameter values [check this pictorially for $\text{Bin}(2, \theta)$]. If $T$ is odd, then there is one representation in terms of $(T + 1)/2$ interior components and one involving $(T - 1)/2$ interior components and the two extreme components. From a statistical point of view, the latter representation has one more component than the former.

Fortunately, there is a simple rule of thumb that describes the identifiability of the parameters in a finite mixture model. If you specify a mixture model with $m$ interior components and if the total number of free parameters $2m - 1$ is less than or equal to the number of free parameters in the multinomial, here $T$, then in fact the parameters are identifiable.

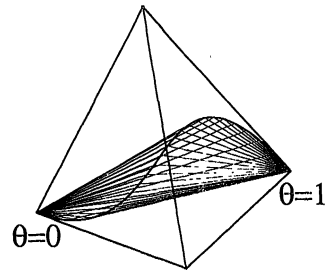


FIG. 2.7. *The unicomponent density curve for the* $\text{Bin}(3, \theta)$ *model.*

The following example illustrates the geometric structures corresponding to this index result. We consider the $\text{Bin}(3, \theta)$ model, with sample space $\{0, 1, 2, 3\}$. The probability simplex $\mathbf{P}^3$ can be represented as a tetrahedron, shown in Figure 2.7.

The curve $\mathbf{C} = \{\mathbf{f}_\theta\}$ starts at $(1, 0, 0, 0)$, where $\theta = 0$, and moves sinuously around to $(0, 0, 0, 1)$, where $\theta = 1$. The mixture set $\mathbf{M}$ has two smooth two-dimensional boundary surfaces, corresponding to the two different types of mixtures of index 1.5. One boundary corresponds to latent distributions with two support points, one of which is $\theta = 0$; the other is two point mixtures with one mass point at $\theta = 1$. The one component curve $\mathbf{C}$ corresponds to one seam along these two surfaces; the other seam is along one edge of the tetrahedron and is formed by mixtures of index 1 that are mixtures with support at $\theta = 0$ and 1. These boundary mixtures are identifiable, whereas no interior points have identifiable latent distributions.

In Figure 2.8 we show a cross section through the set $\mathbf{M}$, where we cut through the tetrahedron in the plane of all density vectors $\mathbf{p}$ on $\{0, 1, 2, 3\}$ with mean 1.5; that is, $\sum p(x) \cdot x = 1.5$. The two edges of the set $\mathbf{M}$ are the boundaries corresponding to the mixtures with index 1.5.

2.4.6. *Related problems.* When one leaves the i.i.d. case and considers other independent but not identically distributed structures, the analysis can be considerably more difficult and the results less simple. See, for example, Follman and Lambert (1991). Another complication arises in the i.i.d. case if one has nonlatent parameters in the model, so that one must address joint identifiability. Lindsay, Clogg and Grego (1991) managed to solve one such joint identifiability question in the Rasch model.

We conclude by noting that there are many interesting relationships between binomial mixture models and other natural distributions generated by sequences of binary variables. We point out two such cases.

Consider the distribution $\text{Bin}(1, p_1) * \text{Bin}(1, p_2)$, the convolution of two Bernoulli trials with success probabilities $p_1$ and $p_2$, respectively. This is a distribution on $\{0, 1, 2\}$ and so has density vector in $\mathbf{P}^2$. When $p_1 = p_2 := p$, the distribution is $\text{Bin}(2, p)$. As an *exercise*, show that the set consisting of all

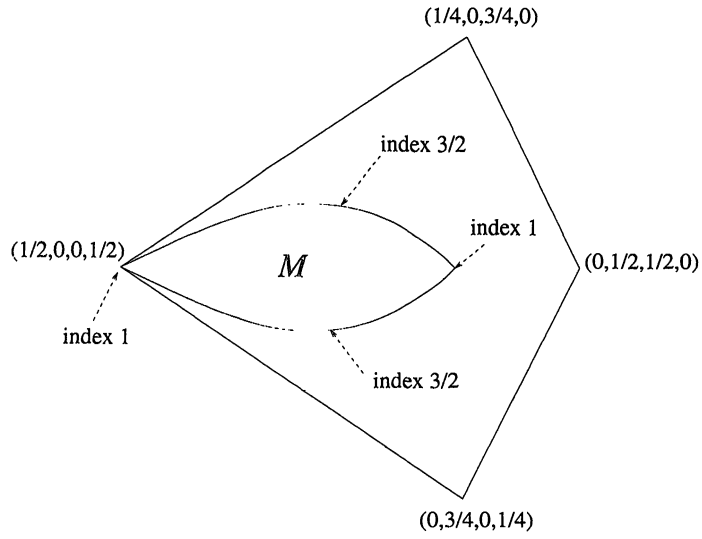


FIG. 2.8. *Cross section of the binomial mixture density set.*

Bernoulli convolutions with $p_1 \neq p_2$ is exactly the complement of the set of Bin(2, $p$) mixtures.

We can think of the variation of the value of $p$ *between* the Bernoulli trials as being a distinct and nonoverlapping form of distributional assumption to the type of variation which occurs in the mixture model. There are a number of interesting papers on the structure of the space of Bernoulli convolutions, for example, Hoeffding (1956), but most pertinent for data analysis is that these models contrast with the mixtures in that the resulting convolution distributions are *lighter* tailed than the binomial distributions. For example, it can be shown that the binomial distribution is a dilation of any convolution distribution with the same mean.

Another interesting relationship arises from de Finetti's theorem. If we have an *infinite* sequence of exchangeable binary variates, say $X_1, X_2, \ldots$, then this theorem indicates that the sequence has a distribution that can be represented as a mixture, over latent variable $p$, of sequences of Bernoulli trials with success probability $p$. Diaconis (1977) investigated the implications of this result for *finite* sequences of exchangeable binary variates. These results indicate that if a finite sequence of binary variates is exchangeable, then the sum either has the distribution of a mixture of binomials or very nearly so, where Diaconis makes the "nearly so" statement precise.

**2.5. Moment representations.** We have now found out a great deal about the structure of the exponential family mixture, but a fundamental question remains: Given a density vector $\mathbf{p} \in \mathbf{P}^T$, is it possible to determine in a straightforward way whether $\mathbf{p}$ is in our mixture model? That is, whether or not $\mathbf{p} \in \mathbf{M}$. One simple test is already available to us from (2.4). That is,

we can check to see if

$$\ln\left(\frac{p(t)}{f(t;\phi_0)}\right) \text{ is convex.}$$

If it is not, then $\mathbf{p}$ cannot be a set of mixture probabilities. Indeed, this kind of plot is very similar to ratio plots earlier in this chapter.

However, we can greatly sharpen this result in the special case when the density is exponential family on the lattice $\{0, 1, 2, \ldots, T\}$. Recall that in this circumstance, the ratios

$$R(t) = f(t;Q)/f(t;\phi_0) = \int \exp((\Phi - \phi_0)t)\, d\Gamma(\Phi)$$

are values of the moment generating function of some positive measure $\Gamma$. When the values of $t$ are on the lattice, this implies that

$$R(0), R(1), \ldots, R(T)$$

are the moments of the nonnegative measure corresponding to the distribution of $\exp(\Phi - \phi_0)$. The argument can be used in reverse to show that $\mathbf{p}$ is a set of mixture probabilities if and only if

$$\frac{p(0)}{f(0;\phi_0)}, \ldots, \frac{p(T)}{f(T;\phi_0)}$$

are the moments of some measure on $(0, \infty)$. [For simplicity, we are here assuming the parameter space is $(-\infty, +\infty)$.]

Thus our question is equivalent to the following: When is a sequence of $T+1$ numbers, say $m_0, \ldots, m_T$, equal to the sequence of moments $\int x^k\, d\nu(x)$ for some positive measure $\nu$ with full mass on $(0, \infty)$?

The answer can be specified most easily through the use of moment matrices. We form a sequence of *moment matrices* $M_p$ as follows. If $p$ is even, say $p = 2k$, then

$$M_{2k} := \begin{bmatrix} m_0 & m_1 & m_2 & \cdots & m_k \\ m_1 & m_2 & m_3 & \cdots & m_{k+1} \\ m_2 & m_3 & m_4 & \cdots & \cdot \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m_k & m_{k+1} & \cdot & \cdot & m_{2k} \end{bmatrix}.$$

If $p$ is odd, say $2k+1$, then

$$M_{2k+1} := \begin{bmatrix} m_1 & m_2 & m_3 & \cdots & m_{k+1} \\ m_2 & m_3 & m_4 & \cdots & \cdot \\ m_3 & m_4 & m_5 & \cdots & \cdot \\ \vdots & \vdots & \vdots & \vdots & \cdot \\ m_{k+1} & \cdot & \cdot & \cdot & m_{2k+1} \end{bmatrix}.$$

There are many results known about the nature of these matrices and their relationship to the number of mass points in a positive measure, with the special case of $m_0 = 1$ being that of a probability measure. For our purposes we note the following:

- If both $M_{2k}$ and $M_{2k-1}$ are positive definite, then $m_0, m_1, \ldots, m_{2k}$ are the moments for some nonnegative measure $d\gamma(\phi)$ on $[0, \infty)$. (In fact, there exist infinitely many measures with these moments.)
- If $m_0, m_1, \ldots, m_{2p}, \ldots$ are the moments of a measure with exactly $K$ support points, then the sequence of matrices $M_0, M_1, M_2, \ldots, M_p, \ldots$ has the property that they are positive definite for $p \leq 2K-2$ and are nonnegative definite with rank $K$, thereafter.

These results enable us to test if a set of points is a moment sequence. In particular, if we are testing if a sequence of ratios $R(t)$ is in the interior of the mixture space, then the fact that they lie in a region of nonidentifiable distributions means that the highest order moment matrices that can be constructed from the sequence must be strictly positive definite. To show that a sequence of ratios is in the boundary of the mixture space corresponding to some $K$ point mixture, it suffices to show that the matrices are nonnegative definite, with rank $K$. (Note that the entire moment sequence is determined uniquely after maximal rank is reached, using the fact that the determinant of a rank deficient matrix is zero.)

This approach to testing for the presence of mixture structure was discussed in Lindsay, Clogg and Grego (1991).

**2.6. Certain nested mixture models.** Many times we will wish to apply the theory of mixtures to models that have auxiliary (nonlatent) parameters $\theta$ and so fall into the class of semiparametric mixture models. An important statistical question, about which we know relatively little in general, is the nature of the identifiability of the auxiliary parameters in the presence of the latent distribution. However, there are certain important cases, including the following normal example, in which it is indisputable that a fundamental loss of identifiability occurs and that we must be aware of the consequences of this loss. In particular, when it occurs, it may be impossible to sensibly estimate the parameters in the presence of the nonidentifiability.

As a first example, we consider the *normal mixture* problem. We assume that the component densities are $N(\mu, \sigma^2)$ and that there is an unknown latent distribution $Q$ on the mean parameter $\mu$, together with an unknown variance parameter $\sigma^2$ that is common to all the component densities. We write this model as $N(Q, \sigma^2)$.

We first prove that if $\mu$ has a latent distribution that is $N(\alpha, \tau^2)$, then the marginal distribution of $X$ is $N(\alpha, \sigma^2 + \tau^2)$. Notice that in this convolution model, $X =_{\text{dist}} \Phi + Z$, where $\Phi$ and $Z$ are independent normal variables with means $\alpha$ and 0 and variances $\tau^2$ and $\sigma^2$, respectively, so this claim is just a standard result about the convolution of two normals, provable using moment generating functions:

$$E(\exp(tX)) = E(\exp(t(\Phi + Z))) = E(\exp(t\Phi))E(\exp(tZ))$$
$$= \exp\left(\alpha t + \frac{\tau^2 t^2}{2} + \frac{\sigma^2 t^2}{2}\right).$$

It follows that any mixture $N(Q, \sigma^2)$ can *also* be represented as a normal mixture by $N(Q^*, \sigma^2 - \delta)$, where $Q^*$ is the convolution of $Q$ and $N(0, \delta)$. Thus the class of mixture distributions, as $\sigma$ varies, are *nested*,

$$\{N(Q, \sigma^2 + \tau^2):\ Q \in \text{p.m.}\} \subset \{N(Q; \sigma^2):\ Q \in \text{p.m.}\}$$

and the joint parameters $(Q, \sigma)$ are *not* identifiable.

We do note that if we restrict attention to finite *discrete* latent distributions $Q$, then the pair $(Q, \sigma)$ is identifiable. This can be shown by the fact that if $Q$ has $p$ points of support, then the pair can be recovered from the first $2p$ moments of $X$ using the methods of Lindsay (1989b).

One might ask what happens to the nonparametric maximum likelihood approach to estimating $(Q, \sigma)$. It is easily seen that there exists a nonparametric MLE for $Q$, say $\hat{Q}_\sigma$, for each fixed value of $\sigma$, because the likelihood for a sample is then bounded. However, the profile likelihood function $L(\hat{Q}_\sigma, \sigma)$ is clearly decreasing in $\sigma$ by the above nesting property, since increasing the value of $\sigma$ shrinks the class of eligible models that can be maximized over. In fact, as $\sigma$ gets small, the solution $\hat{Q}_\sigma$ converges weakly to the empirical distribution for the data, and the profile likelihood becomes infinite. See, for example, Hathaway (1985).

Although this means that the likelihood method fails in this example, we do note that this is related to the fact that the class of normal mixtures is very flexible, providing smooth approximations to many other distributions. Roeder (1990) exploited this feature to generate a method of density estimation based on using estimators of the form $N(Q, \sigma^2)$, with parameter selection based on a goodness-of-fit criterion.

There are other two parameter exponential families that have a nested mixing structure. Jewell (1982) showed that the Weibull mixture families had such a structure, with the shape parameter playing the role of $\sigma$. As an illustration of his approach, we show how the technique is readily extended to the gamma family.

Consider the two parameter gamma density

$$f(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathcal{I}[x > 0].$$

We ask if for any given $\alpha$ and $\lambda_0$, and any $\delta$, with $\alpha - \delta > 0$, there exists a latent distribution $Q$ such that

$$\int f(x; \alpha, \lambda)\, dQ(\lambda) = f(x; \alpha - \delta, \lambda_0). \tag{2.9}$$

We can write this equation symbolically as

$$\text{Gam}(\alpha, Q) = \text{Gam}(\alpha - \delta, \lambda_0).$$

This equation can be rearranged so that we seek a solution $Q$ to

$$\int \lambda^\alpha \exp(-\lambda x)\, dQ(\lambda) = x^{-\delta} \exp(-\lambda_0 x) c,$$

where $c$ is a finite constant.

First we ask if we should expect a solution. Recall that a necessary condition for a density $p(x)$ to be a mixture of a one parameter exponential family $\exp(\phi x - \kappa(\phi))$ is that $\ln(p(x)/f(x;\phi))$ is convex. Note that convexity holds in this example because $-\delta \ln(x)$ is convex in $x$. To go further, it is natural to exploit certain results on Laplace transforms, such as found in Feller [(1971, Vol. II, pages 439–441)]. A function $f(x)$ is said to be *completely monotone* on $[0,\infty)$ if its derivatives have the alternating sign property $(-1)^n f^{(n)}(x) \geq 0$. It is known that the function $f$ is completely monotone if and only if it is the Laplace transform of a measure $\mu$ on $[0,\infty)$; that is, $f(x) = \int e^{-\lambda x}\, d\mu(\lambda)$.

It is easy to see that the function $x^{-\delta}e^{-x\lambda_0}$ is the product of two completely monotone functions and as such, it must be completely monotone. Let $\mu$ be its generating measure. We then have, if we set $dQ(\lambda) = c\lambda^{-\alpha}\, d\mu(\lambda)$, a formal solution to our problem. The one remaining point to check is that the measure so generated is finite, in the sense that $\int \lambda^{-\alpha}\, d\mu(\lambda) < \infty$. This can be checked as follows: Since

$$\int x^{\alpha-1} \exp(-\lambda x)\, d\mu(\lambda) = x^{\alpha-1-\delta} \exp(-\lambda_0 x),$$

we integrate both sides over $x$ in $[0,\infty)$. The left-hand side is proportional to $\int \lambda^{-\alpha}\, d\mu(\lambda)$ and the right-hand side has a finite integral provided that $\alpha - \delta > 0$, which we have assumed.

Now that (2.9) is established, it follows that the gamma mixture models are also nested:

PROPOSITION 7. $\{\text{Gam}(\alpha+\delta, Q)\} \subset \{\text{Gam}(\alpha, Q)\}$.

[*Exercise*: As a more direct proof, show that the latent distribution $Q$ with density proportional to $(\lambda - \lambda_0)^{\delta}\, \lambda^{-\alpha}\, \mathcal{I}[\lambda > \lambda_0]\, d\lambda$ does the job in (2.9).]

**2.7. Concluding remark.** This chapter has given an introduction to some of the key ideas regarding the mathematical structures of mixture models. For much more on these properties, see the papers referenced in the text. For more general information on the identifiability question, in addition to the standard books on mixture models, there is the recent book by Prakasa Rao (1992), that has a chapter devoted to the identifiability of mixtures.