*The Annals of Statistics*
1999, Vol. 27, No. 5, 1706–1721

# BINOMIAL MIXTURES: GEOMETRIC ESTIMATION OF THE MIXING DISTRIBUTION

BY G. R. WOOD

*Massey University*

Given a mixture of binomial distributions, how do we estimate the unknown mixing distribution? We build on earlier work of Lindsay and further elucidate the geometry underlying this question, exploring the approximating role played by cyclic polytopes. Convergence of a resulting maximum likelihood fitting algorithm is proved and numerical examples given; problems over the lack of identifiability of the mixing distribution in part disappear.

**1. Introduction.** Over the past thirty years there has been considerable interest in the problem of estimating the mixing distribution underlying a mixture of binomial distributions. Formally, the problem is stated as follows. Let $h_p(x)$ be the binomial probability mass function, for a fixed number of trials $n$ and probability of success $p$. Let $G$ be a cumulative distribution function on [0, 1] and let

$$f(x) = \int_0^1 h_p(x)\, dG(p) \quad \text{for } x = 0, 1, \ldots, n$$

be the mixture of binomials determined by $G$. Given observations $x_1, \ldots, x_N$ of the mixture, the challenge is to estimate $G$. This problem has often been regarded as intractable, for when $f$ is known exactly, $G$ can be determined only up to its first $n$ moments; binomial mixtures do not determine the mixing distribution uniquely, or in the terminology of the subject, the mixing distribution is not identifiable.

Our purpose here is to add to work begun more than a decade ago by Lindsay [11, 12, 13, 14, 15] and to uncover more of the geometry in this binomial setting. As we do this we find that the identifiability difficulty in part disappears and that a range of geometric estimators of the mixing distribution, and associated fitting algorithms, become evident. Each estimate is supported by at most $(n + 2)/2$ points in [0, 1].

The geometry underlying this situation is analogous in many ways to the geometry underlying linear models [24]. The model determines a convex subset of Euclidean space, and the data a point in that space; choice of a distance measure allows us to smooth the data vector to a nearest point in the convex model space. When the data vector lies outside the model space the identifiability problem vanishes, and we argue that this is the norm for small sample

Received December 1997; revised September 1999.

*AMS* 1991 *subject classifications*. Primary 62G99; secondary 62P15, 52B12.

*Key words and phrases*. Binomial, mixture, mixing distribution, geometry, moment curve, cyclic polytope, nearest point, least squares, weighted least squares, maximum likelihood, Kullback–Leibler distance.

sizes or situations where the number of trials $n$ is greater than approximately ten. For lower $n$ values the identifiability difficulties become tractable.

We pause now to review the landmarks in the history of this problem. It appears to have been considered first in the late 1960s by Lord [17] who addressed the problem in the context of psychological testing. Each student in a psychological test is assumed to have a "true score" of $p \in [0, 1]$ and sits an $n$ question test, with outcome $x$ determined by a binomial distribution with parameters $n$ and $p$. Given the results from a large number of students, the task is to estimate the distribution $G$ of true scores in the population. Assuming smoothness of the mixing distribution, Lord used a calculus of variations technique to estimate $G$. The difficulties encountered by Lord are explained by the geometric viewpoint: his "negative probabilities" and discreteness of the mixing distribution are to be expected ([17], pages 268 and 269).

In 1975 Cressie joined Lord for a different attack on the problem [18]. They obtained an interval estimate for the mean of the posterior true score distribution, the Bayes estimator of the true score, by finding the extremes of its value over a "most likely" set of mixing distributions. This set was determined using a $\chi^2$ criterion. With the aid of a theorem of Markov, the authors showed that the extremes occur for finitely supported mixing distributions and used an optimization algorithm to find them. More recently, Sivaganesan and Berger [25] have employed a Bayesian, moment focused approach to the problem. As in [18], the fact that the posterior mean is a function of the lower order moments of the mixing distribution is used. In contrast to [18], however, a prior distribution is placed on this moment space and an interval estimate for the posterior mean obtained by using a "most likely" set (a highest posterior density region) in the moment space.

Nonparametric maximum likelihood estimation of a mixing distribution was addressed by Laird in [8] under, however, the assumption of identifiability. She was able to show that the estimator is "self-consistent," a property which permits an iterative method for evaluating the mixing distribution. Turnbull [26] had showed this to be a special case of the EM algorithm. Under certain conditions, Laird showed the nonparametric maximum likelihood estimator to be a step function and under further conditions, that it is finitely supported.

Of greatest interest to us here is the series of works by Lindsay and others [11, 12, 13, 14, 15] on the geometry of mixture likelihoods. This work is quite general and recognizes for the first time that the maximum likelihood estimator of the mixing distribution is found by maximizing the loglikelihood function over the convex hull of the "likelihood curve." Lindsay's work relates properties of the maximum likelihood estimator of the mixing distribution to properties of the convex hull of the likelihood curve. Our development is linked to that of Lindsay as the paper progresses. We conclude this brief summary by drawing the attention of the reader to [9], where a method for computing the nonparametric maximum likelihood estimator of the mixing distribution, based on directional derivatives, is presented, and to [2], [16], two recent mixture model reviews.

Our results stem from the observation that the convex set of mixtures of binomial distributions is affinely isomorphic to the convex hull of the moment curve, $\{(x, x^2, \ldots, x^n): 0 \le x \le 1\} \subseteq \mathbf{R}^n$. This curve and the associated cyclic polytopes used in our study have been thoroughly investigated in [5] and [20]. We exploit these results here.

The paper is structured as follows. In Section 2 we review the known geometry underlying the problem and introduce some new geometrical ideas. Maximum likelihood estimation of $G$ is discussed in Section 3 for the identifiable case, where the observed mixture lies outside the model space. Section 4 discusses estimation of the mixing distribution in the nonidentifiable case, where the observed mixture lies inside the model space. Two general convergence results (Theorems 1 and 2), depending on ideas from Choquet theory, are given in Section 5, while in Section 6 numerical examples are presented. In Section 7 we assess the probability that we will have identifiability, using a theoretical argument and also simulation evidence. A discussion and summary complete the paper in Section 8.

**2. The geometric setting.** Here we review the geometric structure underlying the problem, so providing the framework for estimating the mixing distribution $G$ in the next section. For fixed $n$ and $p$ we may view $h_p(x)$, the binomial probability mass function, as a point in $\mathbf{R}^{n+1}$, the $(n+1)$-tuple,

$$(h_p(0), h_p(1), \ldots, h_p(n))$$

which we denote $h_p$. As $p$ runs from 0 to 1 such points trace out the binomial curve $B_n$ in the simplex,

$$T_n = \left\{ x = (x_0, x_1, \ldots, x_n): \sum_{i=0}^{n} x_i = 1 \text{ and } x_i \ge 0 \text{ for all } i \right\}$$

of all probability mass functions on $\{0, 1, \ldots, n\}$. The mixing distribution $G$ can then be viewed as a probability measure on the binomial curve, with center of mass (or barycenter),

$$f(x) = \int_0^1 h_p(x)\, dG(p) \quad \text{for } x = 0, 1, \ldots, n,$$

the mixture of binomials determined by $G$. Necessarily $f$ lies in the convex hull of $B_n$, denoted $\mathrm{co}B_n$. The binomial curve is precisely the likelihood curve used by Lindsay [11] when all the values $0, 1, \ldots, n$ appear in the sample. Figure 1 pictures the relationship between $T_n$, $B_n$ and $\mathrm{co}B_n$ for the case $n = 2$. Here $B_2 = \{((1-p)^2, 2p(1-p), p^2): 0 \le p \le 1\}$ with the set of all possible mixtures being the convex hull of $B_2$.

We assume throughout that we have a large number of observations, $x_1, \ldots, x_N$ from $f$, the density histogram of which provides us with an approximation, $\hat{f}$, to $f$. Recall that $\hat{f}$ is the maximum likelihood estimate of $f$ under the multinomial model; we must use $\hat{f}$ to estimate $G$.

We now make a critical observation. Only when $\hat{f}$ lies in the model space,

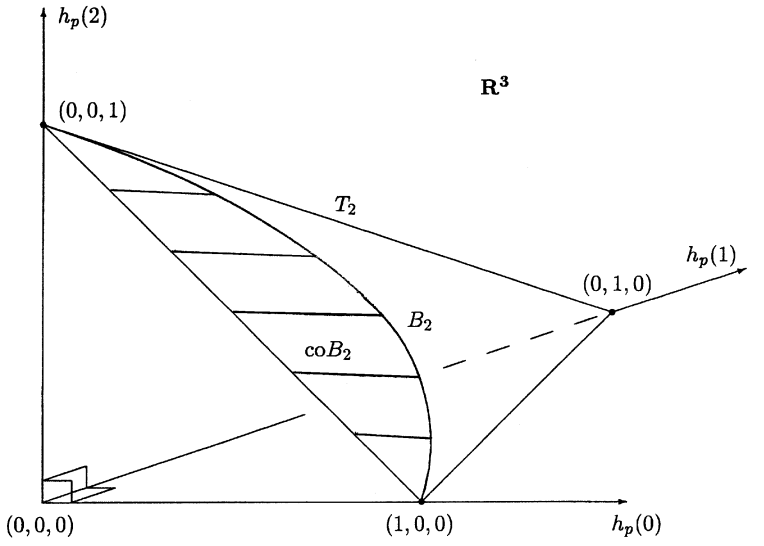


FIG. 1. *The simplex $T_2$ of probability mass distribution functions on $\{0, 1, 2\}$, the curve of binomial probability mass distribution functions $B_2$ and the mixtures of binomial distributions, $\mathrm{co}B_2$, shaded.*

$\mathrm{co}B_n$, does the nonidentifiability problem rear its head. In this case there are many possible mixing distributions $G$ which may have given rise to $f$. When $\hat{f}$ lies outside $\mathrm{co}B_n$ we smooth it, using a distance measure, to a unique point $\hat{f}_\infty$ in the model space, associated with which there is a (fortunately unique) mixing distribution. In the next section we view $\hat{f}_\infty$ as the limit of a sequence of discrete approximations $\hat{f}_k$, hence the notation. We shall show in Section 7 that for $n \geq 10$ it is usual for $\hat{f}$ to lie outside the model space.

Practical estimation of $G$ described here begins with a discrete approximation to $B_n$. Specifically, for $k \in \mathbf{N}$, the natural numbers, we define a $k$-segment piecewise linear approximation to $B_n$, namely,

$$B_n(k) = \bigcup_{i=1}^{k} \mathrm{co}\{h_{(i-1)/k}, h_{i/k}\}$$

We shall use $\mathrm{co}B_n(k)$ to estimate $G$. In Theorem 1, given in Section 5, we show that as $k$ increases the estimate of $G$ converges in a weak* sense to the unique finitely supported mixing distribution which gives rise to $\hat{f}_\infty$.

Our results hinge on the following fact, demonstrated in [27], proof of Theorem 1. The binomial curve $B_n$ is affinely isomorphic to the moment (or cyclic) curve,

$$C_n = \{c_x = (x, x^2, \ldots, x^n) : 0 \leq x \leq 1\} \subseteq \mathbf{R}^n.$$

We discretize the moment curve as we did the binomial curve, so defining

$$C_n(k) = \bigcup_{i=1}^{k} \text{co}\{c_{(i-1)/k}, c_{i/k}\}$$

and take its convex hull to form the cyclic polytope $\text{co}C_n(k)$. Fortunately, much is known about cyclic polytopes. Coming full circle, $\text{co}C_n(k)$ is affinely isomorphic to $\text{co}B_n(k)$, so we can exploit the properties of cyclic polytopes in our study of binomial mixtures.

Cyclic polytopes came into prominence in the 1960s as the solution to the upper bound conjecture: the maximum number of faces (of any dimension) of an $n$-dimensional polytope with $k > n$ vertices is attained by $\text{co}C_n(k)$. An account of this epic result is contained in [20]. Karlin and Shapley [5] earlier studied the convex hull of the moment curve and will provide us with further results.

Faces of any dimension on cyclic polytopes are simplexes ([20], Chapter 2, Proposition 17) so the same property holds for $\text{co}B_n(k)$. In the event that $\hat{f}$ lies outside the model space this structure suggests that we smooth $\hat{f}$ to produce a "nearest point" estimator $\hat{f}_k$ of the mixture in $\text{co}B_n(k)$. Since faces of $\text{co}B_n(k)$ are simplexes it follows that $\hat{f}_k$, on a face of $\text{co}B_n(k)$, has a unique realization as a convex combination of the vertices of the face. These vertices and their weights provide us with an estimator of $G$. We discuss this procedure in the next section.

In researching this procedure, we first used Euclidean distance, then successively refined it to a "weighted least squares" distance, an iteratively reweighted least squares procedure and finally Kullback–Leibler distance. The weighted least squares procedure acknowledges the statistical importance of the $\chi^2$ measure of goodness-of-fit. In turn, the weighted least squares distance measure approximates Kullback–Leibler distance, which yields the maximum likelihood estimator. Thus we explored a spectrum of fitting methods and found that the more elaborate the distance, the more statistically satisfactory was the solution. For this reason, in the sequel only Kullback–Leibler distance is used.

This estimation approach is available when the maximum likelihood estimator of $G$ is identifiable; the software we use during the estimation process is capable of indicating when this occurs. In the next two sections we discuss separately the identifiable and the nonidentifiable cases.

**3. Estimating the mixing distribution: the identifiable case.** Given $\hat{f}$, how do we estimate the mixing distribution $G$? In Figure 2 we illustrate a three stage process:

1. *Approximation*. Use $\text{co}B_n(k)$ to approximate $\text{co}B_n$, for some $k \in \mathbf{N}$.
2. *Smoothing*. Find the nearest point $\hat{f}_k$ in $\text{co}B_n(k)$ to the empirical mixture $\hat{f}$.

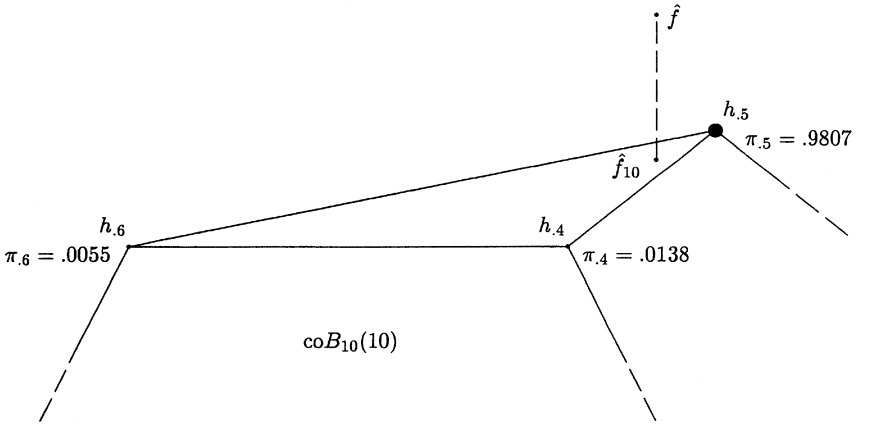


FIG. 2. *An example with binomial parameter* $n = 10$ (*Example* 1 *in Section* 6) *and discretization parameter* $k = 10$, *in which the empirical mixture* $\hat{f}$ *lies outside the polytope of mixtures,* $\text{co}B_{10}(10)$. *The nearest point* $\hat{f}_{10}$ *to* $\hat{f}$ *under Kullback–Leibler distance lies on a triangular face of the polytope with vertices the binomial probability mass functions* $h_{0.4}$, $h_{0.5}$ *and* $h_{0.6}$. *Unique barycentric weights* $\pi_{0.4} = 0.0138$, $\pi_{0.5} = 0.9807$ *and* $\pi_{0.6} = 0.0055$ *at these vertices represent* $\hat{f}_{10}$ *and provide* $g_{10}$, *an estimate of the mixing distribution.*

3. *Representation*. Express (or "represent") $\hat{f}_k$ as the unique convex combination of the extreme points of the simplicial face of $\text{co}B_n(k)$ on which $\hat{f}_k$ lies.

In stage one we approximate the binomial curve using $k+1$ evenly spaced values of $p$, as described in the previous section. In the second stage we smooth $\hat{f}$ by projection onto $\text{co}B_n(k)$, using Kullback–Leibler distance. The strict convexity of the contours of Kullback–Leibler distance ensures that the projection $\hat{f}_k$ is unique. The final stage exploits the simplicial nature of $\text{co}B_n(k)$ to express $\hat{f}_k$ as a convex combination of the extreme points of the simplicial face on which it lies; that is, we form

$$\hat{f}_k = \pi_{p_1} h_{p_1} + \cdots + \pi_{p_m} h_{p_m},$$

where $h_{p_1}, \ldots, h_{p_m}$ are the $m$ vertices of the face. Then the discrete distribution on [0, 1] with weights $\pi_{p_1}$ at $p_1$ and so on, forms a unique estimate $g_k$ of the unknown mixing distribution.

Standard software readily handles these operations efficiently, as will be described in Section 6. That these discrete estimates $g_k$ which represent $\hat{f}_k$ converge to a unique mixing distribution $g_\infty$ which represents $\hat{f}_\infty$ is shown in the first convergence theorem of Section 5.

We now briefly review Kullback–Leibler distance. Given the data $(n_0, n_1, \ldots, n_n) = N\hat{f}$, where $N$ is the total number of observations, the likelihood of a particular distribution $(\theta_i)$ in $\text{co}B_n(k)$ is given by $\prod_{i=0}^{n} \theta_i^{n_i}$. The likelihood ratio, contrasting the likelihood under the observed probability dis-

tribution $(\hat{\theta}_i)$ to that for $(\theta_i)$ is thus

$$\prod_{i=0}^{n} \left( \frac{\hat{\theta}_i}{\theta_i} \right)^{n_i},$$

whence the loglikelihood ratio is

$$\sum_{i=0}^{n} n_i \log \frac{\hat{\theta}_i}{\theta_i}.$$

This in turn is proportional to

$$\sum_{i=0}^{n} \hat{\theta}_i \log \frac{\hat{\theta}_i}{\theta_i},$$

the so-called Kullback–Leibler distance between $(\hat{\theta}_i)$ and $(\theta_i)$. It follows that we can find the maximum likelihood estimator of $(\theta_i)$ by minimizing the Kullback–Leibler distance between $(\theta_i)$ in $\mathrm{co}B_n$ and the fixed point $(\hat{\theta}_i)$. An example of such a calculation is presented in Section 6.

**4. Estimating the mixing distribution: the nonidentifiable case.** In the event that $\hat{f}$ belongs to $\mathrm{co}B_n$ the mixing distribution can be identified only up to its first $n$ moments. This lack of identifiability cannot be side-stepped and was discussed by Lindsay in [10]. If $\hat{f} \in \mathrm{co}B_n(k)$ for some $k > n+1$ then all the mixing distributions $g = (\pi_0, \pi_{1/k}, \ldots, \pi_1)$ associated with $\hat{f}$ occur as the solution space to the underdetermined system of linear equations

$$\begin{bmatrix} h_0(0) & h_{1/k}(0) & \ldots & h_1(0) \\ \vdots & \vdots & & \vdots \\ h_0(n) & h_{1/k}(n) & \ldots & h_1(n) \\ 1 & 1 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \pi_0 \\ \pi_{1/k} \\ \vdots \\ \pi_1 \end{bmatrix} = \begin{bmatrix} \hat{f}(0) \\ \vdots \\ \hat{f}(n) \\ 1 \end{bmatrix} \quad \text{with } \pi_{i/k} \geq 0 \text{ for } i = 0, 1, \ldots, k.$$

The solution space is an affine slice through the simplex of probability mass functions $T_k$ in $\mathrm{R}^{k+1}$, so forms a polytope. We term it $P_k$, the "slice of mixing distributions," or priors in a Bayesian context. Without additional information we cannot reduce the estimation of $g$ to any set smaller than $P_k$. Note that our aim should be to find the mixing distributions $P$ associated with $\hat{f}$ in $\mathrm{co}B_n$, rather than the $P_k$ determined by $\mathrm{co}B_n(k)$. The manner in which $P_k$ approximates $P$ is described in Theorem 2 in Section 5.

In [18], in the context of psychological testing, we are given a new test score $x$ and must estimate the true score $p$ of the student. The best estimate under

squared error loss is the posterior mean, shown in [18] and [25] to take the form

$$\frac{\text{linear combination}\{\text{moments about the origin of } g \text{ of order} \le n+1\}}{\text{linear combination}\{\text{moments about the origin of } g \text{ of order} \le n\}}.$$

Members of $P_k$ share the same moments up to those of order $n$. Thus the posterior mean is a linear function on a compact convex set, so attains its extreme values at extreme points of $P_k$. Example 2 in Section 6 illustrates this observation.

**5. Convergence results.** A natural question which now arises is the following. If $B_n(k)$ yields $g_k$ as the maximum likelihood estimate of the unknown mixing distribution, what can be said about the behavior of $g_k$ as $k$ increases? We answer this in the following theorem, the statement and proof of which draw upon ideas and terminology from Choquet theory. This theory deals with the representation of points in compact convex sets by means of measures on the extreme points of the set; we refer the reader to [1], [3] or [22] for background.

For the present it suffices to recall that a boundary probability measure $g$ on $\text{co}B_n$ is one which has support on the extreme points of the set; in the case of $\text{co}B_n$ the extreme points are $B_n$ itself ([5], Theorem 7.3). We say that a boundary probability measure $G$ represents $f$, or the resultant of $G$ is $f$, if $f(x) = \int_0^1 h_p(x)\, dG(p)$ for $x = 0, 1, \ldots, n$. It is usual to denote the weak*-compact convex set of all boundary probability measures on $\text{co}B_n$ by $Z_1^+(\text{co}B_n)$, while the subset of $Z_1^+(\text{co}B_n)$ representing $f \in \text{co}B_n$ is denoted $Z_f^+(\text{co}B_n)$ ([1], page 86).

THEOREM 1. *Let $B_n$ denote the binomial curve in the simplex of probability mass functions $T_n$ in $\mathbf{R}^{n+1}$ and take $\hat{f} \in T_n \backslash \text{co}B_n$. Let $T_n$ be equipped with Euclidean distance, a weighted Euclidean distance or Kullback–Leibler distance. Then, in each case*:

(i) *There is a unique nearest point to $\hat{f}$ in $\text{co}B_n$, denoted $\hat{f}_\infty$.*

(ii) *There is a unique boundary probability measure $g_\infty$ on $\text{co}B_n$ representing $\hat{f}_\infty$, supported by at most $(n+2)/2$ points in $B_n$.*

*Now let $k \in \mathbf{N}$ and $B_n(k)$ be the associated discrete approximation to $B_n$. Then, again in each case,*

(iii) *There is a unique nearest point in $\text{co}B_n(k)$ to $\hat{f}$, denoted $\hat{f}_k$, and $\hat{f}_k \to \hat{f}_\infty$ in $\mathbf{R}^{n+1}$ as $k \to \infty$.*

(iv) *There is a unique boundary probability measure $g_k$ in $Z_1^+(\text{co}B_n)$, representing $\hat{f}_k$, supported by at most $n+1$ points and such that $g_k \to g_\infty$ as $k \to \infty$ in the weak* topology on $Z_1^+(\text{co}B_n)$.*

PROOF. (i) The contours of the Euclidean, weighted Euclidean and Kullback–Leibler distance functions are strictly convex. It follows that a point in the convex model space $\text{co}B_n$ nearest to $\hat{f}$ will exist and be unique.

(ii) In [27] it was shown that $B_n$ is affinely isomorphic to $C_n$, the moment curve in $\mathbf{R}^n$. Given this affine equivalence, it follows immediately that $\mathrm{co}B_n$ is affinely isomorphic to $\mathrm{co}C_n$, the $n$th moment space (the set of $n$-tuples which are first to $n$th moments for cumulative distribution functions defined on $[0, 1]$).

In [5], Theorem 11.1, it is shown that a point in the (topological) boundary of $\mathrm{co}C_n$ has a unique representation via a boundary probability measure. Moreover, in [6], Chapter 2, Theorem 2.1, it is shown that such a measure is supported by at most $(n+2)/2$ points. We can conclude that the same result holds for $\mathrm{co}B_n$, so (ii) follows.

(iii) That $\hat{f}_k$ exists and is unique follows as in (i). Suppose that $\hat{f}_k$ does not converge to $\hat{f}_\infty$ in $\mathbf{R}^{n+1}$. Then there exists a subsequence $(\hat{f}_{k_j})$ of $(\hat{f}_k)$ which remains outside an $\varepsilon$-ball centered on $\hat{f}_\infty$, for some $\varepsilon > 0$. Now $\mathrm{co}B_n$ is compact so there is a subsequence $(\hat{f}_{k_{j_i}})$ of $(\hat{f}_{k_j})$ which converges to a point in $\mathrm{co}B_n$, $l$ say. Since $\hat{f}_\infty$ is unique, $\|\hat{f} - l\| > d$. As $k$ increases, $\mathrm{co}B_n(k)$ is an improving approximation to $\mathrm{co}B_n$, so $\|\hat{f} - \hat{f}_k\|$ converges to $d$. This contradicts the existence of the convergent subsequence for which $\lim_i \|\hat{f} - \hat{f}_{k_{j_i}}\|$ is greater than $d$. Thus $\hat{f}_k$ converges to $\hat{f}_\infty$ in $\mathbf{R}^{n+1}$.

(iv) Every face of $\mathrm{co}B_n(k)$ is simplicial ([20], Chapter 2, Proposition 17). Since $\hat{f}_k$ lies on such a face, it is uniquely represented by a boundary probability measure $g_k$. By Carathéodory's theorem (see, e.g., [20], Chapter 1, Theorem 11) $g_k$ is supported by at most $n+1$ points.

In order to show that $g_k \to g_\infty$ as $k \to \infty$, in the weak$^*$ topology, we remark that $\mathrm{co}B_n$ is stable ([21], Example 3.4). This is sufficient to ensure that the resultant map $r\colon Z_1^+(\mathrm{co}B_n) \to \mathrm{co}B_n$ is open ([21], Satz 2.6). The resultant map is also (weak$^*$) continuous. This follows from [1], Proposition I.2.2 applied to $a|_{\mathrm{co}B_n}$ with $a$ a linear functional on $\mathbf{R}^{n+1}$.

Let $R$ be the equivalence relation on $Z_1^+(\mathrm{co}B_n)$ induced by the resultant map $r$. Since $r$ is continuous and open, $\mathrm{co}B_n$ has the quotient topology relative to $r$ and the weak$^*$ topology of $Z_1^+(\mathrm{co}B_n)$ ([7], Chapter 3, Theorem 8). Thus the map which takes an element $f \in \mathrm{co}B_n$ to $[Z_f^+(\mathrm{co}B_n)]$, the equivalence class of $Z_f^+(\mathrm{co}B_n)$, from $\mathrm{co}B_n$ to $Z_1^+(\mathrm{co}B_n)/R$, is continuous ([7], page 96).

From (iii) we have that $\hat{f}_k \to \hat{f}_\infty$ in $\mathrm{co}B_n$ whence $[Z^+_{r(g_k)}(\mathrm{co}B_n)] \to [Z^+_{r(g_\infty)}(\mathrm{co}B_n)]$, as $k \to \infty$. Suppose that $g_k \not\to g_\infty$. Then given a sufficiently small neighborhood $U$ of $g_\infty$ there exists a subsequence $(g_{k_j})$ of $(g_k)$ such that $g_{k_j}$ is not in $U$ for all $j$. But $Z_1^+(\mathrm{co}B_n)$ is compact in the weak$^*$ topology, so there exists a convergent subsequence of $(g_{k_j})$, $(g_{k_{j_i}})$ with limit $g'_\infty (\neq g_\infty)$, say. The identification map from $Z_1^+(\mathrm{co}B_n)$ to $Z_1^+(\mathrm{co}B_n)/R$ is continuous so $[Z^+_{r(g_{k_{j_i}})}(\mathrm{co}B_n)] \to [Z^+_{r(g'_\infty)}(\mathrm{co}B_n)]$, while the quotient space $Z_1^+(\mathrm{co}B_n)/R$ is Hausdorff so $[Z^+_{r(g'_\infty)}(\mathrm{co}B_n)] = [Z^+_{r(g_\infty)}(\mathrm{co}B_n)]$, whence $r(g'_\infty) = r(g_\infty)$. Now $g'_\infty$ and $g_\infty$ are boundary probability measures on $\mathrm{co}B_n$ representing the same point, so again using the uniqueness result ([5], Theorem 11.1), we have that $g'_\infty = g_\infty$, a contradiction. Thus $g_k \to g_\infty$ as $k \to \infty$. □

In the nonidentifiable case we can produce a sequence of representing measures for $\hat{f}$ in which the $k$th representing measure is supported by no more than $k+1$ extreme boundary points. Our next result reassures us that such a sequence will possess a subsequence which converges to a representing measure for $\hat{f}$. For the proof of this result, the author is most grateful to Robert Phelps.

THEOREM 2. *Let $B_n$ denote the binomial curve in the simplex of probability mass functions $T_n$ in $\mathbf{R}^{n+1}$ and take $\hat{f} \in \mathrm{co}B_n$. Let $g_k$, supported on the vertices of $\mathrm{co}B_n(k)$, represent $\hat{f}$ for all sufficiently large $k$. Then*:

(i) *There exists a subsequence $g_{k_j}$ of $g_k$ which converges in the weak* topology to a boundary probability measure $g$ on $\mathrm{co}B_n$.*
(ii) *The boundary probability measure $g$ represents $\hat{f}$.*

PROOF. The boundary probability measures $Z_1^+(\mathrm{co}B_n)$ on the compact convex set $\mathrm{co}B_n$ are weak* compact and metrizable, so there exists a subsequence $(g_{k_j})$ of $(g_k)$ which converges in the weak* topology to a boundary probability measure $g$ on $\mathrm{co}B_n$. Thus $\int l\,dg_{k_j} \to \int l\,dg$ as $j \to \infty$, for every real-valued continuous function $l$ on $\mathrm{co}B_n$, with the integrals taken over $B_n$. In particular, this holds for the restriction to $\mathrm{co}B_n$ of every continuous affine function on $\mathbf{R}^{n+1}$. For such a function $e$ we have $e(\hat{f}) = \int e\,dg_{k_j} \to \int e\,dg$ as $j \to \infty$. Thus $e(\hat{f}) = \int e\,dg$ for all such $e$ so $g$ represents $\hat{f}$ (projection onto a coordinate direction is a continuous affine function on $\mathrm{co}B_n$). Note that $g$ is also finitely supported ([6], Chapter 2, Theorem 2.1). □

**6. Numerical examples.** We now illustrate the estimation procedure with an example from the literature in which $\hat{f}$ lies outside the model space. A second example illustrates the nonidentifiable case. Software to determine whether the mixing distribution is identifiable, and to estimate $\hat{f}_k$ and the associated mixing distribution $g_k$, was developed in MATLAB, using the `nnls` and `constr` macros in the optimization toolbox. This software is available from the author on request.

EXAMPLE 1. The simulated data set analyzed here is from Cressie ([4], pages 102–104). In this example, $n = 10$ and $N = 10{,}000$ values are drawn from a mixture of binomials, with $G$ a cumulative distribution function on [0, 1] with mean 0.5 and negligible standard deviation. These values give an empirical mixture density of

$$\hat{f} = (7, 102, 477, 1140, 2053, 2476, 2027, 1173, 437, 98, 10)/10{,}000.$$

Using $B_{10}(10)$, for example, we find $\hat{f}_{10}$ lies on a face of $B_{10}(10)$ with just three vertices, corresponding to values of $p$ of 0.4, 0.5 and 0.6. The associated weights are $\pi_{0.4} = 0.0138$, $\pi_{0.5} = 0.9807$ and $\pi_{0.6} = 0.0055$. Figure 2 illustrated the situation.

Figure 3 numerically displays the convergence of the mixing distribution estimator $g_k$ [Theorem 1(iv)] toward the limiting mixing distribution $g_\infty$ as $k$ increases, using Kullback–Leibler distance. Evidently the process has closely recovered the mixing distribution.

A comparison of the posterior means determined by the Kullback–Leibler estimator, using $k = 100$, and those given in [4], Table 1, is presented in Table 1; the maximum likelihood estimator correctly leads to a value near 0.5 no matter what the observed value.

Two large data sets are given in [17]: a sample of size 3135 from a 15-item test (page 268) and a sample of size 21310 from a 20-item test (page 290). For both data sets, $\hat{f}$ lies outside the binomial mixture model space.

We turn now to the situation where $\hat{f}$ lies inside the model space, illustrating the ideas with a simple example.

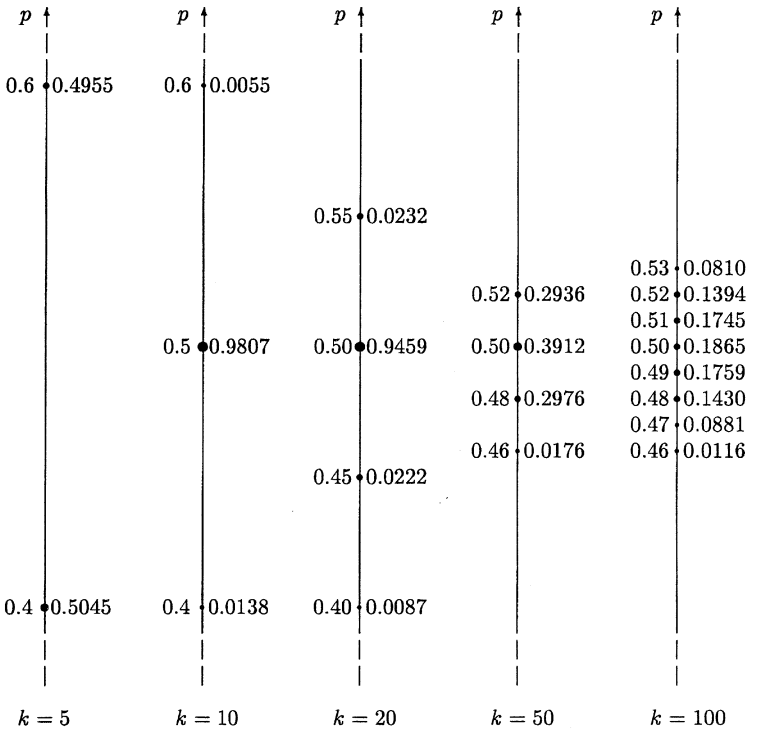


FIG. 3. *Maximum likelihood fitting of the mixing distribution as the approximation to the binomial curve improves (i.e., k increases) from left to right. By the time that* $k = 100$, *the mixing distribution, a small perturbation around* 0.5, *is captured well.*

TABLE 1

*A comparison of the posterior means based on the "quick and easy" empirical Bayes estimator in* [4] *and the maximum likelihood estimator of the mixing distribution. Even for extreme observations the Kullback–Leibler based estimator pushes the estimate to the correct value, very close to* 0.5

| | **Conditional posterior mean** | |
|---|---|---|
| **Observed value** | **"Quick and easy"** | **Based on ML mixture** |
| 0 | 0.100 | 0.4930 |
| 1 | 0.387 | 0.4942 |
| 2 | 0.434 | 0.4955 |
| 3 | 0.485 | 0.4967 |
| 4 | 0.498 | 0.4980 |
| 5 | 0.487 | 0.4993 |
| 6 | 0.503 | 0.5005 |
| 7 | 0.518 | 0.5018 |
| 8 | 0.543 | 0.5030 |
| 9 | 0.624 | 0.5042 |
| 10 | 0.900 | 0.5054 |

EXAMPLE 2. Consider the problem where we approximate $B_2$ using binomial probability values $p$ of 0, $0.\dot{3}$, 0.5, $0.\dot{6}$ and 1, and $\hat{f} = (5/16, 6/16, 5/16)$, a mixture lying inside the convex model space. Figure 4a illustrates the situation. Note that the results of Section 4 remain valid when the values of $p$ are unequally spaced.

As $\hat{f}$ moves within the model space, the slice of mixing distributions $P_4$ is variously a triangle, a quadrilateral or a pentagon. For the $\hat{f}$ of this example, $P_4$ is a pentagon, as shown in Figure 4b. The values of the posterior mean, as $x$ runs through 0, 1 and 2 and $g$ runs over the five vertices, are shown in Table 2, with the range of these values shown on the right. For a given $x$, any prior distribution in $P_4$ would produce a posterior mean in this interval.

Note how vertices $v_2$ and $v_4$ determine the extremes of the interval. These are the distributions with smallest and largest third moment about the origin, respectively. We caution the reader that the intervals presented in the table are not confidence intervals in the usual sense. Two stages of widening would have to take place for this to be the case: we should use $\text{co}B_n$ as the model space and we should allow $\hat{f}$ to range over a confidence region in $\text{co}B_n$. Methods for finding the vertices of the slice of mixing distributions are described in [19].

**7. How likely is identifiability?** How likely is it that $\hat{f}$ will lie outside $\text{co}B_n$? There are two immediate reasons why the empirical mixture may not lie in the model space: first, sampling variation, particularly in the tails, can

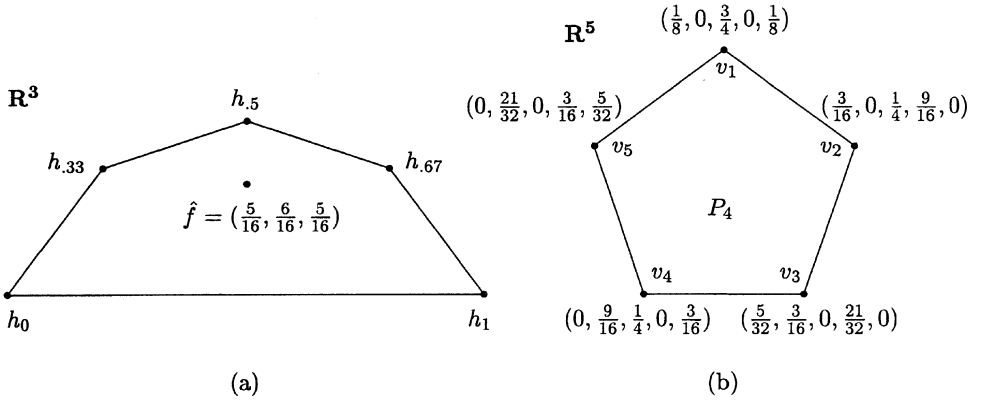


FIG. 4. *An example in which the empirical mixture* $\hat{f}$ *lies inside the polytope of mixtures. In* (a) *the polytope is seen to be the convex hull of five points on the binomial curve* $B_2$ *and* $\hat{f} = (5/16, 6/16, 5/16)$ *lies inside the polygon of mixtures. The associated set of mixing distributions* $P_4$ *is shown in* (b) *as a pentagon in* $\mathbf{R}^5$.

push $\hat{f}$ outside co$B_n$ and second, the binomial mixture model may be wrong. We now explore the effect of sampling variation, but should bear in mind that in practice an incorrect model will also favor identifiability.

In [27] an initial investigation of this question was begun, examining the probability that a randomly chosen distribution in $T_n$, the simplex of all distributions on $\{0, 1, \ldots, n\}$, is a mixture of binomial distributions. We considered $T_n$ equipped with Lebesgue measure. For $n = 2$ the probability is 2/3, for $n = 3$ it is 3/10, for $n = 4$ it is 0.0914 while for $n = 5$ it drops to 0.0189. By the time $n = 10$ the probability is $2.14 \times 10^{-8}$ (see [27], Table 1). The volume of the model space of binomial mixtures relative to the volume of the simplex of all probability mass functions on $\{0, 1, \ldots, n\}$ shrinks rapidly to zero.

This was very much a first view of the problem and did not take into account the fact that $\hat{f}$, under the model, is centered on a point inside co$B_n$ and will

TABLE 2
*Values of the posterior mean, for all combinations of data* $x$ *and vertex* $v_i$ *of the slice of mixing distributions,* $P_4$

| | | Vertex | | | | | Range of posterior mean |
|---|---|---|---|---|---|---|---|
| | | $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ | |
| $x$ | 0 | 0.30 | 0.23 | 0.24 | 0.37 | 0.36 | [0.23, 0.37] |
| | 1 | 0.50 | 0.61 | 0.59 | 0.39 | 0.41 | [0.39, 0.61] |
| | 2 | 0.70 | 0.63 | 0.64 | 0.77 | 0.76 | [0.63, 0.77] |

TABLE 3

*Simulation results showing that as n increases the probability that $\hat{f}$ lies inside the model space decreases rapidly. Low sample sizes N also decrease the chance that $\hat{f}$ lies in the model space. In each case* 1000 $\hat{f}$ *points were generated*

| $n$ | Number of values N generated from $f$ to produce an $\hat{f}$ | Proportion of the $\hat{f}$ inside co$B_n(100)$ |
|---|---|---|
| 2 | 50 | 0.9840 |
| | 100 | 0.9999 |
| | 500 | 1.0000 |
| 5 | 100 | 0.3820 |
| | 500 | 0.8490 |
| | 1000 | 0.9530 |
| 10 | 1000 | 0.0030 |
| | 5000 | 0.0390 |
| 15 | 10000 | 0.0000 |

certainly not generate empirical distributions uniformly on the simplex. A simulation was conducted in order to shed more light on this issue.

For $n = 2, 5, 10$ and 15 and $f$ centrally located in co$B_n$ [in fact a uniform convex combination of the vertices of co$B_n(100)$], 1000 $\hat{f}$ points were generated, each based upon a predetermined number $N$ of observations from $f$. Table 3 reports the proportion of these $\hat{f}$ distributions lying in co$B_n(100)$. Note that as $n$ increases or $N$ decreases, it becomes less likely that $\hat{f}$ will lie in co$B_n(100)$. The software developed yields the minimum Euclidean distance between $\hat{f}$ and co$B_n(100)$, so indicating whether $\hat{f}$ lies in the mixture set.

To conclude, for typical sample sizes, the "thinness" of the model space within the simplex, coupled with real deviations from the model, make it very likely that $\hat{f}$ will lie outside co$B_n$ for values of $n$ of ten and beyond.

**8. Summary and discussion.** Given a mixture of binomial distributions, the problem of estimating the mixing distribution has long been considered intractable, since the mixture determines the mixing distribution only up to the first $n$ moments.

In this paper we have looked closely at the geometry underlying this problem and revealed two helpful aspects. First, sampling variation and model inadequacy often render the maximum likelihood estimator of the mixing distribution identifiable. Second, very convenient geometry makes the maximum likelihood estimator of the mixing distribution estimable in the identifiable case.

Our investigations are summarized in Table 4. If $\hat{f}$ lies outside co$B_n$ then a unique geometrically motivated maximum likelihood estimator of $G$ is available, with support on no more than $(n + 2)/2$ points in [0, 1]. An improving sequence of approximations to this estimator has been described. On the other

TABLE 4
*A summary of the estimation procedure, for small and large n*

| | | **Location of $\hat{f}$** | |
|---|---|---|---|
| | | **Outside co$B_n$** | **Inside co$B_n$** |
| $n$ | Small | Estimate unique | Estimate not unique (problem tractable) |
| | Large | Estimate unique | Estimate not unique (problem intractable) |

hand, if $\hat{f}$ belongs to co$B_n$ then $G$ is not identifiable. Finding the vertices of the polytope of priors then facilitates estimation of the posterior mean; this is a tractable task when $n$ is small.

It is interesting in light of this insight to reflect on the "negative probabilities" found by Lord in [17], page 268 and Figure 1. Representing a point outside the model space inevitably will produce negative weights on points in the binomial curve. Lord's procedure also correctly leads to the discreteness which we have seen is inherent in the maximum likelihood estimate of the mixing distribution.

We conclude by remarking that the simplex of probability mass functions on $n+1$ points, $T_n$, is a foliation of exponential families (see, for example, [23], Section 3.1); the binomial distributions form just one strand. The methods of this paper can be used to unravel mixtures of any such family, for example, mixtures of truncated Poisson distributions.

**Acknowledgments.** For introduction to the problem and encouragement, the author is most grateful to John Deely. Ian Coope, Phillip Wolfe and Zelda Zabinsky are sincerely thanked for optimization advice, Persi Diaconis for encouragement and many references, Bent Fuglede for pointing out the importance of stable convex sets, Vic Klee for introducing the author to cyclic polytopes and Michael Perlman for his kind and constructively critical advice.

## REFERENCES

[1] ALFSEN, E. M. (1971). *Compact Convex Sets and Boundary Integrals*. Springer, Berlin.
[2] BOHNING, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *J. Statist. Plann. Inference* **47** 5–28.
[3] CHOQUET, G. (1969). *Lectures on Analysis* (J. Marsden, T. Lance and S. Gelbart, eds.). Benjamin, New York.
[4] CRESSIE, N. (1979). A quick and easy empirical Bayes estimate of true scores. *Sankhyā Ser. B* **41** 101–108.
[5] KARLIN, S. and SHAPLEY, L. S. (1953). Geometry of moment spaces. *Mem. Amer. Math. Soc.* **12**.

[6] KARLIN, S. and STUDDEN, W. J. (1966). Tchebycheff Systems: with applications in analysis and statistics. In *Pure and Applied Mathematics* (R. Courant, L. Bers, J. J. Stoker, eds.) **15**. Interscience, New York.
[7] KELLEY, J. L. (1955). *General Topology*. Van Nostrand, Princeton.
[8] LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc*. **73** 805–811.
[9] LESPERANCE, M. L. and KALBFLEISCH, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *J. Amer. Statist. Assoc*. **87** 120–126.
[10] LINDSAY, B. G. (1981). Properties of the maximum likelihood estimator of a mixing distribution. In *Statistical Distributions in Scientific Work* (C. Taillie, ed.) **5** 95–109. Reidel, Dordrecht.
[11] LINDSAY, B. G. (1983). The geometry of mixture likelihoods: a general theory. *Ann. Statist*. **11** 86–94.
[12] LINDSAY, B. G. (1983). The geometry of mixture likelihoods II: the exponential family. *Ann. Statist*. **11** 783–792.
[13] LINDSAY, B. G. (1986). Exponential family mixture models (with least squares estimators). *Ann. Statist*. **14** 124–137.
[14] LINDSAY, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. IMS, Hayward, CA.
[15] LINDSAY, B. G., CLOGG, C. C. and GREGO, J. (1991). Semi-parametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc*. **86** 96–107.
[16] LINDSAY, B. G. and LESPERANCE, M. L. (1995). A review of semiparametric mixture models. *J. Statist. Plann. Inference* **47** 29–39.
[17] LORD, F. M. (1969). Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). *Psychometrika* **34** 259–299.
[18] LORD, F. M. and CRESSIE, N. (1975). An empirical Bayes procedure for finding an interval estimate. *Sankhyā Ser. B* **37** 1–9.
[19] MATTHEISS, T. H. and RUBIN, D. S. (1980). A survey and comparison of methods for finding all vertices of convex polyhedral sets. *Math. Oper. Res*. **5** 167–185.
[20] MCMULLEN, P. and SHEPHARD, G. C. (1971). *Convex Polytopes and the Upper Bound Conjecture*. Cambridge Univ. Press.
[21] PAPADOPOULOU, S. (1982). Stabile konvexe Mengen. *Jahresber. Deutsch. Math.- Verein*. **84** 92–106.
[22] PHELPS, R. R. (1966). *Lectures on Choquet's Theorem*. Van Nostrand, Princeton.
[23] SAMPSON, A. R. and SMITH, R. L. (1982). Assessing risks through the determination of rare event probabilities. *Oper. Res*. **30** 839–866.
[24] SAVILLE, D. J. and WOOD, G. R. (1991). *Statistical Methods: The Geometric Approach*. Springer, New York.
[25] SIVAGANESAN, S. and BERGER, J. (1993). Robust Bayesian analysis of the binomial empirical Bayes problem. *Canad. J. Statist*. **21** 107–119.
[26] TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38** 290–295.
[27] WOOD, G. R. (1992). Binomial mixtures and finite exchangeability. *Ann. Probab*. **20** 1167–1173.

MASSEY UNIVERSITY
INSTITUTE OF INFORMATION SCIENCES AND TECHNOLOGY
COLLEGE OF SCIENCES
PRIVATE BAG 11 222
PALMERSTON NORTH
NEW ZEALAND
E-MAIL: g.r.wood@massey.ac.nz