The Annals of Statistics
1989, Vol. 17, No. 2, 722–740

MOMENT MATRICES: APPLICATIONS IN MIXTURES[1]

BY BRUCE G. LINDSAY

The Pennsylvania State University

The use of moment matrices and their determinants are shown to elucidate the structure of mixture estimation as carried out using the method of moments. The setting is the estimation of a discrete finite support point mixing distribution. In the important class of quadratic variance exponential families it is shown for any sample there is an integer $\hat{\nu}$ depending on the data which represents the maximal number of support points that one can put in the estimated mixing distribution. From this analysis one can derive an asymptotically normal statistic for testing the true number of points in the mixing distribution. In addition, one can construct consistent nonparametric estimates of the mixing distribution for the case when the number of points is unknown or even infinite. The normal model is then examined in more detail, and in particular the case when σ^2 is unknown is given a comprehensive solution. It is shown how to estimate the parameters in a direct way for every hypothesized number of support points in the mixing distribution, and it is shown how the structure of the problem yields a decomposition of variance into model and error components very similar to the traditional analysis of variance.

1. Introduction and summary. Although the method of moments has long been in disfavor because of its inefficiency relative to maximum likelihood, there are times that its simple form can be an instrument of convenience. The objective of this article is to demonstrate that in the technically difficult problem of determining an unknown mixing distribution there is an elegant and useful mathematical structure behind the method of moments. With this as a tool one can explore in a straightforward way a number of problems which are considerably more formidable in a likelihood analysis. For example, questions concerning the number of support points to the distribution can be answered by considering the determinants of certain matrices of moments. Estimators of discrete mixing distributions will be unique, when they exist, with easy to solve equations. In the normal mixture problem, one can directly estimate the normal component variance associated with an unknown p-point mixing distribution. The methods are easily programmed in any language that offers direct matrix manipulation. As such, these estimators also provide a computationally fast way to find consistent initial values for a likelihood maximization algorithm. Further comments on computation will be relegated to the discussion in Section 6; the main objective here is to lay out the theory. For a general background on the discrete mixing distribution problem the book by Titterington, Smith and Makov (1985) is recommended.

Received January 1987; revised May 1988.

[1]This research was supported by National Science Foundation Grant DMS-88-01514.

AMS 1980 *subject classifications*. Primary 62E10, 62G05; secondary 62H05.

Key words and phrases. Moment matrix, Hankel determinant, method of moments, quadratic variance exponential family, mixing distribution, mixture model.

We start with the formal problem of interest: The random variable X is said to have a *mixture distribution* relative to a parametric family of distributions $\{F_\theta\colon \theta \in \Omega\}$ if X has the distribution function

$$F_Q(x) = \int F_\theta(x)\, dQ(\theta).$$

Here Q, the *mixing distribution*, is a distribution on the parameter space Ω. This article focusses upon the so-called finite mixtures problem [e.g., Everitt and Hand (1981) and Titterington, Smith and Makov (1985)], in which Q is assumed to be a discrete distribution with a finite number of points of support. That number will be denoted $\nu = \nu(Q)$. We will write the mixing distribution as

$$Q(\theta) = \sum \pi_j \delta(\theta_j),$$

with $\theta_1, \ldots, \theta_\nu$ being the unknown support points and $\pi_1, \ldots, \pi_\nu$ being the unknown masses. The discussion will start by assuming that ν is known, but inference on this number will be discussed, as will inference on Q in the presence of an unknown scale parameter σ in the normal model.

The primary tool used in this analysis is the moment matrix, which we now define. Let G be a distribution with $2p$ moments, say $m_1 = m_1(G) = E[X]$, $m_2(G), \ldots, m_{2p}(G) = E[X^{2p}]$. The pth *moment matrix* of G is

$$\mathbf{M}_p(G) = \begin{bmatrix} 1 & m_1 & m_2 & \cdots & m_p \\ m_1 & m_2 & m_3 & \cdots & m_{p+1} \\ m_2 & m_3 & m_4 & \cdots & m_{p+2} \\ \vdots & & & & \vdots \\ m_p & \cdot & \cdot & \cdot & m_{2p} \end{bmatrix}. \tag{1.1}$$

Of crucial interest to us is the way that the structure of $\mathbf{M}_p$ reveals information about the number and location of the support points for a discrete distribution G. A companion article [Lindsay (1989)] discusses a number of features of these matrices. Of particular relevance to this article is the following representation of $\det M_p(G)$: If $X_0, X_1, \ldots, X_p$ are independent and identically distributed with distribution G, then

$$\det M_p(G) = E\left[\prod_{i>j} (X_i - X_j)^2\right].$$

We will repeatedly draw upon representations of this form to gain insight into the form of certain related determinants; all these results can be proved by the method described in the Appendix of Lindsay (1989).

These results relate to the method-of-moments mixture problem as follows: One method to estimate an unknown mixing distribution Q might be to estimate a set of its moments $m_p(Q) = \int \theta^p\, dQ(\theta)$ and then, from these estimated moments, determine the corresponding distribution function. If we wish to estimate Q with a p-point distribution, then it is clear that we will need to estimate $2p - 1$ moments in order to have enough constraints to determine the $2p - 1$ dimensional parameter set $\theta_1, \ldots, \theta_\nu, \pi_1, \ldots, \pi_\nu$.

Implicit so far in this discussion is the potential to use the moments of a function of θ, say $g(\theta)$, chosen so that the moment system can be easily and consistently estimated. There may, in fact, be more than one way to construct the function $g(\theta)$. One construction described here applies if the parametric family $\{F_\theta\}$ is a quadratic variance exponential family [Morris (1982, 1983)]. For them it is possible to construct unbiased estimators $\hat{m}_p$ of the moments $m_p(Q)$ of the mixing distribution Q on the mean-value parameter; these estimators use a linear combination of the sample moments of X of order p and lower. This and other methods of construction will be discussed in Section 2.

However, when we construct estimates of moments, the sequence of estimated values $1, \hat{m}_1, \ldots, \hat{m}_{2p-1}$ need not correspond to any distribution Q. Section 2 presents simple methods to determine from the estimated moment matrices whether a solution exists. When it exists, the construction is straightforward. In the process it will be shown that there is a random number $\hat{\nu}$ such that for every $p \leq \hat{\nu}$ a p-point estimate $\hat{Q}_p$ exists, whereas for $p > \hat{\nu}$ there does not exist a method-of-moments estimate. The theory can easily be modified so that the estimated support points satisfy constraints placed on the parameter space.

The above analysis leads in a straightforward way to the consideration of inferential methods for testing hypotheses about the value of $\nu(Q)$. In Section 3 it is shown how to construct an asymptotically normal test statistic for the adequacy of any particular number of points. In contrast to the likelihood ratio test for this hypothesis, the test statistic is explicitly defined and the limiting distribution is known and simple. The logical next step is to consider nonparametric estimation of Q by $\hat{Q}_{\hat{\nu}}$. The consistency of this method is verified in Section 3 also.

The above results apply in particular to the normal mixtures model in which there is a mixing distribution Q on μ and the variance σ^2 is known. In Sections 4 and 5, the method-of-moments analysis is extended to the case where σ is unknown. It will be shown that for every value of $p = \nu(Q)$ there exists a consistent estimator $\hat{\sigma}_p$ for σ and a corresponding uniquely defined p-point mixing distribution $\hat{Q}_p$ such that the first $2p$ sample moments of X are equal to the first $2p$ moments of the estimated distribution. The structure of the solution as a function of p yields a decomposition similar to the analysis of variance in a sequence of nested linear models, the nesting in this case being upon the number of components in the mixing distribution.

2. The moment estimators.

2.1. *The construction of moment estimators.* The objective of this section is to provide a concise description of the method-of-moments estimators of mixing distributions in certain important families of distributions. Some of the results have been presented before in Titterington, Smith and Makov (1985). We provide here a unified treatment, together with new results concerning the existence of the solution and appropriate corrections for solutions outside the parameter space.

Morris (1982, 1983) identified an important useful common element the normal, binomial, negative binomial, gamma and Poisson families of distributions. He called them *quadratic variance natural exponential families* because within each family the variance of the random variable X is a quadratic function of the mean-value parameter $E[X]$. Morris demonstrated that there are just six such natural exponential families (modulo certain transformations), the one not yet mentioned being the generalized hyperbolic secant. This "quadratic variance property" has a large range of statistical implications, one of which we now exploit.

That is, for each family in the class of quadratic variance exponential families (QVEF) there is a polynomial of degree p in X which is an unbiased estimator of the power μ^p of the mean value parameter μ. In particular, if $f(x; \mu)$ is the density function in its mean-value parametrization and if μ_0 is a particular value of μ, then there exists a constant c_p depending on the family such that for $\gamma_p(x) := c_p\{d^p f(x; \mu)/d\mu^p\}/f(x; \mu)$, evaluated at μ_0, one has [Morris (1982), (8.8)]

$$E\big[\gamma_p(X); \mu\big] = (\mu - \mu_0)^p.$$

It follows that if Q is a mixing distribution on the mean-value parameter μ of a QVEF family and X has the corresponding mixture distribution F_Q, then $\gamma_p(X)$ is an unbiased estimator of the pth moment of μ about μ_0, $m_p(Q)$. For example, in the Poisson model we have $\gamma_p(x) = x(x-1)\ldots(x-p+1)$ as an unbiased estimator of μ^p. If we observe a sample of size n from the mixed distribution F_Q, then $\hat{m}_p = \bar{\gamma}_p$, the mean value of γ_p, is an unbiased estimator of $m_p(Q)$,

$$E_{F_Q}(\bar{\gamma}_p) = \int \gamma_p(x_i)\, dF_\mu(x_i)\, dQ(\mu)$$

$$= \int (\mu - \mu_0)^p\, dQ(\mu).$$

When the number of support points of the unknown mixing distribution Q is given to be "p" we define the method-of-moments estimator of Q to be any p-point distribution $\hat{Q}_p$ which satisfies the $2p-1$ equations

$$\hat{m}_j = m_j\big(\hat{Q}_p\big), \qquad j = 1, \ldots, 2p-1. \tag{2.1}$$

We note that Q has $2p-1$ unknown parameters, so this appears to be a well-determined system.

The choice of the point μ_0 will have no impact on the solution because this method of moments is translation-equivariant. It can therefore be chosen so that the estimators have a simple structure. In fact, the system (2.1) could equivalently have been derived by setting the first $2p-1$ sample moments of X equal to their expectations under F_Q. Therefore this estimator is the usual method-of-moments estimator. The advantage of formulation in terms of the moments of Q,

as in (2.1), comes in determining whether there exists a feasible solution to the equations, as will be shown in Corollary 2B below.

The following results will apply as well to moment estimators constructed from other sequences of moments than those of the mean-value parameter. For example, we could use the following relationships:

Normal:

$$E\left[e^{p(tX)-\sigma^2t^2p^2/2}\right] = \left[e^{t\mu}\right]^p.$$

Exponential:

$$E\left[I[X > pt]\right] = \left[e^{-t\lambda}\right]^p \quad \text{[Brockett (1977)]}.$$

Poisson:

$$E\left[n!I[X = n]\right] = \mu^p e^{-\mu} \quad \text{[Tucker (1963)]}.$$

The last example illustrates a case where one estimates a system of weighted moments $\int \theta^p w(\theta)\, dQ(\theta)$. In this case, provided the weight function $w(\theta)$ is strictly positive, one solves for an estimate of Q by first estimating the weighted distribution Q^* defined by $dQ^*(\theta) = w(\theta)\, dQ(\theta)$, then performing the appropriate transformation.

2.2. *Solving the moment equations.* The following theorem establishes the relevant properties of moment sequences to determine whether a moment solution exists. This result can also be found in Mammana (1954), and has been applied to the mixture problem by Tucker (1963) and Brockett (1977).

THEOREM 2A. (a) *A sequence of numbers* $1, m_1, m_2, \ldots, m_{2p}$ *are the moments of a distribution with exactly p points of support if and only if* $\det \mathbf{M}_1 > 0, \det \mathbf{M}_2 > 0, \ldots, \det \mathbf{M}_{p-1} > 0$ *and* $\det \mathbf{M}_p = 0$.

(b) *If the sequence of numbers* $1, m_1, m_2, \ldots, m_{2p-2}$ *satisfies* $\det \mathbf{M}_1 > 0, \ldots, \det \mathbf{M}_{p-1} > 0$ *and* m_{2p-1} *is any scalar, then there exists a unique p-point distribution with exactly those initial* $2p - 1$ *moments.*

PROOF. Part (a) is implicitly given in Uspensky (1937). Part (b) comes from (a) as follows: Since $\det \mathbf{M}_p$ is linear in m_{2p}, with coefficient $\det \mathbf{M}_{p-1}$ being strictly positive, one can choose m_{2p} so that the conditions for part (a) are satisfied, regardless of the value of m_{2p-1}. (The uniqueness of the distribution will be clear from the method of reconstruction given below.) □

This leads in an obvious fashion to the following corollary. Let $\hat{\mathbf{M}}_p$ have the form of a moment matrix but with estimated moments $\hat{m}_p$. Let $\hat{d}_p$ be its determinant. Define $\hat{\nu} = 1 + \sup\{p\colon \det \hat{\mathbf{M}}_p > 0\}$; that is, we have $\hat{d}_1 > 0, \ldots, \hat{d}_{\hat{\nu}-1} > 0$ but $\hat{d}_{\hat{\nu}} \leq 0$.

COROLLARY 2B. *If* $p > \hat{\nu}$, *then there does not exist a solution to the moment equations. If* $p \leq \hat{\nu}$, *then there exists a unique solution* $\hat{Q}_p$.

Reconstruction of the mixing distribution from its moments is presented in Titterington, Smith and Makov (1985); a brief recapitulation is offered here because it will be useful in discussing the boundary problem.

The first step in reconstructing the distribution from its moments is to determine the points of support. We start by defining a polynomial,

$$S_p(t) = \det \begin{bmatrix} 1 & \hat{m}_1 & \cdots & \hat{m}_{p-1} & 1 \\ \hat{m}_1 & \hat{m}_2 & \cdots & \hat{m}_p & t \\ \vdots & & & \vdots & \vdots \\ \hat{m}_p & \cdots & & \hat{m}_{2p-1} & t^p \end{bmatrix}. \tag{2.2}$$

THEOREM 2C. *If a solution $\hat{Q}_p$ to the moment equations exists, then it has support points equal to the roots of $S_p(t) = 0$.*

This is again a standard result which can be found in Uspensky (1937). We do note, however, that by using the representation methods of Lindsay (1989) one can prove

$$S_p(t) = E\left[\prod_{j<k}(Y_j - Y_k)^2 \prod_i (t - Y_i)\right]/p!,$$

where $Y_1, Y_2, \ldots, Y_p$ are a random sample from the distribution having the required moments. This representation makes the result clear, and in particular, we see that if the support points of the distribution are $r_1, \ldots, r_p$, then $S_p(t) = \Pi(t - r_i)\det \mathbf{M}_{p-1}$.

An interesting and computationally useful feature of the support points is presented in the following lemma.

LEMMA 2D. *If $\hat{\nu} \geq p$ and the roots to $S_p(t)$ and $S_{p-1}(t)$ are $r_1 < \cdots < r_p$ and $s_1 < \cdots < s_{p-1}$, respectively, then the roots are interwoven:*

$$r_1 < s_1 < r_2 < s_2 < \cdots < s_{p-1} < r_p.$$

PROOF. In Uspensky (1937) it is shown (page 369) that

$$S_p'(t)S_{p-1}(t) - S_{p-1}'(t)S_p(t)$$

is a positive number. Hence at zeros of S_p, S_{p-1} has the same sign as S_p'. Since S_p has a full complement of zeros, its derivative must alternate sign as we proceed through them left to right, which in term implies that S_{p-1} has a zero between each of them. Since it has exactly $p - 1$ zeros, this describes the location of all of them. □

Given the roots $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_p$ to the polynomial $S_p(t)$, it is straightforward to solve for the masses $\hat{\pi}_i$ at each support point $\hat{\mu}_i$ by solving the linear system of

equations

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \hat{\mu}_1 & \hat{\mu}_2 & \cdots & \hat{\mu}_p \\ \vdots & \vdots & & \vdots \\ \hat{\mu}_1^{p-1} & \hat{\mu}_2^{p-1} & \cdots & \hat{\mu}_p^{p-1} \end{bmatrix} \begin{bmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \vdots \\ \hat{\pi}_p \end{bmatrix} = \begin{bmatrix} 1 \\ \hat{m}_1 \\ \vdots \\ \hat{m}_{p-1} \end{bmatrix}. \tag{2.3}$$

The matrix on the left is nonsingular, being a Vandermonde matrix, so that there is a unique solution to these equations.

A final important point is that the moment equations will have a solution with the right number of points, at least in the limit.

THEOREM 2E. *If the mixing distribution has ν points of support, where $\nu = \infty$ is allowed, then on a set of realizations with probability* 1,

$$\liminf_{n\to\infty} \hat{\nu}_n \geq \nu.$$

PROOF. The strong law of large numbers implies that the estimated moments and hence the determinants $\hat{d}_1, \ldots, \hat{d}_p$, for any value of p, converge almost surely. If ν is finite, then it follows that on a set of realizations ω with probability 1, there exists an $N(\omega)$ such that for all $n > N(\omega)$ the determinants $\hat{d}_1, \ldots, \hat{d}_{\nu-1}$ are strictly positive, and so $\hat{\nu}_n(\omega) \geq \nu$. If $\nu = \infty$, the same argument shows that $\liminf \hat{\nu}_n \geq p$ for every p. □

From this result we can safely conclude that the estimator is consistent: We know that the first $2p - 1$ moments of $\hat{Q}_n$ converge to those of Q, but since p-point distributions are completely determined by those $2p - 1$ moments through (2.2) and (2.3), we have almost sure convergence of the masses and support points to their true values.

2.3. *Satisfying constraints on the parameter space.* If a solution exists, then the estimated weights in (2.3) are necessarily positive. However, it is not necessarily true that the support points [the roots of $S_p(t) = 0$] are within the parameter space of the mean-value parameter. In this section we describe modifications to the moment estimators which are designed so that the mixing distribution will be consistently estimated even when it does place mass on the boundary of the parameter space.

The appropriate modification to the solution when support points fall outside the parameter space requires careful thought. Consider for illustration the Poisson model, for which the mean-value parameter space is $[0, \infty)$. Suppose that $\mu = 0$ is a support point in the true mixing distribution. As $n \to \infty$ it can be shown that there will exist nonvanishing probability that the leftmost support in the estimated distribution will be negative. A simple repair to use on a method-of-moments estimator when one support point lies outside the boundary of the parameter space would be to simply replace that support point with the bound-

ary value, in this case 0, using the same mass. The arguments of the previous section indicate that this would be a consistent procedure.

Simple as this method is, it has some undesirable features. The estimated distribution no longer matches any moments, and in particular, the described procedure clearly shifts the mean of Q and reduces the variance. We propose instead a procedure which puts mass on the boundary in such a way that the initial moments of order 1 to $2p - 2$ are matched. In addition to preserving the natural moments of the distribution, it eliminates the dependence of the estimator on the highest-order moment, which is often the one with the largest variability.

We suppose that the parameter space is $[0, \infty)$, and we start by describing the key features of the moment sequence which indicate a violation of the boundary. First, define the shifted moment matrix $\mathbf{M}_p^*$ as that $p + 1 \times p + 1$ matrix with (i, j)th entry m_{i+j+1}. Thus $\mathbf{M}_0^* = m_1$, $\mathbf{M}_1^* = \begin{bmatrix} m_1 & m_2 \\ m_2 & m_3 \end{bmatrix}$ and so forth. Let $d_p^* = \det \mathbf{M}_p^*$, and use $\hat{}$ to denote the corresponding matrices and determinants using estimated moments. Note that if $Y_0, Y_1, \ldots, Y_p$ are a random sample from the distribution generating $\mathbf{M}_p^*$, then we have the representation [à la Lindsay (1989)]

$$d_p^* := \det \mathbf{M}_p^* = E\left[\prod_i Y_i \prod_{k>j} (Y_k - Y_j)^2\right] / (p+1)!.$$

From this it is clear that the determinant sequence for a ν-point distribution on $[0, \infty)$ which has positive mass at 0 will have the form $d_1^* > 0, \ldots, d_{\nu-2}^* > 0$, $d_{\nu-1}^* = d_\nu^* = \cdots = 0$. If it has no mass at 0, d_ν^* will be the first zero determinant. In fact, it can be shown [Shohat and Tamarkin (1943)] that the sequence of numbers $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_{2p}$ are the moments for a *nonnegative* distribution with p points by verifying (in addition to Theorem 2A) that $\hat{d}_0^* > 0, \ldots, \hat{d}_{p-2}^* > 0$ and $\hat{d}_{p-1}^* \geq 0$, where the last determinant is 0 if and only if one of the support points of the distribution is 0.

Based on this, one can define the modified method-of-moments solution as follows. First, we define $\hat{\nu}^* = 1 + \sup\{p\colon d_p^* > 0\}$. We suppose that $\hat{\nu} \geq p$, so that a p-point solution to the original moment problem exists, and we consider the following cases:

Case I: $\hat{\nu}^* \geq p$ implies that the solution to the original moment problem has all support points in the interior of the parameter space.

Case II: $\hat{\nu}^* = p - 1$ implies that the p-point solution to the original moment problem puts one support point in $(-\infty, 0]$.

Case III: $\hat{\nu}^* < p - 1$ implies that the p-point solution has more than one negative mass point.

Case I needs no correction. In case III, there is no sensible way to create a suitable p-point distribution. In Case II, we can create a p-point distribution with positive mass at 0, the remaining mass on positive values, which fits the moments $1, \hat{m}_1, \ldots, \hat{m}_{2p-2}$. (Note that one less moment is needed because of the

forced inclusion of 0.) The nonzero support points of this distribution can be found as the roots of the polynomial of degree $p-1$ [compare with (2.2)]

$$S_p^*(t) = \begin{bmatrix} \hat{m}_1 & \hat{m}_2 & \cdots & \hat{m}_{p-1} & 1 \\ \vdots & & & \hat{m}_p & t \\ & & & \vdots & \vdots \\ \hat{m}_p & \cdots & & m_{2p-2} & t^{p-1} \end{bmatrix}.$$

The key role of this polynomial can be seen in the representation

$$S_p^*(t) = E\left[\prod_i Y_i \prod_{k>j} (Y_k - Y_j)^2 \prod_i (t - Y_i)\right]/p!,$$

where $Y_1, \ldots, Y_{p-1}$ is a sample from the distribution with the given moments.

We next consider the problem of keeping the estimated distribution in a fixed range, which, without loss of generality, we can suppose to be $[0,1]$. An example of this would be the binomial (N, μ) distribution. From the root interlacing property we can see that it is possible that if $\hat{Q}_{p-1}$ has mass in $(0,1)$, then $\hat{Q}_p$ has at most two support points violating the constraints of the parameter space, one to the left of 0 and one to the right. We now augment our skills to determine if a violation to the right has occurred. Define yet another moment matrix by $M_p^{**} = M_p - M_p^*$, with determinant d_p^{**}, and let $\nu^{**} = 1 + \sup\{p\colon d_p^{**} > 0\}$. Again, we have several cases to consider: If $\hat{\nu} \geq p$, then the p-point distribution $\hat{Q}$ has support points

(a) in $(0,1)$ if $\nu^* \geq p$ and $\nu^{**} \geq p$;

(b) one point negative, the rest in $(0,1)$ if $\nu^* = p-1$ and $\nu^{**} \geq p$;

(c) one point greater than 1, the rest in $(0,1)$, if $\nu^* \geq p$ and $\nu^{**} = p-1$;

(d) one negative point, one point greater than 1, if $\nu^* = p-1$ and $\nu^{**} = p-1$.

In other cases it is not feasible to correct the distribution in a way yielding p-points in the space $[0,1]$. In cases (b) and (c), the first attempt is to correct the fitted distribution by fitting just the first $2p-2$ moments, as described above. We do note, however, that in shifting the negative mass point to the right, we may shift the rightmost point into violation of right boundary point, or vice versa. In this case and in case (d), we fit only the first $2p-3$ moments, using a distribution with mass at 0, 1 and the roots of the polynomial

$$\det\begin{Bmatrix} m_1 - m_2 & m_2 - m_3 & \cdots & m_{p-3} - m_{p-2} & 1 \\ m_2 - m_3 & m_3 - m_4 & \cdots & m_{p-2} - m_{p-1} & t \\ \vdots & \vdots & & \vdots & \vdots \\ m_{p-2} - m_{p-1} & m_{p-1} - m_p & \cdots & m_{2p-4} - m_{2p-3} & t^{p-3} \end{Bmatrix}.$$

It is an easy exercise to construct the appropriate representation which shows that this polynomial identifies all non-$\{0,1\}$ mass points.

The polynomial (2.2) which yields the support points for $\hat{Q}_p$ was given for the binomial by Blischke (1964) and for the Poisson by Everitt and Hand (1981). Titterington, Makov and Smith (1985) generalized it to the families of the quadratic variance type (without explicitly identifying this feature). An additional family, the Weibull, is shown there to share the property of having an unbiased estimator of the pth moment of Q which is polynomial of degree p in X. Thus this method applies to it as well.

3. Extended method of moments. We have now completed a description of the classical method-of-moments estimators for an unknown mixing distribution when the number of points p in the distribution is treated as known. An important aspect was the role of $\hat{\nu}$ in determining the existence of a solution. In this section we suppose that $\nu(Q)$, the number of support points in Q, is unknown. The methods developed in Section 2 provide a natural framework for developing methods for inference on $\nu(Q)$ and for defining an estimator for Q which relies on no assumptions concerning its structure, and hence could be called nonparametric.

3.1. *Testing for the number of points in the mixing distribution.* First, consider the hypothesis testing problem

$$\mathcal{H}\colon \nu(Q) = p \quad \text{versus} \quad \mathcal{K}\colon \nu(Q) > p.$$

The results of Section 2 suggest that $\det \hat{\mathbf{M}}_p$ could be used as a test statistic, as it has limiting value 0 under the null hypothesis, and positive limit under the alternative. We suppose that for each value of q up to p we have an unbiased estimator $\gamma_q(X)$ of the qth moment of Q.

It is easiest to develop the theory of the test using the ideas of U-statistics. First, form a sequence $X_0, X_1, X_2, \ldots, X_p$ of i.i.d. replicates from distribution F_Q. We start by defining a kernel K

$$K_p(X_0, X_1, X_2, \ldots, X_p) = \det \begin{bmatrix} 1 & \gamma_1(X_1) & \cdots & \gamma_p(X_p) \\ \gamma_1(X_0) & \gamma_2(X_1) & \cdots & \gamma_{p+1}(X_p) \\ \gamma_2(X_0) & & & \vdots \\ \vdots & & & \\ \gamma_p(X_0) & \cdots & & \gamma_{2p}(X_p) \end{bmatrix}.$$

The first observation is that $E[K_p] = \det \mathbf{M}_p(Q)$. This follows because the construction of the matrix allows us to commute the expectation and determinant operators. Since K_p is not permutation-invariant in its arguments, next define an invariant version $\tilde{K}_p$ to be the average value of K_p over the $(p+1)!$ permutations of its arguments. Finally, define $\tilde{d}_p$, based n i.i.d. observations, to be the mean value of $\tilde{K}_p$ over all n-choose-$(p+1)$ subsets of the sample $X_1, X_2, \ldots, X_n$. Clearly $E[\tilde{d}_p] = d_p$, and indeed it is similar to being a bias-corrected version of $\hat{d}_p = \det \hat{\mathbf{M}}_p$; this latter is, in Serfling's terminology (1980), the corresponding V-statistic.

Although it is easiest to describe the theory from the U-statistic point of view, the results below apply as well to the V-statistic versions. However, we note in aside that small scale simulation results indicate that $\hat{d}_p$ can show considerable bias as an estimator of d_p, so that bias correction can be a significant issue. On the other hand, the above scheme for constructing the bias-corrected version is clearly very computationally intensive unless one can find algebraic shortcuts. One compromise solution with great ease of computation relative to the above scheme, and some further advantages, is the jackknife, which we will discuss further below.

The following theorem is a standard result on U-statistics, such as found in Serfling (1980) and treated originally by Hoeffding (1948).

THEOREM 3A. *Suppose that Q has $4p$ moments. Then $\sqrt{n}(\tilde{d}_p - d_p)$ converges in distribution to a normal distribution with mean 0 and variance*

$$\tau_p = (p+1)^2 \operatorname{Var}\left\{E\left[\tilde{K}_p(X_0, X_1, X_2, \ldots, X_p)|X_0\right]\right\},$$

provided that the latter is nonzero. If $\tau_p = 0$, as is the case if and only if Q has fewer than p-points of support, then $\tilde{d}_p$ is $o_p(1/\sqrt{n})$.

PROOF. First, we note that τ_p is finite if Q has $4p$ moments, as the conditional expected value is a polynomial of degree $2p$ in X_0.

Next, to evaluate τ_p when Q has fewer than p points of support, we show that the indicated conditional expectation is the zero polynomial in X_0. Consider any component summand K as above, generated by a sequence of replicates $X_0, X_1, \ldots, X_p$. Compute the expected values by first computing the conditional expected values given a sequence of replicates $\mu_1, \ldots, \mu_p$ from Q. These conditional expected values are all 0 since linear dependencies in the columns of the matrix show that the determinant is 0. □

We are now in a position to discuss the construction of the test. The hypotheses under question correspond roughly to testing H: M_p is nonnegative-definite versus K: M_p is positive-definite. This suggests that one should decide to reject H if there is strong evidence for positive definiteness. If $\hat{M}_p$ is positive-definite, then $\hat{\nu}$ is greater than p. Provided $\hat{M}_p$ is positive-definite, then a measure of the degree of positive definiteness is its determinant $\hat{d}_p$; as indicated by the Tchebycheff type bound in Lindsay (1989), Theorem 2D, its magnitude is a measure of departure from p-pointedness, so gives power in a desired direction.

COROLLARY 3B. *Let $\alpha < 0.5$. Suppose that $\hat{\tau}_p$ is a consistent estimator of τ_p. Define the set $A_n = \{\hat{\nu} < p\}$ and the set $B_n = \{\sqrt{n}\,\tilde{d}_p/\hat{\tau}_p \geq z_\alpha\}$. The test based on rejecting $\mathscr{H}$ on the set*

$$A_n \cup B_n$$

is asymptotically size α and is consistent for testing $\mathscr{H}$ against $\mathscr{K}$.

PROOF. Under both null and alternative hypotheses, the event A_n has probability tending to 1. □

It should be pointed out that in the important case $p = 1$ this method of testing for the number of support points reduces to another well-known test. That is, for distributions in the exponential family the $C(\alpha)$ test for homogeneity [Neyman and Scott (1956)], just as the proposed test, can be expressed as a normalized contrast between the sample variance and the variance as estimated under the assumption $p = 1$. It has been shown to have certain optimal properties as a test of the parametric model against any location-scale family of mixing distributions. [See Moran (1971) for further comments on optimality.]

REMARK. Consistent estimation of τ_p could be done in an obvious fashion from moments. An alternative track which is easy to program is to use the jackknife both to debias $\det \hat{M}_p$ and to estimate its variance. That is, in the examples described above, $\hat{M}_p$ has as entries linear combinations of sample moments. It is therefore a simple programming matter to delete observations one at a time from the matrix, then compute the determinant. One can then directly apply the standard jackknife methodology to derive the bias correction and the estimated variance.

A more difficult question involves the size of the proposed test under the more general null hypothesis, $\mathscr{H}$: $\nu(Q) \leq p$. Although the U-statistic machinery can be used to derive the distribution of $\hat{d}_p$, the size now also depends in a nonobvious fashion on the estimator $\hat{\tau}_p$, which is estimating 0 when $\nu(Q)$ is strictly less than p. For this more general null hypothesis, a more natural test statistic might be the smallest eigenvalue of $\hat{M}_p$, together with a bootstrap approach to constructing a confidence interval for it.

3.2. *A nonparametric estimator of* Q. Next consider the estimation of $\nu(Q)$ by $\hat{\nu}$. Note that one cannot estimate $\nu(Q)$ consistently in the usual sense, as, regardless of the sample size, there are $p + 1$-point distributions sufficiently close to any p-point distribution so as to be statistically indistinguishable. However, we will show that under mild assumptions $\hat{Q}_n := \hat{Q}_{\hat{\nu}}$ is consistent, in the sense of weak convergence with probability 1, for the true distribution Q. Tucker (1963) derived a result of this type in the case of the Poisson distribution. Brockett (1977) attempted to show the result more generally; however, the proof is flawed. (On page 36, the polynomial $q^*(x)$ is identically 0, not degree d, when $c > d$; this is easily checked for $d = 1$, mass 1 at 0.)

We first need the following lemma: It is an extension of the simple but powerful result that if $E(Y_n) \to a$ and $\mathrm{Var}(T_n) \to 0$, then $Y_n \to_p a$.

LEMMA 3C. *Let* $\{Q_n\}$ *be a sequence of distribution functions. If* $\det \boldsymbol{M}_p(Q_n)$ *converges to 0, and if* $m_j(Q_n) \to m_j(Q)$, $1 \leq j \leq 2p - 1$, *for some p-point distribution* Q, *then* $Q_n \to Q$ *weakly.*

PROOF. From Q_n one can construct a distribution Q_{np}, with p or fewer points of support, which matches the first $2p - 1$ moments of Q_n. The convergence of the moments implies convergence of corresponding moment determinants, and so from some value of n on Q_{np} will have p points. Since all the moments of a p-point distribution are determined by the first $2p - 1$, we have convergence of all the moments of Q_{np} to Q, and hence weak convergence.

Next, one applies the Tchebycheff type result from Lindsay (1989), Theorem 2D,

$$P\left\{\inf_i \{|X - r_i|\} > \varepsilon\right\} \leq d_p/d_{p-1}\varepsilon^{2p},$$

where r_i are the support points of the p-point distribution that matches the first $2p - 1$ moments of X. Letting X correspond to the distribution Q_n, then Q_{np} has support points corresponding to r_i, and it is clear that, since weak convergence of Q_{np} implies convergence of the r_i to the support points of Q, that the above inequality gives weak convergence of Q_n to Q. □

THEOREM 3D. *Suppose that Q is a distribution with all moments existing that is determined uniquely from its moment sequence. Then $\hat{Q}_n \to Q$ weakly with probability* 1.

PROOF. First suppose that $\nu(Q)$ is finite. From the construction of $\hat{\nu}$ we do have (on a set of probability 1) that for n sufficiently large, $\hat{\nu}$ will be greater than or equal to $p = \nu(Q)$. Thus in particular, from this value of n on the first $2p - 1$ moments of $F_{\hat{Q}}$ are matched to those of the empirical distribution. Since the latter converge to the moments of F_Q, we have that the estimated moments $m_j(\hat{Q}_n)$ converge to $m_j(Q)$, for $j = 1, \ldots, 2p - 1$. Moreover, either $\det \hat{\mathbf{M}}_p < 0$, in which case $\hat{Q}$ is a p-point estimator, and so $\det \mathbf{M}_p(\hat{Q}) = 0$, or $\det \hat{\mathbf{M}}_p \geq 0$, in which case $\det \mathbf{M}_p(\hat{Q}) = \det \hat{\mathbf{M}}_p$. Thus it is clear from Theorem 3A above that $\det \mathbf{M}_p(\hat{Q})$ converges to 0. Now apply Lemma 3C.

If Q has an infinite number of support points, then $\hat{\nu}$ diverges to ∞ with probability 1. This implies that in the limit, all moments of the mixing distribution are estimated consistently, which in turn implies weak convergence of the estimated distributions. □

REMARK. Lambert and Tierney (1984) considered the efficiency of the nonparametric estimator of an unknown mixing distribution Q in the Poisson model. One result was: Provided that the true distribution Q has infinitely many points of support, maximum likelihood based estimators of certain functionals of the model could be no more efficient than estimators based on the empirical cumulative distribution function. In particular, this implies that the maximum likelihood estimator of $m_p(Q)$ would be no more efficient at estimating $m_p(Q)$ than $\hat{m}_p = \bar{\gamma}_p$. This suggests that the proposed method-of-moments nonparametric estimator of the mixing distribution might not suffer quite the severe failings in efficiency in a nonparametric model [infinite $\nu(Q)$] that it does in the parametric, fixed finite $\nu(Q)$, model.

4. Normal distribution, known variance. The normal case will now be considered in greater detail. In particular, suppose that X has a distribution F which is a mixture of the form $\int n(x; \mu, \sigma)\, dQ(\mu)$, where $n(x; \mu, \sigma)$ is the normal density function. This distribution will be denoted $N(Q, \sigma^2)$. Since in a location family mixing is equivalent to convolution, the model for X is

$$X = \mu + \sigma Z,$$

where μ and Z are independent with distributions Q and $N(0, 1)$, respectively. In this section it will be assumed that the variance σ^2 is known; the σ^2 unknown case will be considered thereafter.

Since the normal model with fixed σ is a quadratic variance exponential family, the results of Section 2 apply. For the normal, the unbiased estimators $\gamma_p(x)$ of the moments of Q are the Hermite polynomials. Thus, for example, $\gamma_1(x) = x$, $\gamma_2(x) = x^2 - \sigma^2$ and $\gamma_3(x) = x^3 - 3\sigma^2 x$. The following lemma provides us with an algebraically useful representation (4.1) for their form in terms of the moments of a complex-valued random variable. This will eventually result in a number of important simplifications.

LEMMA 4A. *Let Z be a standard normal variate independent of X. Let $i = \sqrt{-1}$.*

(i) *Suppose that the moment generating function for $X \sim N(Q, \sigma^2)$ exists on some domain. Then on that same domain, the moment generating function for Q exists and has the representation*

$$\int \exp(t\mu)\, dQ(\mu) = m_Q(t) = E\left[e^{t(X + i\sigma Z)}\right].$$

(ii) *Define*

$$\gamma_p(x, \sigma) := E\left[(X + i\sigma Z)^p | X = x\right]. \tag{4.1}$$

This polynomial of degree p in X has expectation $m_p(Q)$ when $X \sim N(Q, \sigma^2)$.

PROOF. Straightforward, with (ii) following by differentiation of the representation of the moment generating function in (i). □

Next, for an arbitrary distribution F, with corresponding expectation E, define the matrix $\Gamma_p = \Gamma_p(F, \sigma)$ by letting the (i, j)th entry be $E[\gamma_{i+j}(X, \sigma)]$, for $i = 0, 1, \ldots, p$ and $j = 0, 1, \ldots, p$. As in Section 2, the use of the empirical distribution function for F in Γ_p provides an estimated moment matrix $\hat{\mathbf{M}}_p(\sigma)$ for the unobserved distribution Q.

The next theorem provides a U-statistic representation for $\det \Gamma_p$ which will be used in Section 5.

THEOREM 4B. *Let $X_0, X_1, X_2, \ldots, X_p$ be independent replicates from a distribution F with $2p$ moments. Let $Z_0, Z_1, \ldots, Z_p$ be independent standard*

normal variates. Then

$$\det \Gamma_p(F,\sigma) = E\left[\prod_{k>j}\left[X_k - X_j + i\sigma(Z_k - Z_j)\right]^2\right]. \tag{4.2}$$

PROOF. Lindsay (1989) proved the general result that for any moment matrix $\mathbf{M}_p$, $\det \mathbf{M}_p = E[\prod(Y_k - Y_j)^2]$, where $Y_0, Y_1, \ldots, Y_p$ are independent replicates from the distribution involved. Simply apply this representation to the complex-valued random variable $X + i\sigma Z$ and use (4.1). □

REMARKS. There are two important corollaries to this theorem. First, the determinant is invariant under location changes in the distribution of X, so that later we may with impunity use central moments in the estimation of σ. Second, if one takes the conditional expectation given the X-variates of the argument in (4.2), one obtains a symmetric kernel in the sequence $(X_0, X_1, \ldots, X_p)$ which could be used for U-statistic estimation of the value of the determinant; it is an explicit representation of the symmetric kernel $\tilde{K}$ of Section 3.

5. Normal theory method of moments. Assume now that $X \sim N(Q, \sigma^2)$ but that σ is unknown. For the purposes of deriving estimators, it will first be assumed that the number of support points in the mixing distribution Q is known to be p. The following lemma indicates that with this assumption the value of σ can be identified from the first $2p$ true moments of X. This will directly lead to a consistent method of estimation of σ from the sample moments. The next task will be to verify that substitution of this consistent solution into the moment equations of Section 2 leads to a set of moment equations for which a solution $\hat{Q}_p$ exists, thereby yielding a joint solution $(\hat{\sigma}_p, \hat{Q}_p)$ to the set of $2p$ moment equations.

5.1. *Consistent estimation of the parameter sigma.* For simplicity, we will let $d_p(\sigma)$ represent $\det \Gamma_p(F,\sigma)$ whenever the distribution F of random variable X has been clearly defined by context. The symbol $\hat{d}_p(\sigma)$ then refers to the use of the empirical distribution of the data in the place of F. Viewed as functions of σ^2, with F fixed, these are both polynomials of degree $p(p+1)/2$, and the first lemma indicates some useful behavior of the roots when F is a mixed normal distribution.

LEMMA 5A. *Suppose that $X \sim N(Q, \sigma_0^2)$ and that Q is a distribution with exactly p points of support $[\nu(Q) = p]$. Then*:

(i) *For every integer $m \geq 0$, $d_{p+m}(\sigma_0) = 0$.*
(ii) *For any positive $\sigma < \sigma_0$ we have $d_m(\sigma) > 0$, for all $m \geq 0$.*
(iii) *At its first positive root, $d_p(\sigma)$ undergoes a strict sign change,*

$$\partial\left[d_p(\sigma)\right]/\partial\sigma^2|_{\sigma=\sigma_0} < 0.$$

PROOF. The result (i) holds simply because $\Gamma_q(F, \sigma_0)$ is the moment matrix for the distribution Q which has p points of support. Part (ii) will be verified if we show that the matrix $\Gamma_q(F, \sigma)$ is the moment matrix for a distribution Q^* with infinitely many points of support. However, we may here use the identity $N(Q, \sigma_0^2) = N(Q^*, \sigma^2)$, where Q^* is the convolution of Q with a normal $(0, \sigma_0^2 - \sigma^2)$ random variable, and note that $\Gamma_q(F, \sigma)$ must therefore be the moment matrix for Q^*.

Proving (iii) requires some algebra, which we outline here. Using the above Q^* representation, as a function of $\tau^2 = (\sigma_0^2 - \sigma^2)$ the determinant has the representation

$$d_p(\sigma) = E\left[\prod\{\mu_i - \mu_j + \tau(Z_i - Z_j)\}^2\right]/(p+1)!,$$

where the μ's are a sample from Q and the Z's are an independent standard normal sample. Expand the square and analyze the terms which have coefficients τ^2. Some of them have the form

$$E\left[Z_i - Z_j\right]^2 E\left[\Pi(\mu_a - \mu_b)^2\right],$$

where the product includes all $a > b$ not equal to (i, j). It is easily checked that this expectation is strictly positive for a p-point distribution.

It remains to check that all other τ^2 terms are 0. These will have the form

$$E\left[(Z_i - Z_j)(Z_k - Z_l)(\mu_i - \mu_j)(\mu_k - \mu_l)\prod(\mu_a - \mu_b)^2\right].$$

The expectation over the Z's is 0 unless one member of the pair (i, j) equals one member of the pair (k, l). But under this last condition the expectation over the μ's is 0. To show this, consider an expectation of the form

$$E\left[(\mu_1 - \mu_0)(\mu_2 - \mu_0)(\mu_2 - \mu_1)(\mu_2 - \mu_1)\prod(\mu_\alpha - \mu_\beta)^2\right],$$

where the product indicates all pairs $\alpha > \beta$ not among $(1,0)$, $(2,0)$ or $(2,1)$. Taking the expectation conditionally on $\mu_3, \ldots, \mu_p$ and the ordered values of μ_0, μ_1, μ_2, a simple calculation shows that the result is 0. □

The above theorem indicates that if F is known to be a normal mixture with $\nu(Q) = p$, then σ can be identified as the smallest nonnegative root of $d_p(\sigma) = 0$. We might therefrom identify a method for estimating σ from the sample moments by

$$\hat{\sigma}_p := \text{smallest nonnegative root of } \hat{d}_p(\sigma).$$

We first consider the consistency of this method.

Certainly the coefficients of the polynomial $\hat{d}_p$ are continuous functions of the consistent sample moments, being products of sums of moments, and so $\hat{d}_p(\sigma) \to d_p(\sigma)$ almost surely.

The smallest root of a polynomial, however, is not an everywhere continuous function of the coefficients; for example, if the smallest root is a double root, then a slight change in coefficients may make turn these roots into complex

values, and the smallest real root could jump to somewhere else. However, part (iii) of the above lemma shows that $d_p(\sigma)$ undergoes a strict sign change at its first root; continuity of the function $d_p(\sigma)$ as a function of the moments then shows $\hat{\sigma}_p$ to be consistent when $\nu(Q) = p$. Moreover, the usual Taylor expansion argument shows it to be asymptotically normal, since $d_p(\sigma_0)$ is, with asymptotic variance $E[d_p^2]/-E[d_p']$.

As a more practical matter, one wishes to know if $\hat{d}_p(\sigma)$ necessarily has a first root. This matter is nicely resolved below, in Theorem 5C. However, it is clear that there can be several nonnegative roots, as the polynomial $d_p(\sigma)$ is positive at 0 and for some values of p has a positive coefficient for its highest power, and hence goes to $+\infty$ as $\sigma \to \infty$. (For $p = 2$ there is a unique nonnegative root.)

Before proceeding to the estimation of Q, we consider the consequences of this approach when $\nu(Q)$ is unknown. The normal mixture model, with discrete mixing distribution Q, is conceptually similar to a one-way analysis of variance model in which the individual observations are known, but the group identification for each observation is missing. Here $\nu(Q)$ represents the number of groups. Thought of in this way, the parameter σ^2 represents the "error" variance and the variance of μ under Q the "model" variance. As we increase the number of parameters being estimated in a one-way ANOVA, more of the total variance is allocated to the model, and less to the error. The same structure is now shown to hold for the method-of-moments estimators of the error variance σ^2.

THEOREM 5B. *Suppose that the empirical distribution $\hat{F}$ has n points of support. Let s^2 be the sample variance. Then every polynomial $\hat{d}_p(\sigma)$ has first nonnegative root, for $p = 1, \ldots, n$, and the roots satisfy*

$$s = \hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_n = 0.$$

PROOF. We note that $\hat{\mathbf{M}}_p(0)$ is the moment matrix for the empirical distribution, and so it has determinant 0 if $\hat{F}$ has p or fewer points of support, thus verifying the last equality. Otherwise, it is positive-definite. In this case let $\lambda_0(\sigma), \ldots, \lambda_p(\sigma)$ be the ordered eigenvalues of $\hat{\mathbf{M}}_p(\sigma)$. These eigenvalues are strictly positive at $\sigma = 0$. Since they are continuous functions of σ, and $\hat{d}_p(\sigma) = \det \hat{\mathbf{M}}_p(\sigma) = \prod \lambda_i(\sigma)$ has no zeros for σ between 0 and $\hat{\sigma}_p$, we can conclude that $\hat{\mathbf{M}}_p(\sigma)$ is positive-definite on this range. In particular, all its principal minors must be positive-definite there also. It follows that for any $q < p$, $\hat{d}_q(\sigma) = \det \hat{\mathbf{M}}_q(\sigma)$ must be positive on that range. Therefore this polynomial must have its first root to the right of $\hat{\sigma}_p$, as was to be shown. In particular, since we know that $\hat{d}_1$ has root s^2, all the polynomials have roots. □

5.2. *Estimation of the mixing distribution.* The final step is to verify that the consistent root to the determintal equation, $\hat{\sigma}_p$, leads to a solution for the p-point mixing distribution, $\hat{Q}_p$.

THEOREM 5C. *Let $p \leq n$. If $\hat{\sigma}_p$ is a root of multiplicity one to $\hat{d}_p(\sigma) = 0$, then there exists a unique p-point distribution $\hat{Q}_p$ such that $N(\hat{Q}_p, \hat{\sigma}_p^2)$ has as its*

first $2p$ *moments* $1, \hat{m}_1(\hat{\sigma}_p), \ldots, \hat{m}_{2p}(\hat{\sigma}_p)$, *the moments of* Q *as estimated from the Hermite polynomials using the estimated value* $\hat{\sigma}_p$ *for* σ.

PROOF. We know that the determinant of $\hat{\mathbf{M}}_p(\hat{\sigma}_p)$ is 0. Moreover, as in Theorem 5B, it can be shown that all of the principal minors have positive determinant, strictly positive by the multiplicity of the root, and the result follows from Theorem 2A. □

Inference on the number of points $\nu(Q)$ in the mixing distribution when there is a nuisance parameter σ in the model is more difficult than in the cases considered in Section 2. However, there is clearly inferential information in the values of $\hat{\sigma}_p$, as under a p-point model they converge to the same value, namely σ^2, for every $p \geq \nu(Q)$. Appropriate use of these statistics cannot be commented upon here due to the need for considerable more investigation.

6. Discussion. A full discussion of the numerical issues and illustrations of the effectiveness of the methods discussed here are beyond the scope of this article. However, some preliminary comments can be made here. These techniques are most easily programmed in computing languages which allow the direct use of matrix manipulations. It is particularly felicitous if the determinant is an explicit function of the language, as in GAUSS.

Constructing the estimated moments is expedited if there is a simple recursion for obtaining $\gamma_p(X)$ from its predecessors; it is then an easy matter to program for arbitrarily sized matrices. For example, in the normal case, we have the recursion for Hermite polynomials,

$$\gamma_p(x) = x\gamma_{p-1}(x) - (p-1)\sigma^2\gamma_{p-2}(x).$$

Finding roots of polynomials like $S_p(t)$, equation (2.2), is easy for $p = 2$. For higher-order polynomials, it is natural to use the nesting property of the roots, Lemma 2D, to identify a region which contains exactly one root, and then use a simple algorithm, such as bisection, to find the root in each region. (Bisection simply divides the interval repeatedly in half, selecting at each stage the half-interval in which a sign change occurs. Given the known sign-change behavior, it is simple, effective, easy to program and guaranteed convergent.)

To find the smallest root to $\hat{d}_p(\sigma)$, it is again safest to proceed sequentially in p, using now the nesting property of Theorem 5B. A simple bisection algorithm on the interval $[0, \hat{\sigma}_{p-1}]$ has been found to be a fast and effective way to locate $\hat{\sigma}_p$.

Finally, to give a small illustration of the use of the normal method, we consider applying it to Darwin's data:

$$-67, -48, 6, 8, 14, 16, 23, 23, 28, 29, 41, 49, 56, 60, 75.$$

These data were used by Aitkin and Wilson (1980) to illustrate the use of mixture maximum likelihood. We give the parameter estimates for $p = 2$ below,

together with the mode of the likelihood (there are at least two) selected by Aitkin and Wilson based on the use of several starting values in the EM algorithm.

	μ_1	μ_2	π	σ
Maximum likelihood	33.0	−57.3	0.87	19.63
Method of moments	34.0	−46.7	0.84	21.17

Thus it appears in this case that method of moments would have given good initial values for maximum likelihood. It is anticipated that if the method of moments could be extended into higher dimensions without loss of computational ease, it would become a dramatically important tool.

Acknowledgments. I would like to express my appreciation to all the participants in the editorial process for their helpful and enlightening comments.

REFERENCES

AITKEN, M. A. and WILSON, G. T. (1980). Mixture models, outliers and the EM algorithm. *Technometrics* **22** 325–331.

BLISCHKE, W. R. (1964). Estimating the parameters of mixtures of binomial distributions. *J. Amer. Statist. Assoc.* **59** 510–528.

BROCKETT, P. L. (1977). Approximating moment sequences to obtain consistent estimates of distribution functions. *Sankhyā Ser. A* **39** 32–44.

EVERITT, B. S. and HAND, D. J. (1981). *Finite Mixture Distributions*. Chapman and Hall, London.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293–325.

LAMBERT, D. and TIERNEY, L. (1984). Asymptotic efficiency of estimators of functions of mixed distributions. *Ann. Statist.* **12** 1380–1387.

LINDSAY, B. G. (1989). On the determinants of moment matrices. *Ann. Statist.* **17** 711–721.

MAMMANA, C. (1954). Sul problema algebraico dei momenti. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (*4*) **8** 133–140.

MORAN, P. A. P. (1973). Asymptotic properties of homogeneity tests. *Biometrika* **60** 79–85.

MORRIS, C. N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10** 65–80.

MORRIS, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *Ann. Statist.* **11** 515–529.

NEYMAN, J. and SCOTT, E. L. (1966). On the use of $C(\alpha)$ optimal tests of composite hypotheses. *Bull. Inst. Internat. Statist.* **41** 477–497.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SHOHAT, J. A. and TAMARKIN, J. D. (1943). *The Problem of Moments*. Amer. Math. Soc., New York.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

TUCKER, H. (1963). An estimate of the compounding distribution of a compound Poisson distribution. *Theory Probab. Appl.* **8** 195–200.

USPENSKY, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill, New York.

DEPARTMENT OF STATISTICS
THE PENNSYLVANIA STATE UNIVERSITY
219 POND LABORATORY
UNIVERSITY PARK, PENNSYLVANIA 16802