DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

ON THE NUMERICAL EVALUATION OF THE MAXIMUM LIKELIHOOD ESTIMATE OF MIXTURE MEANS

BY HOMER F. WALKER

REPORT #44 JULY, 1975

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-12777

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

On the Numerical Evaluation of the Maximum-Likelihood Estimate of Mixture Means

July, 1975

by

Homer F. Walker
Department of Mathematics
University of Houston

Report 44
NAS-9-12777

On the Numerical Evaluation of the Maximum-Likelihood Estimate of Mixture Means

by

Homer F. Walker

Department of Mathematics, University of Houston

Houston, Texas

1. Introduction.

Let x be an n-dimensional random variable whose density function p is a convex combination of normal densities, i.e.,

$$p(x) = \sum_{i=1}^{m} \alpha_i^o \, p_i(x) \quad \text{for} \quad x \in \mathbb{R}^n,$$

where

$$\alpha_i^o > 0, \quad \sum_{i=1}^{m} \alpha_i^o = 1,$$

and

$$p_i(x) = \frac{1}{(2\pi)^{n/2} \, |\Sigma_i^o|^{1/2}} \, e^{-1/2 (x-\mu_i^o)^T \Sigma_i^{o\,-1} (x-\mu_i^o)}$$

If $\{x_k\}_{k=1,\ldots,N} \subseteq \mathbb{R}^n$ is an independent sample of observations on $\mathbf{x}$, then a maximum-likelihood estimate of the parameters $\{\alpha_i^o, \mu_i^o, \Sigma_i^o\}_{i=1,\ldots,m}$ is a choice of parameters $\{\alpha_i^o, \mu_i^o, \Sigma_i^o\}_{i=1,\ldots,m}$ which locally maximizes the log-likelihood function

$$L = \sum_{k=1}^{N} \log p(x_k),$$

in which p is evaluated with the true parameters $\{\alpha_i^o, \mu_i^o, \Sigma_i^o\}_{i=1,\ldots,m}$ replaced by the estimate $\{\alpha_i, \mu_i, \Sigma_i\}_{i=1,\ldots,m}$. (In the following, it is clear from the context which parameters are used in evaluating the density functions p_i and p. Therefore, these parameters are not explicitly pointed out.)

Clearly, L is a differentiable function of the parameters to be estimated Equating to zero the partial derivatives of L with respect to these parameters, one obtains, after a straightforward calculations the following necessary conditions for a maximum-likelihood estimate:

$$\left.\begin{aligned}
&(1.a) \qquad \alpha_i = \frac{\alpha_i}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)} \\
&(1.b) \qquad \mu_i = \left\{\frac{1}{N} \sum_{k=1}^{N} x_k \frac{p_i(x_k)}{p(x_k)}\right\} \Big/ \left\{\frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\right\} \\
&(1.c) \qquad \Sigma_i = \left\{\frac{1}{N} \sum_{k=1}^{N} (x_k - \mu_i)(x_k - \mu_i)^T \frac{p_i(x_k)}{p(x_k)}\right\} \Big/ \left\{\frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\right\}
\end{aligned}\right\} i=1,\ldots,m.$$

These are known as the likelihood equations, and we shall assume that the parameters under consideration here are restricted to sets in which these equations are sufficient, as well as necessary, for a maximum-likelihood estimate.

The likelihood equations suggest the following iterative procedure for obtaining a solution: Beginning with some set of starting values, obtain successive approximations to a solution by inserting the preceding approximations in the expressions on the right-hand sides of (1.a), (1.b), and (1.c). This scheme is attractive for its relative case of implementation, and it has been investigated by a number of authors. Empirical studies of Day [1], Duda and Hart [2], and Hasselblad [3] suggest that this scheme is convergent and that convergence is particularly fast when the component normal densities in p are "widely separated" in a certain sense. No proof of convergence is given in these papers, although Peters and Walker [8] have shown that, with probability approaching 1 as N approaches infinity, a related procedure (which includes this one as a special case) converges locally to the consistent maximum-likelihood estimate whenever a certain "step-size" is sufficiently small. (An iterative procedure is said to converge locally to a limit if the iterates converge to that limit whenever the starting values are sufficiently near that limit.)

Peters and Coberly [7] have proved that, if all of the parameters μ_i and Σ_i are held fixed, then the iterative procedure suggested by the equations (1.a) alone converges locally to a maximum-likelihood estimate of the parameters α_i, $i=1,\ldots,m$. They also report on numerical studies in which the computational feasibility of this procedure is demonstrated. In this note, we provide sufficient conditions for the iterative procedure suggested by the equations (1.b) alone, for fixed parameters α_i and Σ_i, to converge locally to a maximum-likelihood estimate of the means μ_i, $i=1,\ldots,m$. These conditions are, roughly, that either $m = 2$ or the component normal densities in p be "widely separated" in a certain sense.

2. Preliminary discussion.

We denote by $\mathcal{M}$ the m-fold direct sum of $\mathbb{R}^n$ with itself, and we represent its elements as columns

$$\bar{\mu} = \begin{pmatrix} \mu_1 \\ \cdot \\ \cdot \\ \cdot \\ \mu_m \end{pmatrix} \in \mathcal{M}.$$

(Of course, $\mathcal{M}$ is isomorphic to $\mathbb{R}^{mn}$.) We also find it convenient to represent parameter sets $\{\alpha_i\}_{i=1,\ldots,m}$ and $\{\Sigma_i\}_{i=1,\ldots,m}$ as columns

$$\bar{\alpha} = \begin{pmatrix} \alpha_1 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_m \end{pmatrix} \quad \text{and} \quad \bar{\Sigma} = \begin{pmatrix} \Sigma_1 \\ \cdot \\ \cdot \\ \cdot \\ \Sigma_m \end{pmatrix} ,$$

and, in the following, we use the fact that $\bar{\alpha}$ and $\bar{\Sigma}$ belong to normed vector spaces without explicitly introducing these spaces or their norms.

Setting

$$M_i(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = \{\frac{1}{N}\sum_{k=1}^{N} x_k \frac{p_i(x_k)}{p(x_k)}\} \Big/ \{\frac{1}{N}\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\}, \quad i=1,\ldots,m,$$

we define

$$M(\bar{\alpha},\bar{\mu},\bar{\Sigma}) = \begin{pmatrix} M_1(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \\ \cdot \\ \cdot \\ \cdot \\ M_m(\bar{\alpha},\bar{\mu},\bar{\Sigma}) \end{pmatrix} ,$$

which we regard as a function from $\mathcal{M}$ to itself depending on parameters $\bar{\alpha}$ and $\bar{\Sigma}$. The equations (1.b) can now be written as

$$\bar{\mu} = M(\bar{\alpha},\bar{\mu},\bar{\Sigma}), \tag{2}$$

and the iterative procedure under consideration is the following: Beginning with some starting value $\bar{\mu}^{(1)}$, define successive iterates inductively by

$$\bar{\mu}^{(k+1)} = M(\bar{\alpha},\bar{\mu}^{(k)},\bar{\Sigma}) \tag{3}$$

for $k = 1,2,\ldots$.

In our results concerning the convergence of the procedure (3), the Fréchet derivative of M with respect to $\bar{\mu}$, which we denote by $\nabla_{\bar{\mu}} M$, is of central importance. (For questions concerning the definition and properties of Fréchet derivatives, see Luenberger [6].) Indeed, if $\bar{\alpha},\bar{\mu}$, and $\bar{\Sigma}$ satisfy (2) and if $||\ ||$ is any norm on $\mathcal{M}$, then one can write

$$M(\bar{\alpha},\bar{\mu}',\bar{\Sigma}) - \bar{\mu} = \nabla_{\bar{\mu}} M(\bar{\alpha},\bar{\mu},\bar{\Sigma})(\bar{\mu}'-\bar{\mu}) + o(||\bar{\mu}'- \bar{\mu}||^2)$$

for $\bar{\mu}'$ near $\bar{\mu}$. Consequently, if there exists a norm $||\ ||$ on $\mathcal{M}$ with respect to which $\nabla_{\bar{\mu}} M(\bar{\alpha},\bar{\mu},\bar{\Sigma})$ has operator norm less than 1, then M is <u>locally contractive</u> in that norm near $\bar{\mu}$, i.e., there is a number λ, $0 \le \lambda < 1$, such that

$$||M(\bar{\alpha},\bar{\mu}',\bar{\Sigma}) - \bar{\mu}|| \le \lambda\ ||\bar{\mu}' - \bar{\mu}|| \tag{4}$$

whenever $\bar{\mu}'$ is sufficiently near $\bar{\mu}$. Since an inequality of the form (4) implies the local convergence of the iterative procedure (3) to $\bar{\mu}$, our objectives will be met by giving sufficient conditions for $\nabla_{\underline{\mu}} M(\bar{\alpha},\bar{\mu},\bar{\Sigma})$ to have operator norm less than 1 (with respect to some norm on $\mathfrak{M}$) at parameter vectors $\bar{\alpha},\bar{\mu}$, and $\bar{\Sigma}$ which satisfy (2).

We now calculate $\nabla_{\underline{\mu}} M$ at a set of parameter vectors $\hat{\underline{\alpha}},\hat{\underline{\mu}}$, and $\hat{\underline{\Sigma}}$ (with components $\hat{\alpha}_i,\hat{\mu}_i$, and $\hat{\Sigma}_i$, i=1,...,m) which satisfy the likelihood equations. We first define inner products $\langle\cdot,\cdot\rangle_i$ on $\mathbb{R}^n$ by

$$\langle x,y\rangle_i = x^T(\hat{\alpha}_i\hat{\Sigma}_i^{-1})y \quad \text{for} \quad x,y \in \mathbb{R}^n,\ i=1,\ldots,m.$$

Then, denoting the Fréchet derivative of M_i with respect to μ_j by $\nabla_{\mu_j} M_i$, one verifies with the aid of the likelihood equations that

$$\nabla_{\mu_j} M_i(\hat{\underline{\alpha}},\hat{\underline{\mu}},\hat{\underline{\Sigma}}) = \begin{cases} -\frac{1}{N}\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}(x_k-\hat{\mu}_i)\left\langle \frac{p_j(x_k)}{p(x_k)}(x_k-\hat{\mu}_j),\cdot\right\rangle_j & \text{if } i \neq j \\ I - \frac{1}{N}\sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}(x_k-\hat{\mu}_i)\left\langle \frac{p_i(x_k)}{p(x_k)}(x_k-\hat{\mu}_i),\cdot\right\rangle_i & \text{if } i = j. \end{cases}$$

This yields the following expression, in the form of a matrix of Fréchet derivatives, for $\nabla_{\underline{\mu}} M$ at a solution of the likelihood equations:

$$\nabla_{\underline{\mu}} M(\hat{\underline{\alpha}},\hat{\underline{\mu}},\hat{\underline{\Sigma}}) = \begin{pmatrix} \nabla_{\mu_1} M_1(\hat{\underline{\alpha}},\hat{\underline{\mu}},\hat{\underline{\Sigma}}) & \cdots & \nabla_{\mu_m} M_1(\hat{\underline{\alpha}},\hat{\underline{\mu}},\hat{\underline{\Sigma}}) \\ \vdots & & \vdots \\ \nabla_{\mu_1} M_m(\hat{\underline{\alpha}},\hat{\underline{\mu}},\hat{\underline{\Sigma}}) & \cdots & \nabla_{\mu_m} M_m(\hat{\underline{\alpha}},\hat{\underline{\mu}},\hat{\underline{\Sigma}}) \end{pmatrix} \tag{5}$$

$$= I - \{\frac{1}{N}\sum_{k=1}^{N} \begin{pmatrix} \frac{p_1(x_k)}{p(x_k)}(x_k-\hat{\mu}_1) \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}(x_k-\hat{\mu}_m) \end{pmatrix} \begin{pmatrix} <\frac{p_1(x_k)}{p(x_k)}(x_k-\hat{\mu}_1),\cdot>_1 \\ \vdots \\ <\frac{p_m(x_k)}{p(x_k)}(x_k-\hat{\mu}_m),\cdot>_m \end{pmatrix}^T \}.$$

The inner products $<\cdot,\cdot>_i$ induce an inner product $<\cdot,\cdot>$ on $\mathcal{M}$. In the following, $||\ ||$ will denote both the vector norm and the operator norm defined by this inner product. It is apparent from (5) that, at a solution of the likelihood equations, $\nabla_{\underline{\mu}}M$ is of the form I-Q, where Q is positive semi-definite and symmetric with respect to the inner product $<\cdot,\cdot>$. In fact, we prove in an appendix that Q is positive-definite with probability 1 whenever $N \geq mn$. It follows that, with probability 1 for $N \geq mn$, $||\nabla_{\underline{\mu}}M|| < 1$ at a solution of the likelihood equations if and only if $||Q|| < 2$. We conclude these preliminary remarks with the following

Lemma: $||Q|| < m$.

Proof: Since Q is symmetric with respect to $<\cdot,\cdot>$, one has

$$||Q|| = \sup_{\substack{\bar{v}\in\mathcal{M} \\ ||\bar{v}||\leq 1}} <\bar{v}, Q\bar{v}>.$$

If $\{v_i\}_{i=1,\ldots,m} \subseteq \mathbb{R}^n$ is such that

$$\bar{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \in \mathcal{M} \text{ satisfies } ||\bar{v}|| \leq 1, \text{ then}$$

$$\langle \bar{v}, Q\bar{v} \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{1}{N} \sum_{k=1}^{N} \langle v_i, \frac{p_i(x_k)}{p(x_k)}(x_k - \hat{\mu}_i) \rangle_i \langle v_j, \frac{p_j(x_k)}{p(x_k)}(x_k - \hat{\mu}_j) \rangle_j$$

$$\leq \sum_{i=1}^{m} \sum_{j=1}^{m} \{\frac{1}{N} \sum_{k=1}^{N} \langle v_i, \frac{p_i(x_k)}{p(x_k)}(x_k - \hat{\mu}_i) \rangle_i^2\}^{1/2} \{\frac{1}{N} \sum_{k=1}^{N} \langle v_j, \frac{p_j(x_k)}{p(x_k)}(x_k - \hat{\mu}_j) \rangle_j^2\}^{1/2}$$

$$< \sum_{i=1}^{m} \sum_{j=1}^{m} \langle v_i, [\frac{1}{N} \sum_{k=1}^{N} (x_k - \hat{\mu}_i)(x_k - \hat{\mu}_i)^T \frac{p_i(x_k)}{p(x_k)}] \hat{\Sigma}_i^{-1} v_i \rangle_i^{1/2} \cdot$$

$$\cdot \langle v_j, [\frac{1}{N} \sum_{k=1}^{N} (x_k - \hat{\mu}_j)(x_k - \hat{\mu}_j)^T \frac{p_j(x_k)}{p(x_k)}] \hat{\Sigma}_j^{-1} v_j \rangle_j^{1/2}$$

since $\frac{\hat{\alpha}_i p_i(x)}{p(x)} < 1$ for $i=1,\ldots,m$. From the likelihood equations, one concludes that

$$\langle \bar{v}, Q\bar{v} \rangle < \sum_{i=1}^{m} \sum_{j=1}^{m} \langle v_i, v_i \rangle_i^{1/2} \langle v_j, v_j \rangle_j^{1/2} = (\sum_{i=1}^{m} \langle v_i, v_i \rangle_i^{1/2})^2 \leq m,$$

and the lemma is proved.

3. Sufficient conditions for local convergence.

Sufficient conditions will now be given for local convergence of the procedure (3) to a solution of (2). Our first condition is given by the theorem below.

Theorem 1: Suppose that $m = 2$ and $N \geq 2n$, and let $\underline{\hat{\alpha}}, \underline{\hat{\mu}}, \underline{\hat{\Sigma}}$ be vectors of parameters which satisfy the likelihood equations. If $\bar{\alpha}, \bar{\mu}$, and $\bar{\Sigma}$, satisfy (2) and lie sufficiently near $\underline{\hat{\alpha}}, \underline{\hat{\mu}}$, and $\underline{\hat{\Sigma}}$, then the iterative procedure (3) converges locally to $\bar{\mu}$ with probability 1.

Proof: From the preliminary discussion, we know that the procedure (3) converges locally to $\bar{\mu}$ if $\nabla_{\mu} M(\bar{\alpha},\bar{\mu},\bar{\Sigma})$ has operator norm less than 1 with respect to some vector norm on $\mathfrak{M}$. Then, since $\nabla_{\mu} M$ depends continuously on $\bar{\alpha},\bar{\mu}$, and $\bar{\Sigma}$, it suffices to find a norm on $\mathfrak{M}$ with respect to which $\nabla_{\mu} M(\hat{\bar{\alpha}},\hat{\mu},\hat{\bar{\Sigma}})$ has operator norm less than 1 in order to prove the theorem.

Now $\nabla_{\mu} M(\hat{\bar{\alpha}},\hat{\mu},\hat{\bar{\Sigma}}) = I-Q$, where Q is the operator introduced in the preliminary discussion. With probability 1, Q is positive-definite as well as symmetric with respect to $\langle\cdot,\cdot\rangle$, and, from the Lemma, $||Q|| < m = 2$. Consequently, $||\nabla_{\mu} M(\hat{\bar{\alpha}},\hat{\mu},\hat{\bar{\Sigma}})|| < 1$ with probability 1, and the proof is complete.

We now define an operator Q^o on $\mathfrak{M}$ by

$$Q^o = \int_{\mathbb{R}^n} \begin{pmatrix} \frac{p_1(x)}{p(x)}(x-\mu_1^o) \\ \vdots \\ \frac{p_m(x)}{p(x)}(x-\mu_m^o) \end{pmatrix} \begin{pmatrix} \left\langle \frac{p_1(x)}{p(x)}(x-\mu_1^o), \quad \right\rangle_1 \\ \vdots \\ \left\langle \frac{p_m(x)}{p(x)}(x-\mu_m^o), \cdot \right\rangle_m \end{pmatrix}^T p(x)dx,$$

in which the true parameters (whose vectors we denote by $\bar{\alpha}^o,\bar{\mu}^o$, and $\bar{\Sigma}^o$) are used in evaluating the functions p_i and p and the inner products $\langle\cdot,\cdot\rangle_i$. The operator Q^o can be thought of as an $m\times m$ array of operators on $\mathbb{R}^n$, the $ij^{\underline{th}}$ operator of which is

$$\int_{\mathbb{R}^n} \frac{p_i(x)}{p(x)}(x-\mu_i^o)\left\langle \frac{p_j(x)}{p(x)}(x-\mu_j^o), \cdot \right\rangle_j p(x)dx.$$

If the component normal densities in p are "widely separated" in the sense that each pair of parameters μ_i^o and Σ_i^o differs greatly from every other

pair, then the off-diagonal operators in this array are near zero. On the other hand, regardless of the "separation" of the component densities, the diagonal operators define an operator on $\mathcal{H}$ which lies strictly between the zero operator and the identity operator in the ordering on symmetric operators defined by the inner product $\langle \cdot, \cdot \rangle$. Consequently, if the component normal densities in p are sufficiently "widely separated" in this sense, then the operator $I-Q^o$ has spectral radius less than 1, and, hence, there exists a norm on $\mathcal{H}$ with respect to which $I-Q^o$ has operator norm less than 1. (See Householder [4].) This motivates our second condition.

Theorem 2: Suppose that the component normal densities in p are sufficiently "widely separated" that the spectral radius of $I-Q^o$ is less than 1. Then the probability is 1 that, for sufficiently large N, there exist neighborhoods of $\bar{\alpha}^o, \bar{\mu}^o$, and $\bar{\Sigma}^o$ such that, if $\bar{\alpha}, \bar{\mu}$, and $\bar{\Sigma}$ lie in these neighborhoods and satisfy (2), then the iterative procedure (3) converges locally to $\bar{\mu}$.

Proof: A straightforward calculation and an application of the Strong Law of Large Numbers (see Loéve [5]) yields that $\nabla_{\mu} M(\bar{\alpha}^o, \bar{\mu}^o, \bar{\Sigma}^o)$ converges with probability 1 to $I-Q^o$ as N approaches infinity. Since $I-Q^o$ is assumed to have spectral radius less than 1, it follows that, with probability 1, if N is sufficiently large, then $\nabla_{\mu} M(\bar{\alpha}, \bar{\mu}, \bar{\Sigma})$ has operator norm less than 1 with respect to some norm on $\mathcal{H}$ whenever $\bar{\alpha}, \bar{\mu}$, and $\bar{\Sigma}$ lie near $\bar{\alpha}^o, \bar{\mu}^o$, and $\bar{\Sigma}^o$. If $\nabla_{\mu} M(\bar{\alpha}, \bar{\mu}, \bar{\Sigma})$ has operator norm less than 1 and $\bar{\alpha}, \bar{\mu}$, and $\bar{\Sigma}$ also satisfy (2), then the iterative procedure (3) converges locally to $\bar{\mu}$. This completes the proof.

Appendix

We now prove that the operator

$$Q = \frac{1}{N}\sum_{k=1}^{N} \begin{pmatrix} \frac{p_1(x_k)}{p(x_k)}(x_k-\mu_1) \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}(x_k-\mu_m) \end{pmatrix} \begin{pmatrix} \frac{p_1(x_k)}{p(x_k)}(x_k-\mu_1), \cdot\rangle_1 \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}(x_k-\mu_m), \cdot\rangle_m \end{pmatrix}^T$$

is positive-definite on $\mathcal{M}$ with probability 1 whenever $N \geq mn$. Clearly, it suffices to show that the vectors

$$V(x_k) \equiv \begin{pmatrix} \frac{p_1(x_k)}{p(x_k)}(x_k-\mu_1) \\ \vdots \\ \frac{p_m(x_k)}{p(x_k)}(x_k-\mu_m) \end{pmatrix}, \quad k = 1,\ldots,N,$$

span $\mathcal{M}$ with probability 1 whenever $N \geq mn$. This follows from the more general result below.

Lemma. Let $\{x_k\}_{k=1,\ldots,N}$ be an independent sample of observations on a random variable x in $\mathbb{R}^s$ which is distributed with a probability density function p. If V is a real-analytic function from $\mathbb{R}^s$ to $\mathbb{R}^t$ whose component functions are linearly independent, then the vectors $V(x_k)$, $k=1,\ldots,N$, span $\mathbb{R}^t$ with probability 1 whenever $N \geq t$.

Proof: Denoting the j^{th} component function of V by v_j, we define a real-analytic function V_j from $\mathbb{R}^s$ to $\mathbb{R}^j$ by

$$V_j(x) = \begin{pmatrix} v_1(x) \\ \vdots \\ v_j(x) \end{pmatrix}$$

for $j = 1,\ldots,t$. Our proof of the lemma consists of showing inductively that, for $j = 1,\ldots,t$, the set $\{V_j(x_k)\}_{k=1,\ldots,j}$ spans $\mathbb{R}^j$ with probability 1. We make the preliminary observation that, since the real-analytic functions v_j are assumed to be linearly independent, any non-zero linear combination of them vanishes only on a set of Lebesgue measure zero in $\mathbb{R}^s$.

From the observation above, $V_1(x_1)$ is non-zero with probability 1; hence $V_1(x_1)$ spans $\mathbb{R}^1$ with probability 1. Suppose now that, for some j, $1 \le j < t$, the set $\{V_j(x_k)\}_{k=1,\ldots,j}$ spans $\mathbb{R}^j$ with probability 1. Then, with probability 1, the set $\{V_{j+1}(x_k)\}_{k=1,\ldots,j+1}$ fails to span $\mathbb{R}^{j+1}$ if and only if

$$(*) \qquad V_{j+1}(x_{j+1}) = \sum_{k=1}^{j} c_k \, V_{j+1}(x_k)$$

for some set of constants $\{c_k\}_{k=1,\ldots,j}$. If (*) holds, the constants c_k are determined by

$$\begin{pmatrix} c_1 \\ \vdots \\ c_j \end{pmatrix} = \tilde{V}_j^{-1} \, V_j(x_{j+1})$$

with probability 1, where $\tilde{V}_j$ is the $j \times j$ matrix whose k^{th} column is $V_j(x_k)$. Thus, with probability 1, (*) holds if and only if

$$[\tilde{V}_j^{-1} V_j(x_{j+1})]^T \begin{pmatrix} v_{j+1}(x_1) \\ \cdot \\ \cdot \\ \cdot \\ v_{j+1}(x_j) \end{pmatrix} - v_{j+1}(x_{j+1}) = 0.$$

Now

$$[\tilde{V}_j^{-1} V_j(x)]^T \begin{pmatrix} v_{j+1}(x_1) \\ \cdot \\ \cdot \\ \cdot \\ v_{j+1}(x_j) \end{pmatrix} - v_{j+1}(x)$$

is a non-zero linear combination of the functions $v_1, \ldots, v_{j+1}$ and, hence, vanishes only on a set of Lebesque measure zero in $\mathbb{R}^s$. One concludes that $\{V_{j+1}(x_k)\}_{k=1,\ldots,j+1}$ fails to span $\mathbb{R}^{j+1}$ with probability zero. This completes the induction, and the lemma is proved.

REFERENCES

1. N.E. Day, "Estimating the components of a mixture of normal distributions," Biometrika 56 (1969), pp. 463-474.
2. R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, Inc., New York, 1973.
3. V. Hasselblad, "Estimation of parameters for a mixture of normal distributions," Technometrics 8 (1966), pp. 431-446.
4. A. Householder, Theory of Matrices and Numerical Analysis, Blaisdell Publishing Co., New York, 1964.
5. M. Loéve, Probability Theory, D. Van Nostrand Co., New York, 1963.
6. D.G. Luenberger, Optimization by Vector Space Methods, John Wiley and Sons, Inc., New York, 1969.
7. B.C. Peters and W.H. Coberly, "The numerical evaluation of the maximum-likelihood estimate of mixture proportions," to appear.
8. B.C. Peters and H.F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions," to appear.