A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions
Author(s): Richard J. Hathaway
Source: *The Annals of Statistics*, Vol. 13, No. 2 (Jun., 1985), pp. 795-800
Published by: Institute of Mathematical Statistics
Stable URL: http://www.jstor.org/stable/2241213
Accessed: 03/11/2014 16:08

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve and extend access to *The Annals of Statistics.*

STOR®
http://www.jstor.org

This content downloaded from 128.193.152.162 on Mon, 3 Nov 2014 16:08:54 PM
All use subject to JSTOR Terms and Conditions

The Annals of Statistics
1985, Vol. 13, No. 2, 795–800

A CONSTRAINED FORMULATION OF MAXIMUM-LIKELIHOOD ESTIMATION FOR NORMAL MIXTURE DISTRIBUTIONS[1]

BY RICHARD J. HATHAWAY

University of South Carolina

The method of maximum likelihood leads to an ill-posed optimization problem in the case of a mixture of normal distributions. Estimation in the univariate case is reformulated using simple constraints into an optimization problem having a strongly consistent, global solution.

1. Introduction. A univariate normal density with mean μ and standard deviation σ is denoted by $P(x; \mu, \sigma)$. A mixture of m univariate normal densities, denoted by $p_m(x; \gamma)$, is defined by $p_m(x; \gamma) = \sum_{i=1}^m \alpha_i P(x; \mu_i, \sigma_i)$ for the parameter γ in the parameter space

$$\Gamma = \{\gamma = (\alpha_1, \cdots, \alpha_m, \mu_1, \cdots, \mu_m, \sigma_1, \cdots, \sigma_m)^T \in \mathbb{R}^{3m} \mid$$
$$\textstyle\sum_{i=1}^m \alpha_i = 1, \alpha_i \geq 0, \sigma_i > 0 \quad \text{for} \quad i = 1, \cdots, m\}.$$

In this context, the normal densities are sometimes referred to as component densities. The log-likelihood function $L(\gamma)$, corresponding to a random sample $\{x_1, \cdots, x_n\}$, is defined by $L(\gamma) = \sum_{k=1}^n \log(p_m(x_k; \gamma))$.

The first problem associated with maximum-likelihood estimation arises from the unboundedness of $L(\gamma)$ on Γ (Day, 1969). A global maximum-likelihood estimate always fails to exist. In addition, the unboundedness of $L(\gamma)$ causes failures of optimization algorithms of both the EM (Redner and Walker, 1984) and quasi-Newton (Fowlkes, 1979) types.

In spite of the unboundedness of $L(\gamma)$, statistical theory (Kiefer, 1978) guarantees that a particular local maximizer of $L(\gamma)$ is strongly consistent and asymptotically efficient. Several local maximizers can exist for a given sample, and the other major maximum-likelihood difficulty is in determining when the correct one has been found. Day (1969) noted that spurious maximizers, corresponding to parameter points having some component standard deviations very small relative to others, are generated by any small number of sample points grouped sufficiently close together. The spurious maximizers, like the unboundedness of $L(\gamma)$, can create difficulties when using the EM or quasi-Newton algorithms.

The constrained maximum-likelihood formulation presented in the next section avoids the unconstrained problems of singularities and spurious maximizers of the type noted by Day. Section 3 contains the statement and proof of a consistency result for the corresponding constrained estimator.

Received August 1983; revised October 1984.

[1] This research was supported in part by National Science Foundation Grant IST-8407860.

AMS 1980 *subject classifications.* Primary 62F10; secondary 62F12.

Key words and phrases. Consistency, maximum-likelihood estimation, normal mixture densities.

This content downloaded from 128.193.152.162 on Mon, 3 Nov 2014 16:08:54 PM
All use subject to JSTOR Terms and Conditions

2. The constraints. Day (1969) recommended the use of ML when it is known (and appropriate constraints added) that the variances for the component densities are equal. The crucial ingredient is not equality of component variances; it is knowledge of their relative sizes so that appropriate constraints of the form $\sigma_i = c_{ij}\sigma_j$ can be added (see Quandt and Ramsey, 1978).

The constraints chosen here have the flavor of those above, but exact knowledge is not required. Instead, imprecise knowledge usually available can be imposed imprecisely through the use of the set of linear inequality constraints $\sigma_i \geq c\sigma_j$, which can be written as

$$\min_{i,j}(\sigma_i/\sigma_j) \geq c > 0. \tag{2.1}$$

The first mention of these types of constraints is found in Dennis (1981) and should be attributed independently to E. M. L. Beale and J. R. Thompson. For reasonable choices of c, the constraints in (2.1) rule out the spurious local maximizers corresponding to greatly differing component standard deviations described by Day. The following result shows that imposing the constraints in (2.1) leads to a well-posed optimization problem having a (constrained) global solution. For notational convenience, the set of parameter values in Γ satisfying (2.1) is denoted by Γ_c.

THEOREM 2.1. *Let $\{x_1, \cdots, x_n\}$ be a set of observations containing at least $m + 1$ distinct points. Then for c in $(0, 1]$, there exists a constrained global maximizer of $L(\gamma)$ over Γ_c.*

PROOF. Let $a = \max\{|x_1|, \cdots, |x_n|\}$ and suppose that $\bar{\gamma} \in \Gamma_c$ satisfies $\bar{\mu}_i > a$. Then $L(\bar{\gamma}) \leq L(\gamma')$, where $\gamma' \in \Gamma_c$ is obtained from $\bar{\gamma}$ by setting the ith mean component equal to a. If $\bar{\mu}_i < -a$, then an analogous result holds by setting the ith mean component equal to $-a$. Also, if $\{\gamma^p\} \in \Gamma_c$ is a sequence satisfying $\lim_{p\to\infty}\sigma_i^p = 0$ or ∞, then $\lim_{p\to\infty}L(\gamma^p) = -\infty$.

It follows from the above that $\sup_{\gamma\in\Gamma_c}L(\gamma) = \sup_{\gamma\in S}L(\gamma)$, where $S = \{\gamma \in \Gamma_c \mid |\mu_i| \leq a < \infty, 0 < b \leq \sigma_i \leq d < \infty \text{ for } i = 1, \cdots, m\}$ for some constants a, b, and d. By the compactness of S and the continuity of $L(\gamma)$, there exists a parameter $\tilde{\gamma}^n$ in Γ_c satisfying $L(\tilde{\gamma}^n) = \sup_{\gamma\in S}L(\gamma) = \sup_{\gamma\in\Gamma_c}L(\gamma)$. □

The constraints in (2.1) yield an optimization problem having a global solution and a constrained parameter space with no singularities and at least a smaller number of spurious maximizers. Consistency of the constrained estimator is proved in the next section, but it is mentioned here that other ways of regularizing the maximum-likelihood problem for normal mixtures exist. Redner (1981) has shown the consistency of constrained global maximizers of $L(\gamma)$ in the case that the constraints define a compact subset of Γ, and Policello (1981) obtains consistency and avoids singularities by conditioning on there being at least two observations from each component population present in the sample. Redner's compactness assumption is not necessary, while Policello's approach leads to an iteration which is more complicated and computationally expensive than the modified EM approach based on the constrained formulation presented here (Hathaway, 1983). Additionally, the constrained formulation presented here

This content downloaded from 128.193.152.162 on Mon, 3 Nov 2014 16:08:54 PM
All use subject to JSTOR Terms and Conditions

warrants consideration because of existing computational results. Those results, found in Hathaway (1983), indicate that the corresponding constrained algorithm produces good estimates, even when started with poor initial guesses.

3. Consistency. Strong consistency of the constrained estimator is shown by applying existing maximum-likelihood theory due to Kiefer and Wolfowitz (1956). Some details of the proof will be omitted whenever it is possible instead to reference their work.

Let Γ denote the parameter space for a family of distributions with densities $f(x;\gamma)$ for $\gamma \in \Gamma$. It is assumed that Γ is a measurable subset of Euclidean r-space, with the true parameter $\gamma^0 \in \Gamma$. The operator E will always denote expectation under γ^0. In the set Γ, define the metric

$$\delta(\gamma, \gamma') = \textstyle\sum_{s=1}^{r} |\text{arc tan } \gamma_s - \text{arc tan } \gamma_s'|$$

where $|\cdot|$ is Euclidean distance. Let $\bar{\Gamma}$ denote the set Γ along with the limits of its Cauchy sequences in the sense of $\delta(\cdot, \cdot)$.

Kiefer and Wolfowitz (1956) proved that their Assumptions 1–5 imply the results found in Theorems 1 and 2 in Wald (1949). The following extensions of these results for the case of nonidentifiable distributions are stated next. (Redner (1981) extends Wald's results in a similar way using the assumptions found in Wald (1949) rather than those of Kiefer and Wolfowitz, but Redner's theory does not easily apply to the constrained formulation presented here.)

Let

$$C = \left\{ \gamma \in \bar{\Gamma} \;\middle|\; \int_{-\infty}^{y} f(x;\gamma)\, d\mu = \int_{-\infty}^{y} f(x;\gamma^0)\, d\mu \text{ for all } y \right\},$$

and let $\hat{\Gamma}$ be the quotient topological space obtained from $\bar{\Gamma}$ by identifying C to a point denoted $\hat{\gamma}^0$. The proofs of Theorems 3.1 and 3.2 for nonidentifiable distributions follow immediately from the proofs in Kiefer and Wolfowitz for the identifiable case.

THEOREM 3.1. *Let Assumptions* 1, 2, 3, *and* 5 *of Kiefer and Wolfowitz hold and let S be any closed subset of* $\bar{\Gamma}$ *not intersecting C, then*

$$P\{\lim_{n\to\infty} \sup_{\gamma\in S} \textstyle\prod_{i=1}^{n} f(x_i;\gamma)/\prod_{i=1}^{n} f(x_i;\gamma^0) = 0\} = 1.$$

THEOREM 3.2. *Let Assumptions* 1, 2, 3, *and* 5 *of Kiefer and Wolfowitz hold and let* $\hat{\gamma}^n = \hat{\gamma}^n(x_1, \cdots, x_n)$ *be any function of the observations* $x_1, \cdots, x_n$ *such that*

$$\textstyle\prod_{i=1}^{n} f(x_i;\hat{\gamma}^n)/f(x_i;\gamma^0) \ge \rho > 0 \quad \textit{for all} \quad n,$$

then

$$P\{\lim_{n\to\infty} \hat{\gamma}^n = \hat{\gamma}^0\} = 1.$$

The preceding theory is used below to prove consistency for the constrained mixture problem, but first an important (and necessary) device from Section 6 of Kiefer and Wolfowitz (1956) is noted. This device, discussed in more detail in

This content downloaded from 128.193.152.162 on Mon, 3 Nov 2014 16:08:54 PM
All use subject to JSTOR Terms and Conditions

Perlman (1972), is simply to work with the joint density of k observations instead of the density corresponding to a single observation; here for a mixture of m univariate normals, this amounts to showing consistency by verifying Assumptions 1, 2, 3, and 5 for the joint density of $m + 1$ observations.

THEOREM 3.3. *Let $c \in (0, 1]$ be such that the true parameter γ^0 is in Γ_c, and let $\hat{\gamma}^n$ globally maximize $L(\gamma)$ over Γ_c for the sample $x_1, \cdots, x_n$ of* i.i.d. *observations of $X \sim p_m(x; \gamma^0)$. Then $\hat{\gamma}^n$ is strongly consistent.*

PROOF. The assumptions are verified for the joint density of $m + 1$ observations. First, the set Γ_c is completed (in the sense of $\delta(\cdot, \cdot)$) to obtain

$$\bar{\Gamma}_c = \{\gamma = (\alpha, \mu, \sigma)^T \mid \textstyle\sum_{i=1}^m \alpha_i = 1,\ \alpha_i \geq 0,\ \min \sigma_i/\sigma_j \geq c,\ 0 \leq \sigma_i \leq \infty,$$

$$-\infty \leq \mu_i \leq \infty \text{ for } i = 1, \cdots, m\}.$$

The definition of $p_m(x; \gamma)$ is extended by setting $p_m(x; \gamma) = \sum_{i=1}^m A_i \alpha_i P(x; \mu_i, \sigma_i)$ with $A_i = 1$ if $|\mu_i| < \infty$ and $0 < \sigma_i < \infty$ for $i = 1, \cdots, m$, and $A_i = 0$ otherwise. With this definition, Assumptions 1, 2, and 3 are easily verified for the joint density $\bar{P}(x_1, \cdots, x_{m+1}; \gamma) = \bar{P}(\bar{x}; \gamma)$ of $m + 1$ observations. The proof now rests on a verification of Assumption 5.

Kiefer and Wolfowitz noted that Assumption 5 follows from

$$E \log f(x; \gamma^0) > -\infty \tag{3.1}$$

and

$$E \sup\nolimits_{\gamma \in \bar{\Gamma}} \log f(x; \gamma) < \infty. \tag{3.2}$$

Upon multiplying out the factors, it is easily seen that $\bar{P}(\bar{x}; \gamma)$ is itself a mixture of m^{m+1} components, each having the form $\phi(\bar{x}; \bar{\imath}, \gamma) = \phi(\bar{x}; i_1, i_2, \cdots, i_{m+1}, \gamma) = \prod_{j=1}^{m+1} P(x_j; \mu_{i_j}, \sigma_{i_j})$ for some choices $i_j \in \{1, 2, \cdots, m\}$, $j = 1, \cdots, m$ of the components of $\bar{\imath}$. Now (3.1) for $\bar{P}(\bar{x}; \gamma^0)$ is implied by $E \,|\log \bar{P}(\bar{x}; \gamma^0)| < \infty$, but using the last inequality in the proof of Redner's (1981) Theorem 5, this is true if

$$E \,|\log \phi(\bar{x};\ \bar{\imath}\ \gamma^0)| < \infty \tag{3.3}$$

holds for all components $\phi(\bar{x}; \bar{\imath}, \gamma^0)$. The proof of (3.1) is now finished by noting that (3.3) follows from the easily shown fact

$$\int_{-\infty}^{\infty} |\log P(x; \mu_i^0, \sigma_i^0)| \, P(x; \sigma_j^0, \sigma_j^0)\, dx < \infty \quad \text{for} \quad i, j = 1, \cdots, m.$$

Now (3.2) is true for $\bar{P}(\bar{x}; \gamma)$ if

$$E \sup\nolimits_{\gamma \in \bar{\Gamma}_c} \log \phi(\bar{x}; \bar{\imath}, \gamma) < \infty \tag{3.4}$$

holds for all components $\phi(\bar{x}; \bar{\imath}, \gamma)$. (See Theorem 5, Redner, 1981). To show

This content downloaded from 128.193.152.162 on Mon, 3 Nov 2014 16:08:54 PM
All use subject to JSTOR Terms and Conditions

(3.4), we note that for each component $\phi(\bar{x};\ \bar{\iota},\ \gamma)$, there exists a choice of j, $k \in \{1, \cdots, m+1\}$ and $\ell \in \{1, \cdots, m\}$ such that

$$(3.5)\quad \sup_{\gamma\in\bar{\Gamma}_c} \log \phi(\bar{x};\ \bar{\iota},\ \gamma) \leq \sup_{\gamma\in\bar{\Gamma}_c} \log(\beta(\sigma_\ell)P(x_j;\ \mu_\ell,\ \sigma_\ell)P(x_k;\ \mu_\ell,\ \sigma_\ell))$$

for all $(x_1, \cdots, x_{m+1})$, where $\beta(\sigma_\ell) = (2\pi)^{-(m-1)/2}(c\sigma_\ell)^{1-m}$. The proof is completed by noting that the expectation of the right-hand side of (3.5) equals

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \log(\eta/|\,x_k - x_j\,|^{m+1})p_m(x_k;\ \gamma^0)p_m(x_j;\ \gamma^0)\ dx_k\ dx_j \quad \text{for some} \quad 0 < \eta < \infty,$$

and this latter integral is easily shown to be less than $+\infty$. □

This section is concluded by noting that all the good asymptotic properties associated with the consistent local maximizer (Kiefer, 1978) are shared by the global maximizer $\hat{\gamma}^n$ in the case that $\alpha_i^0 \neq 0$ for $i = 1, \cdots, m$ and $(\mu_i,\ \sigma_i) \neq (\mu_j,\ \sigma_j)$ for $i \neq j$. Also in this special case, the set C consists of only a finite number of points.

4. Concluding remarks. The imposition of simple constraints of the form (2.1) yields a maximum-likelihood problem which is well posed optimizationally. The constrained formulation is statistically well posed in that the global solutions are strongly consistent. Problems associated with singularities do not exist, and those associated with spurious maximizers are at least lessened. The only restriction is to choose a value of c for which the true parameter satisfies (2.1).

It is worth mentioning again that many choices of constraints yield an optimization problem having a global solution. However, the constraints in (2.1) are among the simplest constraints to implement in a numerical algorithm. Using a particular subset of these constraints, a constrained EM algorithm has been developed, and preliminary numerical tests indicate the constrained approach is effective in producing good maximum-likelihood estimates. In one numerical test in Hathaway (1983), 10 samples, each of size 200, were generated from a mixture of 2 normal distributions, with the true parameter defined componentwise by $\alpha_1 = \alpha_2 = .5$, $\mu_1 = 0$, $\mu_2 = 2.5$, and $\sigma_1 = \sigma_2 = 1$. Out of 80 trials (10 samples, 8 poor initial guesses), the constrained algorithm which forced each iterate to satisfy $\alpha_1 \geq .1$, $\alpha_2 \geq .1$, $\sigma_1 \geq .1\sigma_2$, and $\sigma_2 \geq .1\sigma_1$ converged to the most accurate maximizer 65 times, while the (unconstrained) EM algorithm was successful in only 49 cases. This type of robustness is desirable even though good initial guesses are usually available. Even when no prior information as to the choice of c is available, it is possible to calculate good estimates by varying c dynamically during the constrained EM iterations. A complete discussion of the constrained algorithm can be found in Hathaway (1983).

Theorems 3.1 and 3.2 can be applied to many other families of mixture distributions as long as constraints are imposed (if necessary) to insure that the assumptions of Kiefer and Wolfowitz, particularly Assumption 5, hold. For example, for a mixture of m q-variate normals, constraining all characteristic roots of $\sum_i \sum_j^{-1}$ $(1 \leq i \neq j \leq m)$ to be greater than or equal to some minimum

This content downloaded from 128.193.152.162 on Mon, 3 Nov 2014 16:08:54 PM
All use subject to JSTOR Terms and Conditions

value $c > 0$ (satisfied by the true parameter) leads to a constrained (global) maximum-likelihood formulation. Another example is the case of a mixture of m univariate Cauchy distributions with common scale parameter. Assumption 5 is verified using the joint density of $m + q$ observations in the normal case and $m + 1$ for the Cauchy mixture.

Finally, several questions concerning the normal mixture problem remain. Is it possible to let c decrease to zero as the sample size increases to infinity while maintaining consistency? If the answer is yes, then at what rate can c be decreased to zero? For a fixed c, do all spurious maximizers of $L(\gamma)$ in Γ_c disappear as the sample size increases to infinity? Under some restrictions, empirical studies indicate the answer to the last question could be yes. It is not known now how concave $L(\gamma)$ ultimately gets.

Acknowledgement. The author is obliged to the Editor for many helpful suggestions, including one of more naturally basing the proof of Theorem 3.3 on the work of Kiefer and Wolfowitz (1956) rather than that of Wald (1949).

REFERENCES

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56** 463–474.

DENNIS, J. E., JR. (1981). Algorithms for nonlinear fitting. Proceedings of the NATO Advanced Research Symposium, Cambridge University, Cambridge, England.

FOWLKES, E. B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *J. Amer. Statist. Assoc.* **74** 561–575.

HATHAWAY, R. J. (1983). Constrained maximum-likelihood estimation for a mixture of m univariate normal distributions. Unpublished Ph.D. Thesis, Dept. of Mathematical Sciences, Rice University.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum-likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 888–906.

KIEFER, N. M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica* **46** 427–434.

PERLMAN, M. D. (1972). On the strong consistency of approximate maximum likelihood estimators. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 263–282, Univ. of California Press.

POLICELLO, G. E., II (1981). Conditional maximum likelihood estimation in Gaussian mixtures. In *Statistical Distributions in Scientific Work* 5, (C. P. Taillie et al., eds.) 111–125. International Co-operative Publishing House, Maryland.

QUANDT, R. E. and RAMSEY, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc.* **73** 730–738.

REDNER, R. A. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.* **9** 225–228.

REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood, and the EM algorithm. *SIAM Rev.* **26** 195–239.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20** 595–601.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF SOUTH CAROLINA
COLUMBIA, SOUTH CAROLINA 29208

This content downloaded from 128.193.152.162 on Mon, 3 Nov 2014 16:08:54 PM
All use subject to JSTOR Terms and Conditions