*The Annals of Statistics*
1997, Vol. 25, No. 1, 378–386

# MOMENT-BASED OSCILLATION PROPERTIES OF MIXTURE MODELS

BY BRUCE LINDSAY[1] AND KATHRYN ROEDER[2]

*Pennsylvania State University and Carnegie Mellon University*

Consider finite mixture models of the form $g(x; Q) = \int f(x; \theta)\, dQ(\theta)$, where $f$ is a parametric density and $Q$ is a discrete probability measure. An important and difficult statistical problem concerns the determination of the number of support points (usually known as components) of $Q$ from a sample of observations from $g$. For an important class of exponential family models we have the following result: if $P$ has more than $p$ components and $Q$ is an appropriately chosen $p$-component approximation of $P$, then $g(x; P) - g(x; Q)$ demonstrates a prescribed sign change behavior, as does the corresponding difference in the distribution functions. These strong structural properties have implications for diagnostic plots for the number of components in a finite mixture.

**1. Introduction.** Consider a family of univariate probability densities $f(x; \theta)$, with respect to some $\sigma$-finite measure $d\gamma(x)$, parameterized by $\theta \in \Omega$. Frequently, interest lies in mixtures of such densities. The random variable $X$ is said to have a mixture distribution $G(\cdot; Q)$ if it has density

$$g(x; Q) = \int f(x; \theta)\, dQ(\theta), \tag{1}$$

and the mixing distribution $Q$ is a probability measure on $\Omega$. If $Q$ has a finite number of support points $\nu \equiv \nu(Q)$, then we say $Q$ is a finite mixing distribution and we write $Q_\nu = \sum \pi_j \delta(\theta_j)$, with $\theta_1, \ldots, \theta_\nu$ being the support points (often called components) and $\pi_1, \ldots, \pi_\nu$ being the weights.

A problem of longstanding interest in such models is inference on the unknown value of $\nu(Q)$. At the simplest level, this is the problem of determining if $\nu = 1$, the one-component model, or if $\nu > 1$, the multicomponent model. Shaked (1980) presented important results for this problem when the component densities $f(x; \theta)$ are from a one-parameter exponential family. We build on his results in two ways, generalizing to the discrimination between $\nu = p$ versus $\nu > p$, and moving beyond the one-parameter exponential family to the normal mixture model in which each component has a different mean, but the same unknown variance.

Here we summarize Shaked's sign crossings results. Suppose we wish to contrast a multicomponent model $g(x; Q)$ with a plausible one-component

Received October 1992; revised March 1996.

[1]Research supported by NSF Grant DMS-91-06895 and the Population Issues Research Center of Pennsylvania State University.

[2]Research supported by NSF Grants DMS-92-01211 and DMS-94-96219.

*AMS* 1991 *subject classifications*. Primary 62E10, 62G05; secondary 62H05.

*Key words and phrases*. Mixtures, exponential family, total positivity, sign changes, diagnostic plots.

model $f(x;\theta)$. Choose $\theta = \theta^*$ for the one-component model so that the observed variable $X$ has the same mean under both densities:

$$\int x g(x;Q)\, d\gamma(x) = \int x f(x;\theta^*)\, d\gamma(x).$$

Our notation for this last equation will be $E[X;Q] = E[X;\theta^*]$. Shaked showed that $g(x;Q) - f(x;\theta^*)$ has exactly two sign changes, in the order $(+,-,+)$, as $x$ traverses the sample space. That is, $g(x;Q)$ has heavier tails than $f(x;\theta^*)$. Moreover, the difference in the corresponding distribution functions $G(x;Q) - F(x;\theta^*)$ has exactly one sign change, in the order $(+,-)$.

We extend his results as follows: let $P$, the nominal true mixing distribution, satisfy $\nu(P) > p$; choose $Q_p$, a candidate $p$-point probability measure, such that it satisfies

$$E[X^k;P] = E[X^k;Q_p], \qquad k = 0,1,\ldots,2p-1. \tag{2}$$

(In Section 2, we show how to solve for $Q_p$.) Then, in Theorem 3.2, we show that $g(x;P) - g(x;Q_p)$ has *exactly* $2p$ sign changes in the order $(+,-,\ldots,-,+)$, unless it is identically 0 (the case of nonidentifiable $P$). An exact sign change result for the difference in distribution functions is also given in Section 3. In Section 4, these results are extended to normal densities with unknown variance.

Before proceeding to the mathematical verification of these results, we offer a few brief comments on their potential application. In Figure 1(a), we plot $[g(x;P) - g(x;Q_2)]/\sqrt{g(x;P)}$ for the case when $f(x;\theta)$ is Poisson, $P$ puts mass 1/3 each at (1, 3 and 5) and $Q_2$ is constructed to match moments as specified in (2). We note the clear trimodality of this function, in contrast to the unimodality of the density $g(x;P)$ [Figure (1b)].

Shaked demonstrated that his sign change results could be used for diagnostic checks to determine if the data were from a mixture of specified exponential family densities rather than a one-component model. These ideas were further developed in Lindsay and Roeder (1992). When interest lies in assessing the number of components in a finite mixture, the oscillation results obtained in this article have clear implications for diagnostic plots. In Section 5, the number of accidents per year is modeled as a Poisson mixture to illustrate the diagnostic potential of the results. In a companion paper, these results are used to develop diagnostic plots for the case of normal mean mixtures with unknown variance [Roeder (1994)].

## 2. Background.

2.1. *The models under investigation.* We will be interested in component densities $f(x;\theta)$, where both $x$ and $\theta$ have ranges in the real numbers, say $x \in \mathcal{T} \subset R$ and $\theta \in \Omega$, and $f(\cdot;\cdot)$ satisfies regularity conditions which will be expounded in this section.

A real function of two variables, $K(x,\theta)$, ranging over linearly ordered sets $\mathcal{T}$ and $\Omega$ is said to be *strictly totally positive* (STP) of order $r$, if, for all

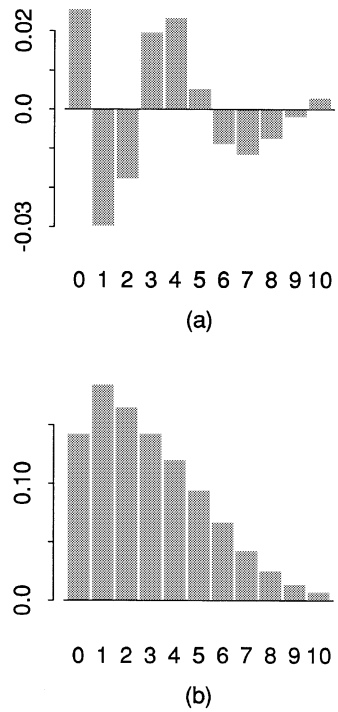


FIG. 1. (a) *Plot of* $[g_3 - g_2]/\sqrt{g_2}$, *where* $g_3$ *is the density of a mixture of three Poissons* $(1/3[f(y;1)+f(y;3)+f(y;5)])$ *and* $g_2$ *is a mixture of two Poissons, selected to have three moments in common with* $g_3$. (b) *Density of* $g_3$.

$x_1 < x_2 < \cdots < x_m$ and $\theta_1 < \theta_2 < \cdots \theta_m$; $x_i \in \mathcal{I}$, $\theta_j \in \Omega$; $1 \le m \le r$, we have the inequalities

$$\begin{vmatrix} K(x_1, \theta_1) & \cdots & K(x_1, \theta_m) \\ \vdots & & \vdots \\ K(x_m, \theta_1) & \cdots & K(x_m, \theta_m) \end{vmatrix} > 0$$

[Karlin (1968), pages 11 and 15]. Many density functions occurring in statistical theory are STP. The list includes the one-parameter exponential family with density function $f(x;\theta) = \exp\{\theta x - \psi(\theta)\}$, the noncentral-$t$ and the noncentral-$\chi^2$ densities.

2.2. *Background on moments and exponential families.* In order to apply our results in a particular model, we need to establish an important structural feature for the component densities $f(x;\theta)$ beyond total positivity. Suppose that $P$ is a mixing distribution with $p$ or more support points. Then we need to be able to construct a $p$-point distribution $Q_p$ such that the first $2p-1$ moments of $g(x;P)$ and $g(x;Q_p)$ match, satisfying (2). Fortunately, there exists an important class of exponential families (the quadratic variance class) in

which $Q_p$ satisfying (2) can be shown to exist. This class includes the normal, gamma, Poisson and binomial distributions. The following is a brief review of techniques found in Lindsay (1989).

In the quadratic variance class of exponential family models [Morris (1983)], for each $k$, there exists a polynomial of degree $k$, call it $\xi_k(x)$, such that

$$\int \xi_k(x) f(x;\theta)\, d\gamma(x) = (\mu - \mu_0)^k \tag{3}$$

for mean value parameter $\mu$. The choice of $\mu_0$ is arbitrary so we set it to 0. For example, in the Poisson with mean $\mu$, $E[X] = \mu$, $E[X(X-1)] = \mu^2$, $E[X(X-1)(X-2)] = \mu^3$ and so forth. In addition, a classical moment result indicates that for a given distribution $P$ with no fewer than $p$-points of support, there exists a unique distribution $Q_p$ with exactly $p$-points of support such that

$$\int \mu^k dQ_p(\mu) = \int \mu^k dP(\mu), \qquad k = 1, \ldots, 2p-1. \tag{4}$$

Thus integrating both sides of (3) with respect to $dQ_p(\mu)$ and $dP(\mu)$ and using (4) yields

$$E[\xi_k(X); P] = E[\xi_k(X); Q_p], \qquad k = 1, \ldots, 2p-1. \tag{5}$$

Finally, the linear transformation taking $(1, x, \ldots, x^{2p-1}) \to (\xi_0(x), \xi_1(x), \ldots, \xi_{2p-1}(x))$ is invertible, so (5) implies (2).

More details on solving (5) for $Q_p$ are given in Lindsay (1989). The solutions can be obtained algebraically for $p = 2$. For arbitrary $p$, the problem involves solving a degree $p$ polynomial for its $p$ real roots.

**3. One-parameter models.** In this section, we obtain sign change results for one-parameter models. The following notation [Karlin (1968), page 20] will be used. Let $a(x)$ be defined on $I$, where $I$ is a subset of the real line. The number of sign changes of $a$ in $I$ is defined by

$$S^-(a) = \sup S^-[a(x_1), \ldots, a(x_m)], \tag{6}$$

where $S^-[a(x_1), \ldots, a(x_m)]$ is the number of sign changes of the indicated sequence, zero terms being discarded, and the supremum is extended over all sets

$$x_1 < x_2 < \cdots < x_m, \qquad x_i \in I,\ m < \infty. \tag{7}$$

We assume throughout that $f(x;\theta)$ is STP and that $P$ and $Q_p$ satisfy (2). The following notation will be used throughout this section: $g_+ \equiv g(x;P)$, $g_p \equiv g(x;Q_p)$, $G_+ \equiv G(x;P)$ and $G_p \equiv G(x;Q_p)$.

REMARK. In the following result, we will give exact sign change results for $g_+ - g_p$ with the proviso "the difference $g_+ - g_p$ is not identically 0" with the possible exception of a $\gamma$-null set. If such an equality in densities occurs, it is clear that there is an identifiability problem; both $P$ and $Q_p$ are generating the same distribution. The results of Lindsay and Roeder (1993) can be used

to determine exactly when this will occur. If the sample space is infinite, it will not occur. If the sample space has $N$ points, then $p$-point distributions $Q_p$ are identifiable when $p \le (N-1)/2$, and so $g_+ - g_p$ cannot be identically 0. If both $P$ and $Q_p$ have more than $(N-1)/2$ points, then $g_+ - g_p$ cannot have exactly $2p$ sign changes, since we can have at most $N-1$ sign changes as we traverse the sample space. Thus our result proves that $P$ and $Q_p$ generate the same density.

LEMMA 3.1. *Provided* $g_+ - g_p$ *is not identically* 0, $S^-(g_+ - g_p) \le 2p$.

PROOF. Define the measure $d\chi(\theta)$ by

$$d\chi(\theta) = d(P + Q_p)(\theta).$$

Let

$$p^*(\theta) = \begin{cases} P(\{\theta\})/[P(\{\theta\}) + Q_p(\{\theta\})], & \text{if } \theta \in \{\theta_1, \ldots, \theta_p\}, \\ 1, & \text{otherwise} \end{cases}$$

and

$$q^*(\theta) = \begin{cases} Q_p(\{\theta\})/[P(\{\theta\}) + Q_p(\{\theta\})], & \text{if } \theta \in \{\theta_1, \ldots, \theta_p\}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta_1, \ldots, \theta_p$ are the support points of $Q_p$.

Then $p^*$ and $q^*$ are versions of the Radon-Nikodym derivatives $dP/d\chi$ and $dQ_p/d\chi$, so that $g_+ - g_p = \int f(x;\theta)[p^*(\theta) - q^*(\theta)]\, d(P + Q_p)(\theta)$.

We now apply Theorem 3.1(b) of Karlin (1968), page 21, noting that $p^*(\theta) - q^*(\theta) = 1$, except possibly at the support of $Q_p$, where it can be negative. Hence it undergoes a maximum of $2p$ sign changes. Karlin's result then implies that integration with respect to the STP kernel $f(x;\theta)$ will result in a function, $g_+ - g_p$, with no more sign changes in $x$ than $p^*(\theta) - q^*(\theta)$ has in $\theta$ relative to $d\chi$. This establishes an upper bound of $2p$ sign changes in $g_+ - g_p$. □

THEOREM 3.2. *Provided* $g_+ - g_p$ *is not identically* 0, $S^-(g_+ - g_p) = 2p$, *with sign changes in the order* $(+, -, \ldots, -, +)$.

PROOF. From Lemma 3.1, we obtain an upper bound on the number of sign changes of $2p$. Because $\int x^k (g_+ - g_p)(x)\, d\nu(x) = 0$ for $k = 1, \ldots, 2p-1$, any polynomial $A(x)$ of degree less than or equal to $2p-1$ satisfies

$$\int A(x)(g_+ - g_p)(x)\, d\gamma(x) = 0.$$

Suppose $S^-(g_+ - g_p) \le 2p-1$. Then we can construct a polynomial $A(x)$ that matches $g_+ - g_p$ in sign (i.e., it has single roots exactly at the roots of $g_+ - g_p$). It follows that $A(x)(g_+ - g_p)(x) \ge 0$, and, since it has 0 integral, it must be 0 except for a set of $\gamma$-measure 0. Hence either $g_+ = g_p$ or $g_+ - g_p$ has $2p$ sign changes. □

REMARK. As is clear from the proof for this result, our oscillation results still hold if we replace $x^k$ in (2) with any system of functions $\alpha_k(x)$, such as $x^k e^{-x}$, provided that one can construct a polynomial $A(x) = \sum a_k \alpha_k(x)$ which has any prespecified set of $2p-1$ 0's. Such an approach could be useful in improving on the robustness of the sample moments in applications by using bounded variables such as $\alpha_k(x) = x^k e^{-x}$. The next theorem, however, uses the special form of $x^k$.

THEOREM 3.3. *Provided $G_+ - G_p$ is not identically 0, $S^-(G_+ - G_p) = 2p-1$, with sign changes in the order $(+, -, \ldots, +, -)$. The roots occur between the roots of $g_+ - g_p$.*

PROOF. An upper bound is obtained on the number of sign changes by appealing to the sign change behavior of $g_+ - g_p$. The function $G_+ - G_p$ is increasing on the intervals $[a, b]$, where $g_+ - g_p \geq 0$:

$$G_+(b) - G_p(b) - (G_+(a) - G_p(a)) = \int I[a < x \leq b]\,(g_+ - g_p)(x)\,d\gamma(x) \geq 0.$$

From this it follows that $G_+ - G_p$ has at most one crossing in each interval where $g_+ - g_p$ is constant in sign, but has none in the first or last interval. Hence $S^-(G_+ - G_p) \leq 2p-1$. Integration by parts gives

$$0 = \int x\,d(G_+ - G_p)(x) = \int [G_+ - G_p](x)\,dx,$$

and, more generally,

$$0 = \int x^k d(G_+ - G_p)(x) = \int x^{k-1}[G_+ - G_p](x)\,dx,$$

up to $k = 2p-1$. Now, follow the proof of Theorem 3.2. If $G_+ - G_p$ had $2p-2$ or fewer sign changes, a polynomial $A(x)$ of degree $2p-2$ could be constructed with matching signs. Hence $A(x)[G_+ - G_p](x) \geq 0$, but has zero integral. The result follows. □

**4. Normal mean mixtures with unspecified variance.** In this section, we consider a mixture model of great interest—the normal mean mixture. We use the following notation: let $f(x; \mu, \tau)$ denote the density of an $N(\mu, \tau)$ random variable and let $g(x; Q, \tau) = \int f(x; \mu, \tau)\,dQ(\mu)$ denote a mixture of normals with corresponding distribution function $G(x; Q, \tau)$. If $\tau$ is known, then this is just a special case of the previous section; however, in practice, $\tau$ will typically be unknown and hence we treat it as a free parameter. In this section, we extend our results to this case. We first present an existence theorem, due to Lindsay (1989), which extends the classic moment results presented in Section 2 to normal mixtures.

THEOREM 4.1. *If $Q$ is a distribution with more than $p$ points of support, then there exists a unique $p$-point distribution, $Q_p$, and variance $\tau_p > \tau$ such*

*that*

$$\int x^k \, dG(x; Q_p, \tau_p) = \int x^k \, dG(x; Q, \tau) \quad \textit{for } k = 0, 1, \ldots, 2p. \tag{8}$$

PROOF. While this is not explicitly stated in Lindsay (1989), it is a consequence of Lemma 5A and Theorem 5C. In the latter, replace the empirical moments with the moments of $X$ under $G(\cdot; Q, \tau)$. □

THEOREM 4.2. *If* $(Q_p, \tau_p)$ *satisfies* (8) *for* $Q = Q_{p+1}$, *a* $p+1$*-point distribution*, *then*

$$g(x; Q_{p+1}, \tau) - g(x; Q_p, \tau_p)$$

*has exactly* $2p+2$ *sign changes*, *occurring in the order* $(-, +, \ldots, +, -)$.

PROOF. Since $\tau_p > \tau$, we can represent the above difference as

$$g(x; Q, \tau) - g(x; Q_p^*, \tau),$$

where $Q_p^*$ is the convolution of $Q_p$ with a normal distribution with mean 0 and variance $\tau_p - \tau$. By the same argument as in Lemma 3.1, this means there are a maximum of $2p+2$ sign changes. The polynomial argument used in the proof of Theorem 3.2 can now be used together with (8) to show that there are at least $2p+1$ sign changes. Moreover, since $Q_p^*$ has more mass in the tails than the discrete $Q_{p+1}$, the difference $g(x; Q, \tau) - g(x; Q_p^*, \tau)$ will have a negative sign in both tails, and so must have an even number of sign changes, hence $2p+2$. □

THEOREM 4.3. $G(x; Q, \tau) - G(x; Q_p, \tau_p)$ *has exactly* $2p+1$ *sign changes, in the order* $(-, +, \ldots, +)$.

PROOF. An argument similar to Theorem 3.3.

This result indicates that

$$g(x; Q_2, \tau) - g(x; \mu, \sigma^2)$$

has four sign changes in the order $(-, +, -, +, -)$ provided $\mu$ is the mean of $Q_2$ and $\sigma^2 = \mathrm{Var}(X) = \tau + \mathrm{Var}(Q_2)$. For this case a supplementary result is available from Roeder (1994). If we instead examine the ratio $R(x) = g(x; Q_2, \tau)/g(x; \mu, \sigma^2)$, we obtain a function proportional to a bimodal normal density. By combining the two results we can see that $R(x)$ is bimodal and that both modes are greater than 1.

In the normal model, with $\pi_1 = \pi_2 = 1/2$, the density $g(x; Q_2, \tau)$ is bimodal if and only if the two separate supports $\mu_1$ and $\mu_2$ satisfy $|\mu_1 - \mu_2| > 2\tau$ [Robertson and Fryer (1969)]. Thus the ratio function is much more sensitive to the existence of two support points than is the density itself. This sensitivity continues to exist even for very small support weights $\pi_i$.

Graphical techniques, such as the normal scores plot [Harding (1948) and Cassie (1954)] and the modified percentile plot [Fowlkes (1979)], have played an important role in identifying whether data follow a mixture of two normal distributions. The geometric characterizations obtained herein extend the arsenal of potential diagnostic plots for normal mixtures.

**5. Applications.** The data appearing in Table 1 of Thyrion (1961) are purported to be a mixture of Poissons. The data consist of observed counts of accidents per year for 9461 Belgian drivers. Fitting the data to a two-point mixture using both the method of moments and maximum likelihood estimation, we obtained estimates for $(\theta_1, \theta_2, \pi)$ equal to (0.162, 1.64, 0.965) and (0.147, 1.23, 0.938), respectively. Using the maximum likelihood estimates, we performed a chi-square goodness-of-fit test which indicates that the two-point mixture model does not provide an adequate fit ($\mathscr{X}^2 = 25.21$). In order to apply Theorem 3.2, we used the empirical density $\hat{f}(y) = (1/n)\sum I\{X_i = y\}$ to estimate $g_+$ and the method of moments solution to estimate $g_p$. In Figure 2, to determine if the mixture has more than two components, we plot

$$\frac{\sqrt{n}[\hat{f}(y) - g_2(y)}{\sqrt{g_2(y)}}$$

and obtain a sign sequence $+ - + - +$. In light of Theorem 3.2 and the approximate standard errors, it can be conjectured that the distribution of the number of accidents per year is a mixture of Poissons with at least three points of support.

For continuous $X$, believed to be a mixture of one-parameter exponential family densities, a diagnostic plot based on a nonparametric empirical analog of $G_+ - G_p$ can be constructed directly. Let $F_n$, the empirical distribution

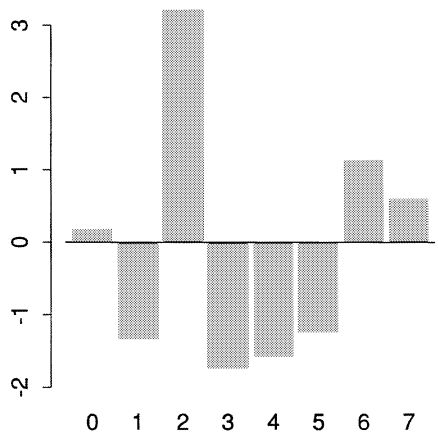


FIG. 2. *Plot of* $\sqrt{n}[\hat{f} - g_2]/\sqrt{g_2}$, *where* $\hat{f}$ *is the empirical density of the data presented in Table* 1 *and* $g_2$ *is the estimated density of these data, fitted to a two-point Poisson mixture density using method of moments.*

function, be an estimate of the alleged distribution $G_+$ and let $\hat{G}_p$ be an estimate of $G_p$ constructed by using the method of moments estimates of the $p$-component model. Naturally, $F_n$ and $\hat{G}_p$ have $2p-1$ moments in common. It follows that if $F_n - \hat{G}_2$ has the sign change behavior specified in Theorem 3.3, then the data provide some support for using more than $p$ components. On the other hand, if a $p$-point mixture is the correct model, then the asymptotic properties of $F_n - \hat{G}_p$ can be obtained from empirical process theory.

Theorems 3.2 and 4.2 can be applied to continuous random variables if $g(x; P)$ and $g(x; Q_{p+1}, \tau)$, respectively, are estimated using nonparametric density estimation techniques. Details of implementation for the normal model are specified in Roeder (1994).

## REFERENCES

CASSIE, R. M. (1954). Some uses of probability paper in the analysis of size frequency distributions. *Australian Journal of Marine Fisheries and Freshwater Research* **5** 513–522.

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

HARDING, J. P. (1949). The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of Marine Biology Association* **28** 141–153.

KARLIN, S. (1968). *Total Positivity* **1**. Stanford Univ. Press.

LINDSAY, B. G. (1989). Moment matrices: applications in mixtures. *Ann. Statist.* **17** 722–740.

LINDSAY, B. G. and ROEDER, K. (1992). Residual diagnostics for mixture models. *J. Amer. Statist. Assoc.* **87** 785–794.

LINDSAY, B. G. and ROEDER, K. (1993). Uniqueness of estimation and identifiability in mixture models. *Canad. J. Statist.* **21** 139–147.

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

ROBERTSON, C. A. and FRYER, J. G. (1969). Some descriptive properties of normal mixtures. *Skand. Aktuarietidskr.* **52** 137–146.

ROEDER, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* **89** 487–495.

SHAKED, M. (1980). On mixtures from exponential families. *J. Roy. Statist. Soc. Ser. B* **42** 192–198.

THYRION, P. (1961). Contribution a l'etude des bonus pour non sinistre en assurance automobile. *Astin Bull.* **1** 142–162.

DEPARTMENT OF STATISTICS
CLASSROOM BUILDING
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802
E-MAIL: lindsay@stat.psu.edu

DEPARTMENT OF STATISTICS
232 BAKER HALL
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15232-3890
E-MAIL: roeder@stat.cmu.edu