BIOMETRICS 54, 525–536
June 1998

Recent Developments in Computer-Assisted Analysis of Mixtures

Dankmar Böhning,* Ekkehart Dietz, and Peter Schlattmann

Department of Epidemiology, Free University Berlin,
Fabeckstrasse 60-62, Haus 562, 14195 Berlin, Germany

SUMMARY

This paper reviews recent developments in the area of computer-assisted analysis of mixture distributions (C.A.MAN). Given a biometric situation of interest in which, under homogeneity assumptions, a certain parametric density occurs, such as the Poisson, the binomial, the geometric, the normal, and so forth, then it is argued that this situation can easily be enlarged to allow a variation of the scalar parameter in the population. This situation is called unobserved heterogeneity. This naturally leads to a specific form of nonparametric mixture distribution that can then be assumed to be the standard model in the biometric application of interest (since it also incorporates the homogeneous situations as a special case). Besides developments in theory and algorithms, the work focuses on developments in biometric applications such as meta-analysis, fertility studies, estimation of prevalence under clustering, and estimation of the distribution function of survival time under interval censoring. The approach is nonparametric for the mixing distribution, including leaving the number of components (subpopulations) of the mixing distribution unknown.

1. Introduction

The importance of mixture distributions, their enormous developments, and their frequent applications over recent years is due to the fact that mixture models offer natural models for unobserved population heterogeneity. What does this mean? Suppose we are dealing with the case that a one-parameter density $f(x, \lambda)$ can be assumed for the phenomenon of interest. Here λ denotes the parameter of the population, whereas x is in the sample space X, a subset of the real line. We call this the homogeneous case. However, this model is often too strict to capture the variation of the parameter over a diversity of subpopulations. In this case, we have that the population consists of various subpopulations, denoted by $\lambda_1, \lambda_2, \ldots, \lambda_k$, where k denotes the number (possibly unknown) of subpopulations. We call this situation the heterogeneous case.

In contrast to the homogenous case, we have the same type of density in each subpopulation j, but a potentially different parameter: $f(x, \lambda_j)$ is the density in subpopulation j. In the sample $x_1, x_2, \ldots, x_n$, it is not observed from which subpopulation the observations are coming. Therefore, we speak of unobserved heterogeneity. Let a latent variable Z describe the population membership. Then the joint density $f(x, z)$ can be written as $f(x, z) = f(x/z)f(z) = f(x, \lambda_z)p_z$, where $f(x/z) = f(x, \lambda_z)$ is the density conditionally on membership in subpopulation z. Therefore, the unconditional density $f(x)$ is the marginal density

$$f(x, P) = \sum_{z=1}^{k} f(x/z)f(z) = \sum_{j=1}^{k} f(x, \lambda_j)p_j, \tag{1}$$

where the margin is taken over the latent variable Z. Note that p_j is the probability of belonging to the jth subpopulation having parameter λ_j. Therefore, the p_j have to meet the constraints $p_j \geq 0, p_1 + \cdots + p_k = 1$. Note that (1) is a mixture distribution with kernel $f(x, \lambda)$ and mixing

* *Corresponding author's email address:* boehning@zedat.fu-berlin.de

Key words: C.A.MAN; Fertility studies; Meta-analysis; Unobserved heterogeneity.

distribution P, in which weights $p_1, \ldots, p_k$ are given to parameters $\lambda_1, \ldots, \lambda_k$. Estimation is done conventionally by maximum likelihood; that is, we have to find the $\hat{P}$ that maximizes the log-likelihood $l(P) = \Sigma_{i=1}^{n} \log f(x_i, P)$. $\hat{P}$ is called the nonparametric maximum likelihood estimator (NPMLE) (Laird, 1978). The software package C.A.MAN (Böhning, Schlattmann, and Lindsay, 1992) provides the NPMLE for P, $\hat{P}$, giving weight $\hat{p}_1, \ldots, \hat{p}_{\hat{k}}$ to $\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{k}}$. Note also that the number of subpopulations k is unknown and estimated.

Many applications are of the following type: Under standard assumptions, the population is homogeneous, leading to a simple, one-parameter and natural density. Examples include the binomial, the Poisson, the geometric, the exponential, and the normal distribution (with additional variance parameter). If these standard assumptions are violated because of population heterogeneity, mixture models can easily capture these additional complexities. Therefore, C.A.MAN offers most of the conventional densities such as normal (common and known different variances), Poisson, Poisson for standardized mortality ratio data, binomial, binomial for rate data, geometric, and exponential, among others. To demonstrate these ideas, we start with a simple example that has recently found its entry into the textbook *Advanced Methods of Marketing Research* (Bagozzi, 1995).

An Introductory Example (Marketing Research)

Data are from a new product and concept test, leading to a variable of interest X = number of individual packs of hard candy purchased within the past 7 days. Figure 1 shows its distribution. Frequently, the assumption of a Poisson distribution, e.g., $f(x, \lambda) = Po(x, \lambda) = e^{-\lambda}\lambda^x/x!$ is done for count data, assuming homogeneity conditions. The heterogeneity analysis provided by C.A.MAN delivers a five-component mixture distribution as shown in Figure 2. These components can easily be interpreted. The two low components correspond to stores with no or almost no sale of the new product, together about 30% of all stores. There are about 50% with a mean sale of 3 packages, 15% with about 7.5 packages, and 10% with the large number of 13 packages.

Developments in the area of computer-assisted analysis of mixtures have been taking place in various areas. There have been theoretical developments, algorithmic developments, developments in direct and indirect applications, the latter meaning developments deviating from the natural genesis of the mixing distribution as capturing population heterogeneity. In the following, we will present some of these developments.

2. Some Pieces of Historic Theory

The strong results of nonparametric mixture distributions are based on the fact that the log-likelihood l is a concave functional on the set of all discrete probability distributions Ω. It is very important to distinguish between the set of all discrete distributions and the set Ω_k of all distributions with a fixed number of k support points (subpopulations). The latter set is not convex. (For details, see Böhning et al., 1992.) The major tool for achieving characterizations and algorithms is the directional derivative at P in the direction Q for both P and Q in Ω:

$$\Phi(P, Q) = \lim_{\alpha \to 0} \frac{l((1-\alpha)P + \alpha(Q)) - l(p)}{\alpha} = \sum_{i=1}^{n} \frac{f(x_i, Q) - f(x_i, P)}{f(x_i, P)}.$$

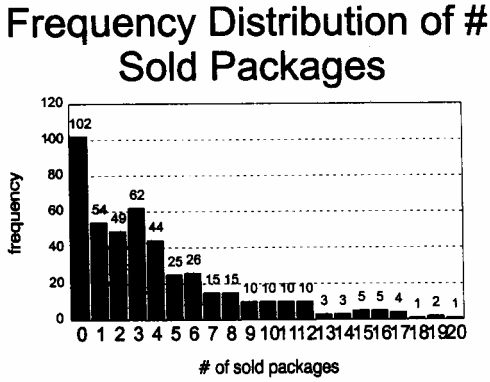

Figure 1. Distribution of sold packages.

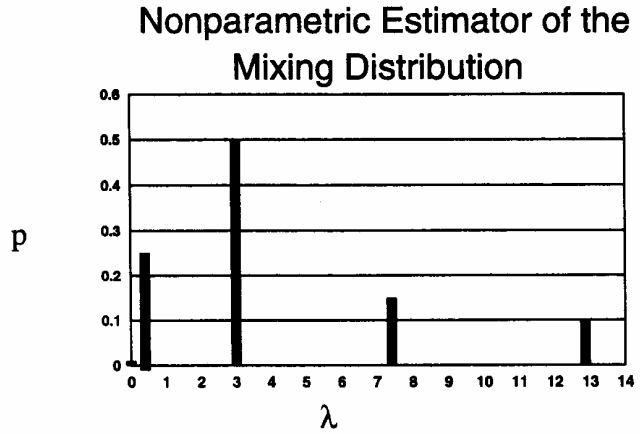

Figure 2. NPMLE of mixing distribution for candy data.

In particular, for one-point mass Q_λ at λ (the vertex of the simplex), the directional derivative is:

$$D_P(\lambda) = \Phi(P, Q_\lambda) = \sum_{i=1}^{n} \frac{f(x_i, \lambda) - f(x_i, P)}{f(x_i, P)} = \sum_{i=1}^{n} \frac{f(x_i, \lambda)}{f(x_i, P)} - n.$$

The determining part in this directional derivative, namely $(1/n)\,\Sigma_{i=1}^{n} f(x_i, \lambda)/f(x_i, P)$, is called the gradient function and denoted by $d(\lambda, P)$. We have the general mixture maximum likelihood theorem (Lindsay, 1983a,b; Böhning, 1982): (a) $\hat{P}$ is NPMLE if and only if $D_{\hat{P}}(\lambda) \leq 0$ for all λ or if and only if $(1/n)\,\Sigma_{i=1}^{n} f(x_i, \lambda)/f(x_i, \hat{P}) \leq 1$ for all λ; (b) $D_{\hat{P}}(\lambda) = 0$ for all support points λ of $\hat{P}$.

Second, the concept of the directional derivative is important in developing reliable converging algorithms (for a review, see Böhning, 1995). Historically, the vertex direction method (VDM) is of interest. In this method, convex combinations $(1 - \alpha)P + \alpha Q_\lambda$ are considered, for which the log-likelihood increase (as a function of the step length α and the vertex direction Q_λ) $l((1-\alpha)P + \alpha Q_\lambda) - l(P)$ is desired to be made as large as possible. A first-order approximation $\alpha D_P(\lambda)$ of this difference leads to the maximization of $D_P(\lambda)$ in λ. Having found a vertex direction with maximum increase, one can choose a monotone or optimal step length α to achieve an update $(1-\alpha)P + \alpha Q_\lambda$.

The VDM is usually slow in its convergence behavior. A faster method (also now used as the standard method in C.A.MAN) is the vertex exchange method (VEM). The basic idea here is to exchange good vertex directions against bad ones already in support of the current mixing distribution. The VEM is defined by $P + \alpha P(\lambda^*)\{Q_\lambda - Q_{\lambda^*}\}$, where $P(\lambda^*)$ is the weight of the bad support point λ^* and α in $[0, 1]$ is a step length. Good and bad support points are identified again by means of the directional derivative. Again, one tries to optimize the gain in the log-likelihood $l(P + \alpha P(\lambda^*)\{Q_\lambda - Q_{\lambda^*}\}) - l(P)$. Now, consider a first-order approximation of this difference: $\alpha P(\lambda^*)\{D_P(\lambda) - D_P(\lambda^*)\}$. Clearly, this is maximized if $D_P(\lambda)$ is maximized in λ and $D_P(\lambda^*)$ is minimized in the support of P. Choosing an optimal or monotonic step length completes the VEM. (For details or different methods, see Böhning (1989, 1995) or Lesperance and Kalbfleisch (1992).)

For the practical realization in C.A.MAN, we recall that the goal is to maximize $l(P)$ in the simplex Ω of all probability distributions P on parameter space λ. We call the solution of this problem the fully iterated nonparametric maximum likelihood estimator, and this solution is achieved in C.A.MAN in two phases (in which phase II is new). In phase I, an approximating grid $\lambda_1, \ldots, \lambda_L$ ($L \leq 50$) is chosen and $l(P)$ is maximized in the simplex Ω_{GRID} of all probability distributions P on grid $\{\lambda_1, \ldots, \lambda_L\}$ with one of the algorithms described above. For example, we can choose the observed data values as an approximating grid. In the introductory example, $L = 21$ different values of sold number of packages were observed: $\{\lambda_1, \ldots, \lambda_L\} = \{0, 1, 2, \ldots, 20\}$. As potential choice of initial weights, the observed relative frequencies or uniform weights $p_i = 1/21$ could be used.

In phase II, all grid points that are left with positive weights as a result of the optimization process in phase I are used as initial values for the EM algorithm (Dempster, Laird, and Rubin, 1977) to produce the fully iterated NPMLE.

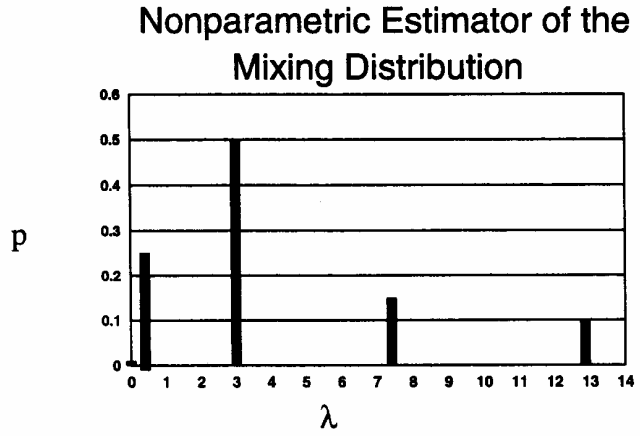

Figure 2. NPMLE of mixing distribution for candy data.

In particular, for one-point mass Q_λ at λ (the vertex of the simplex), the directional derivative is:

$$D_P(\lambda) = \Phi(P, Q_\lambda) = \sum_{i=1}^{n} \frac{f(x_i, \lambda) - f(x_i, P)}{f(x_i, P)} = \sum_{i=1}^{n} \frac{f(x_i, \lambda)}{f(x_i, P)} - n.$$

The determining part in this directional derivative, namely $(1/n)\,\Sigma_{i=1}^{n} f(x_i, \lambda)/f(x_i, P)$, is called the gradient function and denoted by $d(\lambda, P)$. We have the general mixture maximum likelihood theorem (Lindsay, 1983a,b; Böhning, 1982): (a) $\hat{P}$ is NPMLE if and only if $D_{\hat{P}}(\lambda) \leq 0$ for all λ or if and only if $(1/n)\,\Sigma_{i=1}^{n} f(x_i, \lambda)/f(x_i, \hat{P}) \leq 1$ for all λ; (b) $D_{\hat{P}}(\lambda) = 0$ for all support points λ of $\hat{P}$.

Second, the concept of the directional derivative is important in developing reliable converging algorithms (for a review, see Böhning, 1995). Historically, the vertex direction method (VDM) is of interest. In this method, convex combinations $(1 - \alpha)P + \alpha Q_\lambda$ are considered, for which the log-likelihood increase (as a function of the step length α and the vertex direction Q_λ) $l((1-\alpha)P + \alpha Q_\lambda) - l(P)$ is desired to be made as large as possible. A first-order approximation $\alpha D_P(\lambda)$ of this difference leads to the maximization of $D_P(\lambda)$ in λ. Having found a vertex direction with maximum increase, one can choose a monotone or optimal step length α to achieve an update $(1-\alpha)P + \alpha Q_\lambda$.

The VDM is usually slow in its convergence behavior. A faster method (also now used as the standard method in C.A.MAN) is the vertex exchange method (VEM). The basic idea here is to exchange good vertex directions against bad ones already in support of the current mixing distribution. The VEM is defined by $P + \alpha P(\lambda^*)\{Q_\lambda - Q_{\lambda^*}\}$, where $P(\lambda^*)$ is the weight of the bad support point λ^* and α in $[0, 1]$ is a step length. Good and bad support points are identified again by means of the directional derivative. Again, one tries to optimize the gain in the log-likelihood $l(P + \alpha P(\lambda^*)\{Q_\lambda - Q_{\lambda^*}\}) - l(P)$. Now, consider a first-order approximation of this difference: $\alpha P(\lambda^*)\{D_P(\lambda) - D_P(\lambda^*)\}$. Clearly, this is maximized if $D_P(\lambda)$ is maximized in λ and $D_P(\lambda^*)$ is minimized in the support of P. Choosing an optimal or monotonic step length completes the VEM. (For details or different methods, see Böhning (1989, 1995) or Lesperance and Kalbfleisch (1992).)

For the practical realization in C.A.MAN, we recall that the goal is to maximize $l(P)$ in the simplex Ω of all probability distributions P on parameter space λ. We call the solution of this problem the fully iterated nonparametric maximum likelihood estimator, and this solution is achieved in C.A.MAN in two phases (in which phase II is new). In phase I, an approximating grid $\lambda_1, \ldots, \lambda_L$ ($L \leq 50$) is chosen and $l(P)$ is maximized in the simplex Ω_{GRID} of all probability distributions P on grid $\{\lambda_1, \ldots, \lambda_L\}$ with one of the algorithms described above. For example, we can choose the observed data values as an approximating grid. In the introductory example, $L = 21$ different values of sold number of packages were observed: $\{\lambda_1, \ldots, \lambda_L\} = \{0, 1, 2, \ldots, 20\}$. As potential choice of initial weights, the observed relative frequencies or uniform weights $p_i = 1/21$ could be used.

In phase II, all grid points that are left with positive weights as a result of the optimization process in phase I are used as initial values for the EM algorithm (Dempster, Laird, and Rubin, 1977) to produce the fully iterated NPMLE.

Introductory Example Continued (Marketing Research)

In the following, some of the results that can be achieved with C.A.MAN are described. If in phase I all the observed data points are used as an approximating grid, seven points with positive support are identified, namely 0, 1, 3, 4, 7, 8, and 13. In phase II, these seven parameter values are used as initial values of the EM algorithm (along with the associated weights iterated in phase I). From these initial values, five components are estimated using the EM iteration (after collapsing equal components). Figure 3 shows that the iterated solution is indeed the NPMLE.

3. Meta-Analysis

Meta-analysis can be defined as the quantitative analysis of a variety of single-study results with the intention of an integrative presentation. Often in epidemiology or clinical trials, we have as a measure of interest the odds ratio Ψ or, equivalently, the log(odds ratio) $\lambda = \log(\Psi)$. Then the following situation forms the basis for any meta-analysis. We have n independent studies (cohort, case–control) with estimates $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$, from which a pooled estimate $\hat{\lambda}_{\text{pool}} = w_1\hat{\lambda}_1 + \cdots + w_n\hat{\lambda}_n$ is computed. The weights w_j are frequently chosen proportional to $1/\text{var}(\hat{\lambda}_j)$. There exists an extensive debate on the pros and cons of meta-analysis (see the review article of Dickersin and Berlin, 1992). Besides many arguments in favour of meta-analysis, most importanly it seems that it is becoming more and more part of the scientific method to provide evidence in favour or against a certain hypothesis or argument.

Example 1. As one example, consider the meta-analysis provided by Sillero-Arenas et al. (1992) on the relationship of hormone replacement therapy and the occurrence of breast cancer. Figure 4 presents the effect estimates with pointwise 95% confidence intervals for 36 studies (case–control and cohort). One question of debate in meta-analyis is whether individual study estimates of effect can be validly pooled into a common estimate of effect. This is conveniently put as the question of homogeneity or heterogeneity of study results. Homogeneity is conventionally investigated by diagnostic tests such as the χ^2-test of homogeneity. If there is evidence for heterogeneity, then the problem remains on how to proceed. The mixture approach provides an elegant solution for this problem in that it models the heterogeneity distribution in a nonparametric way.

The underlying assumption in most forms of meta-analysis—expressed also graphically in Figure 4—is that of a normal distribution for the effect estimate $\hat{\lambda}_i \sim \text{N}(\lambda_i, \sigma_i^2)$, with $\sigma_i^2 = \text{var}(\hat{\lambda}_i)$. Note that it is important to allow for different variances because the samples sizes will differ from study to study. In the simplest case of homogeneity ($\lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda$), the MLE of λ corresponds to the pooled estimator. If the population is heterogeneous, we must assume the existence of subpopulations with parameter λ_j receiving weight p_j for the jth subpopulation. Consequently, the density of $\hat{\lambda}_i$ corresponding to (1) is

$$\sum_{j=1}^{k} f(x, \lambda_j)p_j = \frac{1}{\sigma_i}\sum_{j=1}^{k} \phi((x - \lambda_j)/\sigma_i)p_j. \tag{2}$$

Here, ϕ is the standard normal density. Note also that k is not assumed to be known.

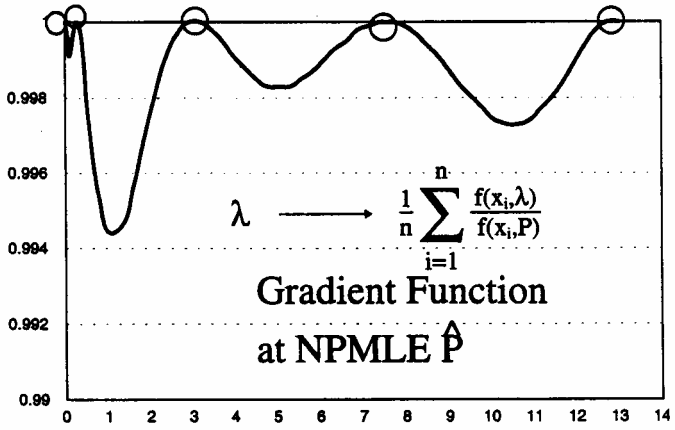

Figure 3. Gradient function at fully iterated NPMLE for introductory example.

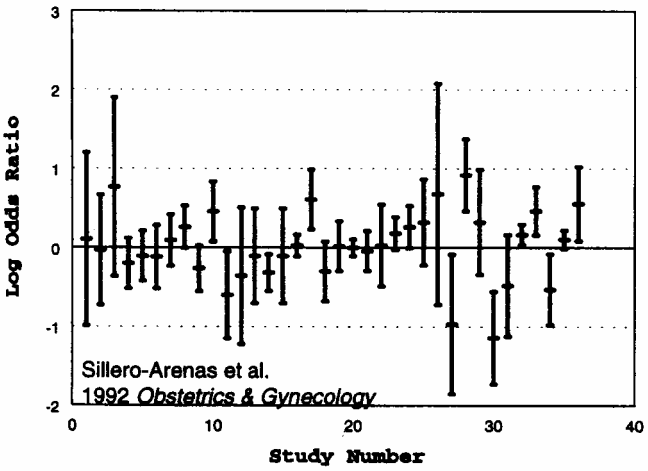

Figure 4. Effect estimates of 36 studies on relationship of hormone replacement therapy and breast cancer.

Example 1 (*continued*). We demonstrate the application of C.A.MAN to meta-analysis with the data of example 1. In phase I (with an approximating grid consisting of the observed study values), two components are found to receive positive weight, namely $\lambda_1 = -0.008333$ ($p_1 = 0.2893$) and $\lambda_2 = 0.83334$ ($p_2 = 0.7107$). This approximating solution is improved in phase II, leading to just one support point $\lambda = 0.046702$ (weight 1), leading to an odds ratio of $\Psi = e^{\lambda} = 1.0478$. Here the NPMLE coincides with the conventional pooled estimator $\hat{\lambda}_{\text{pool}}$. We can conclude from this analysis that there is no evidence for unobserved heterogeneity. An example providing evidence for heterogeneity is a meta-analysis on oral contraception in relation to breast cancer (Malone et al., 1993).

4. Modelling Heterogeneity in Fecundability Studies

In fecundability studies (see Ridout and Morgan, 1991), the situation is as follows. If X represents the cycle number in which pregnancy is reached, then X follows the geometric distribution with density $f(x, \lambda) = (1-\lambda)^{x-1}\lambda$, for $x = 1, 2, 3, \ldots$. The parameter λ is called the fertility parameter. In these studies, we have to cope with the problem of censoring, that is, no pregnancy occurs during the study period. If x denotes the last observed cycle, we ask for the probability that pregnancy occurs in some later cycle, that is,

$$\Pr\{X > x\} = \sum_{y=x+1}^{\infty} (1-\lambda)^{y-1}\lambda = \lambda(1-\lambda)^x \sum_{y=0}^{\infty}(1-\lambda)^y = \lambda(1-\lambda)^x \frac{1}{1-(1-\lambda)} = (1-\lambda)^x.$$

Therefore, $f(x, z, \lambda) = (1-\lambda)^x$ if x is censored ($z = 1$) or $f(x, z, \lambda) = (1-\lambda)^{x-1}\lambda$ if x is not censored ($z = 0$). In other words,

$$f(x, z, \lambda) = (1-\lambda)^{xz}(1-\lambda)^{(x-1)(1-z)}\lambda^{(1-z)}. \tag{3}$$

To demonstrate the modelling we look at a data set originally discussed by Weinberg and Gladen (1986) and later by Ridout and Morgan (1991). (See Table 1, part a.) Of the total of 486 couples, 12 remain unpregnant at the end of the study period. The MLE is, in this case, $1/\hat{\lambda} = 3.455$, with a log-likelihood of -1336.26.

If we allow for heterogeneity corresponding to (1), we achieve the nonparametric geometric mixture

$$f(x, z, P) = \sum_{j=1}^{k} f(x, z, \lambda_j)p_j. \tag{4}$$

Fitting model (4) with C.A.MAN provides a two-component structure for heterogeneity ($\hat{k} = 2$), namely $\lambda_1 = 1$ (receiving weight 0.0418) and $\lambda_2 = 3.6324$ (receiving weight 0.9582). Basically, the result is that we have a homogeneous population of couples (the likehood-ratio test is borderline

Table 1
Observed cycles to pregnancy

(a) Data (nonsmokers) according to Weinberg and Gladen (1986)													
Cycle	1	2	3	4	5	6	7	8	9	10	11	12	>12
No. of pregnancies	198	107	55	38	18	22	7	9	5	3	6	6	12
(b) Data (contraceptive pill users) according to Harlap and Baras (1984)													
Cycle	1	2	3	4	5	6	7	8	9	10	11	12	>12
No. of pregnancies	383	267	209	86	49	122	23	30	14	11	2	43	35

if the NPMLE is compared to the homogenous MLE). However, there is some inflation of couples with success in the first cycle (4%). A second group of 1274 couples is presented in Table 1, part b). They have in common that the women were using the contraceptive pill before trying to become pregnant. It turns out that this population is more heterogenous than the one considered before. The homogeneity MLE is $1/\hat{\lambda} = 3.5$, with a log-likelihood of -2635.939. Analyzing potential heterogeneity with C.A.MAN suggests that the population can be partitioned into two equally sized subpopulations with parameters $1/\hat{\lambda}_1 = 2.4$ and $1/\hat{\lambda}_2 = 4.8$. In addition, the difference in the twc likelihoods of about 10 supports the notion that we are dealing with a heterogenous population (for details on the likelihood ratio test in mixture models, see Böhning et al. (1994), McLachlan (1992), Feng and McCulloch (1996)).

5. Estimation of a Prevalence Rate

In prevalence studies, the parameter of interest is usually the prevalence rate λ, a number between 0 and 1 that, if multiplied with 100, can be interpreted as the percentage of infected humans or animals. The prevalence rate is again usually determined by choosing a sample size N from which the number of infected x out of N is counted, leading to an estimate $\hat{\lambda} = x/N$ for the prevalence rate. It follows from the conventional formulas that the variance of this estimate $\hat{\lambda}$ is given by $\text{var}(\hat{\lambda}) = \lambda(1-\lambda)/N$, which again can be estimated by $x/N^2(1-x/N)$.

In the veterinary sciences, as an example, however, this procedure is often not completely adequate because the animal population occurs in herds. The effect of this is called the clustering effect. Sampling takes this into account by sampling from m herds or farms with potentially different sample sizes N_i, $i = 1, \ldots, m$. The number of infected animals is denoted by x_i for herd or farm $i = 1, \ldots, m$.

It is common practice in epidemiology to use the pooled estimator $\hat{\lambda}_{\text{pool}} = (x_1 + \cdots + x_m)/(N_1 + \cdots + N_m)$ as an estimate of the common prevalence rate λ. The variance of $\hat{\lambda}_{\text{pool}}$ is readily provided as $\text{var}(\hat{\lambda}_{\text{pool}}) = ((1-\lambda)\lambda N_1 + \cdots + (1-\lambda)\lambda N_m)(N_1 + \cdots + N_m)^2 = (1-\lambda)\lambda/N$, with $N = N_1 + \cdots + N_m$. Note that this variance is the identical formula for the variance as in the unstratified sampling.

Problems occur if the clustering effect cannot be ignored, that is, if a common prevalence rate cannot be assumed. If instead a heterogeneous herd population with possible different prevalence parameters is more likely to be the case, then it can be shown that the variance of $\hat{\lambda}_{\text{pool}}$ is inflated by a term corresponding to the variance of the population prevalence rate p (Böhning and Sarol, unpublished manuscript). This variance is denoted by τ^2. In formula form,

$$\text{var}(\hat{\lambda}_{\text{pool}}) = \lambda(1-\lambda)/N + \tau^2[N_1(N_1-1) + \cdots + N_m(N_m-1)]/N^2.$$

Here, λ is the overall prevalence rate (the mean of the population prevalence rates) and τ^2 the variance of the population prevalence rate (λ is the expected value and τ^2 is the variance with respect to P). The above formula demonstrates clearly that, if population heterogeneity is ignored, the variance of the prevalence estimator is underestimated by the term $\tau^2(N_1^2 + \cdots + N_m^2)/N^2$. Also, if there is population homogeneity ($\tau^2 = 0$), both approaches and formulas coincide. To demonstrate these ideas, we look at a data set on herd infection with trypanosomosis discussed in Böhning and Greiner (unpublished manuscript). The data were collected in Mukono County, which is located in the southeastern part of Uganda and covers an area of approximately 200 km^2. The data (listed in Table 2) stem from a cross-sectional pilot study launched and accomplished in June/July 1994 for a project on trypanocide resistance in the peri-urban dairy production near

Table 2

Number of cattle infected with Trypanosoma spp., sample size and infection rate for 50 *dairy farms in Mukono County, Uganda (data from June* 1994, *total sample size* 487)

Farm	Cases	Sample size	Infection rate	Farm	Cases	Sample size	Infection rate
1	4	9	0.44	26	1	7	0.14
2	0	5	0.00	27	1	3	0.33
3	3	9	0.33	28	1	11	0.9
4	14	32	0.44	29	1	3	0.33
5	2	17	0.12	30	1	3	0.33
6	0	3	0.00	31	1	9	0.11
7	1	4	0.25	32	4	9	0.44
8	3	17	0.18	33	0	9	0.00
9	0	7	0.00	34	0	7	0.00
10	0	15	0.00	35	3	19	0.16
11	0	8	0.00	36	1	13	0.8
12	0	12	0.00	37	0	12	0.00
13	0	9	0.00	38	5	18	0.28
14	0	16	0.00	39	2	11	0.18
15	6	16	0.38	40	0	12	0.00
16	2	5	0.40	41	0	2	0.00
17	0	9	0.00	42	2	7	0.29
18	0	6	0.00	43	2	7	0.29
19	2	8	0.25	44	4	10	0.40
20	0	6	0.00	45	3	10	0.30
21	0	3	0.00	46	1	3	0.33
22	1	7	0.14	47	1	15	0.7
23	1	8	0.13	48	0	6	0.00
24	0	10	0.00	49	1	6	0.17
25	12	28	0.43	50	1	6	0.17

Kampala. The sampling frame consisted of 187 dairy farms existing in the region (information from a census taken in April 1994) from which 50 farms were selected at random using random number tables, stratified for 3 categories of herd size: small (1–10 cattle), medium (11–30), and large (more than 30). A total of 487 cattle was sampled on the identified farms.

Frequently, population heterogeneity is so striking that simple graphical methods are already successful, as in this case. It is evident from a histogram of the rate data that present in the distribution is not one center but about three, one at 0, the second at 0.15, and the third at 0.35. This indicates the presence of population heterogeneity due to clustering. Estimation of heterogeneity can be done again with C.A.MAN. The mixture model corresponding to (1) is

$$f(x_i, N_i, P) = \sum_{j=1}^{k} f(x_i, N_i, \lambda_j) p_j, \tag{5}$$

with the binomial mixture kernel $f(x, N, \lambda) = \binom{N}{x}\lambda^x(1-\lambda)^{N-x}$. The heterogeneity analysis with C.A.MAN results in three subpopulations: 17% of the herds are infection free, 48% have an infection rate of 12%, and 36% of the herds show an infection rate of 32%. From this heterogeneity distribution, mean and variances can easily be calculated, leading to

$$\hat{\lambda} = \hat{\lambda}_1\hat{p}_1 + \hat{\lambda}_2\hat{p}_2 + \hat{\lambda}_3\hat{p}_3 \quad \text{and} \quad \hat{\tau}^2 = \hat{p}_1(\hat{\lambda}_1 - \hat{\lambda}^2) + \hat{p}_2(\hat{\lambda}_2 - \hat{\lambda}^2) + \hat{p}_3(\hat{\lambda}_3 - \hat{\lambda}^2).$$

It turns out for this data set that incorporating the heterogeneity leads to a variance for the pooled estimator $\hat{\lambda}_{\text{pool}} = (x_1 + \cdots + x_k)/(N_1 + \cdots + N_k)$, about twice as large as the variance in the case where homogeneity is assumed. The difference between the approaches is visualized in Figure 5. Note that the left confidence interval uses the variance formula $\text{var}(\hat{\lambda}_{\text{pool}}) = \lambda(1-\lambda)/N$, whereas the right uses $\text{var}(\hat{\lambda}_{\text{pool}}) = \lambda(1-\lambda)/N + \tau^2[N_1(N_1-1) + \cdots + N_m(N_m-1)]/N^2$.

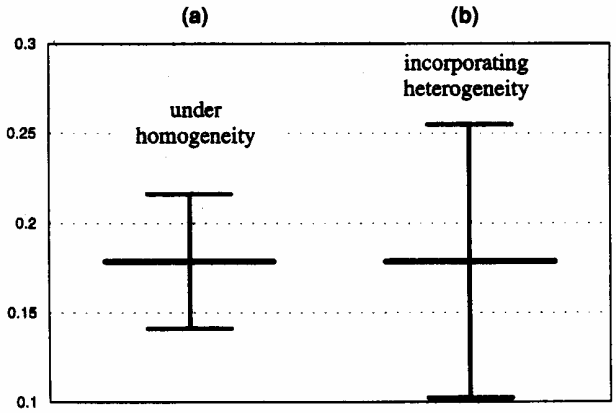

Figure 5. Estimation of prevalence rate with 95% CI (**a**) under assumption of homogeneity and (**b**) incorporating heterogeneity.

6. Interval-Censored Data

In this kind of mixture application, the mixture does not come in as an unobserved heterogeneity distribution; it occurs as a specific side condition. To demonstrate the details, let T be the time until a certain event occurs and $\Pr(T \leq t) = F(t)$ its distribution function. T is allowed to be interval censored, e.g., $T \in (L, R]$; that is, it is only known that the event has occurred between time L and time R. This situation occurs, for example, in repeated testing for occult events (as in tumor genesis). The contribution of the ith interval $(L_i, R_i]$ to the likelihood is $\Pr\{T_i \in (L_i, R_i]\} = \Pr\{L_i < T_i \leq R_i\} = F(R_i) - F(L_i)$. Let $s_0, \ldots, s_m$ be the uniquely ordered different values of $\{L_1, \ldots, L_n, R_1, \ldots, R_n\}$. The contribution to the likelihood of any interval, for example $(s_4, s_7]$, can be written uniquely as the sum of all contributions to the likelihood of neigbouring intervals. To demonstrate, $F(R_i) - F(L_i) = F(s_7) - F(s_4) = 0 \times (F(s_8) - F(s_7)) + 1 \times (F(s_7) - F(s_6)) + 1 \times (F(s_6) - F(s_5)) + 1 \times (F(s_5) - F(s_4)) + 0 \times (F(s_4) - F(s_3)) + 0 \times (F(s_3) - F(s_2)) + 0 \times (F(s_2) - F(s_1)) + 0 \times (F(s_1) - F(s_0))$, with $F(s_0) = 0$. In general, any interval $(L_i, R_i]$ can be written in the form

$$F(R_i) - F(L_i) = \sum_{j=1}^{m} \alpha_{ij}[F(s_j) - F(s_{j-1})],$$

with $F(s_0) = 0$ and

$$\alpha_{ij} = \begin{cases} 1 & \text{if } (s_{j-1}, s_j] \subseteq (L_i, R_i] \\ 0 & \text{otherwise.} \end{cases}$$

This leads to the following full likelihood:

$$\prod_{i=1}^{n} [F(R_i) - F(L_i)] = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij}[F(s_j) - F(s_{j-1})] = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} p_j$$

or log-likelihood $l(p) = \Sigma_{i=1}^{n} \log(\Sigma_{j=1}^{m} \alpha_{ij} p_j)$, where $p_j = F(s_j) - F(s_{j-1})$. Note that $p_j \geq 0$ for all $j = 1, \ldots, m$ and $p_1 + \cdots + p_m = 1$. Thus, the NPMLE $\hat{p}$ is maximizing $l(p) = \Sigma_{i=1}^{n} \log(\Sigma_{j=1}^{m} \alpha_{ij} p_j)$ under the restrictions $p_j \geq 0$ for all $j = 1, \ldots, m$ and $p_1 + p_2 + \cdots + p_m = 1$. This likelihood is easily identified as a mixture likelihood, though here mixing is not on densities but on indicator functions. (For details, see Böhning, Schlattmann, and Dietz (1996) and Gentleman and Geyer (1994).)

Final example. To give a final example, suppose $n = 6$ intervals have been observed $(L_i, R_i]$: (0,1], (1,3], (1,3], (0,2], (0,2], (2,3]. The different observation times are $m + 1 = 4, s_j : 0, 1, 2, 3$. Thus, we have $m = 3$ neighboring intervals $(s_j, s_{j+1}]$: $(0, 1], (1, 2], (2, 3]$. The matrix of indicator

values is

$$A = (\alpha_{ij}) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

leading to the likelihood

$$\begin{aligned} & 1p_1 + 0p_2 + 0p_3 && p_1 \\ \times\ & 0p_1 + 1p_2 + 1p_3 && p_2 + p_3 \\ \times\ & 0p_1 + 1p_2 + 1p_3 && p_2 + p_3 \\ \times\ & 1p_1 + 1p_2 + 0p_3 && p_1 + p_2 \\ \times\ & 1p_1 + 1p_2 + 0p_3 && p_1 + p_2 \\ \times\ & 0p_1 + 0p_2 + 1p_3 && p_3 \\ & = p_1(p_2 + p_3)^2(p_1 + p_2)^2 p_3 \end{aligned}$$

C.A.MAN finds the NPMLE in this case to be $p_1 = 1/3, p_2 = 1/3, p_3 = 1/3$.

Discussion

We have demonstrated the use of mixture models in various application areas and pointed out various developments in the area of computer-assisted analysis of mixtures.

Meta-analysis and heterogeneity. The problem of heterogeneity is frequently debated in the area of meta-analysis. Conventionally, a χ^2-test for heterogeneity is suggested. However, it remains frequently unclear how to proceed if this test is significant (Dickersin and Berlin, 1992). One approach, suggested by DerSimonian and Laird (1986), corrects the weights in the pooling process using variances increased by the amount of estimated variance due to population heterogeneity. (See also Biggerstaff and Tweedie (1997) for a review.) However, it does not provide an estimate of the heterogeneity distribution itself. The mixture approach offers a constructive solution in that an estimation of heterogeneity is provided. It is also possible to classify studies into the various subpopulations using the maximum posterior distribution as a classification rule.

Disease mapping. Disease mapping can be defined as a method for displaying the spatial distribution of disease occurrence, the most prominent forms being the variety of existing cancer atlases. A conventional biometric method used frequently to construct disease atlases (for example, see Cartwright et al., 1990) is based on the SMR (standardized mortality ratio) $= O/E$, where O is the observed number of death cases and E is the expected number of death cases for a given region. E is computed on the basis of an external standard population. Then it is conventionally assumed that, in area i, the observed number of deaths O_i follows a Poisson distribution with parameter λE_i: $f(x_i, \lambda) = \mathrm{Po}(o_i, \lambda E_i) = \exp(-\lambda E_i)(\lambda E_i)^{o_i}/o_i!$. Here, $x_i = o_i/E_i$ is the observed SMR, whereas λ is the theoretical SMR. The conventional display is based on a classification using the P-value under the homogeneous Poisson distribution, where λ is either set to 1 (no increased risk) or replaced by the MLE under homogeneity

$$\hat{\lambda} = \sum_{i=1}^{n} O_i \Big/ \sum_{i=1}^{n} E_i.$$

Another classical method uses the percentiles as a basis for classification. These conventional methods have been criticized by various authors, including Schlattmann and Böhning (1993), because they can lead to various artifacts in the disease map. Also, crude rate estimators have been criticized for some time in connection with disease mapping for their lack of stability (see Clayton and Kaldor, 1987). Stabilized estimators, usually in the context of empirical Bayes estimation, have been proposed and used. However, the use of the nonparametric estimator of the underlying heterogeneity as the basis for the construction of the map appears to be of recent novelty. In recent years, a WINDOWS program by the name DISMAP has been developed out of C.A.MAN (Schlattmann and Böhning, 1993), solely for the purpose of disease mapping. A detailed introduction to disease mapping based on mixtures can be found in Schlattmann, Dietz, and Böhning (1996).

Likelihood ratio test and number of components. Although the nonparametric estimation of the heterogeneity distribution provides an estimate of the number of components itself, it is sometimes

requested to use the likelihood ratio test for testing whether a reduced number of components is likewise sufficient. It is well known (Titterington, Smith, and Makov, 1985; McLachlan and Basford, 1988) that conventional asymptotic results for the null distribution of the likelihood ratio statistic do not hold since the null hypothesis lies on the boundary of the alternative hypothesis. In some cases, theoretical results are available (Böhning et al., 1994), but in other cases, simulation results must be used. In general, a parametric bootstrap procedure can be used (McLachlan, 1992), and it was pointed out recently that this approach leads to valid statistical inference (Feng and McCulloch, 1996).

Interval censoring. In this contribution, emphasis was put on direct applications of mixture modelling, in which the mixture distribution arises as the natural model for (latent) population heterogeneity. The problem of finding the nonparametric maximum likelihood estimate for the distribution function of a survival time under interval censoring is an example of indirect application of mixture modelling, where mixing is on indicator variables instead of densities.

Covariates. Currently, there is no option for handling covariates in C.A.MAN. Although some variables considered here are inherently adjusted for covariates (such as the standardized mortality ratio), analysis of additional covariates is often desirable. Mixture modelling with covariates leads to the area of mixed generalized linear models. One of the authors has developed a variety of macros in GLIM that allow the fitting of mixed generalized linear models when the number of components is fixed in advance (see Dietz, 1992; Dietz and Böhning, 1996). If the number of components is estimated itself, a class of nonparametric mixed generalized linear models will emerge, which we will consider in forthcoming work.

Availability. The package C.A.MAN is available from the authors free of charge. It may be down loaded from the website www.medizin.fu-berlin/sozmed/caman.html.

Acknowledgements

This paper is an extended version of an invited paper that was presented at *Social Science & Statistics: A Conference in Honor of the late Clifford C. Clogg*, held at the Pennsylvania State University, September 26–28, 1996. This version of the paper was prepared for an invited visit to the Department of Medical Statistics at the University of Freiburg, Germany. This research was done with the support of the German Research Foundation. The work of Dankmar Böhning and Peter Schlattmann had the additional support of a BIOMED2 grant on disease mapping and risk assessment.

Résumé

Cet article passe en revue des développements récents dan le domaine de l'analyse assistée par ordinateur de mélanges de distributions (C.A. MAN). Etant donnée une situation biométrique d'intérêt dans laquelle, sous des suppositions d'homogénéité, une densité paramétrique donnée est constatée telles la Poisson, la binomiale, la géométrique, la normale, ... alors ils est avancé que cette situation peut être aisément généralisée de façon à prendre en compte une variation du paramètre d'échelle dans la population. Cette situation est appelée "hétérogénéité non observée." Ceci conduit naturellement à une forme spécifique de la distribution de mélange non paramétrique qui peut être considérée comme étant le modèle standard de l'application biométrique d'intérêt (dès lors qu'il inclus aussi le cas d'homogénéité comme un cas particulier). En plus des développements théoriques et algorithmiques le travail met en avant des développements d'applications biométriques, comme par exemple les meta-analyses, les études de fertilité, les estimations de prévalence sous clusters, et l'estimation de la distribution de fonction de survie avec censure par intervalle. L'approche est non paramétrique pour la distribution mixte, allant jusqu'à laisser le nombre de composants (sous populations) de la distribution mixte non spécifié.

References

Bagozzi, R. P. (1995). *Advanced Methods of Marketing Research.* Cambridge, Massachusetts: Blackwell.

Biggerstaff, B. J. and Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* **16,** 753–768.

Böhning, D. (1982). Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Annals of Statistics* **10,** 1006–1008.

Böhning, D. (1989). Likelihood inference for mixtures: Geometrical and other constructions of monotone step-length algorithms. *Biometrika* **76,** 375–383.

Böhning, D. (1994). A note on test for Poisson overdispersion. *Biometrika* **81,** 418–419.

Böhning, D. (1995). A review of reliable maximum likelihood algorithms for the semi-parametric mixture maximum likelihood estimator. *Journal of Statistical Planning and Inference* **47,** 5–28.

Böhning, D., Schlattmann, P., and Lindsay, B. G. (1992). Computer assisted analysis of mixtures (C.A.MAN): Statistical algorithms. *Biometrics* **48,** 283–303.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parametric exponential family. *Annals of the Institute of Statistical Mathematics* **46,** 373–388.

Böhning, D., Schlattmann, P., and Dietz, E. (1996). Interval censored data: A note on the nonparametric maximum likelihood estimator of the distribution function. *Biometrika* **83,** 462–466.

Cartwright, R. A., Alexander, F. E., McKinney, P. A., and Ricketts, T. J. (1990). *Leukaemia and Lymphoma. An Atlas of Distribution within Areas of England and Wales 1984–1988.* London: Leukaemia Research Fund.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates for age-standardized relative risks. *Biometrics* **43,** 671–681.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

DerSimonian, R. and Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7,** 177–188.

Dickersin, K. and Berlin, J. A. (1992). Meta-analysis: State-of-the-science. *Epidemiologic Reviews* **14,** 154–176.

Dietz, E. (1992). Estimation of heterogeneity—A GLM approach. In *Advances in GLIM and Statistical Modeling, Lecture Notes in Statistics,* L. Fahrmeir, F. Francis, R. Gilchrist, and G. Tutz (eds), 66–72. Berlin: Springer Verlag.

Dietz, E. and Böhning, D. (1996). Statistical inference based on a general model of unobserved heterogeneity. In *Advances in GLIM and Statistical Modeling, Lecture Notes in Statistics,* L. Fahrmeir, F. Francis, R. Gilchrist, and G. Tutz (eds), 75–82. Berlin: Springer Verlag.

Feng, Z. D. and McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society, Series B* **58,** 609–617.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika* **81,** 618–623.

Harlap, S. and Baras, H. (1984). Conception-waits in fertile women after stopping oral contraceptives. *International Journal of Fertility* **29,** 73–80.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73,** 805–811.

Lesperance, M. and Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association* **87,** 120–126.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods, part I: A general theory. *Annals of Statistics* **11,** 783–792.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods, part II: The exponential family. *Annals of Statistics* **11,** 86–94.

Malone, K.E., Daling, J. R., and Weiss, N. S. (1993). Oral contraceptives in relation to breast cancer. *Epidemiologic Reviews* **15,** 80–97.

McLachlan, G. J. (1992). Cluster analysis and related techniques in medical research. *Statistical Methods in Medical Research* **1,** 27–49.

McLachlan, G. F. and Basford, K. E. (1988). *Mixture Models. Inference and Applications to Clustering.* New York: Marcel Dekker.

Ridout, M. S. and Morgan, B. J. T. (1991). Modelling digit preference in fecundability studies. *Biometrics* **47,** 1423–1433.

Schlattmann, P. and Böhning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine* **12,** 943–50.

Schlattmann, P., Dietz, E., and Böhning, D. (1996). Covariate adjusted mixture models with the program DismapWin. *Statistics in Medicine* **15,** 919–929.

Sillero-Arenas, M., Delgado-Rodriguez, M., Rodigues-Canteras, R., Bueno-Cavanillas, A., and Galvez-Vargas, R. (1992). Menopausal hormone repacement therapy and breast cancer: A meta-analysis. *Obstetrics and Gynecology* **79,** 286–294.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

Weinberg, C. R. and Gladen, B. C. (1986). The beta-geometric distribution applied to comparative fecundability studies. *Biometrics* **42,** 547–560.

Received April 1997; *revised July* 1997; *accepted August* 1997.