Stat Comput (2010) 20: 75–86
DOI 10.1007/s11222-009-9117-z

Maximum likelihood computation for fitting semiparametric mixture models

Yong Wang

Received: 12 April 2008 / Accepted: 25 February 2009 / Published online: 14 March 2009

Abstract Three general algorithms that use different strategies are proposed for computing the maximum likelihood estimate of a semiparametric mixture model. They seek to maximize the likelihood function by, respectively, alternating the parameters, profiling the likelihood and modifying the support set. All three algorithms make a direct use of the recently proposed fast and stable constrained Newton method for computing the nonparametric maximum likelihood of a mixing distribution and employ additionally an optimization algorithm for unconstrained problems. The performance of the algorithms is numerically investigated and compared for solving the Neyman-Scott problem, overcoming overdispersion in logistic regression models and fitting two-level mixed effects logistic regression models. Satisfactory results have been obtained.

Keywords Constrained optimization · Maximum likelihood computation · Mixed effects · Neyman-Scott problem · Profile likelihood · Semiparametric mixture

1 Introduction

A semiparametric mixture model has two parameters of different types. One is a mixing distribution function, denoted G throughout the paper, that is completely unspecified, and the other a finite-dimensional structural parameter, denoted $\boldsymbol{\beta}$, that is common to all mixture component distributions. Semiparametric mixture models form a rich and flexible family, which entails nonparametric mixture models and conventional parametric models as its two extreme cases. Semiparametric mixture models have a wide range of applications and provide neat solutions to many nasty problems, e.g., the classical Neyman-Scott problems, overdispersion in generalized linear models, distributional misspecification in mixed effects models, and errors in covariates; see, e.g., Lindsay and Lesperance (1995), Aitkin (1999), Murphy and van der Vaart (2000) and the references therein.

The study of semiparametric mixture models was pioneered by Kiefer and Wolfowitz (1956), who successfully established the consistency of the semiparametric MLE (short for maximum likelihood estimator/estimate/estimation) of $(G, \boldsymbol{\beta})$ under suitable conditions. The computation of a semiparametric MLE, however, appeared to be so difficult that practical applications were only to be seriously investigated almost thirty years later. Some computational suggestions are given by, e.g., Heckman and Singer (1984), Follmann and Lambert (1989) and Aitkin (1999). Nevertheless, how to compute a semiparametric MLE rapidly and reliably still remains a challenge. No simulation study has been found in the literature that involves computing a large number of semiparametric MLEs.

The main obstacle in computing a semiparametric MLE is clearly the existence of the nonparametric distribution G. Finding the nonparametric MLE of G, for a mixture without $\boldsymbol{\beta}$ or with $\boldsymbol{\beta}$ fixed, has also been a challenge in the past. This subproblem, however, can now be solved efficiently by the constrained Newton method with multiple support point inclusion (CNM) that is recently proposed by Wang (2007). The availability of the fast and stable CNM method thus provides several potentials for rapid computation of a semiparametric MLE. In this paper, we study these potentials and propose three general algorithms for semiparametric MLE computation. Each of the proposed algorithms

Y. Wang (✉)
Department of Statistics, The University of Auckland,
Private Bag 92019, Auckland 1142, New Zealand
e-mail: yongwang@stat.auckland.ac.nz

makes a direct use of the CNM method and requires additionally only an unconstrained optimization algorithm, e.g., a quasi-Newton method.

The remainder of the paper is organized as follows. In Sect. 2, we briefly introduce the problem of maximum likelihood estimation of semiparametric mixture models, along with some notation. The CNM method for nonparametric MLE computation is reviewed in Sect. 3. Section 4 presents the three algorithms, describes how the BFGS quasi-Newton method is used internally in these algorithms and discusses a few practical issues. Section 5 gives the results of three numerical studies, which compares the performance of the three algorithms in different scenarios. Summary and final remarks are given in Sect. 6.

2 Semiparametric maximum likelihood estimation

In this section, we briefly describe the problem of maximum likelihood estimation of a semiparametric mixture model. The notation used here is a slight extension to that in Wang (2007).

Consider k independent multivariate samples $\mathbf{y}_1, \ldots, \mathbf{y}_k$, where each $\mathbf{y}_i \equiv (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}} \in \mathcal{Y}_i \subset \mathbb{R}^{n_i}$ is drawn from a distribution with density $f(\mathbf{y}_i|\mathbf{x}_i; \theta, \boldsymbol{\beta})$, $\theta \in \Omega \subset \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^r$. Here $\mathbf{x}_i$ is used to denote the vector of the covariates or even vectorized matrices as in the mixed effects models. The space of $\boldsymbol{\beta}$ may be relaxed to a convex subset of $\mathbb{R}^r$, in which every maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ is an interior point. Let the parameter θ be random (or treated as random) and have some unknown distribution G. Then $\mathbf{y}_i$ has the marginal distribution with density

$$f(\mathbf{y}_i|\mathbf{x}_i; G, \boldsymbol{\beta}) = \int f(\mathbf{y}_i|\mathbf{x}_i; \theta, \boldsymbol{\beta})\, \mathrm{d}G(\theta), \tag{1}$$

and, given $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_k, \mathbf{y}_k)$, the log-likelihood function is

$$\ell(G, \boldsymbol{\beta}) = \sum_{i=1}^{k} \log \left\{ \int f(\mathbf{y}_i|\mathbf{x}_i; \theta, \boldsymbol{\beta})\, \mathrm{d}G(\theta) \right\}. \tag{2}$$

One is thus looking for the semiparametric MLE $(\widehat{G}, \hat{\boldsymbol{\beta}})$, which maximizes $\ell(G, \boldsymbol{\beta})$.

Since for every fixed $\boldsymbol{\beta}$ there must exist a discrete nonparametric MLE of G (Laird 1978; Lindsay 1983), it is sufficient to only consider discrete Gs for computing a semiparametric MLE. For a discrete G that is supported at the distinct points $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_m)^{\mathrm{T}}$ with corresponding probability masses $\boldsymbol{\pi} \equiv (\pi_1, \ldots, \pi_m)^{\mathrm{T}}$, we can also write density (1) as

$$f(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{j=1}^{m} \pi_j f(\mathbf{y}_i|\mathbf{x}_i; \theta_j, \boldsymbol{\beta})$$

and log-likelihood (2) as

$$\ell(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{i=1}^{k} \log \left\{ \sum_{j=1}^{m} \pi_j f(\mathbf{y}_i|\mathbf{x}_i; \theta_j, \boldsymbol{\beta}) \right\}. \tag{3}$$

Of critical importance for computing a nonparametric or semiparametric MLE is the gradient function, which is defined as

$$\begin{aligned} d(\theta; G, \boldsymbol{\beta}) &\equiv \left. \frac{\partial \ell\{(1-\epsilon)G + \epsilon\delta_\theta, \boldsymbol{\beta}\}}{\partial \epsilon} \right|_{\epsilon=0} \\ &= \sum_{i=1}^{k} \frac{f(\mathbf{y}_i|\mathbf{x}_i; \theta, \boldsymbol{\beta})}{f(\mathbf{y}_i|\mathbf{x}_i; G, \boldsymbol{\beta})} - k, \end{aligned} \tag{4}$$

where δ_θ denotes the degenerate distribution function at θ. For computing a semiparametric MLE, one may consider the profile log-likelihood function:

$$\tilde{\ell}(\boldsymbol{\beta}) \equiv \ell(\widehat{G}_{\boldsymbol{\beta}}, \boldsymbol{\beta}) \equiv \max_{G} \ell(G, \boldsymbol{\beta}). \tag{5}$$

For each fixed value of $\boldsymbol{\beta}$, $\tilde{\ell}(\boldsymbol{\beta})$ can be evaluated by computing the nonparametric MLE $\widehat{G}_{\boldsymbol{\beta}}$. This profile log-likelihood may however have multiple semiparametric local maxima; see Sect. 5.1 for an example. If the log-likelihood is bounded for all β and G, then each of these semiparametric local maxima $(G, \boldsymbol{\beta})$ must necessarily satisfy the following conditions:

$$\frac{\partial \ell(G, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}, \tag{6}$$

$$\sup_{\theta \in \Omega} \{d(\theta; G, \boldsymbol{\beta})\} = 0. \tag{7}$$

Any point that satisfies these conditions is guaranteed to be a semiparametric local maximum if it also satisfies the second-order sufficient condition:

$$\frac{\partial^2 \ell(G, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} \text{ is negative definite.}$$

These conditions hold simply owing to the fact that $\hat{\boldsymbol{\beta}}$ is an interior point and, for every fixed $\boldsymbol{\beta}$, $\ell(G, \boldsymbol{\beta})$ is maximized by a nonparametric MLE of G.

In most applications, y_{ij}'s can be assumed to be drawn independently from a distribution with density $f(y_{ij}; \boldsymbol{\beta}, \theta)$, i.e.,

$$f(\mathbf{y}_i|\mathbf{x}_i; \theta, \boldsymbol{\beta}) = \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{x}_{ij}; \theta, \boldsymbol{\beta}). \tag{8}$$

In fact, all numerical studies presented in Sect. 5 use this type of data. It is worth mentioning, however, that the proposed algorithms depend only on the knowledge of the joint density $f(\mathbf{y}_i|\mathbf{x}_i; \theta, \boldsymbol{\beta})$, not on the availability of factorization (8).

3 Computing the nonparametric MLE

This section describes briefly the CNM method for computing the nonparametric MLE (with $\boldsymbol{\beta}$ fixed), as well as a modified implementation of it. The reader is referred to Wang (2007) for detailed description and study of the algorithm. All three algorithms that are proposed in Sect. 4 for computing the semiparametric MLE need to use the CNM method, either as a whole piece or one iteration at a time.

The CNM method consists of two critical ingredients: updating the mixing proportions $\boldsymbol{\pi}$, and expanding and contracting the support set $\boldsymbol{\theta}$. In order to update $\boldsymbol{\pi}$ to $\boldsymbol{\pi}'$, it utilizes the second-order Taylor series expansion of the log-likelihood function in the neighborhood of $\boldsymbol{\pi}$. In the setting of semiparametric mixtures here, let us denote

$$\mathbf{s}_i(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\beta}) \equiv \frac{\partial \log\{f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\beta})\}}{\partial \boldsymbol{\pi}}, \quad i=1,\ldots,k,$$

$$\mathbf{S}^{\mathrm{T}} \equiv \mathbf{S}(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\beta})^{\mathrm{T}} \equiv (\mathbf{s}_1(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\beta}),\ldots,\mathbf{s}_k(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\beta})).$$

By noting that

$$\frac{\partial \ell(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\beta})}{\partial \boldsymbol{\pi}} = \mathbf{S}^{\mathrm{T}}\mathbf{1}, \tag{9}$$

$$\frac{\partial^2 \ell(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\beta})}{\partial \boldsymbol{\pi}\partial \boldsymbol{\pi}^{\mathrm{T}}} = -\mathbf{S}^{\mathrm{T}}\mathbf{S}, \tag{10}$$

and using the Taylor series expansion, we obtain the following quadratic approximation to $\ell(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\beta}) - \ell(\boldsymbol{\pi}',\boldsymbol{\theta},\boldsymbol{\beta})$:

$$\begin{aligned} Q(\boldsymbol{\pi}'|\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\beta}) &\equiv -\mathbf{1}^{\mathrm{T}}\mathbf{S}(\boldsymbol{\pi}'-\boldsymbol{\pi}) + \frac{1}{2}(\boldsymbol{\pi}'-\boldsymbol{\pi})^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{S}(\boldsymbol{\pi}'-\boldsymbol{\pi}) \\ &= \frac{1}{2}\|\mathbf{S}\boldsymbol{\pi}'-\mathbf{2}\| - \frac{k}{2}. \end{aligned}$$

With $\boldsymbol{\theta}$ (and $\boldsymbol{\beta}$) fixed, $\boldsymbol{\pi}$ can be efficiently updated by solving the least squares linear regression problem over a probability simplex:

$$\min_{\boldsymbol{\pi}'} \|\mathbf{S}\boldsymbol{\pi}'-\mathbf{2}\|^2, \quad \text{subject to } \boldsymbol{\pi}'^{\mathrm{T}}\mathbf{1}=1, \quad \boldsymbol{\pi}' \geq \mathbf{0}. \tag{11}$$

By including a backtracking line search to ensure monotone increase of the likelihood, we call it the constrained Newton method.

The expansion of the support point vector $\boldsymbol{\theta}$ can be efficiently achieved by including all local maxima of the gradient function (4), while its contraction is straightforward by discarding those with zero mass after $\boldsymbol{\pi}$ being updated. We hence obtain the following CNM method for computing the nonparametric MLE (with $\boldsymbol{\beta}$ fixed in the setting of semiparametric mixtures).

Algorithm 1 (CNM) Set $s=0$. From an initial estimate G_0 with finite support and $\ell(G_0,\boldsymbol{\beta}) > -\infty$, repeat the following steps.

Step 1: compute all local maxima $\theta^*_{s1},\ldots,\theta^*_{sp_s}$ of $d(\theta; G_s, \boldsymbol{\beta})$, $\theta \in \Omega$. If $\max_j\{d(\theta^*_{sj}, G_s, \boldsymbol{\beta})\} = 0$, stop.

Step 2: set $\boldsymbol{\theta}_s^+ = (\boldsymbol{\theta}_s^{\mathrm{T}}, \theta^*_{s1},\ldots,\theta^*_{sp_s})^{\mathrm{T}}$ and $\boldsymbol{\pi}_s^+ = (\boldsymbol{\pi}_s^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$. Find $\boldsymbol{\pi}^-_{s+1}$, by minimizing $Q(\boldsymbol{\pi}'|\boldsymbol{\pi}_s^+,\boldsymbol{\theta}_s^+,\boldsymbol{\beta})$ and conducting a line search.

Step 3: discard all support points with zero entries in $\boldsymbol{\pi}^-_{s+1}$, which gives $\boldsymbol{\theta}_{s+1}$ and $\boldsymbol{\pi}_{s+1}$ of G_{s+1}. Set $s = s+1$.

The CNM method is guaranteed to converge to a nonparametric MLE, as established by Theorem 1 in Wang (2007).

In the implementation described in Wang (2007), problem (11) is solved by using the method of Haskell and Hanson (1981), which turns the problem into one with non-negativity constraints only and hence can be solved by the NNLS algorithm of Lawson and Hanson (1974). Although this appears to be numerically stable and works reasonably well in practice, its solution is a finite compromise of a limit result and a proper value has to be chosen for a compromising parameter γ, as used in (12) of Wang (2007). The current implementation uses the method of Dax (1990), which also turns problem (11) into one with only non-negativity constraints and can thus be solved by the numerically stable NNLS algorithm. More appealingly, it gives an algebraically exact solution.

Dax's (1990) method can be described as follows. First, problem (11) is transformed equivalently to the problem:

$$\min_{\boldsymbol{\pi}'} \|\mathbf{P}\boldsymbol{\pi}'\|^2, \quad \text{subject to } \boldsymbol{\pi}'^{\mathrm{T}}\mathbf{1}=1, \quad \boldsymbol{\pi}' \geq \mathbf{0}, \tag{12}$$

where $\mathbf{P} \equiv (\mathbf{s}^{(1)}-\mathbf{2},\ldots,\mathbf{s}^{(m)}-\mathbf{2})$, $\mathbf{s}^{(j)}$ being the jth column of $\mathbf{S}$. Then, to solve a problem of form (12), Dax suggests solving the following least squares problem with only non-negativity constraints:

$$\min_{\tilde{\boldsymbol{\pi}}} \|\mathbf{P}\tilde{\boldsymbol{\pi}}\|^2 + |\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\mathbf{1}-1|^2, \quad \text{subject to } \tilde{\boldsymbol{\pi}} \geq \mathbf{0}. \tag{13}$$

He then establishes, by relating the Karush-Kuhn-Tucker conditions for both problems, that if $\tilde{\boldsymbol{\pi}}$ solves problem (13), then $\tilde{\boldsymbol{\pi}}/\tilde{\boldsymbol{\pi}}^{\mathrm{T}}\mathbf{1}$ solves problem (12), and thus problem (11).

It is worth pointing out that the CNM method makes use of the special relationship between the gradient vector and the Hessian matrix of the log-likelihood function of $\boldsymbol{\pi}$, as given in (9) and (10), from which the linear regression formulation (11) is available. This relationship does not hold for semiparametric mixture models. As a result, it can not be applied directly to this type of models. However, its ability of computing a nonparametric MLE rapidly provides several potentials for extension, as studied in the next section.

4 Computing the semiparametric MLE

In the first three subsections, we present the three general algorithms that maximize the likelihood by, respectively, al-

ternating the parameters, profiling the likelihood function and modifying the support set progressively. The use of the BFGS method in the implementation of these algorithms is described in Sect. 4.4 and a few practical issues are discussed in Sect. 4.5.

4.1 Maximization by alternating the parameters

A conventional wisdom for computing a semiparametric MLE is to alternate the computation between an updating of G with $\boldsymbol{\beta}$ fixed and an updating of $\boldsymbol{\beta}$ with G fixed (Böhning 1995; Lindsay 1995). This parameter-alternating method can be easily implemented, given an algorithm for computing a nonparametric MLE, such as CNM, and an algorithm for computing the unconstrained parameter $\boldsymbol{\beta}$, such as a quasi-Newton method (see Sect. 4.4.1). Since CNM converges rapidly, it is thus unnecessary to find the exact, optimal solution when updating G. Using only one iteration of CNM each time can be much more efficient at reducing the total computational cost. The resulting algorithm can be formally described as follows.

Algorithm 2 (CNM-AP) Set $s = 0$, and choose $\boldsymbol{\beta}_0 \in \mathbb{R}^r$ and G_0 with finite support such that $\ell(G_0, \boldsymbol{\beta}_0) > -\infty$. Repeat the following steps.

Step 1: update G_s to G_{s+1} with $\boldsymbol{\beta}_s$ fixed, by using one iteration of CNM.
Step 2: update $\boldsymbol{\beta}_s$ to $\boldsymbol{\beta}_{s+1}$, by maximizing $\ell(G_{s+1}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$.
Step 3: set $s = s + 1$. If converged, stop.

Despite an easy implementation, the performance of this parameter-alternating method is not always satisfactory and it depends particularly on the level of the correlation between G and $\boldsymbol{\beta}$. Here the correlation between G and $\boldsymbol{\beta}$ can be understood in a local sense, by, e.g., considering the Hessian matrix of the log-likelihood function of $(\boldsymbol{\pi}_{-m}, \boldsymbol{\theta}, \boldsymbol{\beta})$, where $\boldsymbol{\pi}_{-m}$ is the $\boldsymbol{\pi}$ without $\pi_m \equiv 1 - \sum_{j=1}^{m-1} \pi_j$. Generally speaking, larger values in magnitude in the block of the Hessian between $(\boldsymbol{\pi}_{-m}, \boldsymbol{\theta})$ and $\boldsymbol{\beta}$, as compared with the diagonal ones, imply higher correlation. In perhaps all practical problems, G and $\boldsymbol{\beta}$ are correlated, but the correlation may be weak in some problems, for example, the Neyman-Scott problem. In this situation, the parameter-alternating method can work very well, as demonstrated in Sect. 5.2. However, when the correlation is high, the parameter-alternating method can be very inefficient and unreliable, as demonstrated in Sect. 5.1. It is not unusual in our experience of fitting mixed effects models that this method fails, due to numerical difficulties, to produce a solution that is close to a point satisfying conditions (6) and (7). Powell (1973) even gave some examples in the general setting of unconstrained optimization, where the alternating method fails, arithmetically, to converge to a stationary point.

4.2 Maximization by profiling the likelihood

Our second algorithm aims to maximize the profile log-likelihood function $\tilde{\ell}(\boldsymbol{\beta})$, which can be evaluated at any fixed $\boldsymbol{\beta}$ by calling the CNM method. This reduces the original problem to an unconstrained optimization problem, which can be easily solved by many algorithms.

Algorithm 3 (CNM-PL) Set $s = 0$, and choose $\boldsymbol{\beta}_0 \in \mathbb{R}^r$ and G_0 with finite support such that $\ell(G_0, \boldsymbol{\beta}_0) > -\infty$. Repeat the following steps.

Step 1: update $\boldsymbol{\beta}_s$ to $\boldsymbol{\beta}_{s+1}$ by increasing $\tilde{\ell}(\boldsymbol{\beta}_s)$ with an unconstrained ascent method.
Step 2: set $s = s + 1$. If converged, stop.

Of many unconstrained optimization algorithms that can be used for updating $\boldsymbol{\beta}$, it is beneficial to use one that evaluates $\tilde{\ell}(\boldsymbol{\beta})$ as few times as possible, so as to reduce the number of calls to the CNM method. The BFGS algorithm has a superlinear order of convergence and is a strong candidate; Sect. 4.4.1 gives a detailed description of how it can be used to update $\boldsymbol{\beta}$. The derivative of the profile log-likelihood, as required by BFGS, can be computed easily, by the relationship

$$\frac{\partial \tilde{\ell}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left.\frac{\partial \ell(G, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{G=\widehat{G}_{\boldsymbol{\beta}}},$$

because of the optimality of $\widehat{G}_{\boldsymbol{\beta}}$. With a proper line search method, the global convergence to a stationary point of the profile log-likelihood function can be easily established, as routinely for an unconstrained optimization algorithm.

4.3 Maximization by modifying the support set

In each iteration, our third algorithm executes one iteration of the CNM method and then updates all parameters $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\beta})$ by using an unconstrained optimization method. Let G_s consist of $\boldsymbol{\pi}_s > \mathbf{0}$ and $\boldsymbol{\theta}_s$.

Algorithm 4 (CNM-MS) Set $s = 0$, and choose $\boldsymbol{\beta}_0 \in \mathbb{R}^r$ and G_0 with finite support such that $\ell(G_0, \boldsymbol{\beta}_0) > -\infty$. Repeat the following steps.

Step 1: update $(\boldsymbol{\pi}_s, \boldsymbol{\theta}_s)$ to $(\boldsymbol{\pi}_s^+, \boldsymbol{\theta}_s^+)$ with $\boldsymbol{\beta} = \boldsymbol{\beta}_s$ fixed, by using one iteration of CNM.
Step 2: update $(\boldsymbol{\pi}_s^+, \boldsymbol{\theta}_s^+, \boldsymbol{\beta}_s)$ to a local maximum $(\boldsymbol{\pi}_{s+1}, \boldsymbol{\theta}_{s+1}, \boldsymbol{\beta}_{s+1})$, by an unconstrained optimization method, as described below.
Step 3: set $s = s + 1$. If converged, stop.

The use of one iteration of CNM in step 1 of the algorithm ensures that, (a) the support set is constantly modified, with both an expansion step by including more support

points and a possible contraction step by discarding support points with zero masses; (b) G_s^+, consisting of $\boldsymbol{\pi}_s^+$ and $\boldsymbol{\theta}_s^+$, is close to being the nonparametric MLE $\widehat{G}_{\boldsymbol{\beta}_s}$ and thus should have a reasonably small support set; and (c) the tentative solution $(\boldsymbol{\pi}_s^+, \boldsymbol{\theta}_s^+, \boldsymbol{\beta}_s)$ increases the likelihood value from $(\boldsymbol{\pi}_s, \boldsymbol{\theta}_s, \boldsymbol{\beta}_s)$, while maintaining $\boldsymbol{\pi}_s^+ > \mathbf{0}$, thus interior to the corresponding probability simplex, with which an unconstrained optimization method can be used safely.

Specifically, the unconstrained optimization method in step 2 could operate as follows. First, to apply an unconstrained optimization method, the equality constraint $\boldsymbol{\pi}_s^{+\mathrm{T}}\mathbf{1} = 1$ needs to be removed, by replacing one element of $\boldsymbol{\pi}_s^+$ with its relationship with the others. Since $\boldsymbol{\pi}_s^+ > \mathbf{0}$ is always maintained, an unconstrained optimization method can always be applied. Second, since the direction and step length found by the unconstrained optimization method does not take into account the constrained probability simplex, a full step forward may move outside the boundary. If this happens, the step length is shortened, so the new solution locates exactly on the boundary of the simplex. Third, a line search, with details given in Sect. 4.4, is then conducted to ensure monotone and sufficient increase of the log-likelihood. Fourth, discard all zero-massed support points and reduce the parameter space accordingly. The above process is repeated until the solution has converged to a local maximum, which has $\boldsymbol{\pi}_{s+1} > \mathbf{0}$.

Note that if any zero-massed support point is removed, the new estimate $(\boldsymbol{\pi}_{s+1}, \boldsymbol{\theta}_{s+1}, \boldsymbol{\beta}_{s+1})$ will have a smaller dimension than $(\boldsymbol{\pi}_s^+, \boldsymbol{\theta}_s^+, \boldsymbol{\beta}_s)$. It is a local maximum in the new parameter space, but not necessarily in the original parameter space defined by $(\boldsymbol{\pi}_s^+, \boldsymbol{\theta}_s^+, \boldsymbol{\beta}_s)$. Nevertheless, since the likelihood has been kept increasing for every new iterate during the computation, the global convergence of the algorithm can be guaranteed.

Theorem 1 *Assume that $\ell(G, \boldsymbol{\beta})$ is bounded above for all $\boldsymbol{\beta}$ and G, and let $\{(G_s, \boldsymbol{\beta}_s)\}$ be any sequence produced by Algorithm 4. Then*

$$\lim_{s\to\infty} \left.\frac{\partial \ell(G_s, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_s} = \mathbf{0} \quad \textit{and}$$

$$\lim_{s\to\infty} \sup_{\theta\in\Omega}\{d(\theta; G_s, \boldsymbol{\beta}_s)\} = 0.$$

The first limit result holds obviously, owing to the optimization in step 2 of Algorithm 4. The second limit result can be established, which is omitted here, as an analogue to that of Theorem 1 in Wang (2007) for the convergence of the CNM method.

4.4 Using the BFGS method

The well-known quasi-Newton BFGS method for unconstrained optimization is used in our implementation of all three algorithms and is described below; see, e.g., Fletcher (1987). It is used to update $\boldsymbol{\beta}$ in CNM-AP and CNM-PL and to update $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\beta})$ in CNM-MS.

4.4.1 *Updating $\boldsymbol{\beta}$*

To update $\boldsymbol{\beta}$, a Newton-like method first finds an ascent direction from the point $(G, \boldsymbol{\beta})$ by

$$\boldsymbol{\eta} = -\mathbf{D}\mathbf{g}, \tag{14}$$

where $\mathbf{D}$ is a negative definite matrix and

$$\mathbf{g} \equiv \mathbf{g}(G, \boldsymbol{\beta}) \equiv \frac{\partial \ell(G, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

By backtracking, the log-likelihood can be guaranteed to satisfy the Armijo rule:

$$\ell(G, \boldsymbol{\beta} + \sigma^k \boldsymbol{\eta}) \geq \ell(G, \boldsymbol{\beta}) + \nu\sigma^k \mathbf{g}^{\mathrm{T}}\boldsymbol{\eta},$$

where $0 < \nu < \frac{1}{2}$, $0 < \sigma < 1$, and k is the least integer in $\{0, 1, 2, \ldots\}$ that satisfies the inequality. The global convergence to a stationary point is guaranteed, provided all eigenvalues of all $\mathbf{D}$s are bounded away from both zero and infinity.

The BFGS method has a superlinear order of convergence and does no matrix inversion. It updates $\mathbf{D}$ directly by the formula:

$$\mathbf{D}' = \mathbf{D} + \left(1 + \frac{\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{D}\boldsymbol{\gamma}}{\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\gamma}}\right)\frac{\boldsymbol{\delta}\boldsymbol{\delta}^{\mathrm{T}}}{\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\gamma}} - \frac{\boldsymbol{\delta}\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{D} + \mathbf{D}\boldsymbol{\gamma}\boldsymbol{\delta}^{\mathrm{T}}}{\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\gamma}}, \tag{15}$$

where $\boldsymbol{\beta}'$ updates $\boldsymbol{\beta}$, $\boldsymbol{\delta} = \boldsymbol{\beta}' - \boldsymbol{\beta}$, $\boldsymbol{\gamma} = \mathbf{g}' - \mathbf{g}$ and $\mathbf{g}' = \mathbf{g}(G, \boldsymbol{\beta}')$. For any negative definite $\mathbf{D}$ and $\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\gamma} < 0$, $\mathbf{D}'$ is guaranteed to be negative definite. Therefore, one should update $\mathbf{D}$, only when $\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\gamma} < 0$ is satisfied. Theoretically speaking, to ensure global convergence $\mathbf{D}$ should be replaced with $\mathbf{D}'$ only if the eigenvalues of $\mathbf{D}'$ are properly bounded as above. In practice, however, checking the eigenvalues of $\mathbf{D}'$ is costly and seemingly unnecessary. By only ensuring $\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\gamma} < 0$ when updating $\mathbf{D}$, we have not experienced any divergence.

The initial matrix of $\mathbf{D}$ is typically chosen to be minus the identity matrix, which can be badly scaled to a given problem. To increase efficiency, one may replace the backtracking with a two-phase line search with bracketing and sectioning, respectively; see Fletcher (1987, Sect. 2.6). During the bracketing phase, one could replace repeatedly the current bracketing interval $[\boldsymbol{\beta} + a\boldsymbol{\eta}, \boldsymbol{\beta} + b\boldsymbol{\eta}]$ with $[\boldsymbol{\beta} + b\boldsymbol{\eta}, \boldsymbol{\beta} + \{b + 2(b - a)\}\boldsymbol{\eta}]$, until either a local optimal point in the given ascent direction is bracketed or the boundary of the domain of definition is reached. For example, for the Neyman-Scott problem studied in Sect. 5.2, the standard deviation should always remain positive and so should the

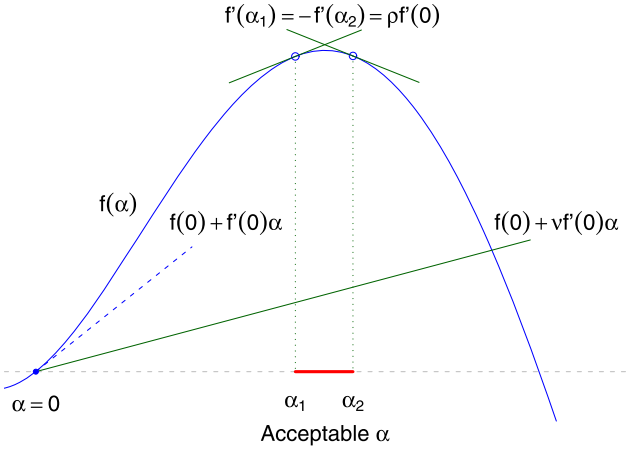

Fig. 1 Criteria for line search

entire bracketing interval. During the sectioning phase, the bracketing interval is then repeatedly bisected and replaced with the half that brackets a local maximum. The sectioning proceeds until the width of the interval is sufficiently small or a test on the slope, in addition to the Armijo rule, is satisfied, e.g.,

$$|\mathbf{g}'^{\mathrm{T}}\boldsymbol{\delta}| \le \rho \mathbf{g}^{\mathrm{T}}\boldsymbol{\delta},$$

for some $\rho \in (\nu, 1)$. We use $\sigma = \frac{1}{2}$, $\nu = \frac{1}{3}$ and $\rho = \frac{1}{2}$ in the implementation.

Figure 1 gives an illustration of the above criterion for line search, based on the function

$$f(\alpha) \equiv \ell(G, \boldsymbol{\beta} + \alpha\boldsymbol{\eta}).$$

An acceptable α, as shown between α_1 and α_2, must satisfy both the Armijo rule

$$f(\alpha) \ge f(0) + \nu f'(0)\alpha,$$

to ensure monotone and sufficient increase of the log-likelihood, and the slope test

$$|f'(\alpha)| \le \rho f'(0),$$

to ensure a nearly optimal solution being found. Generally, at the end of the bracketing phase, it needs to guarantee that some acceptable α's are inside the bracketing interval, while the sectioning is to locate one such α. If the parameter space is restricted, then there may not exist points in the search direction that satisfy the slope test. If the boundary is reached before any acceptable α is bracketed, the bracketing phase should be terminated prematurely, and the slope test will then be abandoned during the sectioning phase.

4.4.2 *Updating* $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\beta})$

The BFGS method can be similarly applied to updating $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\beta})$ as needed in CNM-MS. One difference for updating $\boldsymbol{\pi}$, as from updating the standard deviation in the Neyman-Scott problem, is that the right endpoint of the bracketing interval during the bracketing phase should allow the new $\boldsymbol{\pi}$ to locate on the boundary of the probability simplex. This thus gives the possibility for eliminating redundant support points.

Since $\boldsymbol{\pi} > \mathbf{0}$, we apply the BFGS method to update the parameter vector $(\boldsymbol{\pi}_{-m}, \boldsymbol{\theta}, \boldsymbol{\beta})$, which replaces π_m with $1 - \sum_{j=1}^{m-1} \pi_j$ in the log-likelihood function (3). The method proceeds almost identically to what has been described in Sect. 4.4.1. We also use the two-phase line search here, with the two constraints $\boldsymbol{\pi} \ge \mathbf{0}$ and $\boldsymbol{\pi}^{\mathrm{T}}\mathbf{1} = 1$ being checked for every new step forward during the bracketing phase. If no acceptable point is bracketed before a violation occurs, replace the right endpoint of the bracketing interval with the point in the direction that is on the boundary of the simplex. Then, during the sectioning phase, a right endpoint that satisfies only the Armijo rule is to be located, by bisecting the bracketing interval repeatedly.

In our implementation, the initial $\mathbf{D}$ is always chosen to be minus the identity matrix. It is updated by formula (15), only when $\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\gamma} < 0$ is satisfied. Note that $\boldsymbol{\delta}^{\mathrm{T}}\boldsymbol{\gamma} < 0$ could fail, if the slope test does not hold. Whenever there are zero-massed support points removed, $\mathbf{D}$ is simply reset to minus the identity matrix, but with a reduced dimension. Further improvement is possible, but the performance of the current implementation appears to be very good already.

4.5 Discussion

Several issues are briefly discussed below, including: choosing initial values, dealing with multiple local maxima, testing convergence, collapsing support points that are insignificantly different, modifying the algorithms for fitting finite mixtures, and combining algorithms alternatively.

Starting an algorithm with good initial values certainly helps reduce the computational cost. In our experience, it is generally more efficient to start with a flat mixture density than with a sharp one, since with the former each observation can get a considerable share of the probability mass, thus resulting in a reasonably large likelihood value. To achieve a flat mixture density, one may first consider using a flat component density, which usually means a proper value for $\boldsymbol{\beta}$. A flat component density further implies a small number of support points that are needed initially and thus a reduction of computational cost. One, for example, can compute and use the unicomponent MLE as the initial estimate, as done in our numerical studies. If using the unicomponent MLE or any other unicomponent estimate is somehow insufficient to provide a satisfactory likelihood value, more support points can be included to spread the mixture density.

The proposed three algorithms are guaranteed to find a local maximum of the profile log-likelihood function (5). In

the case of multiple local maxima, the routine strategy is to try different initial values so as to increase the opportunity of locating the global semiparametric MLE. When $\boldsymbol{\beta}$ is of one or two dimensions, one could use CNM to evaluate the profile log-likelihood over a fine grid of $\boldsymbol{\beta}$; see Sect. 5.1 for an example. According to our experience, for mixed effects logistic models as studied in Sect. 5.3, it is not unusual for the profile log-likelihood function to have multiple local maxima. In comparison, for the Neyman-Scott problem as studied in Sect. 5.2, it does appear to be very rare, although it may still occur.

There are many criteria that can be used for terminating an iterative optimization algorithm. In our numerical studies, we simply use

$$\ell(G_s, \boldsymbol{\beta}_s) - \ell(G_{s-1}, \boldsymbol{\beta}_{s-1}) \le \tau, \tag{16}$$

for $\tau = 10^{-6}$, say. This criterion is simple and convenient to use and usually works very well when an algorithm converges rapidly, such as CNM-PL and CNM-MS in all cases. The difference in log-likelihood also has a statistical meaning. In the situation when an algorithm requires many iterations, such as CNM-AP in fitting a mixed effects logistic model, it may be terminated prematurely. Other stopping criteria may also be considered, but they appear to be less convenient and have other disadvantages; see Fletcher (1987, p. 22), for a discussion on testing convergence.

Mathematically speaking, similar support points need not to be collapsed, but they may be produced purely due to numerical reasons. If collapsing similar support points does not decrease the numerical value of the log-likelihood, one certainly should do so. This does not affect the convergence of an algorithm yet gives the advantage of using parsimonious support sets among those that are numerically indifferent.

With a simple modification, the proposed algorithms, especially CNM-MS, can also be used for fitting finite mixtures with a known number of components, a problem that has been studied by Böhning (2003). All that needs to be changed is to restrict the number of support points to be at most the specified number of support points in step 2 of CNM. The end solution must thus be an estimate with the specified number of components, or the semiparametric MLE if it is over-specified.

As suggested by a referee of the paper, in principle the proposed strategies can also be combined with alternative algorithms to CNM for computing a nonparametric MLE. However, the performance gain of CNM over these other algorithms, as studied in Wang (2007), makes it a likely stronger candidate. The performance of the EM algorithm is studied in Sect. 5.1.

5 Numerical studies

In this section, we present in three examples the results of some numerical studies of the performance of the EM (Aitkin 1996, 1999) and three proposed algorithms. The first one concerns fitting a simple logistic regression model with a random intercept to a single simulated data set, which exhibits overdispersion if the intercept is treated as a fixed effects parameter. The performance of all four algorithms is studied and described in details in this example. The other two examples are repeated simulation studies, on, respectively, solving the Neyman-Scott problem and fitting two-level mixed effects logistic regression models. The EM algorithm is excluded from both simulation studies and the CNM-AP method from the second, because of their slow convergence in these cases.

Except that the NNLS algorithm embedded in CNM is given in FORTRAN (Wang 2007), all other computer code is written and executed in R (2006). For the EM algorithm, the R implementation of Einbeck et al. (2007) is used. In order to give a reasonably fair comparison, we always first execute the rapidly convergent CNM-PL method on a given data set and terminate it by criterion (16), with $\tau = 10^{-6}$. Then its achieved maximum likelihood value is used as the critical value to stop the other methods that are executed for the same data set. In the simulation studies of Examples 2 and 3, no algorithm appears to have converged to a different solution. Computation was done on a desktop computer with an Intel Pentium 4 Duo Core 2.80 GHz CPU.

In regard to the initial values, all algorithms use the direct MLE of $\boldsymbol{\beta}$ by treating θ as a fixed-effects variable. For the initial mixing distribution G_0, the EM algorithm uses the Gaussian quadrature with the specified number of support points, while the other three algorithms use a uniform grid of 10 support points with equal mass allocated.

5.1 Example 1

In this example, we study using the semiparametric mixture models to solve the overdispersion problem. A simulated data set is used to illustrate some difficulties that may arise in practice for an algorithm and to study in details the performance of four algorithms: EM, CNM-AP, CNM-PL and CNM-MS.

As given in Table 1, each observation consists of the number of successes (y_i) out of n_i Bernoulli trials and a covariate x_i. Fitting the standard logistic regression model to the data gives the simple MLE, $\hat{\theta} = 0.218$ and $\hat{\beta} = 0.302$, which exhibits a substantial overdispersion, with a residual deviance 225.1 for 18 degrees of freedom. Data sets that exhibits overdispersion for a specified family of models are commonly seen in practice. E.g., in dose-response analysis, overdispersion can arise due to heterogeneous litters, and

Table 1 A simulated data set

i	y_i	n_i	x_i	i	y_i	n_i	x_i
1	3	20	2.22	11	11	30	2.87
2	3	20	0.92	12	15	30	2.94
3	5	20	2.58	13	15	30	0.83
4	5	20	2.22	14	23	30	3.76
5	16	20	5.39	15	25	30	0.40
6	19	20	2.77	16	25	30	1.50
7	20	20	2.77	17	27	30	1.80
8	20	20	1.88	18	28	30	2.13
9	20	20	3.02	19	29	30	3.52
10	20	20	3.28	20	30	30	3.10

failure to account for overdispersion can result in significant under-estimation of the dose effect. In the following, we consider fitting a semiparametric mixture of logistic regression models to the data, by treating the intercept θ as random with a nonparametric distribution G. In this case, the component density at y_i is given by

$$f(y_i|x_i;\theta,\beta) = \{p(\theta+\beta x_i)\}^{y_i}\{1-p(\theta+\beta x_i)\}^{n_i-y_i},$$

where

$$p(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$$

is the probability of success.

For the purpose of illustration, the profile log-likelihood function $\tilde{\ell}(\beta)$ has been evaluated over a uniform grid of 200 β-values on $[-2, 4]$, where each evaluation is one call to the CNM method. Figure 2(a) shows a "smooth" curve of $\tilde{\ell}(\beta)$, obtained by connecting the evaluated points with line segments. The plot shows that the profile log-likelihood has multiple local maxima, and we indicate its global maximum by a circle and a vertical dashed line. Figure 2(b) shows how $\widehat{G}_\beta$ varies with β, in a plot that looks like a bunch of connected tree branches. For each fixed value of β, the center of a branch corresponds to the location of a support point of $\widehat{G}_\beta$ and the width is proportional to the probability mass allocated to it. Note that a support point may emerge at a new location or disappear completely as β varies, and so we have left the connections open when the distances between support points are large. In fact, it appears quite rare that support points are split or join together in a continuous fashion. From the plot, one can also see that the global maximum has four support points. The oblique trend of $\widehat{G}_\beta$ with β suggests a strong correlation between the two. One hence can expect that the parameter-alternating method CNM-AP should not work well in this case.

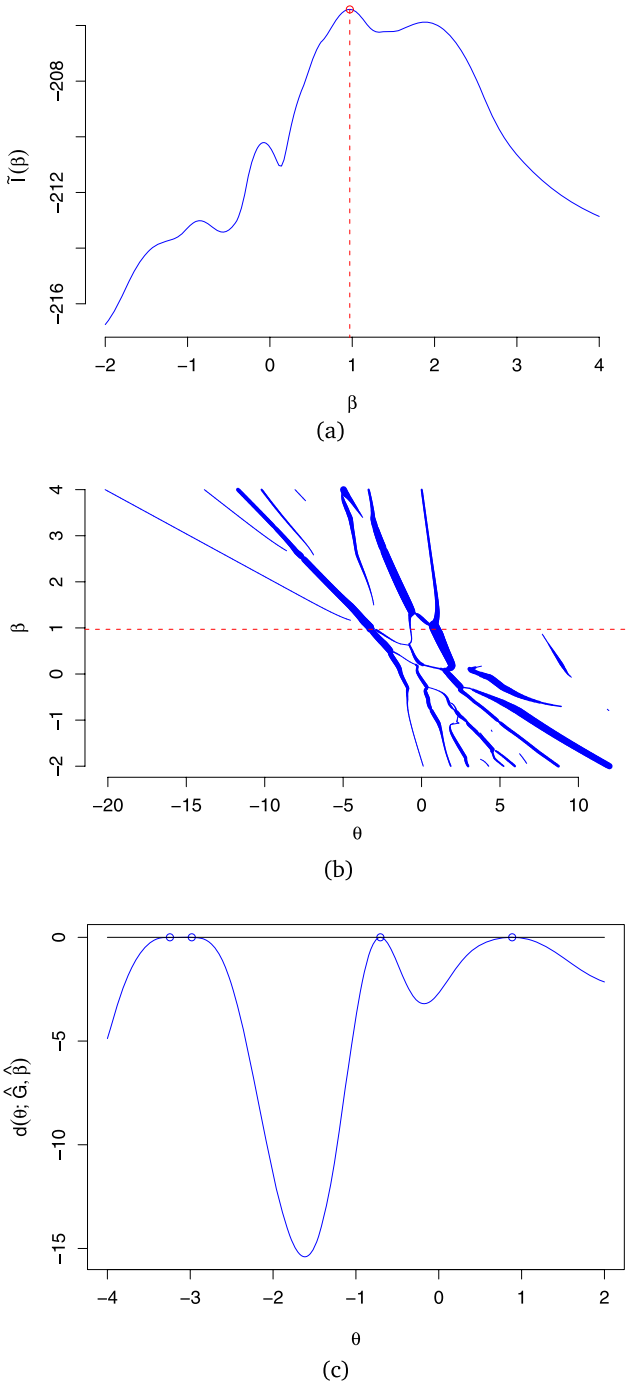

Fig. 2 For the data set in Example 1: (**a**) profile log-likelihood; (**b**) $\widehat{G}_\beta$ that varies with β; (**c**) gradient function evaluated at the global maximum. The *dashed lines* in (**a**) and (**b**) indicate the location of the global maximum, which has four support points, as shown by *circles* in (**c**)

The computational results are given in Table 2, where the global optimum has been found to be

$$\hat{\boldsymbol{\pi}} = (0.270, 0.130, 0.068, 0.532)^{\mathrm{T}},$$
$$\hat{\boldsymbol{\theta}} = (-3.245, -2.981, -0.705, 0.886)^{\mathrm{T}},$$
$$\hat{\beta} = 0.970$$

Table 2 Computing the semiparametric MLE for the data in Table 1

Algorithm	s	m_s	$\lvert\frac{\partial\ell(G_s,\boldsymbol{\beta})}{\partial\beta}\rvert_{\beta=\beta_s}$	$\sup_\theta\{d(\theta; G_s, \beta_s)\}$	Time (s)
EM ($m_0 = 2$)	5	2	3.74×10^{-7}	1.20×10^{3}	0.08
EM ($m_0 = 3$)	30	3	3.49×10^{-5}	5.15×10^{-2}	0.29
EM ($m_0 = 4$)	267	3	2.08×10^{-7}	5.15×10^{-2}	2.96
EM ($m_0 = 5$)	5850	4	4.28×10^{-6}	1.99×10^{-6}	58.97
EM ($m_0 = 6$)	7830	4	4.29×10^{-6}	1.99×10^{-6}	89.79
CNM-AP	197	4	4.64×10^{-5}	1.47×10^{-11}	48.69
CNM-PL	4	4	5.54×10^{-5}	2.13×10^{-10}	7.93
CNM-MS	2	4	4.97×10^{-8}	4.09×10^{-9}	1.88

(rounding to 3 decimal places). The gradient function evaluated at this solution is plotted in Fig. 2(c), which apparently satisfies well condition (7). Of the four algorithms, CNM-PL and CNM-MS are very fast and terminate in four and two iterations, 8.41 and 1.90 seconds, respectively. During the computation, CNM-PL called the full CNM method nine times in total and CNM-MS made only two partial calls. In contrast, the other two algorithms, EM and CNM-AP, have converged quite slowly, in terms of both the number of iterations and execution time. With $m_0 = 2, 3, 4$ initial support points, the EM algorithm converged to sub-optimal solutions and was terminated by criterion (16) instead, with $\tau = 10^{-10}$.

Generally speaking, the EM algorithm has a case-dependent performance and is not always slow. Its slow convergence in this example is due to the flat, terrace-shaped area of the gradient curve around the two leftmost support points, as shown in Fig. 2(c). The existence of such flat areas can make it extremely difficult for EM to allocate probability mass among the support points in the area. It is known that in the worst case the EM algorithm has sublinear convergence, when the Jacobian of the EM step, as a function evaluated at the MLE, is singular. Some techniques have been proposed to accelerate its convergence; see, e.g., Jamshidian and Jennrich (1997).

Another disadvantage of using the EM algorithm for semiparametric MLE computation is its unreliability. To compute a semiparametric (or nonparametric) MLE whose number of support points is unknown beforehand, EM can be either executed repeatedly by gradually incrementing the number of support points, as described by DerSimonian (1986) and Aitkin (1996, 1999), or started with a large number of initial support points, as suggested by Laird (1978). The algorithm may, however, produce a solution with fewer support points (after collapsing virtually identical support points and removing those with virtually zero mass) than initially specified, even if solutions with the specified number of support points do exist. The semiparametric MLE for this data set has four support points, but the EM algorithm, started with $m_0 = 4$, converged to a solution with only three distinct support points. This three-point solution is not even a local maximum of the profile log-likelihood that is shown in Fig. 2(a). It is actually the same solution as obtained by EM with $m_0 = 3$, from which a user may conclude falsely that the semiparametric MLE has been found. Using a large number of initial support points eases the problem, by increasing the opportunity for EM to converge to a local maximum of the profile log-likelihood, but there is no guarantee. We have seen data sets for which EM does not converge to the semiparametric MLE, even when it is started with far more support points than in the semiparametric MLE.

Note that the proposed CNM-based methods do not suffer from the above unreliability problem, due to the steps in CNM that expand and contract the support set. They always converge to a local maximum of the profile log-likelihood, or the global maximum when it is the only maximum.

5.2 Example 2

In this example, we study the famous Neyman and Scott (1948) problem of estimating a common variance from a number of samples, each being drawn independently from a normal distribution with a possibly distinct mean. This problem has a remarkable impact on fundamental issues of statistical inference. In their classical paper, Kiefer and Wolfowitz (1956) suggest using a semiparametric mixture model to solve this problem, by assuming a nonparametric distribution G for all the means. Being perhaps the simplest semiparametric mixture model, its maximum likelihood computation does not appear to have been well studied in the past. A particular finding of our study is that unlike in the situation of Example 1, the variance parameter is nearly "orthogonal" to G, which makes the parameter-alternating method also perform very well.

Let us denote the ith sample by $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$, where y_{ij}'s are drawn independently from the normal distribution with mean $\theta_i \sim G$ and variance β^2. The component density is thus fully determined by $(n_i, \bar{y}_i, r_i)$, as

$$f(\mathbf{y}_i; \theta, \beta) = \frac{1}{(\sqrt{2\pi}\beta)^{n_i}} e^{-\frac{r_i + n_i(\bar{y}_i - \theta)^2}{2\beta^2}},$$

where $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ and $r_i = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$.

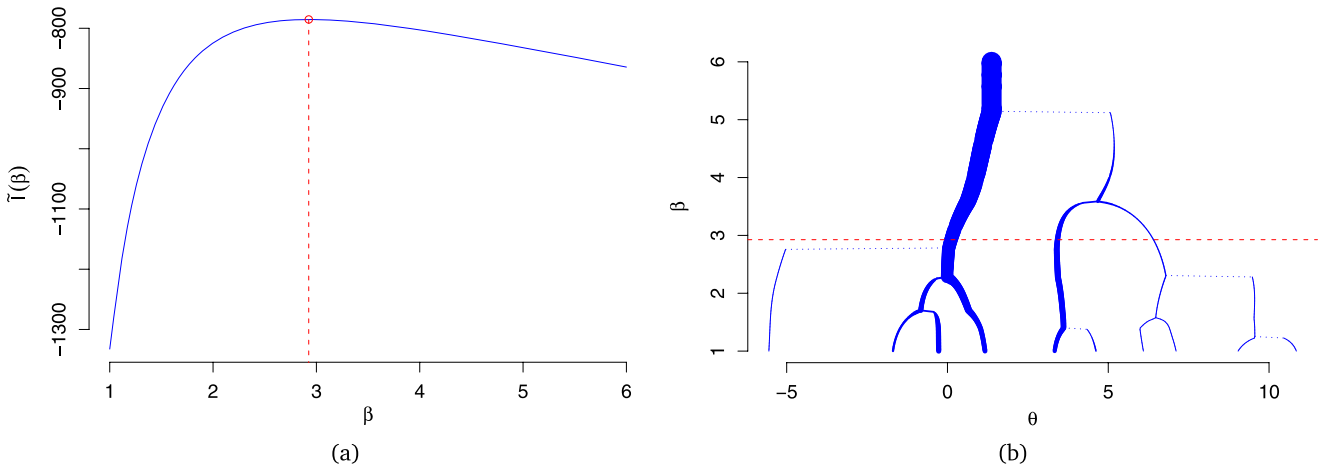

Fig. 3 For a typical Scott-Newman problem: (**a**) profile log-likelihood; (**b**) $\widehat{G}_\beta$ that varies with β. The *dashed line* in each picture indicates the location of the computed semiparametric MLE

Table 3 Five-number summaries of the number of iterations and execution times required by three algorithms over 100 simulated data sets

Algorithm	Number of iterations					Time (s)				
	Minimum	Q1	Median	Q3	Maximum	Minimum	Q1	Median	Q3	Maximum
CNM-AP	8	13	15	16	25	1.19	1.68	1.88	2.21	3.23
CNM-PL	2	3	4	4	5	0.95	1.43	1.66	1.90	2.49
CNM-MS	1	1	1	1	3	0.23	0.35	0.46	0.70	1.85

A simulation study was conducted to compare the proposed three algorithms. Each data set is generated from a semiparametric mixture model with $\boldsymbol{\pi} = (0.7, 0.3)^{\mathrm{T}}$, $\boldsymbol{\theta} = (0, 4)^{\mathrm{T}}$ and $\beta = 3$, with $k = 100$ strata, each of size $n_i \in \{1, 2, 3, 4, 5\}$ with equal proportions. Analogous to Fig. 2, we produced Fig. 3, using a data set that is randomly generated from this model. It is clear from Fig. 3(a) that there is a unique maximum, as is typically true for data sets so generated. How $\widehat{G}_\beta$ varies with β is shown in Fig. 3(b), which looks structurally different from Fig. 2(b), in that the branches here are fairly vertical. This implies that $\widehat{G}_\beta$ is much less sensitive to the change of β than in Example 1, although they are still correlated, of course.

We generated 100 data sets randomly from the above model and applied all three algorithms to them, always started with the uni-component maximum likelihood estimates, i.e., $\beta_0^2 = \sum_{i=1}^k \{r_i + n_i(\bar{y}_i - \bar{y})^2\} / \sum_{i=1}^k n_i$ and G_0 has mass one at the grand mean $\bar{y} \equiv \sum_{i=1}^k n_i \bar{y}_i / \sum_{i=1}^k n_i$. All three algorithms performed well, as summarized in Table 3. CNM-PL generally took fewer number of iterations but a longer time than the other two. Unlike in Example 1, the parameter-alternating method is also very efficient and reliable here, owing to the special structure of the Neyman-Scott problem.

5.3 Example 3

The simulation study in this example is a further sophistication from Example 1, by including a stratum effect and a multidimensional $\boldsymbol{\beta}$. This sophistication certainly prolongs the execution time of each algorithm, but the relative performance among the algorithms appears to differ very little.

In this model, all subjects in each stratum $i \in \{1, \ldots, k\}$ share a common intercept $\theta \sim G$. For subject $j \in \{1, \ldots, n_i\}$ in the ith stratum, let $\mathbf{x}_{ij}$ be the vector of covariates, n_{ij} the number of trials and y_{ij} the number of successes. The joint density at $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$ is thus given by

$$f(\mathbf{y}_i | \mathbf{x}_i; \theta, \boldsymbol{\beta}) = \prod_{j=1}^{n_i} \left\{ p(\theta + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{ij}) \right\}^{y_{ij}} \left\{ 1 - p(\theta + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_{ij}) \right\}^{n_{ij} - y_{ij}}.$$

For data generation, we use $\boldsymbol{\pi} = (0.7, 0.3)^{\mathrm{T}}$, $\boldsymbol{\theta} = (0, 4)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (1, 3)^{\mathrm{T}}$. Each element of $\mathbf{x}_{ij}$ is an independent normal deviate with mean -0.5 and unit variance. Further, we use $k = 100$ strata and choose $(n_i, n_{ij}) \in \{2, 3\} \times \{6, 7, 8, 9, 10\}$ in a cyclic manner.

The CNM-AP algorithm was excluded from this simulation study, since it is usually very slow or even unreliable in this case. The maximum likelihood estimates for the standard logistic regression model are chosen to be the initial

Table 4 Five-number summaries of the number of iterations and execution times required by CNM-PL and CNM-MS over 100 simulated data sets

Algorithm	Number of iterations					Time (s)				
	Minimum	Q1	Median	Q3	Maximum	Minimum	Q1	Median	Q3	Maximum
CNM-PL	5	6	7	7	9	26.9	37.5	43.0	50.2	74.0
CNM-MS	1	2	2	2	4	3.1	8.8	10.1	13.1	26.7

values, as in Example 1, with all stratum information ignored. The performance of CNM-PL and CNM-MS on the same 100 data sets that were randomly generated from the model described above is summarized in Table 4. While they both have performed well, CNM-MS is the apparent winner.

6 Summary and remarks

Three general algorithms, CNM-AP, CNM-PL and CNM-MS, have been proposed above for computing the maximum likelihood estimate of a semiparametric mixture model. All three algorithms make a direct use of the CNM method and employ an additional optimization algorithm for unconstrained problems. They seek to maximize the log-likelihood function by, respectively, alternating the parameters, profiling the likelihood function and modifying progressively the support set. Their performance has been numerically investigated for solving the Neyman-Scott problem and fitting mixed effects logistic regression models. The CNM-MS algorithm performs consistently well in all cases studied. The conventional parameter-alternating strategy as used in CNM-AP has varied efficiency that depends on the correlation between the two parameters. The CNM-PL algorithm is also quite efficient but generally takes a longer time than CNM-MS.

Throughout the paper, we focus on problems in which θ has a continuous space. The proposed algorithms, however, can also be applied when a discrete space of θ. Such situations can arise, e.g., in the presence of censored data (Tsodikov 2003; Zeng and Lin 2007). What needs to be modified in these algorithms is only the support expansion step in the CNM method, which includes all local maxima of the gradient function. Since a gradient function that is defined on a discrete space can be very irregular before convergence and may have a large number of local maxima, including all of them may remarkably increase, unnecessarily, the computational cost per iteration. A suggestion has been given by Wang (2008) in the context of computing a nonparametric MLE for interval-censored data, which is to include at each iteration only those that have the largest gradient values between inclusively every two neighboring support points. This allows for a possible expansion of the support set at an exponential rate, which is usually needed at the early stage of the computation, yet avoids including too many points from a small neighborhood, most of which will become redundant eventually. With this support expansion strategy, the support set maintained after the initial stage is virtually identical to that of the nonparametric MLE. We can adopt this strategy in almost an identical manner for fitting semiparametric mixture.

Semiparametric mixture models are a flexible family and can be used to help solve many nasty problems with heterogeneous populations. A particular advantage of these models is their risk-free specification of a nonparametric distribution, as compared with a parametric assumption. In the past, the difficulty in fitting a semiparametric MLE is a known disadvantage of using them, although neither is it easy to fit a parametric mixed effects model if numerical integration is needed. With the presented algorithms available, we reckon that the semiparametric approach may now also have a computational edge over the parametric approach.

Acknowledgements The author would like to thank the Coordinating Editor and the two referees whose constructive comments have led to many improvements in the manuscript. This research is supported by a Marsden grant of the Royal Society of New Zealand (9145/3608546).

References

Aitkin, M.: A general maximum likelihood analysis of overdispersion in generalised linear models. Stat. Comput. **6**, 251–262 (1996)

Aitkin, M.: A general maximum likelihood analysis of variance components in generalized linear models. Biometrics **55**, 117–128 (1999)

Böhning, D.: A review of reliable algorithms for the semi-parametric maximum likelihood estimator of a mixture distribution. J. Stat. Plan. Inference **47**, 5–28 (1995)

Böhning, D.: The EM algorithm with gradient function update for discrete mixtures with known (fixed) number of components. Stat. Comput. **13**, 257–265 (2003)

Dax, A.: The smallest point of a polytope. J. Optim. Theory Appl. **64**, 429–432 (1990)

DerSimonian, R.: Maximum likelihood estimation of a mixing distribution. J. R. Stat. Soc., Ser. C **35**, 302–309 (1986)

Einbeck, J., Darnell, R., Hinde, J.: NPMLREG: Nonparametric maximum likelihood estimation for random effect models. R Package Version 0.43 (2007)

Fletcher, R.: Practical Methods of Optimization, 2nd edn. Wiley, New York (1987)

Follmann, D.A., Lambert, D.: Generalizing logistic regression by nonparametric mixing. J. Am. Stat. Assoc. **84**, 295–300 (1989)

Haskell, K.H., Hanson, R.J.: An algorithm for linear least squares problems with equality and nonnegativity constraints. Math. Program. **21**, 98–118 (1981)

Heckman, J., Singer, B.: A method for minimizing the impact of distributions assumptions in econometric models for duration data. Econometrica **52**, 271–320 (1984)

Jamshidian, M., Jennrich, R.I.: Acceleration of the EM algorithm by using quasi-Newton methods. J. R. Stat. Soc., Ser. B **59**, 569–587 (1997)

Kiefer, J., Wolfowitz, J.: Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Stat. **27**, 886–906 (1956)

Laird, N.M.: Nonparametric maximum likelihood estimation of a mixing distribution. J. Am. Stat. Assoc. **73**, 805–811 (1978)

Lawson, C.L., Hanson, R.J.: Solving Least Squares Problems. Prentice-Hall, New York (1974)

Lindsay, B.G.: The geometry of mixture likelihoods: A general theory. Ann. Stat. **11**, 86–94 (1983)

Lindsay, B.G.: Mixture Models: Theory, Geometry and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 5. Institute for Mathematical Statistics, Hayward (1995)

Lindsay, B.G., Lesperance, M.L.: A review of semiparametric mixture models. J. Stat. Plan. Inference **47**, 29–39 (1995)

Murphy, S.A., van der Vaart, A.W.: On profile likelihood. J. Am. Stat. Assoc. **95**, 449–465 (2000)

Neyman, J., Scott, E.L.: Consistent estimates based on partially consistent observations. Econometrica **16**, 1–32 (1948)

Powell, M.J.D.: On search directions for minimization algorithms. Math. Program. **4**, 193–201 (1973)

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2006)

Tsodikov, A.: Semiparametric models: a generalized self-consistency approach. J. R. Stat. Soc., Ser. B **65**, 759–774 (2003)

Wang, Y.: On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. J. R. Stat. Soc., Ser. B **69**, 185–198 (2007)

Wang, Y.: Dimension-reduced nonparametric maximum likelihood computation for interval-censored data. Comput. Stat. Data Anal. **52**, 2388–2402 (2008)

Zeng, D., Lin, D.Y.: Maximum likelihood estimation in semiparametric regression models with censored data. J. R. Stat. Soc., Ser. B **69**, 507–564 (2007)