ELSEVIER
Volume 51, issue 6
1 March 2007
ISSN 0167-9473
COMPUTATIONAL STATISTICS & DATA ANALYSIS
Incorporating Statistical Software Newsletter
The official journal of
iasc
The International Association for Statistical Computing
A Section of The International Statistical Institute
Available online at
ScienceDirect
www.sciencedirect.com

Available online at www.sciencedirect.com
ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 2946 – 2957

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

A linearization procedure and a VDM/ECM algorithm for penalized and constrained nonparametric maximum likelihood estimation for mixture models

Ji-Ping Wang

Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, IL 60208, USA

Received 9 July 2006; received in revised form 20 November 2006; accepted 21 November 2006
Available online 15 December 2006

Abstract

Suppose independent observations X_i, $i=1,\ldots,n$ are observed from a mixture model $f(x;Q)\equiv\int f(x;\lambda)\,\mathrm{d}Q(\lambda)$, where λ is a scalar and $Q(\lambda)$ is a nondegenerate distribution with an unspecified form. We consider to estimate $Q(\lambda)$ by nonparametric maximum likelihood (NPML) method under two scenarios: (1) the likelihood is penalized by a functional $g(Q)$; and (2) Q is under a constraint $\mathbf{g}(Q)=\mathbf{g}_0$. We propose a simple and reliable algorithm termed VDM/ECM for Q-estimation when the likelihood is penalized by a linear functional. We show this algorithm can be applied to a more general situation where the penalty is not linear, but a function of linear functionals by a *linearization procedure*. The constrained NPMLE can be found by penalizing the quadratic distance $|\mathbf{g}(Q)-\mathbf{g}_0|^2$ under a large penalty factor $\gamma>0$ using this algorithm. The algorithm is illustrated with two real data sets.

Keywords: Mixture models; Nonparametric maximum likelihood; Computing algorithm; Penalized NPMLE; Constrained NPMLE; VDM/ECM

1. Introduction

A mixture model arises as one models independent observations $x_1,\ldots,x_n$ generated from a parametric density $f(x_i;\lambda_i)$ where the parameter λ_i $(\in\Omega)$ varies in i. To account for the heterogeneity of λ, one assumes that the parameter λ_i, is first independently generated from a *mixing distribution* $Q(\lambda)$, and then $x_i|\lambda_i$ from $f(x_i;\lambda_i)$. Under this setting, x_i's can be regarded as i.i.d. observations from the mixture density as follows:

$$f(x;Q)=\int f(x;\lambda)\,\mathrm{d}Q(\lambda).$$

Mixture models have attracted tremendous attention for their successes in various fields. A mixture model attains great flexibility by allowing λ in $f(x;\lambda)$ to vary to accommodate the heterogeneity that is often inherent in natural populations. For example, the abundance of distinct bird species within a wild ecological community may differ significantly (Fisher et al., 1943). If we model the number of observed birds from each species with a Poisson model $f(x;\lambda)$, then λ varies as a consequence of the heterogeneity in bird abundance. Failure to account for the heterogeneity often results in

E-mail address: jzwang@northwestern.edu.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.11.033

severe under-estimation of the species richness (Bunge and Fitzpatrick, 1993). An extensive review of the applications of mixture models can be found in Titterington et al. (1985) and McLachlan and Peel (2000).

Our attention here is focused on the situation where a parameter $\theta \in \mathbb{R}^K$ determined by the mixture distribution $f(x; Q)$ is of interest, while little knowledge regarding the form of Q is available. Since the distribution Q is unknown, θ defines a functional of Q, mapping from an infinite-dimensional space to the real space, i.e. $\theta : \mathbb{Q} \mapsto \mathbb{R}^K$, where $\mathbb{Q} = \{Q : Q \text{ is a probability measure}\}$. A natural and appealing estimator of θ is the plug-in estimator $\theta(\hat{Q})$, where $\hat{Q}$ is the nonparametric maximum likelihood estimator (NPMLE) of Q (Lindsay, 1983, 1995). The plug-in estimator is simple and robust since no assumption has to be made regarding the form of Q. More importantly $\theta(\hat{Q})$ often is consistent for $\theta(Q)$ by consistency of $\hat{Q}$ given identifiability and some technical conditions (Kiefer and Wolfowitz, 1956; Leroux, 1992; Van de Geer, 2002). Although asymptotic results regarding $\theta(\hat{Q})$ beyond consistency remain unknown in general, in some special situations (e.g. Susko et al., 1998) standard results do exist, and hence one can construct the confidence interval for $\theta(Q)$ or testing hypothesis $H_0 : \theta(Q) = \theta_0$ via standard methods, e.g. profile likelihood method. To build the nonparametric profile likelihood on θ, one needs to find the constrained NPMLE, i.e.

$$\hat{Q} = \arg_{[Q:\theta(Q)=\theta_0]} \sup\{\ell_0(Q; \mathbf{x})\}, \quad Q \in \mathbb{Q}, \quad \theta_0 \in \mathbb{R}^K,$$

where $\ell_0(Q; \mathbf{x})$ is the original log likelihood defined as $\ell_0(Q; \mathbf{x}) := \sum_i \log[f(x_i; Q)]$.

Closely related to the constrained NPMLE, the penalized NPMLE method has been shown effective in stabilizing the plug-in estimator in some important applications (Wang and Lindsay, 2005). Unlike many model-selection oriented penalized likelihood approaches (Chen and Kalbfleisch, 1996; Leroux, 1992; McLachlan and Peel, 2000), the penalty in these applications often concerns a function of the parameter (θ) under estimation. The penalty works as a prior distribution for θ from the Bayesian perspective, bringing in the desirable shrinkage effect to the plug-in estimate $\theta(\hat{Q})$.

Finding the constrained or penalized NPMLE is not trivial. The constrained NPMLE actually can be obtained by penalized NPMLE method from a duality result (Lindsay, 1995; Susko et al., 1998). One existing algorithm is the generalized ISDM by Susko et al. (1998), as an extension of ISDM algorithm by Lesperance and Kalbfleisch (1992). In this paper we propose an alternative, namely, VDM/ECM, and show its simplicity and reliability in implementation. In Section 2, we first introduce a *linearization* procedure for the penalized NPMLE that generalizes the result from Wang and Lindsay (2005). In Section 3, we then discuss the VDM/ECM algorithm in detail for the penalized NPMLE. In Section 4, we introduce some additional optimization results that bridge the penalized and constrained NPMLE, and show how the VDM/ECM algorithm can be applied to the constrained NPML estimation. Some numerical results are presented in Section 5.

2. Penalized NPMLE with linear and non-linear penalties

We first consider the simplest case where the likelihood is penalized by a *linear* functional. A linear functional $h(Q)$ is defined as follows:

$$h(Q) = \int h(\lambda)\, \mathrm{d}Q(\lambda)$$

for some real and continuous function $h(\cdot)$. Let $y_1, \ldots, y_d$ be the distinct data points from the observed data $\mathbf{X}$, and $n_j = \sum_i I(x_i = y_j)$, then the penalized likelihood can be written as

$$\ell(Q) := \ell_0(Q) - h(Q),$$

where $\ell_0(Q) = \sum_j [n_j \log[f(y_j; Q)]$ is the original unpenalized mixture likelihood. Define the *extended likelihood vector curve* as

$$\Gamma = \{\mathbf{L}(\lambda) = [f(y_1; \lambda), f(y_2; \lambda), \ldots, f(y_d; \lambda), h(\lambda)]^{\mathrm{T}} : \lambda \in \Omega\},$$

and the *extended likelihood vector space* as $\mathbf{M} = conv(\Gamma)$ where "*conv*" stands for *convex hull*. Under this notation, the log likelihood $\ell(Q)$ can be regarded as a function defined in a finite-dimensional sub-space $\mathbf{M}$, i.e. $\ell(Q) \equiv \ell(f(y_1; Q), \ldots, f(y_d; Q), h(Q))$. In the following sections, all the discussions on the geometry of the likelihood function $\ell(Q)$ is referred to the extended likelihood vector space. If $\mathbf{M}$ is closed and bounded, we have the following results (Lindsay, 1995; Susko et al., 1998):

Theorem 1. (I) *$\ell(Q)$ is strictly concave in* $\mathbf{M}$. (II) *There exists a solution, $\hat{Q}$, to the penalized likelihood function, with no more than $d+1$ support points.* (III) *Define the extended gradient function as*

$$D_{\hat{Q}}(\lambda)=\sum_{j=1}^{d} n_j\left[\frac{f(y_j;\lambda)}{f(y_j;\hat{Q})}-1\right]-[h(\lambda)-h(\hat{Q})], \tag{1}$$

then $\hat{Q}$ is NPMLE if and only if $D_{\hat{Q}}(\lambda)\leqslant 0$, $\forall\lambda\in\Omega$; furthermore, λ_0 is a support point if and only if $D_{\hat{Q}}(\lambda_0)=0$.

Theorem 1 states that the gradient inequality criterion is still sufficient and necessary for $\hat{Q}$ to be the NPMLE. Now, we consider a more general situation where the log likelihood is penalized by a functional g involving m linear functionals, i.e. $g(\mathbf{h})=g(h_1,h_2,\ldots,h_m)$. The penalized log likelihood is

$$\ell(Q)=\ell_0(Q)-g[\mathbf{h}(Q)]. \tag{2}$$

Note that g is not necessarily a linear functional of Q. In general, finding NPMLE from (2) becomes much harder than when g is linear. Let $\hat{Q}_\alpha=(1-\alpha)\hat{Q}+\alpha\Delta_\lambda$ where Δ_λ denotes the degenerate distribution at λ. Note that $\mathbf{h}(\hat{Q}_\alpha)=(1-\alpha)\mathbf{h}(\hat{Q})+\alpha\mathbf{h}(\lambda)$, therefore the likelihood function $\ell(\hat{Q}_\alpha)$ for a given $\hat{Q}$ and λ becomes a function of α. A $\hat{Q}$ is defined as a *local* solution if

$$D_{\hat{Q}}(\lambda)=\frac{\partial}{\partial\alpha}\{\ell_0(\hat{Q}_\alpha)-g[\mathbf{h}(\hat{Q}_\alpha)]\}|_{\alpha=0}\leqslant 0,\quad \forall\lambda\in\Omega. \tag{3}$$

Since the actual objective function based on $\ell(Q)$ only involves terms defined in the extended likelihood vector space, the evaluation of the *directional derivative*, i.e. $D_{\hat{Q}}(\lambda)$ will also involve terms in the same space. The following *linearization* theorem provides a way to convert the likelihood into an alternative objective function with a linear penalty, such that NPMLE for (2) can be found by iteratively maximizing the alternative likelihood function.

Theorem 2. *Let $g(\mathbf{h})=g(h_1,\ldots,h_m)$ be a double-differentiable function with respect to* $\mathbf{h}$, *where $h_i=h_i(\lambda)$ are continuous functions of λ. Define $h_i(Q)=\int h_i(\lambda)\,\mathrm{d}Q(\lambda)$, and $g_i'(\mathbf{h}_Q)=\frac{\partial}{\partial h_i}g(\mathbf{h})|_{\mathbf{h}=\mathbf{h}(Q)}$. Assume that the extended likelihood vector space $\mathbf{M}=conv\{[f(y_1,\lambda),\ldots,f(y_d,\lambda),h_1(\lambda),\ldots,h_m(\lambda)]^{\mathrm{T}}:\lambda\in\Omega\}$ is compact, $g_i'(\mathbf{h}_Q)$ and $h_i(Q)$ are bounded for $i=1,\ldots,m$, $\forall Q$. Define*

$$\ell^*(Q)=\ell_0(Q)-\sum_{i=1}^{m}\gamma_i^* h_i(Q), \tag{4}$$

then

(I) *a necessary and sufficient condition for $\hat{Q}$ to be a local solution for* (2) *is that $\hat{Q}$ is the NPMLE* (*global*) *for the linearized version* (4) *at $\gamma_i^*=g_i'(\mathbf{h}_{\hat{Q}})$, for $i=1,\ldots,m$.*

(II) *If $g(\mathbf{h})$ is convex, i.e. the Hessian matrix $|\frac{\partial^2 g}{\partial h_i\partial h_j}|$ is semi-positive definite $\forall$ $\mathbf{h}\in\{\mathbf{h}(Q):Q\in\mathbb{Q}\}$, then a sufficient and necessary condition for that $\hat{Q}$ is the global solution for* (2) *is that $\hat{Q}$ is the NPMLE for the linearized version* (4) *at $\gamma_i^*=g_i'(\mathbf{h}_{\hat{Q}})$, for $i=1,\ldots,m$.*

Proof. (I) The gradient function is given by $D_{\hat{Q}}(\lambda)=\sum n_j[\frac{f(y_j,\lambda)}{f(y_j;\hat{Q})}-1]-\sum g_i'(\mathbf{h}_{\hat{Q}})[h_i(\lambda)-h_i(\hat{Q})]$, $\forall\lambda\in\Omega$. Clearly if $g_i'(\mathbf{h}_{\hat{Q}})$ is finite for $i=1,\ldots,m$, $\hat{Q}$ is a local solution for the likelihood in (2) if and only if $D_{\hat{Q}}(\lambda)\leqslant 0$, $\forall\lambda\in\Omega$. Meanwhile this condition is exactly the necessary and sufficient one for $\hat{Q}$ to be the NPMLE for (4) at $\gamma_i^*=g_i'(\mathbf{h}_{\hat{Q}})$.

(II) If $g(\mathbf{h})$ is convex, then $-g(\mathbf{h})$ is concave. It is easy to verify that the log likelihood is strictly concave in $\mathbf{M}$. Hence the local solution is also the global NPMLE. □

Theorem 2 implies that a local solution for the penalized likelihood $\ell_0 - g(\mathbf{h})$ must be the global solution of the *linearized* version. We will show in the next section that this can be achieved by an iterative VDM/ECM algorithm.

The boundedness of $\mathbf{M}$, $g_i'(\mathbf{h}_Q)$ and $h_i(Q)$, for $i = 1, \dots, m$ is required to guarantee that the global maximum likelihood is finite, and the gradient criterion is applicable. Otherwise, even if the likelihood is bounded from above, the gradient criterion in Theorem 1 may fail. For example, if $\ell_0(Q)$ is bounded above, then so is $\ell_0(Q) - h^2$ for any linear functional h such that $h(\lambda) \in (-\infty, \infty)$. However, the gradient criterion in Theorem 1 cannot be satisfied for any $\hat{Q}$ since the directional derivative $D_{\hat{Q}}(\lambda) = \sum_{j=1}^{d} n_j [\frac{f(y_j;\lambda)}{f(y_j;\hat{Q})} - 1] - 2h(\hat{Q})[h(\lambda) - h(\hat{Q})]$ is not bounded from above. This complexity can be viewed from another perspective. Suppose there exists NPMLE $\hat{Q}$ for $\ell(Q) - h^2$ and $0 < h(\hat{Q}) < \infty$. So $\hat{Q}$ is the constrained NPMLE under $h(Q) = h(\hat{Q})$. Let λ^* be such that $h(\lambda^*) \to -\infty$, and $\hat{Q}^* = (1-\varepsilon)\hat{Q} + \varepsilon\Delta_{\lambda^*}$ (where Δ_{λ^*} stands for a degenerate distribution at λ^*) and $\varepsilon \approx 0^+$. One can always force $h(\hat{Q}^*) \approx -h(\hat{Q})$ while $\ell(\hat{Q}^*) \approx \ell(\hat{Q})$ due to the tiny weight ε. Practically, $\hat{Q}$ and $\hat{Q}^*$ will lead to indistinguishable penalized likelihood value.

It is often not practical to have all the boundedness conditions satisfied. In some cases, one might restrain Ω to be in a certain region to obtain the boundedness for the functionals. For example, for a Poisson mixture, if $h(\lambda) = \lambda$, then clearly h is not bounded. However, one might want to force $\lambda \leqslant \max_i(x_i)$ or $\lambda \leqslant C$ for some constant C to bound h.

Sometimes we are interested in a penalized form $g(\mathbf{h}) = \sum_{i=1}^{m} \gamma_i g_i(h_i)$ where γ_is are constant *penalty factors*, by which we can penalize several functionals simultaneously. Consider the following likelihoods:

$$\ell(Q) = \ell_0(Q) - \sum_{i=1}^{m} \gamma_i g_i(h_i), \tag{5}$$

$$\ell^*(Q) = \ell_0(Q) - \sum_{i=1}^{m} \gamma_i^* h_i(Q). \tag{6}$$

Define $g_i'[h_i(\hat{Q})] = \frac{\mathrm{d}g_i}{\mathrm{d}h_i}|_{h_i = h_i(\hat{Q})}$. Clearly given similar boundedness conditions as in Theorem 2, a global solution from (6) given $\gamma_i^* = \gamma_i g_i'[h_i(\hat{Q})]$ must be a local solution of (5). A sufficient condition for a local solution to be the global one is that g_i's are all convex.

To see how the penalty parameter γ_i affects the estimation in (5) when $g(\mathbf{h}) = \sum_{i=1}^{m} \gamma_i g_i(h_i)$, we have the following result:

Theorem 3. *Let $\hat{Q}$ be the NPMLE under $\gamma = \{\gamma_1, \dots, \gamma_m\}$ for* (5) *and $\tilde{Q}$ be that under the same penalties except that the penalty factor for g_i is $\tilde{\gamma}_i \geqslant \gamma_i$ for one fixed i. Then $g_i[h_i(\hat{Q})] \geqslant g_i[h_i(\tilde{Q})]$.*

Proof. Omitted. □

Theorem 3 is an extension of the result obtained by Wang and Lindsay (2005) for one-penalty situation. It states the fact that penalizing a functional g_i in the likelihood reduces the value of $g_i(\hat{Q})$. Hence the stability of a plug-in estimator $\theta(\hat{Q})$, if unstable, can be potentially improved by applying an appropriate penalty.

3. A VDM/ECM algorithm for penalized NPMLE

Various algorithmic methods are available for finding NPMLE in the mixture setting. The well-known algorithms include EM method (for fixed number of components, Laird, 1978), the Vertex Direction Method (VDM, Lindsay, 1983), the Vertex Exchange Method (VEM, Böhning, 1985), the Intra-simplex Direction Method (ISDM, Lesperance and Kalbfleisch, 1992; Susko et al., 1998) and other hybrid methods such as VDM/EM (DerSimonian, 1986; Lindsay and Roeder, 1993) and EM/VEM (Böhning et al., 1992). In particular, the ISDM algorithm in Susko et al., 1998 was designed for the constrained and penalized NPMLE. Here we discuss a hybrid of VDM and Expectation–conditional –maximization (ECM, Meng and Rubin, 1993) for the penalized NPML estimation. This method is simple but reliable. We will apply it to the constrained NPMLE in the next section.

3.1. Linear penalty

The VDM/ECM is a variant of VDM/EM where the EM procedure is replaced by ECM. The VDM step identifies the point on a grid of $\lambda \in \Omega$ that achieves the maximum gradient given the current estimate of Q. This point is then added into the support of Q. In the subsequent ECM step, Q is refined until a local maximum is obtained. In the following, we first discuss the algorithm in detail for the penalized likelihood with one linear penalty, i.e. $\ell(Q) = \ell_0(Q) - h(Q)$, and then extend it to the situation where the penalty is a function of linear functionals.

Since the NPMLE $\hat{Q}$ will be a discrete distribution with a finite number of support points, the log likelihood function can be written in the form of a finite mixture with a Lagrange multiplier δ as follows:

$$\ell_2 = \sum_i \log[f(x_i)] - \sum_k \pi_k h_k + \delta\left(\sum_k \pi_k - 1\right), \tag{7}$$

where λ_k and π_k are the kth component and its weight for $k = 1, \ldots, K$, $f_k(\cdot) \equiv f(\cdot; \lambda_k)$ and $h_k \equiv h(\lambda_k)$. Define the latent data, $z_{ik} = 1$ if x_i was generated from the kth component and 0 otherwise for $i = 1, \ldots, n$ (the augmented data will be denoted as $\mathbf{Z}$ below). Then the complete likelihood is

$$\ell_3 = \sum_i \sum_k z_{ik} \log[f_k(x_i)] + \sum_i \sum_k z_{ik} \log(\pi_k) - \sum_k \pi_k h_k + \delta\left(\sum_k \pi_k - 1\right). \tag{8}$$

Let $\boldsymbol{\psi} = (\boldsymbol{\Lambda}^{\mathrm{T}}, \boldsymbol{\pi}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\Lambda}$ and $\boldsymbol{\pi}$ are the mixture components and corresponding weights. In the following, we use a superscript $^{(t)}$ to index the VDM iteration. The VDM/ECM algorithm alternates among the following steps:

(1) Let G be a fine grid for λ in Ω. Given the current estimate $\boldsymbol{\psi}^{(t)} = (\boldsymbol{\Lambda}^{(t)^{\mathrm{T}}}, \boldsymbol{\pi}^{(t)^{\mathrm{T}}})^{\mathrm{T}}$ find the grid point $\lambda^{(t)} \in G$ that maximizes the gradient function (1).

(2) Let $Q_\alpha^{(t)} = (1-\alpha)\hat{Q}^{(t)} + \alpha\Delta_{\lambda^{(t)}}$, where $\Delta_{\lambda^{(t)}}$ is the degenerate distribution at $\lambda^{(t)}$, find the optimal $\alpha^{(t)}$ that maximizes the following log likelihood function:

$$\sum_i \log[f_{Q_\alpha^{(t)}}(x_i)] - h(Q_\alpha^{(t)})$$

using Newton's method.
Note: The log likelihood function is strictly concave in α, so the solution exists and is guaranteed unique.

(3) Let $\boldsymbol{\Lambda}^* = (\boldsymbol{\Lambda}^{(t)^{\mathrm{T}}}, \lambda^{(t)})^{\mathrm{T}}$,$\boldsymbol{\pi}^* = \{(1-\alpha)\boldsymbol{\pi}^{(t)^{\mathrm{T}}}, \alpha^{(t)}\}^{\mathrm{T}}$. Run the ECM procedure (see below) using $\boldsymbol{\Lambda}^*$ and $\boldsymbol{\pi}^*$ as starting values to obtain an update $\hat{Q}^{(t+1)}$.

(4) Repeat the above steps until the algorithm converges. The criterion used for convergence here is $\mathrm{Sup}_\lambda\, D_{\hat{Q}}(\lambda) < 0.005$. This guarantees that the log likelihood at convergence will be no more than 0.005 below the true maximum (Lindsay, 1995, p. 132).

The ECM step uses $\boldsymbol{\psi}^{(0)} \equiv (\boldsymbol{\Lambda}^{*\mathrm{T}}, \boldsymbol{\pi}^{*\mathrm{T}})^{\mathrm{T}}$ as starting values. A superscript $^{(m)}$ here indexes the iterations within the ECM loop. Let $K^{(t)}$ be the number of components in $\hat{Q}^{(t)}$. The ECM step proceeds as follows:

(1) E step: given the current estimate, define $\boldsymbol{\psi}^{(m)} = (\boldsymbol{\Lambda}^{(m)\ \mathrm{T}}, \boldsymbol{\pi}^{(m)\ \mathrm{T}})^{\mathrm{T}}$,

$$z_{ik}^{(m+1)} := E(z_{ik}|\boldsymbol{\psi}^{(m)}) = \frac{f_k^{(m)}(x_i)\pi_k^{(m)}}{\sum_k f_k^{(m)}(x_i)\pi_k^{(m)}}.$$

(2) CM step: to maximize the complete likelihood function (8), we need to solve the following $2K^{(t)} + 1$ equations simultaneously:

$$\sum_i z_{ik}^{(m+1)} \frac{1}{\pi_k} + \delta - h_k = 0, \quad \text{for } k = 1, \ldots, K^{(t)}, \tag{9}$$

$$\sum_k \pi_k = 1, \tag{10}$$

$$\sum_i z_{ik}^{(m+1)} \frac{f_k'(x_i)}{f_k(x_i)} - \pi_k h_k' = 0, \quad \text{for } k = 1, \ldots, K^{(t)}. \tag{11}$$

The number of free parameters reduces to $2K^{(t)}$ (including δ) due to the constraint on π_k's. To solve these equations simultaneously is difficult as each equation involves multiple parameters and is non-linear. We have developed a sequential conditional–maximization strategy as follows:

- $\mathbf{M}_1$ step: the first step is to update $\boldsymbol{\pi}$ from (9) + (10) by fixing $\boldsymbol{\psi}$ at $\boldsymbol{\psi}^{(m)}$. From (9) we obtain

  $$\pi_k = -\frac{\sum_i z_{ik}^{(m+1)}}{\delta - h_k^{(m)}}. \tag{12}$$

  By (10),

  $$\sum_k -\frac{\sum_i z_{ik}^{(m+1)}}{\delta - h_k^{(m)}} = 1.$$

  Define the function $w(\delta) = \sum_i [-\frac{\sum_i z_{ik}^{(m+1)}}{\delta - h_k^{(m)}}] - 1$. Note δ must satisfy $\delta \leqslant \min_k h_k$ to guarantee $\pi_k > 0$. One can show there is one and only one solution by noting that $w(\delta)$ is a monotone function of δ ranging from -1 to ∞ for $\delta \leqslant \min_k h_k$. The update $\delta^{(m+1)} = \delta(\mathbf{Z}^{(m+1)}, \boldsymbol{\Lambda}^{(m)})$ can be reliably found by Newton's method, and hence $\boldsymbol{\pi}^{(m+1)}$ from (12) given $\delta^{(m+1)}$.
  We call this step from $(\boldsymbol{\Lambda}^{(m)}, \boldsymbol{\pi}^{(m)})$ to $(\boldsymbol{\Lambda}^{(m)}, \boldsymbol{\pi}^{(m+1)})$ the $\mathbf{M}_1$ step, in which the monotonicity of the full log likelihood is guaranteed.
- $\mathbf{M}_2$ step: this step updates $\boldsymbol{\Lambda}$ based on $\boldsymbol{\pi}^{(m+1)}$, $\delta^{(m+1)}$. Given $\boldsymbol{\pi}^{(m+1)}$ and $\delta^{(m+1)}$, Eq. (11) only involves λ_k alone. Newton's method can be used to find the update $\boldsymbol{\Lambda}^{(m+1)}$. Again the log likelihood increases in the $\mathbf{M}_2$ step since we are maximizing the objective function treating $(\mathbf{Z}^{(m+1)}, \boldsymbol{\pi}^{(m+1)}, \delta^{(m+1)})$ as fixed.

3.2. Penalty is a function of linear functionals

Finding the penalized NPMLE with an arbitrary non-linear penalty can be computationally intractable. For a class of non-linear penalties where the penalizing functional can be expressed as a function of linear functionals, e.g. $g(\mathbf{h})$, direct maximization of the likelihood $\ell_0(Q) - g(\mathbf{h})$ can be as simple as the linear case based on the results in Theorem 2. A local maximum is guaranteed. We can iteratively maximize the linearized penalized likelihood using the VDM/ECM as follows:

(1) Initialize $\mathbf{h} = \mathbf{h}_0 = (h_{10}, h_{20}, \ldots, h_{m0})^{\mathrm{T}}$, find the penalized NPMLE $\hat{Q}^{(0)}$ for the linearized penalized likelihood (4) under $\gamma_i^* = g_i'(\mathbf{h}_0)$ for $i = 1, \ldots, m$.
(2) In the tth iteration, update $\mathbf{h}^{(t)} = \mathbf{h}[\hat{Q}^{(t-1)}]$ and find the penalized NPMLE $\hat{Q}^{(t)}$ for the penalized likelihood (4) under $\gamma_i^* = g_i'(\mathbf{h}^{(t)})$.
(3) Repeat step 2 until $|\mathbf{h}^{(t)} - \mathbf{h}^{(t-1)}| \leqslant tol$ for some $tol > 0$.

4. Constrained NPMLE by penalizing

The marked reliability and simplicity of VDM/ECM algorithm make it a good alternative for constrained NPMLE when the constraint functionals are linear or functions of linear functionals. If the constraint is a scalar constraint, i.e. $g(Q) = g_0$, then one can tune the penalty factor γ in the likelihood $\ell_0(Q) - \gamma g(Q)$ until $g(\hat{Q}_\gamma) = g_0$ for a certain γ, and $\hat{Q}_\gamma$ is the constrained NPMLE (see the duality result in Lindsay (1995, p. 143)). In particular, the monotonicity results in Theorem 3 point out the direction in which to tune γ (increase or decrease) towards the target constraint.

For example, if $g(\hat{Q}_\gamma) > g_0$ given the current γ, then one needs to increase γ, and vice versa. This procedure requires that $g(\hat{Q}_\gamma)$ is a continuous function of γ at least around $g(\hat{Q})$ where $\hat{Q}$ is the true constrained NPMLE. If fitting the constrained NPMLE is to build the profile likelihood on $g(Q)$, then the log likelihood value evaluated at the penalized NPMLE $\hat{Q}_\gamma$ for a given γ, i.e. $\ell_0(\hat{Q}_\gamma)$, is exactly the profile likelihood value at $g(Q) = g(\hat{Q}_\gamma)$. Hence one only needs to find the penalized NPMLE at different γ's in order to construct the profile likelihood.

In a more general situation where multiple scalar constraints exist, to obtain the constrained NPMLE by tuning the penalty factors can be much harder than the single constraint case. Here we adopt the approach by Susko et al. (1998), to penalize a quadratic distance function with a large positive penalty factor, but with a different implementation.

Suppose we need to find NPMLE under the constraint $\mathbf{g}(Q) = \mathbf{g}_0$, where $\mathbf{g}(Q) \equiv \mathbf{g}[\mathbf{h}(Q)] = (g_1(\mathbf{h}), \ldots, g_K(\mathbf{h}))^{\mathrm{T}}$, $\mathbf{g}_0 \equiv (g_{10}, g_{20}, \ldots, g_{K0})^{\mathrm{T}}$ and $\mathbf{h}(Q) \equiv (h_1, \ldots, h_M)^{\mathrm{T}}$, where $\mathbf{h}$ are linear functionals. We further assume that $M \geqslant K$ such that the constrained NPMLE solution exists. We consider the log likelihood function with a quadratic penalty as follows:

$$\ell_0(Q) - \gamma[\mathbf{g}(Q) - \mathbf{g}_0]^{\mathrm{T}}[\mathbf{g}(Q) - \mathbf{g}_0], \quad \gamma > 0. \tag{13}$$

Note the concavity of the penalized likelihood function in the extended likelihood vector space does not necessarily hold. A local solution is not necessarily the global one. If we let $\gamma \to \infty$, a local solution under γ eventually will satisfy the constraint $\mathbf{g} = \mathbf{g}_0$. To see this, let $\hat{Q}_1$ and $\hat{Q}_2$ be the NPMLE under penalties $\gamma_1 \geqslant \gamma_2 > 0$. By monotonicity in Theorem 3, $|\mathbf{g}(\hat{Q}_1) - \mathbf{g}_0| \leqslant |\mathbf{g}(\hat{Q}_2) - \mathbf{g}_0|$, where $|\cdot|$ is the Euclidean norm. Intuitively as we increase γ to ∞, eventually $\mathbf{g} \to \mathbf{g}_0$. This result is formally stated in the following theorem.

Theorem 4. *Suppose the same compactness and boundedness conditions as in Theorem 2 hold. We consider to maximize $\ell_0(Q)$ under the constraint $\mathbf{g}(Q) \equiv (g_1(\mathbf{h}), \ldots, g_K(\mathbf{h}))^{\mathrm{T}} = \mathbf{g}_0 \equiv (g_{10}, g_{20}, \ldots, g_{K0})^{\mathrm{T}}$, where $\mathbf{h} = (h_1, \ldots, h_M)$ are linear functionals of Q for $M \geqslant K$. Let $\hat{Q}_\gamma$ be the mixing distribution that solves $Q = \arg_Q \sup\{\ell_0(Q) - \sum_m \gamma_m^* h_m(Q)\}$ for $\gamma_m^* = 2 * \gamma \sum_k (g_k(\mathbf{h}) - g_{k0}) g'_{km}(\mathbf{h}), i = 1, \ldots, m$, where $g'_{km}(\mathbf{h}) \equiv \frac{\partial g_k}{\partial h_m}|_{\mathbf{h}=\mathbf{h}(Q)}$. We assume $|\mathbf{g}(\hat{Q}_\gamma) - \mathbf{g}_0|$ is continuous in γ; furthermore the constrained subspace $\mathbf{M}^* = \{[f(y_1; Q), f(y_2; Q), \ldots, f(y_d; Q)]^{\mathrm{T}} : \mathbf{g}(Q) = \mathbf{g}_0, Q \in \mathbb{Q}\}$, is nonempty such that there exists at least one point in M^* satisfying $\ell_0(Q) > -\infty$. Then*

(I) *if $\mathbf{M}^*$ is convex, then $\hat{Q}_\gamma$ converges to the global solution that satisfies the constraint $\mathbf{g}(Q_\gamma) = \mathbf{g}_0$ as $\gamma \to \infty$;*
(II) *otherwise, $\hat{Q}_\gamma$ converges to a local solution satisfying the constraint $\mathbf{g}(Q_\gamma) = \mathbf{g}_0$ as $\gamma \to \infty$.*

Proof. By Theorem 2, $\hat{Q}_\gamma$ defined above, must be a local solution for $\ell_0(Q) - \gamma(\mathbf{g} - \mathbf{g}_0)^{\mathrm{T}}(\mathbf{g} - \mathbf{g}_0)$. As $\gamma \to \infty$, $|\mathbf{g} - \mathbf{g}_0|_{\hat{Q}_\gamma}$ must go to zero, i.e. $\lim_{\gamma\to\infty} \mathbf{g}(\hat{Q}_\gamma) = \mathbf{g}_0$, otherwise $\hat{Q}_\gamma$ is beat by some $\hat{Q}$ that satisfies $\mathbf{g}(Q) = \mathbf{g}_0$ (such a Q exists by assumption). Let $\hat{Q}_{\gamma=\infty} \equiv \lim_{\gamma\to\infty} Q_\gamma$. Therefore $\hat{Q}_{\gamma=\infty}$ is the constrained local solution for the likelihood $\ell_0(Q)$ under constraint $\mathbf{g}(Q) = \mathbf{g}(\hat{Q}_{\gamma=\infty}) = \mathbf{g}_0$. If $\mathbf{M}^*$ is convex, because of strict concavity of the unpenalized log likelihood $\ell_0(Q)$ over $\mathbf{M}^*$, the local solution must be the global one (Luenberger, 1969, p. 191). If the convexity does not hold, $\hat{Q}_{\gamma=\infty}$ in theory can be a local solution. □

Remark 1. Obviously the convexity of the constrained likelihood vector space holds when $\mathbf{g}$ is a linear function of $\mathbf{h}$. If $\mathbf{g}$ is not a linear function of $\mathbf{h}$, then the convexity may fail. For example, suppose we have $g(h) = h^2$ where $h(Q)$ is linear in Q and $h(Q) \in (-\infty, \infty)$. If $h(Q_1) = -h(Q_2) = h_0 > 0$, then $g[h(Q_1)] = g[h(Q_2)]$. However $-h_0 < h[\varepsilon Q_1 + (1-\varepsilon)Q_2] = (2\varepsilon - 1)h_0 < h_0$ for $0 < \varepsilon < 1$. Therefore given two points $\mathbf{f}_{Q_1}, \mathbf{f}_{Q_2} \in \mathbf{M}^*$ under the constraint $g[h(Q)] = h_0^2$ in Theorem 4, $\varepsilon\mathbf{f}_{Q_1} + (1-\varepsilon)\mathbf{f}_{Q_2} \notin \mathbf{M}^*$. Consequently, the solution from the gradient criterion can be a local one.

5. Numeric results

5.1. Toxicological data

We illustrate the VDM/ECM algorithm by fitting the constrained NPMLE to the toxicological data (Table 1) published in Weil, 1970, which was analyzed by Susko et al. (1998). The random variable x_{ij} is assumed to have a Binomial

Table 1
Toxicological data (Weil, 1970)

Control	x_{1j}	13	12	9	9	8	8	12	11	9	9	8	11	4	5	7	7
	m_{1j}	13	12	9	9	8	8	13	12	17	10	9	13	5	7	10	10
Treatment	x_{2j}	12	11	10	9	10	9	9	8	8	4	7	4	5	3	3	0
	m_{2j}	12	11	10	9	11	10	10	9	9	5	9	7	10	6	10	7

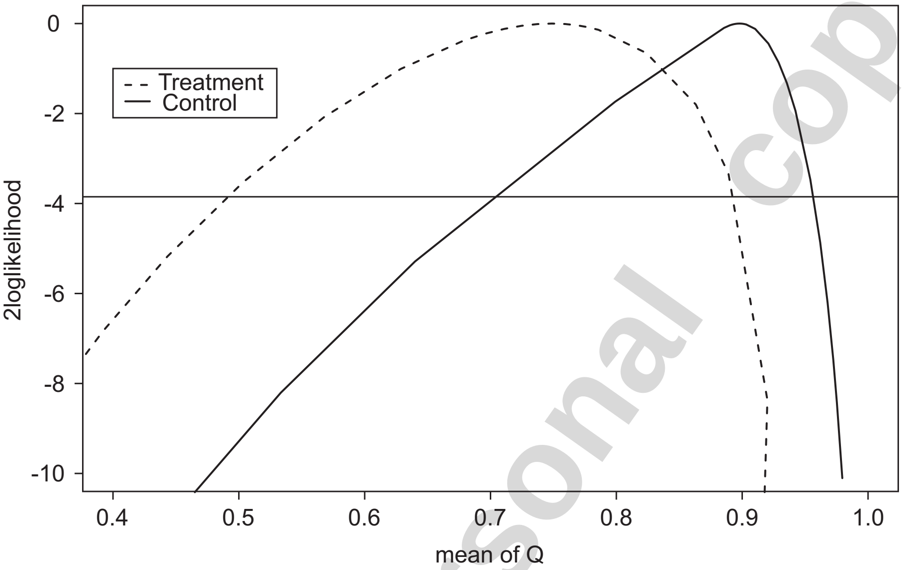

Fig. 1. Profile likelihood on the mean of mixing distribution for the toxicological data.

mixture distribution $Bin(m_{ij}, Q(p))$, where $i = 1, 2$ indexes the control or treatment group, respectively, $j = 1, \ldots, 16$ the replicates. We are interested in how the mean and variance of $Q(p)$ differ between the treatment and control groups.

The unpenalized NPMLE $\hat{Q}$ for the control group is: $\hat{\mathbf{\Lambda}}_1 = (0.85700, 0.94831)$ and $\hat{\boldsymbol{\pi}}_1 = (0.55236, 0.44764)$; for the treatment group, $\hat{\mathbf{\Lambda}}_2 = (0, 0.47180, 0.92250)$ and $\hat{\boldsymbol{\pi}}_2 = (0.05947, 0.26364, 0.67689)$. This gives a contrast of mean (control vs treatment) as 0.898 vs 0.749, and variance as 0.002 vs 0.074.

The mean of the mixing distribution is a linear functional, i.e. $\mu(Q) = \int p \, \mathrm{d}Q(p)$. Since the variance $\sigma^2(Q) = \int p^2 \, \mathrm{d}Q(p) - \mu^2(Q)$, by the linearization theorem, to maximize the penalized likelihood $\ell(Q) = \ell_0(Q) - \gamma\sigma^2(Q)$, we need to maximize the alternative objective function $\ell_1(Q) = \ell_0(Q) - \gamma \int p^2 \, \mathrm{d}Q - \gamma^* \mu(Q)$ iteratively. The NPMLE $\hat{Q}$ for $\ell(Q)$ must also be the NPMLE for $\ell_1(Q)$ at $\gamma^* = \gamma * 2 * \mu(\hat{Q})$. The nonparametric profile likelihood on $\mu(Q)$ and $\sigma^2(Q)$ are presented in Figs. 1 and 2 with a horizontal cutoff line $-\chi^2_{0.95,df=1} = -3.85$. The variance plot (Fig. 2) was essentially the same as reported in Fig. 3 of Susko et al. (1998), however, the profile plot for $\mu(Q)$ was slightly different from Fig. 1 of Susko et al. (1998). We suspect that the likelihood plot in Fig. 1 by Susko et al. (1998) was not 2*loglikelihood as labeled, but loglikelihood.

The two plots provide insights into $\mu(Q)$ and $\sigma^2(Q)$ for the control and treatment groups. For example, the control group appears to have a larger mean of p than the treatment group; the treatment group is more heterogeneous regarding the success rate p than the treatment group. However, because the nonparametric likelihood ratio statistic does not follow a standard χ^2 distribution, conclusions cannot be drawn formally.

5.2. Expressed sequence tag (EST) data

The second example (Table 2) is adapted from a genomic study by Wang and Lindsay (2005) where the total number of expressed genes in a cDNA library needs to be estimated using the expressed sequence tag (EST) data. Briefly a cDNA library is a population that contains millions of cDNA clones derived from N (unknown) distinct gene species.

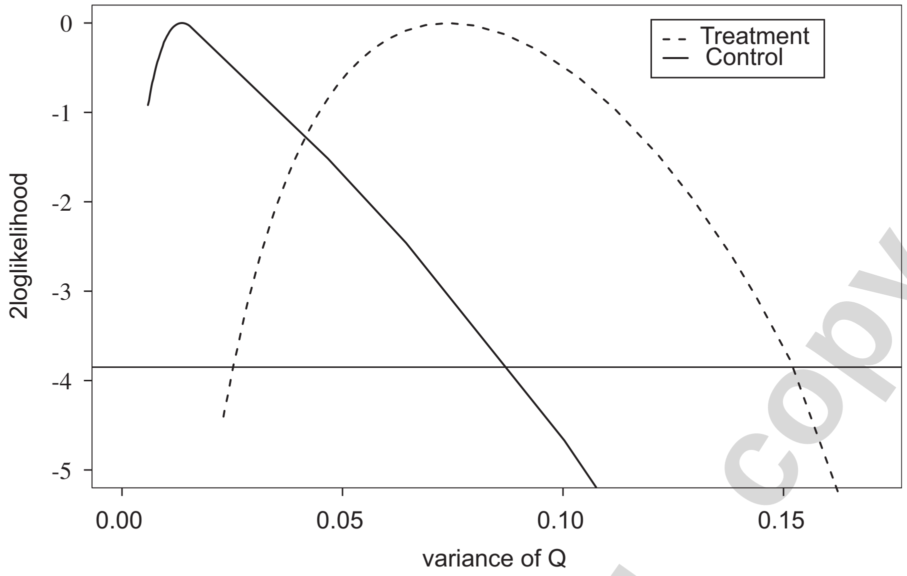

Fig. 2. Profile likelihood on the variance of mixing distribution for the toxicological data.

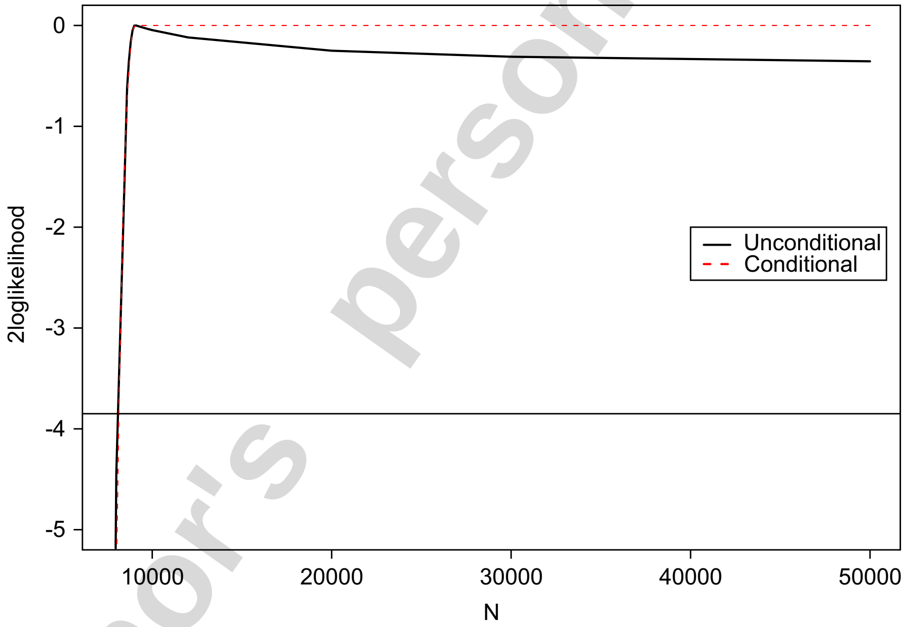

Fig. 3. Profile unconditional and conditional likelihood on *N* for EST data.

Table 2
Arabidopsis thaliana root EST data

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17^+	D
n_j	2187	490	133	121	37	51	22	19	7	8	6	7	6	4	5	5	18	3126

Each particular gene may have thousands of clones in the library. A random sample of S clones is obtained and they are classified into D distinct gene species. Let x_i, $i = 1, \ldots, D$ be the number of clones (individuals) sampled from the ith observed gene. The data then can be summarized by a sufficient statistic $\mathbf{n} = (n_1, \ldots, n_t)$ where $t = \max(x_i)$

and $n_j = \sum_i I\{x_i = j\}$. Assuming X_i follows a Poisson mixture distribution $f[x; Q(\lambda)] = \int \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}\,\mathrm{d}Q(\lambda)$, then the likelihood can be written as

$$\begin{aligned} L(N, Q) &= \binom{N}{n_1, \ldots, n_t} \prod_{j=0}^{t} [f(j; Q)]^{n_j} \\ &\propto \binom{N}{D} f(0; Q)^{N-D} [1 - f(0; Q)]^D \times \prod_{j>0} \left[\frac{f(j; Q)}{1 - f(0; Q)}\right]^{n_j} \\ &\equiv L_m(N, Q) \times L_c(Q). \end{aligned} \tag{14}$$

There exist two likelihood-based approaches in N-estimation, namely, conditional MLE or unconditional MLE (see Sanathanan, 1977). The conditional one is to find the MLE $\hat{Q}$ first from $L_c(Q)$, then estimate N based on the marginal likelihood by

$$\hat{N} = D\theta(\hat{Q}),$$

where $\theta(\hat{Q}) = \frac{f(0;\hat{Q})}{1-f(0;\hat{Q})}$. The log conditional likelihood can be expressed in a different form, involving probabilities of a mixture of zero-truncated Poisson distributions as follows (Wang and Lindsay, 2005):

$$\ell_c(P) = \sum_j n_j \log[g(j; P)], \tag{15}$$

where $g(j; P) = \int \frac{\mathrm{e}^{-\lambda}\lambda^j}{j!(1-\mathrm{e}^{-\lambda})}\,\mathrm{d}P(\lambda)$, and

$$\mathrm{d}P(\lambda) = \frac{(1 - \mathrm{e}^{-\lambda})\,\mathrm{d}Q(\lambda)}{\int (1 - \mathrm{e}^{-\lambda})\,\mathrm{d}Q(\lambda)}. \tag{16}$$

If Q is defined on $\Omega = (0, \infty)$, then

$$\mathrm{d}Q(\lambda) = \frac{(1 - \mathrm{e}^{-\lambda})^{-1}\,\mathrm{d}P(\lambda)}{\int (1 - \mathrm{e}^{-\lambda})^{-1}\,\mathrm{d}P(\lambda)}.$$

The $P - Q$ relationship is a re-weighting of the support points in the mixing distribution. Under P, the conditional likelihood can be regarded as that of i.i.d. observations from a mixture of zero-truncated Poisson distribution, and hence the NPMLE algorithm can be applied. In addition, the functional $\theta(Q)$ can be expressed as a linear functional of P i.e. $\theta(Q) \equiv \theta(P) = \int (\mathrm{e}^{\lambda} - 1)^{-1}\,\mathrm{d}P(\lambda)$. Therefore we can investigate the profile likelihood on $\theta(P)$, or equivalently on $N = D\theta(P)$.

To construct the profile likelihood on θ, we consider the following penalized likelihood:

$$\ell_c^{\gamma}(P) = \ell_c(P) - \gamma\theta(P). \tag{17}$$

Let $\hat{P}$ be the resulting NPMLE. If $\gamma > 0$, then clearly $\theta(\hat{P}) < \infty$. If $\gamma = 0$, then $\hat{P}$ may include a component $\hat{\lambda}_{(1)}$ that can be arbitrarily close to zero (Wang and Lindsay, 2005), such that $\theta(\hat{P}) \approx \infty$. Note that in the Poisson distribution, if one component has $\lambda = 0$, then it only generates 0. However, if $\lambda = 0^+$, then for the zero-truncated probability g, we have $\lim_{\lambda \to 0} g(1; \lambda) = 1$. The algorithm can push the smallest component in the mixture to be arbitrarily close to zero to improve the fit at $j = 1$. The global maximum likelihood is often obtained after including a component that is essentially zero. If $\gamma < 0$, since $\theta(\lambda)$ is unbounded above in Ω, and hence so is the maximum penalized likelihood form (17).

Let $\hat{\theta}_0$ be the plug-in estimate at the unpenalized NPMLE $\hat{P}_0$ (at $\gamma = 0$). If $\hat{\theta}_0 < \infty$ then the profile likelihood at $\theta = \hat{\theta}_0$ is $\ell_c(\hat{P}_0)$. At $\theta > \hat{\theta}_0$, since including a tiny component with a negligible weight can inflate $\theta(\hat{P})$ to any value while the change of likelihood value can be controlled at an arbitrarily tiny level, the profile likelihood will be equal to $\ell_c(\hat{P}_0)$ from a limit point of view. On the other hand, if $\hat{\theta}_0 = \infty$ then the profile likelihood at $\theta < \hat{\theta}_0$ can be obtained by tuning $\gamma > 0$.

Alternatively one can investigate the profile full likelihood on N based on (14). For a given N corresponding to $\theta_N = \frac{N}{D} - 1$, we can fit NPMLE for a Poisson mixture to obtain $\hat{Q}_{\theta_N}$. The $\hat{N}$ that attains the maximum profile likelihood is the *unconditional* estimator of N. For EST data in Table 2, the profile conditional likelihood and full likelihood are compared in Fig. 3.

Fig. 3 displays very similar relationship between $\hat{N}$ and the two profile likelihoods. The profile unconditional likelihood becomes flat rapidly when N increases above $\hat{N}$, and it eventually converges to a constant as $N \to \infty$ (the minimum N that attains the maximum profile conditional likelihood is larger than that from the unconditional likelihood. This is not clearly displayed from Fig. 3 because of the scale of plot. Theoretical justification of this can be found in Wang and Lindsay, 2005). This behavior is driven by the properties of the likelihood function. As N increases, a component $\lambda = 0$ is added in $\hat{\mathbf{\Lambda}}$ and its weight increases. Given a mixing distribution Q_α with $\mathbf{\Lambda} = (0, \mathbf{\Lambda}_1^{\mathrm{T}})$ and $\boldsymbol{\pi}_\alpha = (\alpha, (1-\alpha)\boldsymbol{\pi}_1^{\mathrm{T}})$, one can show that the conditional likelihood does not change with α if $\mathbf{\Lambda}_1^{\mathrm{T}}$ and $\boldsymbol{\pi}_1^{\mathrm{T}}$ are fixed. This implies that if $(\mathbf{\Lambda}, \boldsymbol{\pi}_\alpha)$ gives the maximum conditional likelihood, then Q_{α^*} with $(\mathbf{\Lambda}, \boldsymbol{\pi}_{\alpha^*})$ also does $\forall \alpha^* \in [0, 1)$. For a given N the marginal likelihood is maximized at $f(0) = \frac{N-D}{N}$. Hence as N climbs up, the algorithm tends to keep $(\mathbf{\Lambda}_1^{\mathrm{T}}, \boldsymbol{\pi}_1^{\mathrm{T}})$ unchanged, while tuning α to make $\hat{Q}_\alpha$ maximize the conditional and marginal likelihood simultaneously. Since the log marginal likelihood with $f(0; Q_\alpha) = \frac{N-D}{N}$ converges to $\log(\frac{\mathrm{e}^{-D} D^D}{D!})$ (Wang and Lindsay, 2005, Eq. (9)), the profile unconditional likelihood converges.

If there were a cutoff line for a given confidence level, e.g. -3.85 for 95% for a $\chi^2_{df=1}$ distribution, then the resulting confidence interval from both profiles are very similar. In particular the upper bound is ∞. This phenomenon reflects the theoretical possibility that under the infinite population assumption there can be many extremely rare genes in the cDNA library. On the other hand, since in many real problems $N \ll \infty$, the NPMLEs can be unstable because a small/tiny component is often fit in $\hat{P}$. Imposing some *ad hoc* penalties or priors can greatly improve the plug-in estimator $\hat{N}$ (Wang and Lindsay, 2005).

6. Discussion

The contribution of this paper is VDM/ECM algorithm and a linearization procedure for the penalized or constrained NPML estimation when the penalty or constraint is linear or a function of linear functional(s). This algorithm is EM-based, and thus works very reliably. The number of outer iterations (VDM/ECM) is often roughly equal to the number of components in the NPMLE solution since in each iteration a new component is added into the support and then updated. Within each VDM/ECM iteration, the embedded EM for the finite mixture can be time-consuming, depending on the data set size and fraction of missing information (McLachlan and Krishnan, 1997). The author experienced difficulty in implementing the existing ISDM algorithm by Lesperance and Kalbfleisch (1992) and Susko et al. (1998), and therefore cannot make a quantitative comparison of the two algorithms regarding performance. The codes were written in R and JAVA and available on request. The running time for the JAVA codes can range from several milliseconds to several seconds on a Linux machine with CPU frequency 1.7 GHz. For the EST data example, the average running time for any fixed penalty is about 2 s.

The usefulness of the algorithm depends on the theoretical development the NPML estimation of the mixture models. Unfortunately the knowledge on general theory in asymptotics beyond weak consistency remain limited (if exists). We illustrated the algorithm with two real examples. The profile likelihood plots in both case studies provide insights into the parameter under comparison or estimation, however, rigorous conclusion is undrawable at this moment.

Acknowledgment

The author thanks the reviewer for very helpful and detailed comments.

References

Böhning, D., 1985. Numerical estimation of a probability measure. J. Statist. Planning Inference 11, 57–69.

Böhning, D., Schlattman, P., Lindsay, B.G., 1992. Computer assisted analysis of mixtures (c.a.man): statistical algorithms. Biometrics 48, 283–304.

Bunge, J., Fitzpatrick, M., 1993. Estimating the number of species: a review. J. Amer. Statist. Assoc. 88, 364–373.

Chen, J., Kalbfleisch, J.D., 1996. Penalized minimum distance estimates in finite mixture models. Canad. J. Statist. 2, 167–176.

DerSimonian, R., 1986. Maximum likelihood estimation of a mixing distribution. J. Roy. Statist. Soc. Ser. C 35, 302–309.

Fisher, R.A., Corbet, A.S., Williams, C.B., 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. J. Animal Ecology 12, 42–58.

Kiefer, J., Wolfowitz, J., 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Statist. 27, 887–906.

Laird, N.M., 1978. Nonparametric maximum likelihood estimation of a mixing distribution. J. Amer. Statist. Assoc. 73, 805–811.

Leroux, B.G., 1992. Consistent estimation of a mixing distribution. Ann. Statist. 20, 1350–1360.

Lesperance, M., Kalbfleisch, J.D., 1992. An algorithm for computing the nonparametric MLE of a mixing distribution. J. Amer. Statist. Assoc. 87, 120–126.

Lindsay, B.G., 1983. The geometry of mixture likelihoods: a general theory. Ann. Statist. 11, 86–94.

Lindsay, B.G., 1995. Mixture models: theory, geometry and applications, vol. 5. Institute of Mathematical Statistics.

Lindsay, B.G., Roeder, K., 1993. Uniqueness of estimation and identifiability in mixture models. Canad. J. Statist. 87 (21), 139–147.

Luenberger, G.D., 1969. Optimization by Vector Space Methods. Wiley, New York.

McLachlan, G., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York, NY.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

Meng, X.-L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80, 267–278.

Sanathanan, L., 1977. Estimating the size of a truncated sample. J. Amer. Statist. Assoc. 72, 669–672.

Susko, E., Kalbfleisch, J.D., Chen, J., 1998. Constrained nonparametric maximum-likelihood estimation for mixture models. Canad. J. Statist. 26, 601–617.

Titterington, D., Smith, A., Makov, U., 1985. Statistical Analysis of Finite Mixture Distributions. Wiley, New York.

Van de Geer, S., 2002. Asymptotic theory for maximum likelihood in nonparametric mixture models. Comput. Statist. Data Anal. 41, 453–464.

Wang, J.-P.Z., Lindsay, B.G., 2005. A penalized nonparametric maximum likelihood approach to species richness estimation. J. Amer. Statist. Assoc. 100, 942–959.

Weil, C., 1970. Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. Food Cosmetics Toxicology 31, 177–182.