Ann Inst Stat Math (2016) 68:269–297
DOI 10.1007/s10463-014-0490-9

Parameterizing mixture models with generalized moments

Zhiyue Huang · Paul Marriott

Received: 23 September 2013 / Revised: 12 July 2014 / Published online: 15 October 2014

Abstract This paper considers a new way of parameterizing mixture models where parameters are interpreted as the generalized moments of the mixing distribution. Following a dimensionality reduction approach, approximate models have a finite-dimensional parameter with a corresponding parameter space: a moment space. The geometry of the moment space is studied and we derive the properties of the reconstructed mixing distributions. Links between the reparameterization and estimation methods for mixture models are also briefly discussed.

Keywords Moments · Chebyshev system · Local mixture models · Functional principle component analysis

1 Introduction

Mixture models can be found in a wide variety of statistical applications; a comprehensive introduction can be found in McLachlan and Peel (2000). This paper considers the class of non-parametric mixtures of exponential families

$$f_{\mathrm{Mix}}(x; Q) := \int_a^b f(x; \theta)\mathrm{d}Q(\theta), \quad x \in \mathcal{S},$$

Z. Huang (✉) · P. Marriott
Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West,
Waterloo, ON N2L 3G1, Canada
e-mail: z23huang@uwaterloo.ca

P.Marriott
e-mail: pmarriot@math.uwaterloo.ca

where for each $\theta \in [a, b]$, the component distribution $f(x; \theta)$ is in the exponential family, i.e.

$$f(x; \theta) = h(x) \exp(\eta(\theta) T(x) - A(\theta)),$$

where $T(x)$, $h(x)$, $\eta(\theta)$ and $A(\theta)$ are known functions, and the mixing distribution $Q(\theta)$ is a probability measure over a known, compact set $[a, b]$.

Example 1 Consider a general mixture of Poisson distributions,

$$f_{\text{Mix}}(x; Q) = \int_a^b \text{Pois}(x; \theta) \mathrm{d}Q(\theta),$$

where $\text{Pois}(x; \theta)$ is the probability mass function of a Poisson distribution with mean θ and $Q(\theta)$ is a probability measure over a compact set $[a, b]$. Because the above model takes more variability into account than a single Poisson distribution, it has been used in a wide range of scientific fields for modelling non-homogeneous populations; see Schlattmann (2009).

As an example, consider a Thailand-based cohort study analyzed by Böhoning (1995). To study the health status of 602 pre-school children, the number of times that a child who showed symptoms of fever, a cough, a runny nose, or these symptoms together, is recorded. Previous studies showed the existence of overdispersion. Therefore, a mixture of Poisson is suggested.

Example 2 Consider a general mixture of normal distributions,

$$f_{\text{Mix}}(x; Q) = \int_a^b N(x; \theta, \sigma^2) \mathrm{d}Q(\theta),$$

where $N(x; \theta, \sigma^2)$ is the probability density function of a normal distribution with mean θ and variance σ^2, and $Q(\theta)$ is a probability measure over a compact region $[a, b]$. The above model has wide applications due to its flexibility; see McLachlan and Peel (2000).

One possible application of a mixture of normals is to assess the impact of possible underlying genotypes that display continuous variation in the population. Roeder (1994) gave an example of the sodium–lithium countertransport (SLC) activity in red blood cells. The data set consists of the SLC activity from 190 individuals. It is suspected that the SLC activity is mainly affected by a gene with two alleles. This implies that there are three possible genotypes. And thus, it is reasonable to consider a mixture of normals with at most three components; see Roeder (1994).

A general, or nonparametric, mixture model has an infinite dimensional parameter space. In frequentist statistics, the maximum likelihood estimator with an infinite dimensional parameter may not be consistent and may not be efficient in the sense that the Cramer–Rao bound is not attained even asymptotically; see Neyman and Scott (1948). In Bayesian statistics, a prior on an infinite dimensional space is not easily defined and can be highly informative even with large amounts of data; see Marriott

(2007). To deal with this issue, we can use the modified likelihood (Lindsay 1980) or reduce the dimension of the parameter space (Marriott 2002, 2007). This paper follows the dimensionality reduction idea and makes several contributions.

Firstly, we give a general framework for the reparameterization of a mixture model with a complete orthonormal basis in a Hilbert space. The new parameters are interpreted as the moments of the mixing distribution $Q(\theta)$, which are induced by Chebyshev systems (defined in Sect. 2).

Definition 1 Let $\{u_i(\theta)\}_{i=0}^r$ form a Chebyshev system over $[a, b]$. For $i = 0, 1, \dots, r$, the *ith moment* of a probability measure $Q(\theta)$ induced by $\{u_i(\theta)\}_{i=0}^r$ is defined as:

$$m_i(Q) := E_\theta[u_i(\theta); Q] = \int_a^b u_i(\theta) \mathrm{d}Q(\theta) < \infty.$$

Secondly, after approximating the reparameterized model using a dimensionality reduction technique, we introduce the moment space (defined in Sect. 3) as the parameter space of the approximated model. The moment space is natural from the point of view of the interpretation of the parameters and allows us to reconstruct the mixing distribution from the moments.

Lastly, we study the geometry of the moment space, showing important properties of the reconstructed mixing distributions, including results on existence, uniqueness, number of support points and gradient characterization.

In Sect. 2, we illustrate how to reparameterize a mixture model by a complete orthonormal basis and interpret the new parameters. In Sect. 3, we approximate the reparameterized model by a model with a finite-dimensional parameter and study the quality of the approximation. In Sect. 4, we study the geometry of the moment space in two ways: the positive representation and the gradient characterization. In Sect. 5, we show two real examples to demonstrate the application of the reparameterization method. Lastly, we discuss the links between the reparameterization and two estimation methods: the method of moments and the maximum likelihood estimator.

2 Parameterization in moments

Consider a measure space $(\mathcal{S}, \Sigma, \mu_0)$ and the $L^2(\mathcal{S}, \mu_0)$ space induced by it. We assume μ_0 is a probability measure with support $\mathcal{S}$. We will also denote, where appropriate, $\mu_0(x) = f_0(x)\mu(x)$ with respect to a fixed measure μ, typically Lebesgue or a counting measure. The probability function $f_0(x)$ is either fully known or lies in a parametric family. Let the set $\{e_i(x)\}_{i=0}^\infty$ form a complete orthonormal basis of $L^2(\mathcal{S}, \mu_0)$, i.e.

$$\langle e_i(x), e_j(x)\rangle_{L^2(\mathcal{S},\mu_0)} = \int_{\mathcal{S}} e_i(x) e_j(x) f_0(x) \mathrm{d}x = \delta_{ij},$$

where δ is the Kronecker delta.

Assume that for each $\theta \in [a, b]$, the function $f(x; \theta)/f_0(x)$ belongs to $L^2(\mathcal{S}, \mu_0)$, i.e., $\int_{\mathcal{S}} f(x; \theta)^2 / f_0(x) dx < \infty$. According to standard results in Hilbert spaces

(Debnath and Mikusiński 1999), we have the expansion

$$f(x;\theta) = \sum_{i=0}^{\infty} u_i(\theta) e_i(x) f_0(x), \tag{1}$$

where for each $i \in \{0, 1, \dots\}$,

$$u_i(\theta) = \left\langle e_i(x), \frac{f(x;\theta)}{f_0(x)} \right\rangle_{L^2(\mathcal{S},\mu_0)}. \tag{2}$$

Taking expectations on both sides of (1), with respect to θ under the mixing distribution $Q(\theta)$, gives

$$f_{\mu_0}(x; \boldsymbol{m}_\infty) = \sum_{i=0}^{\infty} m_i(Q) e_i(x) f_0(x),$$

where $\boldsymbol{m}_\infty = (m_1, m_2, \dots)^{\mathrm{T}} \in R^\infty$, and for each i,

$$m_i(Q) = \int_a^b u_i(\theta) \mathrm{d}Q(\theta) = E_\theta[u_i(\theta); Q].$$

When $e_0(x) \equiv 1$, $x \in \mathcal{S}$, we have $u_0(x) = \int_{\mathcal{S}} 1 \times \frac{f(x;\theta)}{f_0(x)} f_0(x) \mathrm{d}x = 1$, and so gives a reparameterization

$$f_{\mu_0}(x; \boldsymbol{m}_\infty) = f_0(x) + \sum_{i=1}^{\infty} m_i(Q) e_i(x) f_0(x). \tag{3}$$

Furthermore, by orthogonality, for $i > 0$, we have $\int_{\mathcal{S}} 1 \times e_i(x) f_0(x) \mathrm{d}x = 0$ so that (3) integrates to one.

Definition 2 The set of functions $\{u_i(\theta)\}_{i=0}^{r}$ is a *Chebyshev system* over $\Theta \subseteq R$, if we have $\det(u_i(\theta_j))_{i,j=0}^{r} > 0$ whenever $\theta_0 < \theta_1 < \cdots < \theta_r$ and $\theta_j \in \Theta$, $j = 0, 1, \dots, r$.

Definition 3 If $\{u_i(\theta)\}_{i=0}^{\infty}$ in (2) forms a Chebyshev system over $[a, b]$ with $u_0(\theta) \equiv 1$, then for a given $f_0(x)$, $x \in \mathcal{S}$, the formula (3) is *the reparameterization of the mixture model $f(x; Q)$ in the moments induced by* $\{u_i(\theta)\}_{i=0}^{\infty}$.

When $e_0(x) \equiv 1$, the reparameterization of the mixture model $f_{\mathrm{Mix}}(x; Q)$ is locally defined by $\{e_i(x)\}_{i=0}^{\infty}$ at $f_0(x)$. The choice of the basis depends on the specific inference problem. Here, we give an example where the mean of the mixed distribution, rather than the whole of the mixing distribution, is of inferential interest. Other examples will be seen in Sect. 3.

2.1 The moments induced by power functions

This subsection considers inference for the mean parameter θ_0 in the mixture model

$$f_{\text{Mix}}(x;\theta_0,Q)=\int_a^b f(x;\theta_0+\eta)\mathrm{d}Q(\eta),$$

where $Q(\eta)$ is a zero mean, for identification reasons, probability measure over $[a,b]$. Furthermore, the component distributions are natural exponential models with quadratic variance functions. This class includes the normal, Poisson, gamma, binomial and negative binomial families; see Morris (1982, 1983), and has the following formal definition.

Definition 4 If $f(x;\theta)$ is a natural exponential family in the mean parameterization, then $V_f(\theta)$, defined by $V_f(\theta){:=}E_X[(X-\theta)^2]$, is called *the variance function*. If the variance function $V_f(\theta)$ is quadratic with the form $V_f(\theta)=v_0+v_1\theta+v_2\theta^2$, then we say $f(x;\theta)$ is a *natural exponential family with quadratic variance function*.

In this problem θ_0 is the parameter of interest and the mixing distribution Q is considered as a nuisance parameter. The mixture model $f_{\text{Mix}}(x;\theta_0,Q)$ could be expanded by a Laplace expansion; see Marriott (2007). Here, we describe this process within our new framework.

Following (Morris 1983), we define, for $i=0,1,2,\ldots,$

$$P_i(x;\theta)=\frac{V_f^i(\theta)}{f(x;\theta)}\frac{\partial^i}{\partial\theta^i}f(x;\theta),$$

$a_i=i!\prod_{j=0}^{i-1}(1+jv_2)\equiv i!b_i$, and let $f_0(x)=f(x;\theta_0)$. Morris (1983) showed that

$$\langle P_i(x;\theta_0),P_j(x;\theta_0)\rangle_{L^2(\mathcal{S},\mu_0)}=\delta_{ij}a_iV_f^i(\theta_0)$$

and

$$\left\langle P_i(x;\theta_0),\frac{f(x;\theta)}{f_0(x)}\right\rangle_{L^2(\mathcal{S},\mu_0)}=\int_{\mathcal{S}}P_i(x;\theta_0)f(x;\theta)\mathrm{d}x=b_i(\theta-\theta_0)^i=b_i\eta^i.$$

For a given $\theta_0\in(a,b)$, a mixture of natural exponential families with quadratic variance functions can be reparameterized as:

$$f(x;\boldsymbol{m}_{\theta_0,\infty})=f(x;\theta_0)+\sum_{i=2}^{\infty}m_i(Q)\frac{1}{i!}\frac{P_i(x;\theta_0)}{V_f^i(\theta_0)}f(x;\theta_0),\tag{4}$$

where $\boldsymbol{m}_{\theta_0,\infty}=(\theta_0,m_2,m_3,\ldots)^{\mathrm{T}}\in R^{\infty}$ and for each $i=1,2,\ldots,$

$$m_i(Q)=\int_a^b\eta^i\mathrm{d}Q(\eta).$$

For each i, $u_i(\eta) = \eta^i$. The set $\{u_i(\eta)\}_{i=0}^{\infty}$ is a power set in a compact region and so forms a Chebyshev system. So, (4) is a reparameterization of mixture models in the moments induced by power functions.

3 Dimensionality reduction in parameter spaces

In the reparameterization of mixture model in Definition 3, we still have a non-finite dimensional parameter space which could be problematic from an inferential point of view; see Marriott (2007). Following Marriott (2002, 2007), we approximate (3) with a finite sum

$$f_{\mu_0}(x; \boldsymbol{m}_c) := f_0(x) + \sum_{i=1}^{r} m_i(Q) e_i(x) f_0(x), \tag{5}$$

where $\boldsymbol{m}_c = (m_1, m_2, \ldots, m_r)^{\mathrm{T}} \in R^r$. The error between $f_{\mu_0}(x; \boldsymbol{m}_c)/f_0(x)$ and $f_{\mu_0}(x; \boldsymbol{m}_\infty)/f_0(x)$ in $L^2(\mathcal{S}, \mu_0)$ is

$$\begin{aligned} e_r(Q) &:= \left\| \left(f_{\mu_0}(x; \boldsymbol{m}_c) - f_{\mu_0}(x; \boldsymbol{m}_\infty) \right) / f_0(x) \right\|^2_{L^2(\mathcal{S}, \mu_0)} \\ &= \int_{\mathcal{S}} \left(\sum_{i=r+1}^{\infty} m_i(Q) e_i(x) \right)^2 f_0(x) \mathrm{d}x \\ &= \sum_{i=r+1}^{\infty} m_i^2(Q) < \infty. \end{aligned} \tag{6}$$

3.1 The moments induced by an integral operator

To find a good finite dimensional approximation, one approach is to find an orthonormal basis $\{\gamma_i(x)\}_{i=0}^{\infty}$ in $L^2(\mu_0, \mathcal{S})$ which minimizes $e_r(Q)$ for a given r. However since the error $e_r(Q)$ depends on the unknown mixing distribution Q, instead we aim to minimize an upper bound of $e_r(Q)$ instead. By the Cauchy–Schwarz inequality, an upper bound is

$$\int_a^b (\mathrm{d}Q(\theta)/\mathrm{d}\theta)^2 \mathrm{d}\theta \times \sum_{i=r+1}^{\infty} \int_a^b u_i^2(\theta) \mathrm{d}\theta. \tag{7}$$

With the assumption that $\int_a^b (\mathrm{d}Q(\theta)/\mathrm{d}\theta)^2 \mathrm{d}\theta$ is finite, minimizing $\sum_{i=r+1}^{\infty} \int_a^b u_i^2(\theta) \mathrm{d}\theta$ is equivalent to minimizing the upper bound.

We can obtain a basis $\{\gamma_i(x)\}_{i=0}^{\infty}$ from the eigenfunctions $\{\tilde{\gamma}_i(x)\}_{i=0}^{\infty}$ to the integral operator

$$(Ag)(x) := \int_{\mathcal{S}} g(y) K(x, y) \mathrm{d}y < \infty, \tag{8}$$

with the kernel function

$$K(x, y) := \int_a^b \frac{f(x;\theta)}{f_0^{1/2}(x)} \frac{f(y;\theta)}{f_0^{1/2}(y)} \mathrm{d}\theta, \quad (x, y) \in \mathcal{S} \times \mathcal{S}.$$

This integral operator is positive and self-adjoint; see Debnath and Mikusiński (1999). If $A(\cdot)$ is also compact, the set $\{\gamma_i(x)\}_{i=0}^{\infty}$ forms the complete orthonormal basis in $L^2(\mu_0, \mathcal{S})$ that minimizes $\sum_{i=r+1}^{\infty} \int_a^b u_i^2(\theta)\mathrm{d}\theta$, where for each i,

$$\gamma_i(x) := \tilde{\gamma}_i(x)/f_0^{1/2}(x), \quad x \in \mathcal{S}; \tag{9}$$

see the results on functional principle component analysis in Horváth and Kokoszka (2012). To show the compactness of $A(\cdot)$, one sufficient condition is that

$$\int_a^b \int_{\mathcal{S}} f^2(x;\theta)/f_0(x)\mathrm{d}x\mathrm{d}\theta < \infty;$$

see (16) in the Appendix.

To have $\gamma_0(x) \equiv 1$ and $\tilde{\gamma}_0(x) = f_0^{1/2}(x)$, we need

$$\begin{aligned} \lambda_0 f_0^{1/2}(y) &= \int_{\mathcal{S}} f_0^{1/2}(x) K(x, y)\mathrm{d}x \\ &= f_0^{-1/2}(y) \int_{\mathcal{S}} \int_a^b f(x;\theta) f(y;\theta)\mathrm{d}\theta\mathrm{d}x \\ &= f_0^{-1/2}(y) \int_a^b f(y;\theta)\mathrm{d}\theta. \end{aligned}$$

It follows that $\lambda_0 = b - a$ and $\gamma_0(x) \equiv 1$ when $f_0(x) = \frac{1}{b-a} \int_a^b f(x;\theta)\mathrm{d}\theta$.

Theorem 1 *For each* $i = 0, 1, \ldots,$ *let*

$$\phi_i(\theta) := \left\langle \gamma_i(x), \frac{f(x;\theta)}{f_0(x)} \right\rangle_{L^2(\mathcal{S},\mu_0)} = \int_{\mathcal{S}} f(x;\theta)\gamma_i(x)\mathrm{d}x < \infty, \quad \theta \in [a, b],$$

where $\gamma_i(x)$ *is defined by* (9). *For each* $r = 1, 2, \ldots$, *the set of functions* $\{\phi_i(\theta)\}_{i=0}^{r}$ *forms a Chebyshev system over* $[a, b]$.

Proof See the Appendix.

Example 1 (continued) Let $[a, b] = [0, 25]$ and

$$f_0(x) = \frac{1}{25} \int_0^{25} \mathrm{Pois}(x;\theta)\mathrm{d}\theta.$$

It can be shown that

$$\sum_{x=0}^{\infty} (\mathrm{Pois}(x;\theta))^2 / f_0(x) < \infty,$$

for each $\theta \in [0, 25]$. Figure 1 shows the largest 10 eigenvalues of the integral operator $A(\cdot)$ in (8), the functions $\gamma_i(x)$ and $\phi_i(\theta)$ corresponding to the largest 4 eigenvalues.

Example 2 (continued) For each fixed $\sigma^2 \geq 0$, let $[a, b] = [0, 0.7]$ and

$$f_0(x) = \frac{1}{0.7} \int_0^{0.7} N(x; \theta, \sigma^2) \mathrm{d}\theta.$$

It can be shown that

$$\int_{-\infty}^{\infty} \left(N(x; \theta, \sigma^2)\right)^2 / f_0(x) \mathrm{d}x < \infty,$$

for each $\theta \in [a, b]$. For $\sigma^2 = 0.07^2$, Fig. 2 shows the largest 10 eigenvalues of the integral operator $A(\cdot)$ in (8), the functions $\gamma_i(x)$ and $\phi_i(\theta)$ corresponding to the largest 4 eigenvalues.

3.2 The quality of the approximation

We need to consider the parameter space for the new parameters, $\boldsymbol{m}_c$, and also reconstruct the mixing distribution Q from $\boldsymbol{m}_c$. Recall that the new parameters are interpreted as the moments of Q. Then, it is natural to consider the moment space in Definition 5 as the parameter space of (5). More nice properties of the moment space will be seen in Sect. 4.

Definition 5 Let $\{u_i(\theta)\}_{i=0}^r$ form a Chebyshev system over $[a, b]$ with $u_0(\theta) \equiv 1$. *The moment space induced by* $\{u_i(\theta)\}_{i=0}^r$ is

$$\mathcal{K}_r = \left\{ \boldsymbol{m}_c = (m_1, m_2, \ldots, m_r)^{\mathrm{T}} \in R^r | \boldsymbol{m}_c = \int_a^b \boldsymbol{u}_c(\theta) \mathrm{d}Q(\theta) \right\},$$

where $\boldsymbol{u}_c(\theta) = (u_1(\theta), u_2(\theta), \ldots, u_r(\theta))^{\mathrm{T}} \in R^r$ and Q is a probability measure over $[a, b]$.

Note that there is no requirement that $f_{\mu_0}(x; \boldsymbol{m}_c)$ be a non-negative function for any $\boldsymbol{m}_c \in \mathcal{K}_r$. To ensure that $f_{\mu_0}(x; \boldsymbol{m}_c)$ behaves like a probability density (or mass) function, the non-negative condition for each $x \in \mathcal{S}$ also should be considered; see Marriott (2002, 2007). However, the non-negative conditions could be omitted in some cases, as we will see in Sect. 4.2.

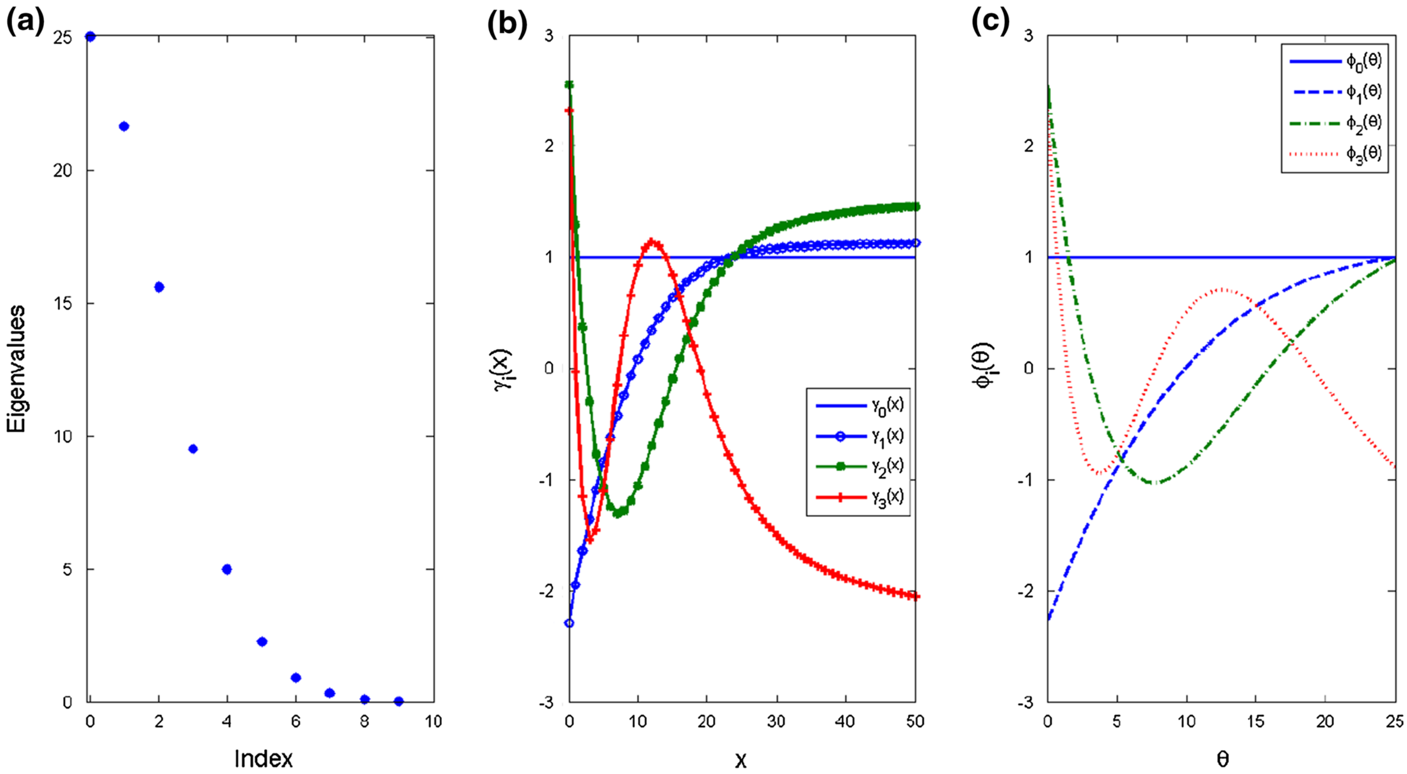

Fig. 1 Plot of **a** the largest 10 eigenvalues, **b** the functions $\gamma_i(x)$, **c** the functions $\phi_i(\theta)$, in the mixture of Poisson for $i = 0, 1, 2, 3$

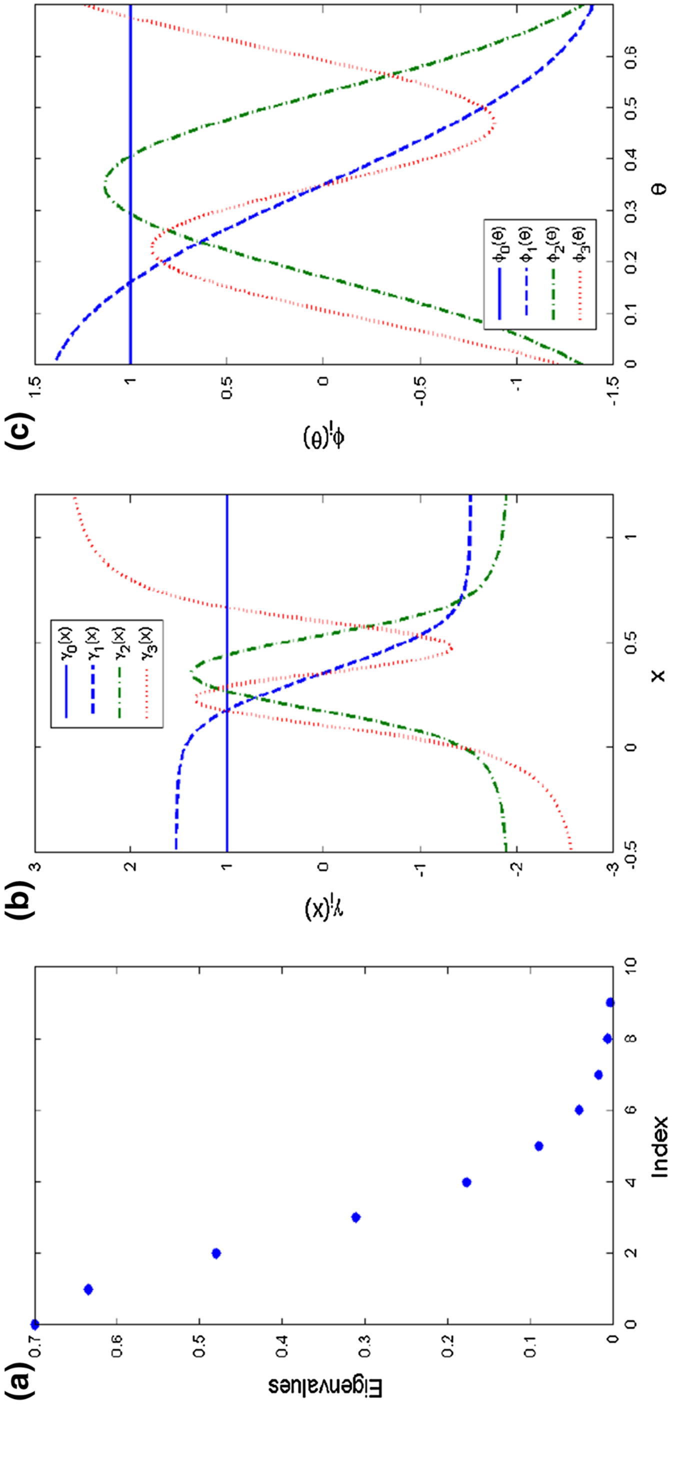

Fig. 2 Plot of **a** the largest 10 eigenvalues, **b** the functions $\gamma_i(x)$, **c** the functions $\phi_i(\theta)$, in the mixture of normal for $i = 0, 1, 2, 3$

By (6), we have, for any mixing distribution Q over $[a, b]$,

$$\min_{\boldsymbol{m}_c \in \mathcal{K}_r} \| \left(f_{\mu_0}(x; \boldsymbol{m}_c) - f_{\mathrm{Mix}}(x; Q) \right) / f_0(x) \|^2_{L^2(\mathcal{S}, \mu_0)} \leq e_r(Q).$$

Furthermore, if $f_{\mu_0}(x; \boldsymbol{m}_c)$ is reparameterized with the moments induced by $\{\phi_i(\theta)\}_{i=0}^{\infty}$, we can have

$$\begin{aligned} e_r(Q) &\leq \int_a^b (\mathrm{d}Q(\theta)/\mathrm{d}\theta)^2 \mathrm{d}\theta \times \sum_{i=r+1}^{\infty} \int_a^b \phi_i^2(\theta)\mathrm{d}\theta \\ &= \int_a^b (\mathrm{d}Q(\theta)/\mathrm{d}\theta)^2 \mathrm{d}\theta \times \sum_{i=r+1}^{\infty} \int_{\mathcal{S}} \tilde{\gamma}_i(x)(A\tilde{\gamma}_i)(x)\mathrm{d}x \\ &= \int_a^b (\mathrm{d}Q(\theta)/\mathrm{d}\theta)^2 \mathrm{d}\theta \times \sum_{i=r+1}^{\infty} \lambda_i. \end{aligned}$$

We return to Examples 1 and 2 to examine the non-negativeness and the quality of the approximation of $f_{\mu_0}(x; \boldsymbol{m}_c)$ over the sample space. Because each $f_{\mu_0}(x; \boldsymbol{m}_c)$ can be written as a convex combination of $f_{\mu_0}(x; \boldsymbol{u}_c(\theta))$, we aim to study the approximation of $f_{\mu_0}(x; \boldsymbol{u}_c(\theta))$ to the component distribution $f(x; \theta)$ at each $(x, \theta) \in \mathcal{S} \times [a, b]$. The quality of the approximation is measured by

$$\epsilon_4(x; \theta) = f_{\mu_0}(x; \boldsymbol{u}_c(\theta)) - f(x; \theta) \tag{10}$$

for each $(x, \theta) \in \mathcal{S} \times [a, b]$.

Example 1 (continued) Let $\boldsymbol{u}_c(\theta) = (u_1(\theta), \ldots u_4(\theta))^{\mathrm{T}} \in R^4$. We consider the cases that $\boldsymbol{u}_c(\theta)$ is induced by $\{(\theta - 12.5)^i\}_{i=0}^4$ in Sect. 2.1 and $\{\phi_i(\theta)\}_{i=0}^4$ in Sect. 3.1. Figure 3 shows the negative region of $f_{\mu_0}(x; \boldsymbol{u}_c(\theta))$ over $\mathcal{S} \times [0, 25]$ in these two cases.

Figure 4 examines the quality of the approximation. Various issues of the reparameterization with power moments are seen from panel (a). Firstly, the quality of the approximation is non-uniform at each point in the sample space. Secondly, the approximation is poor when θ is away from $\theta = \theta_0$. This is due to the nature of the underlying Laplace approximation where a polynomial approximation only behaves well in a small neighbourhood of $\theta = \theta_0$. On the other hand, from the panel (b), we see that the quality of the approximation is almost uniform at each point $(x, \theta) \in \mathcal{S} \times [a, b]$, when the moments are induced by $\{\phi_i(\theta)\}_{i=0}^4$.

Example 2 (continued) Consider a fixed $\sigma^2 = 0.07^2$. Again consider $\boldsymbol{u}_c(\theta) = (u_1(\theta), \ldots u_4(\theta))^{\mathrm{T}} \in R^4$ and $\boldsymbol{u}_c(\theta)$ is induced by either $\{(\theta - 0.35)^i\}_{i=0}^4$ or $\{\phi_i(\theta)\}_{i=0}^4$. Figure 5 shows the negative regions of $f_{\mu_0}(x; \boldsymbol{u}_c(\theta))$ over $\mathcal{S} \times [0, 0.7]$ under these two types of reparameterizations. Also, Fig. 6 gives the contour plots of $\epsilon_4(x; \theta)$ over $\mathcal{S} \times [0, 0.7]$. From the panel (a), we see the non-uniform and local approximation properties of the reparameterization with the power moments. On the other hand,

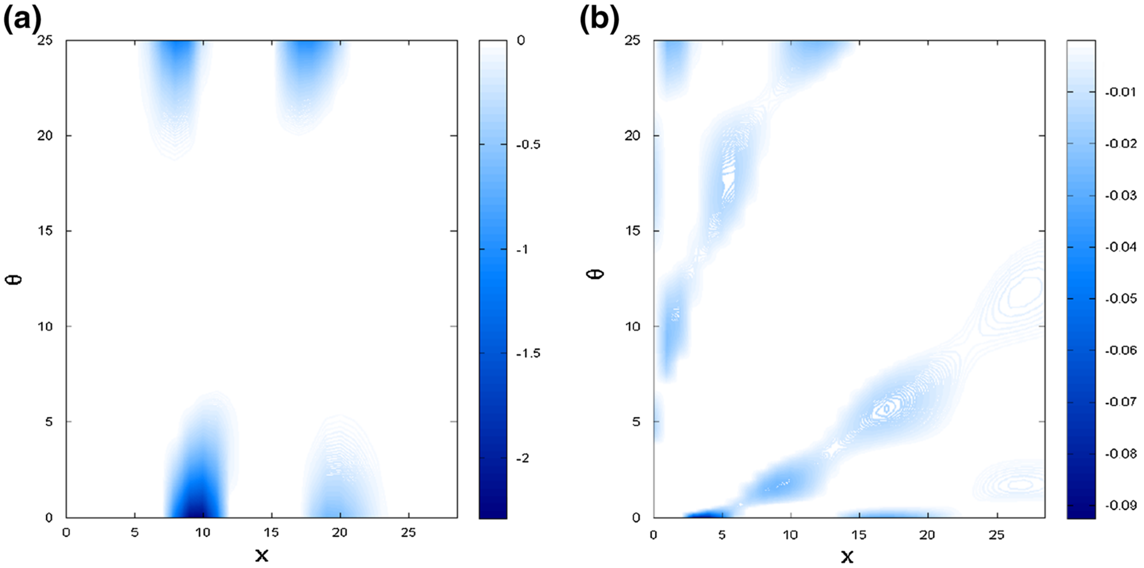

Fig. 3 Plots of the negative regions of $f_{\mu_0}(x; \boldsymbol{u}_c(\theta))$ when the moments are induced by **a** $\{(\theta - 12.5)^i\}_{i=0}^4$, and **b** $\{\phi_i(\theta)\}_{i=0}^4$, for the mixture of $\mathrm{Pois}(x; \theta)$

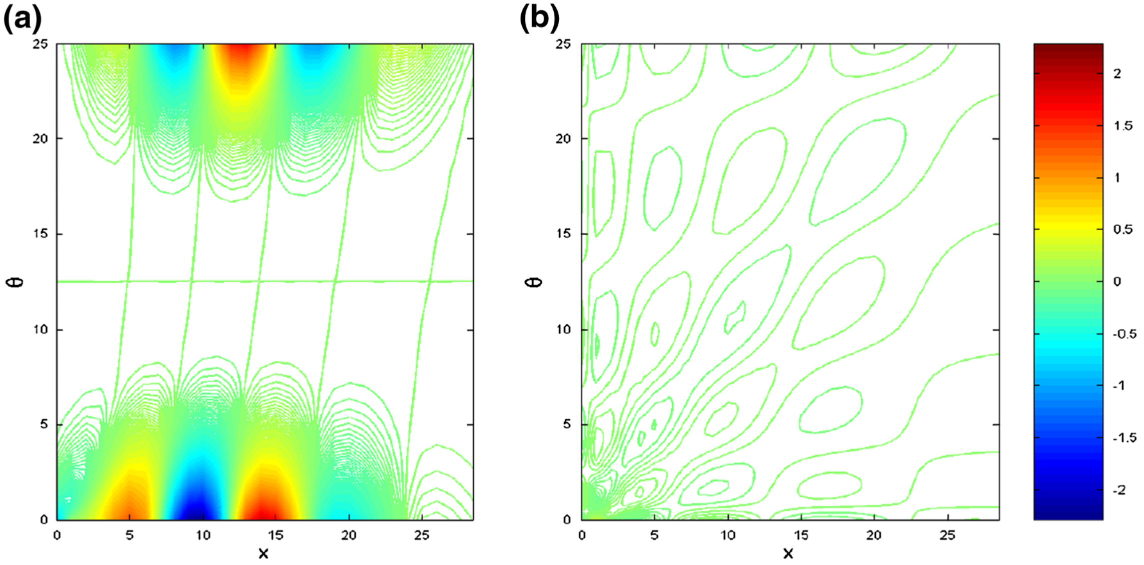

Fig. 4 Contour plots of $\epsilon_4(x;\theta)$ when the moments are induced by **a** $\{(\theta-12.5)^i\}_{i=0}^4$, and **b** $\{\phi_i(\theta)\}_{i=0}^4$, for the mixture of $\mathrm{Pois}(x;\theta)$

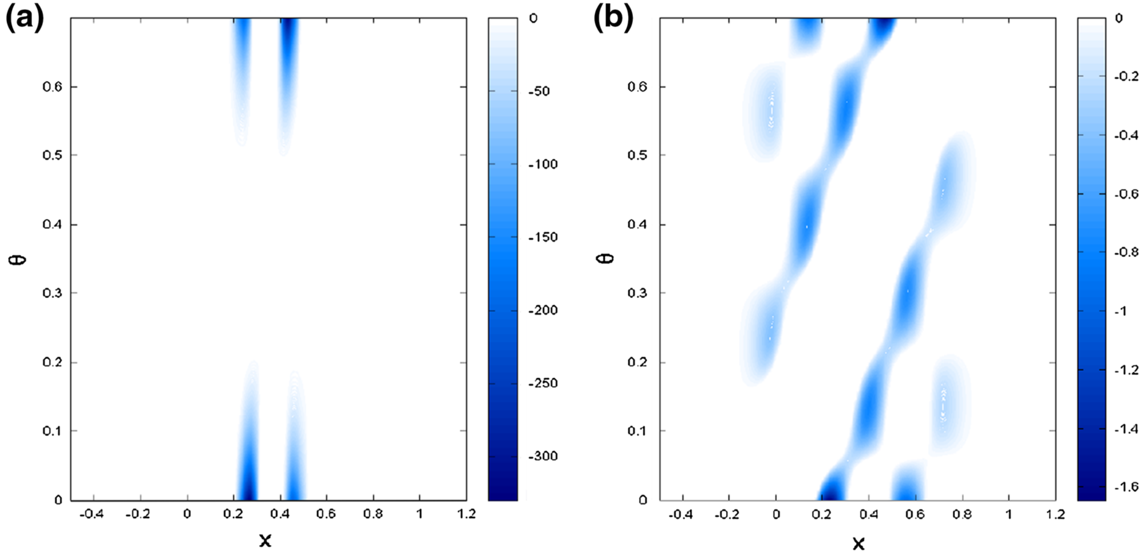

Fig. 5 Plots of the negative regions of $f_{\mu_0}(x; \boldsymbol{u}_c(\theta))$ when the moments are induced by **a** $\{(\theta - 0.35)^i\}_{i=0}^4$, and **b** $\{\phi_i(\theta)\}_{i=0}^4$, for the mixture of $N(x; \theta, \sigma^2)$

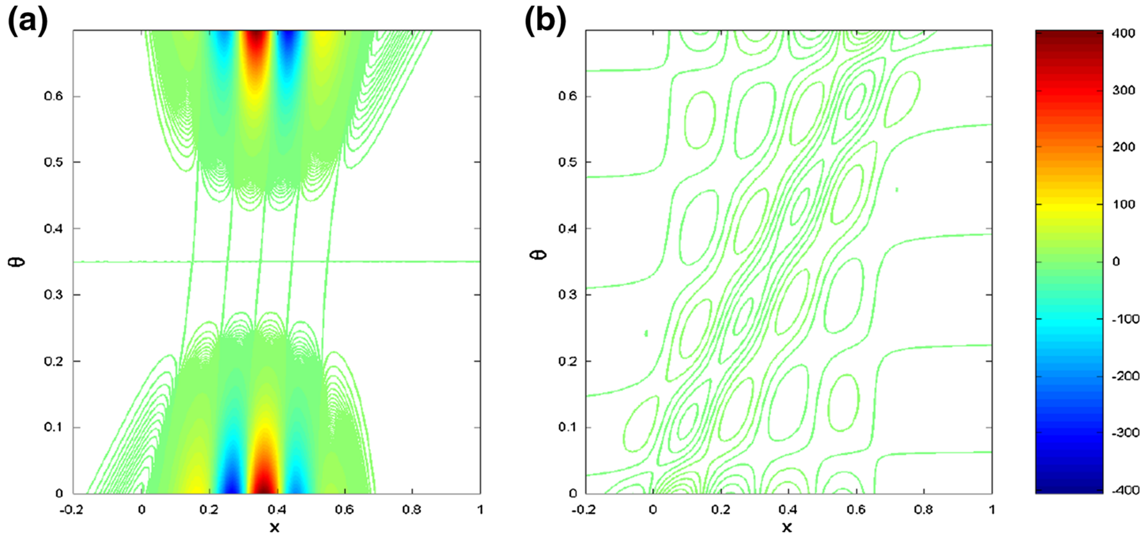

Fig. 6 Contour plots of $\epsilon_4(x;\theta)$ when the moments are induced by **a** $\{(\theta-0.35)^i\}_{i=0}^4$, and **b** $\{\phi_i(\theta)\}_{i=0}^4$, for the mixture of $N(x;\theta,\sigma^2)$

the quality of the approximation is more uniform when the moments are induced by $\{\phi_i(\theta)\}_{i=0}^4$.

4 The geometry of moment spaces

In this section, we study the geometry of the moment-based parameter space. We aim to link the moments $\boldsymbol{m}_c \in \mathcal{K}_r$ to the probability measure $Q(\theta)$ and we do this in two ways: the positive representation and the gradient characterization.

First, we introduce the moment cone induced by a Chebyshev system and its connection to the moment space. Let $\{u_i(\theta)\}_{i=0}^r$ form a Chebyshev system over $[a, b]$ with $u_0(\theta) \equiv 1$ and assume that for each $i = 0, 1, \ldots, r$, the function $u_i(\theta)$ is continuous. When θ moves from a to b, the trace of $\boldsymbol{u}(\theta) = (u_0(\theta), \ldots, u_r(\theta))^{\mathrm{T}} \in R^{r+1}$ forms *the moment curve* Γ_{r+1} in R^{r+1}.

Definition 6 The conical cone of the curve Γ_{r+1} is called *the moment cone* induced by $\{u_i(\theta)\}_{i=0}^r$, that is

$$\mathcal{M}_{r+1} := \left\{ \boldsymbol{c} = (c_0, c_1, \ldots, c_r)^{\mathrm{T}} \in R^{r+1} | \boldsymbol{c} = \int_a^b \boldsymbol{u}(\theta) \mathrm{d}\sigma(\theta) \right\},$$

where $\sigma(\theta)$ is a nondecreasing right continuous function of bounded variation and $\theta \in [a, b]$.

The moment cone contains the convex hull of Γ_{r+1}, denoted by $\mathrm{conv}(\Gamma_{r+1})$, because for each $\boldsymbol{m} \in \mathrm{conv}(\Gamma_{r+1})$, the vector $\boldsymbol{m} = (1, \boldsymbol{m}_c^{\mathrm{T}})^{\mathrm{T}} = \int_a^b \boldsymbol{u}(\theta) dQ(\theta)$, where $Q(\theta)$ is a probability measure over $[a, b]$. Moreover, we have the following result.

Proposition 1 *If* $u_0(\theta) \equiv 1$ *in a Chebyshev system* $\{u_i(\theta)\}_{i=0}^r$ *over* $[a, b]$, *then the boundary of* $\mathrm{conv}(\Gamma_{r+1})$ *is a subset of the boundary of the moment cone* $\mathcal{M}_{r+1}$ *induced by* $\{u_i(\theta)\}_{i=0}^r$.

Proof See the Appendix.

4.1 The positive representations

As will be shown, a positive representation of a vector $\boldsymbol{m} \in \mathrm{conv}(\Gamma_{r+1})$ corresponds to a mixing distribution Q. To illustrate the positive representation of a nonzero vector in $\mathrm{conv}(\Gamma_{r+1})$, we need to first introduce the positive representation and its index.

Definition 7 A nonzero vector $\boldsymbol{c}$ has *a positive representation* in a Chebyshev system $\{u_i(\theta)\}_{i=0}^r$, if it can be written in the form of

$$\boldsymbol{c} = \sum_{j=1}^{J} \beta_j \boldsymbol{u}(\theta_j), \tag{11}$$

where $\boldsymbol{u}(\theta) \in \Gamma_{r+1}$, $a \leq \theta_1 < \theta_2 < \cdots < \theta_J \leq b$ and $\beta_j > 0$, $j = 1, 2, \ldots, J$. If $\sum_{j=1}^J \beta_j = 1$, the positive representation (11) is called a *convex representation.*

Definition 8 Let

$$\mathcal{I}(\theta) := \begin{cases} 1, & \text{if } \theta \in (a, b); \\ 1/2, & \text{if } \theta = a \text{ or } b. \end{cases}$$

If $\boldsymbol{c}$ has the positive representation (11), *the index of* $\boldsymbol{c}$, denoted by $\mathcal{I}(\boldsymbol{c})$, is $\sum_{j=1}^{J} \mathcal{I}(\theta_j)$.

According to Carathéodory's theorem, for each vector $\boldsymbol{m} \in \text{conv}(\Gamma_{r+1})$, there exists a convex representation of $\boldsymbol{m}$ by $\{u_i(\theta)\}_{i=0}^{r}$ with $J < r + 1$. We have the following:

Theorem 2 *For each $\boldsymbol{m}_c \in \mathcal{K}_r$, the moment space, there exists a probability measure $Q(\theta)$ such that $\boldsymbol{m}_c = \int_a^b \boldsymbol{u}_c(\theta) \mathrm{d}Q(\theta)$ and $Q(\theta)$ has at most $r + 1$ support points over $[a, b]$.*

If we further assume $\boldsymbol{m}$ is on the boundary of $\text{conv}(\Gamma_{r+1})$, the upper bound of the number of support points can be sharpened using the following proposition in Karlin and Studden (1966).

Proposition 2 *A nonzero vector $\boldsymbol{c}$ is a boundary point of $\mathcal{M}_{r+1}$, the moment cone, induced by $\{u_i(\theta)\}_{i=0}^{r}$ over $[a, b]$ if and only if $\mathcal{I}(\boldsymbol{c}) < (r + 1)/2$. Moreover, its positive representation is unique with $J \leq (r + 2)/2$.*

With Proposition 2 and the fact that $\hat{\boldsymbol{m}}$ is on the boundary of $\mathcal{M}_{r+1}$, we have the following.

Theorem 3 *If $\boldsymbol{m}_c$ is on the boundary of $\mathcal{K}_r$, there exists one unique probability measure $\hat{Q}(\theta)$ such that $\boldsymbol{m}_c = \int_a^b \boldsymbol{u}_c(\theta) \mathrm{d}\hat{Q}(\theta)$ and $\hat{Q}(\theta)$ has at most $(r + 2)/2$ support points.*

Example 1 (continued) Figure 7 shows the moment cones $\mathcal{M}_3$ induced by $\{(\theta - 12.5)^i\}_{i=0}^{2}$ and $\{\phi_i(\theta)\}_{i=0}^{2}$. In each plot, the curve Γ_3 is induced by the corresponding Chebyshev system and its convex hull $\text{conv}(\Gamma_3)$.

The boundary of $\mathcal{M}_3$ contains the boundary of $\text{conv}(\Gamma_3)$; see Proposition 1. The boundary vectors of $\text{conv}(\Gamma_3)$ are either $\boldsymbol{u}(\theta) \in R^3$ or $(1 - \alpha)\boldsymbol{u}(0) + \alpha \boldsymbol{u}(1)$, where $0 < \alpha < 1$. Therefore, the index of a boundary vector is either 1 or 3/2; see Theorem 3. On the other hand, if the index of a vector is less than 3/2, it must locate on the boundary. Moreover, when $\boldsymbol{m}$ is on the boundary, it uniquely corresponds to a mixing distribution. For example, one point on Γ_3 is the image of Pois(12) in R^3.

4.2 The gradient characterization

The gradient characterization is useful for computational algorithms. In the literature of the non-parametric MLE for mixture models, there exists a class of computational algorithms based on the same convex structure as considered here; see Böhoning (1995) and Wang (2007). This class has more stable computational speeds than the EM algorithm, which is also commonly used for mixture models.

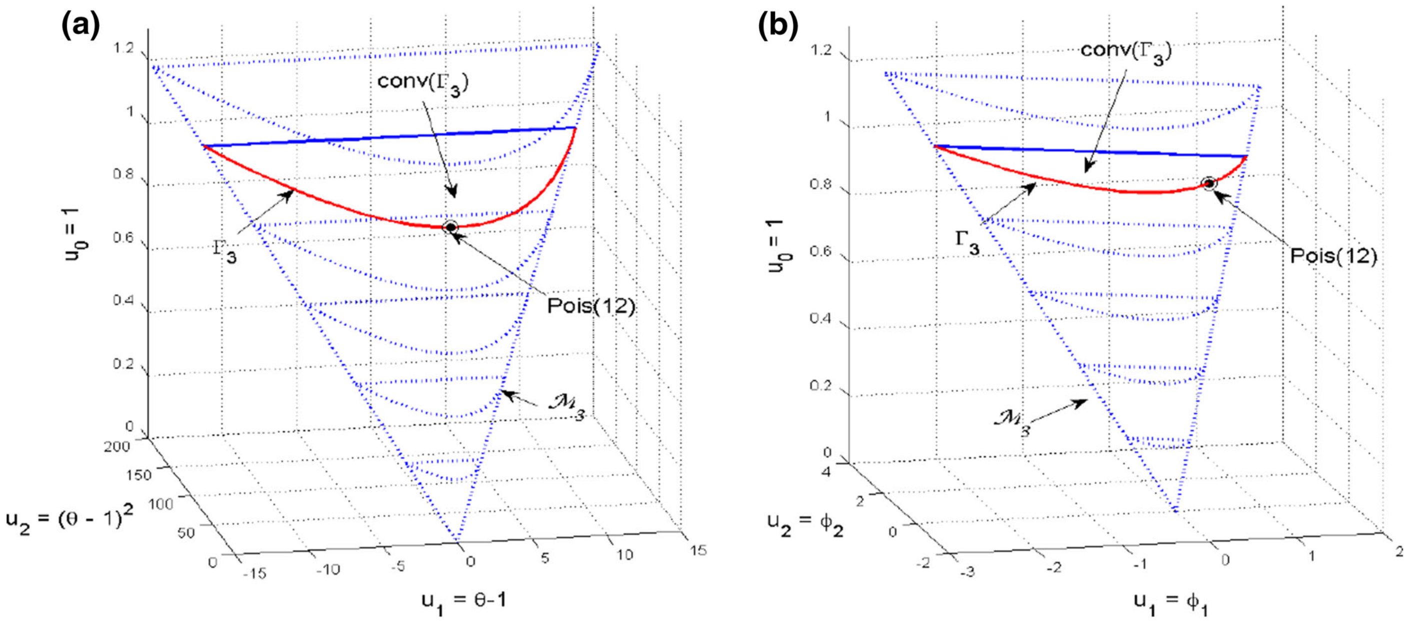

Fig. 7 Plots of the moment cones induced by **a** $\{(\theta - 12.5)^i\}_{i=0}^2$, and **b** $\{\phi_i(\theta)\}_{i=0}^2$ for the mixture of Poisson

In this subsection, we consider the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{m}_c} \quad & \mathcal{L}(\boldsymbol{m}_c) \\ \text{s.t.} \quad & \boldsymbol{m}_c \in \mathcal{K}_r, \end{aligned} \tag{12}$$

where $\mathcal{L}(\boldsymbol{m}_c)$ is an arbitrary loss function and strictly convex with respect to $\boldsymbol{m}_c$. Given a random sample $X_1, X_2, \ldots, X_N$, one example of $\mathcal{L}(\boldsymbol{m}_c)$ is the log-likelihood type function

$$\ell(\boldsymbol{m}_c) := -\sum_{n=1}^{N} \log(f_{\mu_0}(x_n; \boldsymbol{m}_c)). \tag{13}$$

Denote

$$\boldsymbol{s}_n := \triangledown \log(f_{\mu_0}(x_n; \boldsymbol{m}_c)) = \left(\frac{e_1(x) f_0(x)}{f_{\mu_0}(x_n; \boldsymbol{m}_c)}, \ldots, \frac{e_r(x) f_0(x)}{f_{\mu_0}(x_n; \boldsymbol{m}_c)} \right)^{\mathrm{T}},$$

and $\boldsymbol{S}^{\mathrm{T}} := (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N)$. Note that

$$\triangledown \ell(\boldsymbol{m}_c) = -\boldsymbol{S}^{\mathrm{T}} \mathbf{1},$$

and

$$\triangledown^2 \ell(\boldsymbol{m}_c) = \boldsymbol{S}^{\mathrm{T}} \boldsymbol{S},$$

where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}} \in R^N$, $\triangledown$ is the gradient and $\triangledown^2$ is the Hessian. The objective function $\ell(\boldsymbol{m}_c)$ is strictly convex if and only if the matrix $\boldsymbol{S}^{\mathrm{T}} \boldsymbol{S}$ is positive definite. Note that the non-negative condition on $f_{\mu_0}(x; \boldsymbol{m}_c)$ does not need to be explicitly included here because for each observed X, the value of $f_{\mu_0}(x; \boldsymbol{m}_c)$ must be positive to minimize the objective function.

Since the optimization problem (12) is convex, its solution $\hat{\boldsymbol{m}}_c$ is unique and on the boundary of $\mathcal{K}_r$. So there exists a supporting hyperplane of $\mathcal{K}_r$ at $\hat{\boldsymbol{m}}_c$ such that

$$\mathcal{H} = \left\{ \boldsymbol{h} = (h_1, \ldots, h_r)^{\mathrm{T}} \in R^r | (\hat{\boldsymbol{m}}_c - \boldsymbol{h})^{\mathrm{T}} \triangledown \mathcal{L}(\hat{\boldsymbol{m}}_c) = 0 \right\}.$$

The following theorem states the relationship between $\mathcal{H}$ and the support points of $\hat{Q}$ in Theorem 3. Recall that, we have defined $\boldsymbol{u}_c(\theta) = (u_1(\theta), \ldots, u_r(\theta))^{\mathrm{T}} \in R^r$.

Theorem 4 *Let $\hat{\Theta}$ be the set of support points of $\hat{Q}$. Then, if a point $\hat{\theta} \in [a, b]$ is an element of $\hat{\Theta}$, then $\boldsymbol{u}_c(\hat{\theta})$ is on the hyperplane $\mathcal{H}$. The converse also holds.*

Proof See Lindsay (1983a, b).

The above theorem also implies that $\hat{\Theta}$ is the set of zeros of the gradient function of the objective function $\mathcal{L}(\boldsymbol{m}_c)$ which is defined as:

$$\begin{aligned}\mathcal{D}(\hat{\boldsymbol{m}}_c, \boldsymbol{u}_c(\theta)) &:= \frac{\partial}{\partial \epsilon} \mathcal{L}((1-\epsilon)\hat{\boldsymbol{m}}_c + \epsilon \boldsymbol{u}_c(\theta))\bigg|_{\epsilon=0} \\ &= (\boldsymbol{u}_c(\theta) - \hat{\boldsymbol{m}}_c)^{\mathrm{T}} \bigtriangledown \mathcal{L}(\hat{\boldsymbol{m}}_c).\end{aligned}$$

Moreover, we can use the gradient function to characterize $\hat{\boldsymbol{m}}_c$ as follows.

Theorem 5 *The following three statements are equivalent:*

1. $\hat{\boldsymbol{m}}_c$ *minimizes* $\mathcal{L}(\boldsymbol{m}_c)$.
2. $\inf_\theta \mathcal{D}(\hat{\boldsymbol{m}}_c, \boldsymbol{u}_c(\theta)) = 0$.
3. $\hat{\boldsymbol{m}}_c$ *maximizes* $\inf_\theta \mathcal{D}(\boldsymbol{m}_c, \boldsymbol{u}_c(\theta))$.

Proof See Lindsay (1983a, b).

Now, we continue Example 1 to illustrate Theorems 4 and 5.

Example 1 (continued) In each panel of Fig. 8, we see the curve Γ_3 induced by $\{\phi_1(\theta), \phi_2(\theta)\}$ and its convex hull $\mathcal{K}_2$ in the space of $(m_1, m_2)^{\mathrm{T}} \in R^2$. The contours show the identical values of the objective function

$$\mathcal{L}(\boldsymbol{m}_c) = (\boldsymbol{t} - \boldsymbol{m}_c)^{\mathrm{T}}(\boldsymbol{t} - \boldsymbol{m}_c), \tag{14}$$

where $\boldsymbol{m}_c \in R^2$ and $\boldsymbol{t} = (-2, 0)^{\mathrm{T}}$. This objective function is chosen for its nice visual interpretation in Fig. 8. Because $\boldsymbol{t} \notin \mathcal{K}_2$, $\mathcal{L}(\boldsymbol{m}_c)$ is strictly convex with respect to $\boldsymbol{m}_c$. The minimum value of $\mathcal{L}(\boldsymbol{m}_c)$ over $\mathcal{K}_2$ is 0.5517. As we see, the contour $\mathcal{L}(\boldsymbol{m}_c) = 0.5517$ has a unique intersection $\hat{\boldsymbol{m}}_c$ with $\mathcal{K}_2$. Moreover, the intersection $\hat{\boldsymbol{m}}_c$ is on the boundary of $\mathcal{K}_2$. In Fig. 8a, the solid line represents the supporting hyperplane $\mathcal{H}$ of $\mathcal{K}_2$ at $\hat{\boldsymbol{m}}_c$. Here, we have $\hat{\boldsymbol{m}}_c = \boldsymbol{u}_c(\hat{\theta}) \in \mathcal{H}$; see Theorem 4.

Moreover, $\bigtriangledown\mathcal{L}(\hat{\boldsymbol{m}}_c)$ is orthogonal to the supporting hyperplane. For any vector $\boldsymbol{u}_c(\theta) \neq \hat{\boldsymbol{m}}_c$ on Γ_3, we have the vector $\boldsymbol{u}_c(\theta) - \hat{\boldsymbol{m}}_c$. From Fig. 8a, it can be seen that the angle $\psi \in [0, \pi]$ between $\bigtriangledown\mathcal{L}(\hat{\boldsymbol{m}}_c)$ and $\boldsymbol{u}_c(\theta) - \hat{\boldsymbol{m}}_c$ is always acute. Therefore, we have

$$\cos(\psi) = \frac{\mathcal{D}(\hat{\boldsymbol{m}}_c, \boldsymbol{u}_c(\theta))}{\sqrt{(\boldsymbol{u}_c(\theta) - \hat{\boldsymbol{m}}_c)^{\mathrm{T}}(\boldsymbol{u}_c(\theta) - \hat{\boldsymbol{m}}_c)}\sqrt{(\bigtriangledown\mathcal{L}(\hat{\boldsymbol{m}}_c))^{\mathrm{T}}(\bigtriangledown\mathcal{L}(\hat{\boldsymbol{m}}_c))}} > 0;$$

see Theorem 5 (2). It is also obvious that $\cos(\psi) = 0$ if and only if $\boldsymbol{u}_c(\theta) = \hat{\boldsymbol{m}}_c$. In Fig. 8b, we see that for any $\boldsymbol{m}'_c \neq \hat{\boldsymbol{m}}_c$ in $\mathcal{K}_2$, there always exists a $\boldsymbol{u}_c(\theta') \in \Gamma_3$ such that the angle ψ' between $\bigtriangledown\mathcal{L}(\boldsymbol{m}'_c)$ and $\boldsymbol{u}_c(\theta') - \boldsymbol{m}'_c$ is obtuse. It follows that $\inf_\theta \mathcal{D}(\boldsymbol{m}'_c, \boldsymbol{u}_c(\theta)) < 0$; see Theorem 5 (3).

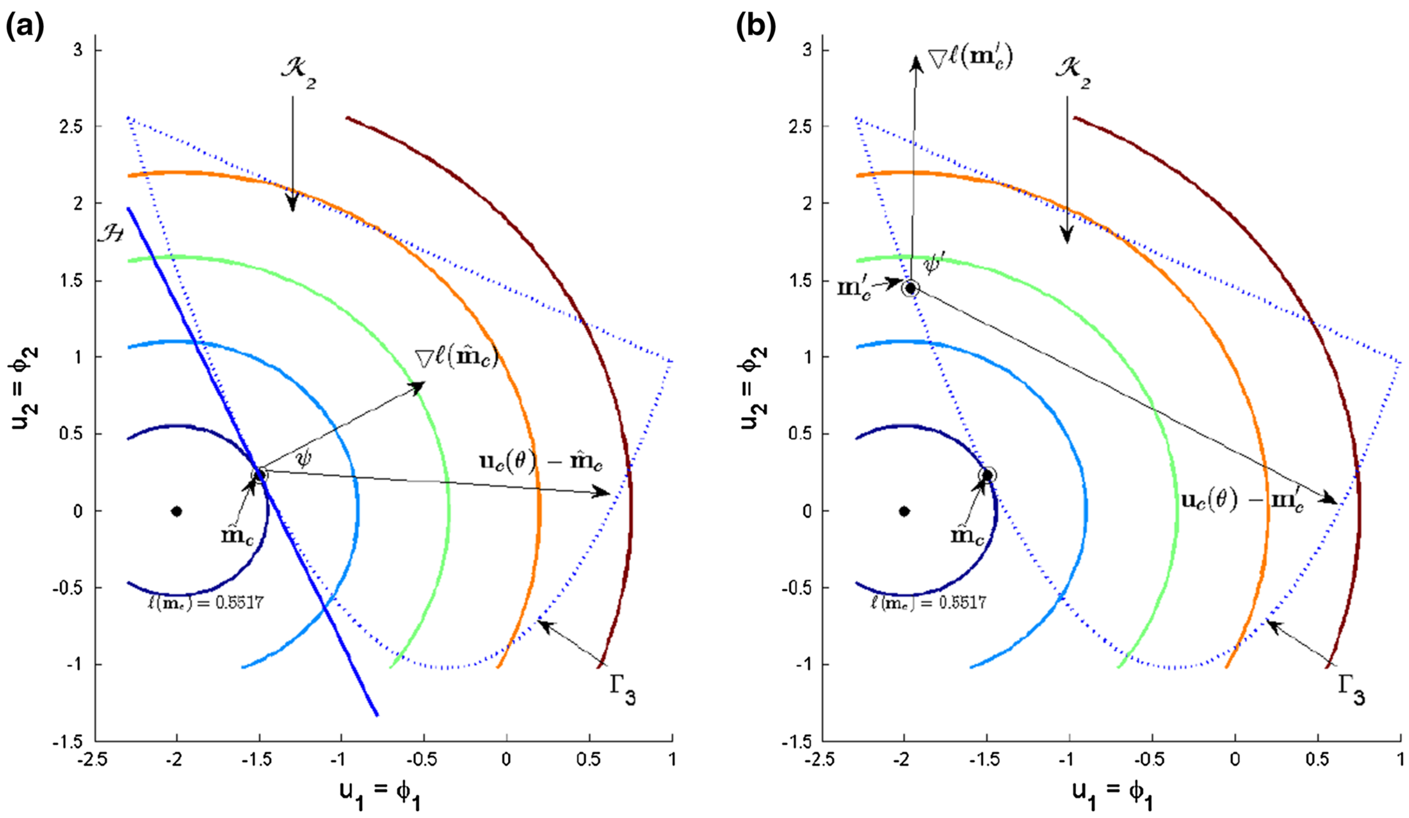

Fig. 8 Plots of $\mathcal{K}_2$ induced by $\{\phi_i(\theta)\}_{i=0}^{2}$ and a visual interpretation of Theorems 4 and 5

5 Real examples

5.1 The Thailand cohort study data

Consider the data on morbidity in northeast Thailand which was analyzed by Böhoning (1995); see also Schlattmann (2009). We fit a mixture of Poisson with $[a, b] = [0, 25]$ by minimizing $\ell(\boldsymbol{m}_c)$ in (13) over $\mathcal{K}_r$, where $f_{\mu_0}(x; \boldsymbol{m}_c)$ is induced by $\{\phi_i(\theta)\}_{i=0}^r$ from Example 1 in Sect. 3.1 and $r = 2, 3, \ldots, 24$. Let $\hat{\boldsymbol{m}}_c^{(r)} \in R^r$ be the solution and $\hat{Q}^{(r)}$ be the probability measure reconstructed from $\hat{\boldsymbol{m}}_c^{(r)}$. Also, let $\hat{Q}^{(\infty)}$ be the non-parametric MLE of the mixing distribution.

The results are summarized in Table 1. When $r \leq 5$, the $\hat{\boldsymbol{m}}_c^{(r)} \in R^r$ is not on the boundary of $\mathcal{K}_r$ and, thus, no unique mixing distribution could be reconstructed from the moments. For each $r \geq 6$, the probability measure $\hat{Q}^{(r)}$ is unique and reported in Table 1. The log-likelihoods of $f_{\text{Mix}}(x; \hat{Q}^{(r)})$ become stable and close to the log-likelihood of $f_{\text{Mix}}(x; \hat{Q}^{(\infty)})$, when r is large; also see Fig. 9b. Furthermore, for large r, the two distributions $\hat{Q}^{(r)}$ and $\hat{Q}^{(\infty)}$ are almost the same; see Table 1. Both of the observations support that the quality of the approximation of $f_{\mu_0}(x; \boldsymbol{m}_c)$ to $f_{\text{Mix}}(x; Q)$ is accurate, where $\boldsymbol{m}_c \in \mathcal{K}_r$ and r is finite.

5.2 The sodium–lithium countertransport data

Consider the SLC data analyzed by Roeder (1994). We fit a mixture of normal with same variance. Let $[a, b] = [0, 0.7]$. For each $r \geq 4$, we minimize $\ell(\boldsymbol{m}_c, \sigma^2)$ over $\mathcal{K}_r \times R^+$, where $\boldsymbol{m}_c \in R^r$ is the moment vector induced by $\{\phi_i(\theta)\}_{i=1}^r$ from Example 2 in Sect. 3.1. Note here $\{\phi_i(\theta)\}_{i=1}^r$ depends on the value of σ^2. Given σ^2, let $\hat{\boldsymbol{m}}_c^{(r)}(\sigma^2)$ minimize $\ell(\boldsymbol{m}_c, \sigma^2)$. The variance is estimated as the $\hat{\sigma}^2 = \arg\min_{R^+} \ell(\hat{\boldsymbol{m}}_c^{(r)}(\sigma^2), \sigma^2)$. Let $\hat{Q}^{(r)}$ be the reconstructed mixing distribution from $\hat{\boldsymbol{m}}_c(\hat{\sigma}^2) \in \mathcal{K}_r$.

Table 1 Some results of the fitted models to the Thailand cohort study data

r	Mixing parameters	Mixing proportions	Log-likelihod (10^3)
6	(0.433, 0.3778, 10.664)	(0.2823, 0.4983, 0.2194)	−1.7943
7	(0, 2.93, 6.516, 13.301)	(0.1552, 0.4350, 0.2981, 0.1117)	−1.5896
8	(0.200, 2.951, 8.328, 15.948)	(0.2119, 0.4776, 0.2568, 0.0538)	−1.5295
9	(0, 2.008, 3.942, 8.587, 16.157)	(0.1535, 0.3128, 0.2487, 0.2349, 0.0500)	−1.5304
10	(0.158, 2.816, 8.102, 16.019)	(0.1995, 0.4758, 0.2699, 0.0548)	−1.5268
11	(0, 0.878, 2.961, 8.226, 16.091)	(0.1363, 0.0964, 0.5966, 0.1175, 0.0533)	−1.5225
12	(0.147, 2.824, 8.173, 16.164)	(0.1981, 0.4794, 0.2689, 0.0537)	−1.5225
13	(0.143, 2.817, 8.164, 16.154)	(0.1968, 0.4800, 0.2693, 0.0538)	−1.5225
⋮	⋮	⋮	⋮
∞	(0.143, 2.817, 8.164, 16.156)	(0.1969, 0.4800, 0.2689, 0.0539)	−1.5225

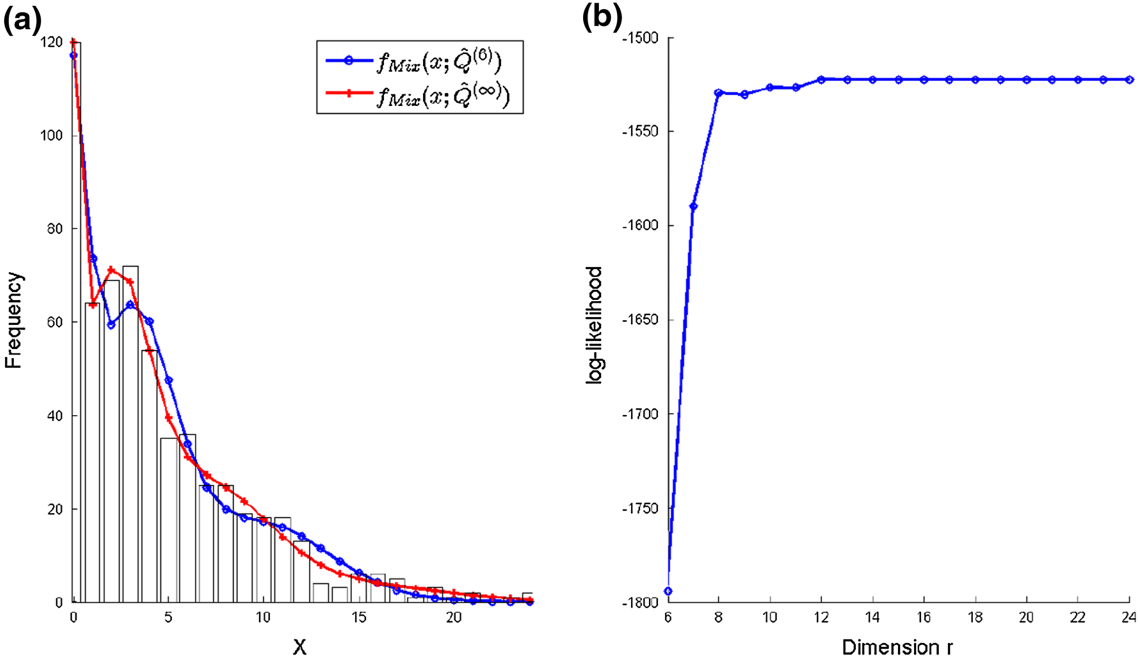

Fig. 9 Plots of **a** the observed frequency, the frequency of $f_{\text{Mix}}(x;\,\hat{Q}^{(6)})$ and $f_{\text{Mix}}(x;\,\hat{Q}^{(\infty)})$; **b** the log-likelihoods against the dimension r of the moment space $\mathcal{K}_r$

Table 2 Some results of the fitted models to the SLC data

r	Mixing parameters	Mixing proportions	Standard deviation	Log-likelihod
5	(0.220, 0.511)	(0.9015, 0.0985)	0.060	164.68
6	(0.195, 0.345)	(0.6429, 0.3571)	0.090	165.59
7	(0.246, 0.393, 0.521)	(0.8907, 0.0677, 0.0417)	0.081	183.31
8	(0.234, 0.339)	(0.8597, 0.1403)	0.069	182.51
9	(0.221, 0.372, 0.563)	(0.7735, 0.2198, 0.0067)	0.064	189.26
10	(0.231, 0.393, 0.567)	(0.8300, 0.1534, 0.0166)	0.069	188.04
11	(0.209, 0.251, 0.392, 0.582)	(0.4460, 0.3576, 0.1737, 0.0226)	0.057	190.37
⋮	⋮	⋮	⋮	⋮

The results are summarized in Table 2 and some of the fitted models are plotted in Fig. 10. From the figure, we can see that $f_{\mathrm{Mix}}(x; \hat{Q}^{(r)})$ may not give sensible fitting until $r \geq 9$. Based on the prior knowledge that the data may consist of three subpopulations, we choose $f_{\mathrm{Mix}}(x; \hat{Q}^{(9)})$ as our final fitted model. This result is close to the three component mixture of normal with same variance fitted by maximizing the log-likelihood, in which the mixing parameters are $(0.223, 0.379, 0.577)$ with the mixing proportions $(0.774, 0.202, 0.024)$ and a standard deviation of 0.058; see Roeder (1994). Lastly, we want to point out that the non-parametric MLE is not appropriate here because it is exactly the empirical distribution of the sample.

6 Discussion and conclusion

The essential idea in this paper is to use a moment space in R^r to approximate an infinite-dimensional parameter space. We end this paper with a brief discussion of the following issues.

The links to the method of moments The method of moments for mixture models has a long history; see McLachlan and Peel (2000). It can be computed easily but suffers from a lack of efficiency; see Lindsay (1989). As a result, it is typically used to find the initial value for the computational algorithms of other estimation methods. According to the discussion in Sect. 2, the moments of mixing distribution are induced locally and correspond to local approximations of non-parametric mixtures. We may expect that the quality of the approximation is related to the efficiency of the method of moments. However, the connection between the two still remains unclear.

The links to the non-parametric MLE The link between the reparameterization in moments and the non-parametric MLE is their geometry. Lindsay (1983a, b) studied the geometry in the likelihood-based embedding spaces and gave the fundamental properties of the non-parametric MLE. Because the set of mixture is a convex hull of a curve in both cases, the properties of the reparameterization in moments are similar to those given by Lindsay.

The two types of geometry are different. The dimension of the likelihood-based embedding space is the number of distinct values in a random sample; see Lindsay

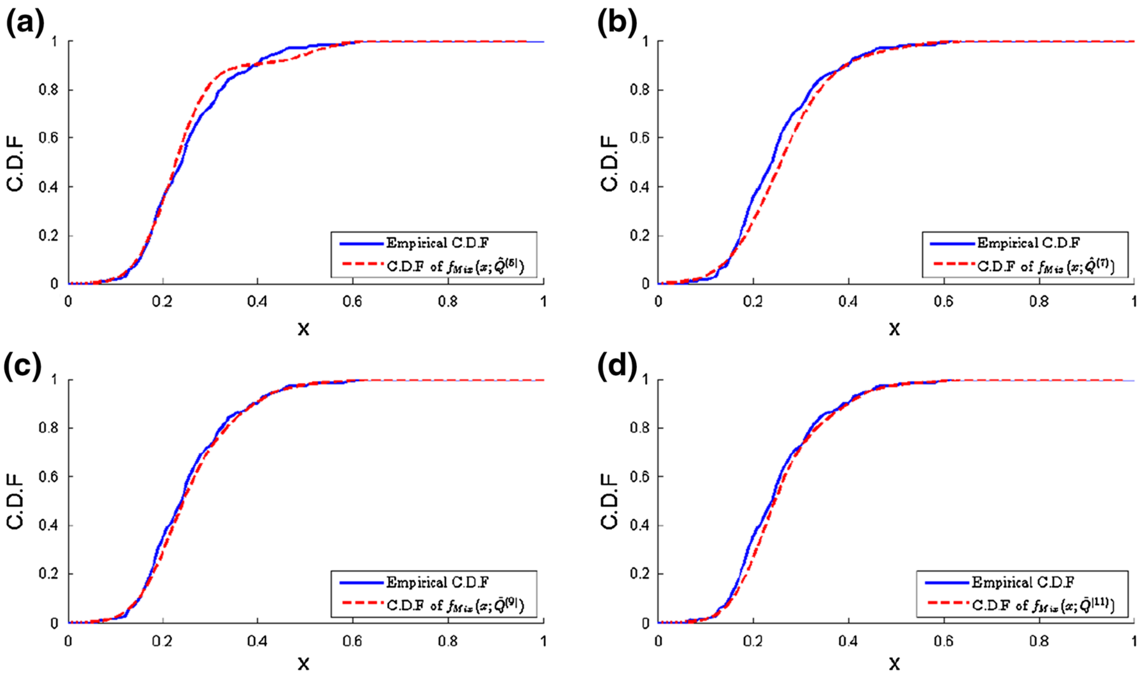

Fig. 10 Plots of the empirical C.D.F of data and the C.D.F of **a** $f_{\text{Mix}}(x;\ \hat{Q}^{(5)})$; **b** $f_{\text{Mix}}(x;\ \hat{Q}^{(7)})$; **c** $f_{\text{Mix}}(x;\ \hat{Q}^{(9)})$; **d** $f_{\text{Mix}}(x;\ \hat{Q}^{(11)})$

(1983a) and Lindsay (1983b), while the dimension of the moment space is chosen depending on the quality of the approximation. In the two examples of the moments discussed in this paper, the dimension of the moment space is usually smaller than that of the likelihood-based embedding space, especially when the mixing region $[a, b]$ is narrow and the decay of the eigenvalues of (8) is fast. This makes the mixing distribution reconstructed from the moments have a sharpened upper bound of the number of support points.

Extension to the generalized linear mixed models As pointed out by the referees, mixture models have a much wider application in the context of regression. These models are known as the generalized linear mixed models or the random effects models; see Diggle et al. (2002). It would be interesting future work to examine the performance of the reparameterization of mixture models to the generalized linear mixed models and in practical applications.

Appendix A

7.1 Strictly totally positive kernel functions

The strictly totally positive kernel functions and Chebyshev systems are defined in the following ways in (Karlin and Studden 1966).

Definition 9 A real-valued kernel function $K(s, \theta)$, $(s, \theta) \in \mathcal{S} \times \Theta \subseteq R^2$, is called *strictly totally positive* of order r, if for each $J = 1, 2, \ldots, r$, we have $\det(K(s_i, \theta_j))_{i,j=0}^{J} > 0$, whenever $s_0 < s_1 < \cdots < s_r$, $\theta_0 < \theta_1 < \cdots < \theta_r$ and $(s_i, \theta_j) \in \mathcal{S} \times \Theta$, $i, j = 0, 1, \ldots, r$.

Consider a kernel function for $(x, y) \in \mathcal{S} \times \mathcal{S}' \subseteq R^2$,

$$K^*(x, y) := \int_a^b L(x, \theta) M(y, \theta) \mathrm{d}\theta, \tag{15}$$

where $L(x, \theta)$, $(x, \theta) \in \mathcal{S} \times [a, b] \subset R^2$ and $M(y, \theta)$, $(y, \theta) \in \mathcal{S}' \times [a, b] \subset R^2$. The following proposition is proved in Karlin and Studden (1966).

Proposition 3 *If the kernel function in* (15) *exists for each* $(x, y) \in \mathcal{S} \times \mathcal{S}'$ *and* $L(x, \theta)$ *and* $M(x, \theta)$ *are strictly totally positive, then* $K(x, y)$ *is strictly totally positive.*

Pinkus (1996) further states that the eigenfuctions from a strictly totally positive kernel function could also form a Chebyshev system.

Proposition 4 *Let*

$$(A'g)(\theta) = \int_a^b g(\theta) K'(\theta, \theta') \mathrm{d}\theta,$$

be a compact, self-adjoint, positive integral operator in the form of (8). *Moreover, the kernel function* $K'(\theta, \theta')$ *is strictly totally positive over* $[a, b] \times [a, b]$. *Then, the* *integral operator $A'(\cdot)$ has the eigenvalues $\lambda_0 > \lambda_1 > \cdots > 0$ and associated eigenfunctions $\phi_0(\theta), \phi_1(\theta), \ldots$, which are continuous over $[a, b]$. For each $r = 1, 2, \ldots$, the set $\{\phi_i(\theta)\}_{i=0}^{r}$ forms a Chebyshev system over $[a, b]$.*

It is known that a Hilbert–Schmidt operator is compact; see Debnath and Mikusiński (1999). To show $K^*(x, y)$ is Hilbert–Schmidt, one sufficient condition is $\int_a^b \int_{\mathcal{S}} L^2(x; \theta)\mathrm{d}x\mathrm{d}\theta < \infty$ and $\int_a^b \int_{\mathcal{S}'} M^2(y; \theta)\mathrm{d}y\mathrm{d}\theta < \infty$, because

$$\int_{\mathcal{S}} \int_{\mathcal{S}'} \left(K^*(x, y)\right)^2 \mathrm{d}x\mathrm{d}y \le \textstyle\int_a^b \int_{\mathcal{S}} L^2(x; \theta)\mathrm{d}x\mathrm{d}\theta \times \int_a^b \int_{\mathcal{S}'} M^2(y; \theta)\mathrm{d}y\mathrm{d}\theta. \quad (16)$$

7.2 Proof of Theorem 1

Firstly, we show that $\{\phi_i(\theta)\}_{i=0}^{\infty}$ is the set of eigenfunctions of the integral operator

$$(A'g)(\theta) = \int_a^b g(\theta)K'(\theta, \theta')\mathrm{d}\theta,$$

where

$$K'(\theta, \theta') = \int_{\mathcal{S}} \frac{f(x; \theta) f(x; \theta')}{f_0(x)} \mathrm{d}x.$$

For each i, we have

$$\begin{aligned}
\lambda_i \phi_i(\theta) &= \lambda_i \int_{\mathcal{S}} \frac{\tilde{\gamma}_i(x)}{f_0^{1/2}(x)} f(x; \theta)\mathrm{d}x \\
&= \int_{\mathcal{S}} \frac{f(x; \theta)}{f_0^{1/2}(x)} \int_{\mathcal{S}} \tilde{\gamma}_i(y) K(x, y)\mathrm{d}y\mathrm{d}x \\
&= \int_a^b \int_{\mathcal{S}} \tilde{\gamma}_i(y) \frac{f(y; \theta')}{f_0^{1/2}(y)} \mathrm{d}y \int_{\mathcal{S}} \frac{f(x; \theta) f(x; \theta')}{f_0(x)} \mathrm{d}x\mathrm{d}\theta' \\
&= \int_a^b \phi_i(\theta') K'(\theta, \theta')\mathrm{d}\theta'.
\end{aligned}$$

Next, note that the one-parameter exponential family $f(x; \theta)$ is strictly totally positive; see Lidnsay and Roeder (1993). According to Proposition 3, the kernel function $K'(\theta, \theta')$ is strictly totally positive. Then, it follows from Proposition 4 that the set of eigenfuctions $\{\phi_i(\theta)\}_{i=0}^{r}$ forms a Chebyshev system over $[a, b]$. □

7.3 Proof of Proposition 1

We want to show that for each boundary vector $\boldsymbol{m}^*$ of $\mathrm{conv}(\Gamma_{r+1})$, there exists a supporting hyperplane of $\mathrm{conv}(\Gamma_{r+1})$ at $\boldsymbol{m}^*$ which is also a supporting hyperplane of $\mathcal{M}_{r+1}$ at $\boldsymbol{m}^*$.

Firstly, the convex hull $\mathrm{conv}(\Gamma_{r+1})$ is the intersection of $\mathcal{M}_{r+1}$ and the hyperplane $\mathcal{H}_1 = \left\{\boldsymbol{h} = (1, h_1, h_2, \ldots, h_r)^{\mathrm{T}} \in R^{r+1}\right\}$. Then, in $\mathcal{H}_1$, there exists a vector $\tilde{\boldsymbol{\beta}}_c = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_r)^{\mathrm{T}} \in R^r$ such that for each $\boldsymbol{m} \in \mathrm{conv}(\Gamma_{r+1})$, we have

$$\sum_{i=1}^{r} m_i \tilde{\beta}_i \geq \sum_{i=1}^{r} m_i^* \tilde{\beta}_i.$$

Let $\tilde{\beta}_0 = -\sum_{i=1}^{r} m_i^* \tilde{\beta}_i$. For each $\boldsymbol{m} \in \mathrm{conv}(\Gamma_{r+1})$, we have

$$\sum_{i=0}^{r} m_i \tilde{\beta}_i \geq \sum_{i=0}^{r} m_i^* \tilde{\beta}_i.$$

Therefore, the vector $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \tilde{\beta}_c^{\mathrm{T}})^{\mathrm{T}} \in R^{r+1}$ determines the hyperplane

$$\left\{\boldsymbol{h} \in R^{r+1} | (\boldsymbol{h} - \boldsymbol{m}^*)^{\mathrm{T}} \tilde{\boldsymbol{\beta}} = 0\right\} \tag{17}$$

as a supporting hyperplane of $\mathrm{conv}(\Gamma_{r+1})$ at $\boldsymbol{m}^*$.

Note that any vector in $\mathcal{M}_{r+1}$ can be written as $\Delta\boldsymbol{m}$, where $\Delta \geq 0$ and $\boldsymbol{m} \in \mathrm{conv}(\Gamma_{r+1})$. We have the inequality:

$$\begin{aligned}\sum_{i=0}^{r}(\Delta m_i - m_i^*)\tilde{\beta}_i &= (\Delta - 1)\tilde{\beta}_0 + \sum_{i=1}^{r}(\Delta m_i - m_i^*)\tilde{\beta}_i \\ &= (1 - \Delta)\sum_{i=1}^{r} m_i^* \tilde{\beta}_i + \sum_{i=1}^{r}(\Delta m_i - m_i^*)\tilde{\beta}_i \\ &= \Delta \sum_{i=1}^{r}(m_i - m_i^*)\tilde{\beta}_i \geq 0,\end{aligned}$$

and thus the hyperplane (17) is also a supporting hyperplane of $\mathcal{M}_{r+1}$. □

Acknowledgments The authors thank the Editor, the Associate Editor and the two referees for their insightful comments and suggestions that lead to substantially improve the paper.

References

Anaya-Izquierdo, K. A., Marriott, P. (2007). Local mixtures of the exponential distribution. *Annals of the Institute of Statistical Mathematics*, *59*, 111–134.

Böhoning, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference*, *47*, 5–28.

Debnath, L., Mikusiński, P. (1999). *Introduction to Hilbert Spaces with Applications* (2nd ed., pp. 87–130). San Diego: Academic Press Inc.

Diggle, P., Heagerty, P., Liang, K. Y., Zeger, S. (2002). *Analysis of longitudinal data*. Oxford: Oxford University Press.

Horváth, L., Kokoszka, P. (2012). *Inference for functional data with applications* (pp. 37–39). New York: Springer.

Karlin, S., Studden, W. J. (1966). *Tchebycheff systems: With applications in analysis and statistics* (pp. 1–49). New York: Wiley.

Lindsay, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. *Philosophical Transactions of the Royal Society of London Series A, Mathematical and Physical Sciences*, *296*, 639–662.

Lindsay, B. G. (1983a). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics*, *11*, 86–94.

Lindsay, B. G. (1983b). The geometry of mixture likelihoods. II. The exponential family. *The Annals of Statistics*, *11*, 783–792.

Lindsay, B. G. (1989). Moment matrices: applications in mixture. *The Annals of Statistics*, *17*, 722–740.

Lidnsay, B. G., Roeder, K. (1993). Uniqueness of estimation and identifiability in mixture models. *Canadian Journal of Statistics*, *21*, 139–147.

Marriott, P. (2002). On the local geometry of mixture models. *Biometrika*, *89*, 77–93.

Marriott, P. (2007). Extending local mixture models. *Annals of the Institute of Statistical Mathematics*, *59*, 95–110.

McLachlan, G. J., Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, *10*, 65–80.

Morris, C. (1983). Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, *11*, 515–529.

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

Pinkus, A. (1996). *Spectral properties of totally positive kernels and matrices. Total positivity and its application* (pp. 477–511). Dordrecht: Kluwer Academic Publishers.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, *89*, 487–495.

Schlattmann, P. (2009). *Medical applications of finite mixture models*. Berlin: Springer.

Wang, Y. (2007). On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *Journal of the Royal Statistical Society: Series B*, *69*, 185–198.