AISM (2007) 59:95–110
DOI 10.1007/s10463-006-0100-6

Extending local mixture models

Paul Marriott

Received: 18 April 2006 / Revised: 18 August 2006 /
Published online: 7 February 2007

Abstract Local mixture models have proved useful in many statistical applications. This paper looks at ways in which the local assumption, which is used in an asymptotic approximation, can be relaxed in order to generate a much larger class of models which still have the very attractive geometric and inferential properties of local mixture models. The tool used to develop this large class of models is the Karhunen–Loève decomposition. Computational issues associated with working with these models are also briefly considered.

Keywords Local mixture models · Information geometry · Karhunen-Loève decomposition · Posterior approximations

1 Introduction

The idea of using a *local mixture model* to simplify the inferential problems associated with general mixture families has proved to be a useful one, see Marriott (2002). Applications can be found in measurement error modelling, Marriott (2003), in Bayesian prediction Marriott (2002), lifetime data analysis and in influence analysis Critchley and Marriott (2003), also related ideas can be found in Eguchi (2005) and Anaya-Izquierdo and Marriott (2005). We also note strong links between the local mixture approach and that of Amari (1990, Chapter 8), which looks at estimation in the presence of very large numbers of

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

nuisance parameters. Both techniques use the geometric construct of a normal fibre bundle, defined below. Amari's results show that the dimension of the fibre of the normal bundle can grow with the sample size while still allowing efficient inference for the interest parameters, and this paper has analogous results when the inferential issue is marginal inference on an interest parameter. This paper extends the discussion of Marriott (2005) which first introduced the forms of fibre bundle discussed here.

The following example shall be used throughout to illustrate the discussion.

Example 1 Consider the problem of inference about the parameter μ for the model

$$X \sim f_X(x|\mu, Q) := \int f_X(x|\mu + \eta) \mathrm{d}Q(\eta) \tag{1}$$

where Q is a distribution which, for identification reasons, is constrained to have mean zero, and $f_X(x|\mu)$ is the exponential distribution with mean parameter μ. We might, for example, consider model (1) in the context of a simple lifetime analysis problem, where an exponential baseline model has been used, but the possibility of an unmeasured frailty, or random effect needs to be investigated. We emphasis that the problem of interest is to learn about μ, which has a population meaning as $E(X)$ regardless of the mixing distribution Q. In Bayesian terms we want to calculate the marginal posterior for μ,

$$\int P(\mu|x_1, \ldots, x_n, Q) \mathrm{d}P(Q) = \int_{\mathcal{Q}} \left\{ \prod_{i=1}^{n} \int f_X(x_i|\mu + \eta) \mathrm{d}Q(\eta) \right\} \mathrm{d}P(Q) \tag{2}$$

for some measure $\mathrm{d}P$ on $\mathcal{Q}$ which will be some subset of the space of distributions. For simplicity we shall assume throughout that the information in the likelihood dominates that of the priors and shall not therefore consider their form explicitly.

The structure of (2) is clearly problematic from an inferential point of view. Apart from the fact (1) itself is not in closed form, it is required to integrate out over an infinite dimensional 'nuisance' parameter $Q \in \mathcal{Q}$. The idea of the local mixture, and the extensions of it in this paper, is to replace this infinite dimensional integral with a finite dimensional one without a great change in marginal inference on μ.

2 Normal bundles

Amari (1990) showed the statistical importance of a normal bundle in the context of both inference in curved exponential families and undertaking inference on an interest parameter when there are, potentially, very large numbers of nuisance parameters. One way of understanding this structure is to consider when $f_X(x|\mu)$ is a parametric family of density functions and we can construct a larger family $f_X(x|\mu, \xi)$ such that (1) $f_X(x|\mu, 0)$ equals $f_X(x|\mu)$ for all μ, (2) for each fixed μ_0 the family $f_X(x|\mu_0, \xi) - f_X(x|\mu_0)$ is Fisher orthogonal to the score

of $f_X(x|\mu)$ at μ_0 and (3) the family $f_X(x|\mu_0, \xi)$ has zero -1-curvature either at $(\mu_0, 0)$ or globally. Here the -1-curvature is defined with the ∇^{-1}-connection, see Amari (1990). A family $f_X(x|\mu, \xi)$ which satisfies conditions (1), (2) and (3) can be called a normal -1-affine fibre bundle and the subfamily parametrized by ξ for a given μ_0 is called the fibre at μ_0.

As shown in Amari (1990) such normal bundles arise naturally when $f_X(x|\mu, \xi)$ is a full exponential family and the fibres are defined by the ancillary family associated with the maximum likelihood estimate of the curved exponential family $f_X(x|\mu)$. In this example it is the geometric structure of the -1-affine fibre bundle which ensures third order asymptotic efficiency in the estimation of the subfamily, $f_X(x|\mu)$, (Amari 1990, Theorem 5.6).

The normal bundle is easily constructed when $f_X(x|\mu)$ is a curved exponential family by using the embedding full exponential family, but there is a much more general way of defining normal -1-affine fibre bundles on general parametric families. This construction is most easily understood in the affine space defined by

$$\langle X_{\text{Mix}}, V_{\text{Mix}}, + \rangle .$$

In this construction the set X_{Mix} is defined as

$$X_{\text{Mix}} = \left\{ f(x) | f \in C^\infty(S, R), f \in L^2(\nu), \int f(x) \mathrm{d}\nu = 1 \right\},$$

a subset of the smooth functions from the fixed support set S to R, and ν is a measure defined to have support on S. Furthermore V_{Mix} is defined as the vector space

$$V_{\text{Mix}} = \left\{ f(x) | f \in C^\infty(S, R), f \in L^2(\nu), \int f(x) \mathrm{d}\nu = 0 \right\}.$$

Finally the addition operator is the usual addition of functions. Note that there is no requirement that an element of X_{Mix} be a positive function and in fact the space of density functions is a convex subset. The affine structure of $\langle X_{\text{Mix}}, V_{\text{Mix}}, + \rangle$ agrees with Amari's -1 connection in the sense that affine subsets are ∇^{-1} flat, and will be referred to as the -1 geometry, see Marriott (2002). The condition on being a subset of $L^2(\nu)$ allows inner products to be defined on the affine space in a natural way.

Using this construction a finite dimensional normal fibre at μ_0 to a family $f_X(x|\mu)$ is spanned by a set of linearly independent functions $g_i(x, \mu_0)$ $i = 1, \ldots, n$ such that $g_i(x, \mu_0) \in V_{\text{Mix}}$ and all of these functions are Fisher orthogonal with the score vector of the parametric family at μ_0. This condition is given by

$$\int_S \frac{g_i(x, \mu_0) \frac{\partial f_X}{\partial \mu}(x|\mu)|_{\mu=\mu_0}}{f_X(x|\mu_0)} \mathrm{d}\nu = 0, \tag{3}$$

where S is the sample space. Care is required in the interpretation of (3) as an orthogonality condition. Amari (1990, Chapter 3) considers various representations of the tangent vector $\frac{\partial}{\partial\theta^i}$ for a general parametric family $f_X(x|\theta)$. The $+1$ representation is the vector space isomorphism given by

$$\frac{\partial}{\partial\theta^i} \to \frac{\partial \log f_X(x|\theta)}{\partial\theta^i}$$

under which the Fisher information matrix has the usual co-ordinate form

$$\int \frac{\partial \log f_X(x|\theta)}{\partial\theta^i}\frac{\partial \log f_X(x|\theta)}{\partial\theta^j} f_X(x|\theta)\mathrm{d}x.$$

However, much more relevant for this paper is the -1 representation defined via

$$\frac{\partial}{\partial\theta^i} \to \frac{\partial f_X(x|\theta)}{\partial\theta^j}.$$

Since it is clear that

$$\int \frac{\partial \log f_X(x|\theta)}{\partial\theta^i}\frac{\partial \log f_X(x|\theta)}{\partial\theta^i} f_X(x|\theta)\mathrm{d}x = \int \frac{\frac{\partial f_X(x|\theta)}{\partial\theta^i}\frac{\partial f_X(x|\theta)}{\partial\theta^j}}{f_X(x|\theta)}\mathrm{d}x$$

the form of (3) is the natural generalisation of this -1-representation of the Fisher information on the vector space V_{Mix} of which both g_i and $\frac{\partial f_X}{\partial\mu}(x|\mu)$ are members.

The fibre is defined as the finite dimensional affine space

$$f_X(x|\mu_0) + \sum_{i=1}^{n} \lambda_i g_i(x, \mu_0),$$

and a normal bundle is then automatic if we have this condition for all μ and the dependence of $g_i(x, \mu)$ on μ is smooth. Thus the construction of rich families of normal bundles in $\langle X_{\mathrm{Mix}}, V_{\mathrm{Mix}}, +\rangle$ depends on being able to find such families $g_i(x, \mu)$. The local mixture model gives one way of constructing these fibres.

2.1 Local mixture models

The idea behind a local mixture model is to approximate (1) by replacing the dependence the infinite dimensional Q with a finite dimensional parameter, where the approximation is good for a particular class of mixing distributions. In particular if $\mathcal{Q}$ is the set of distributions which are 'close to a delta function', then a Laplace approximation gives that

$$\int f_X(x|\mu+\eta)\mathrm{d}Q(\eta) \approx f_X(x|\mu) + \alpha\frac{\partial^2}{\partial\mu^2}f_X(x|\mu) + \beta\frac{\partial^3}{\partial\mu^3}f_X(x|\mu) \qquad (4)$$
$$\doteq f_X(x|\mu,\alpha,\beta)$$

where $\alpha = \frac{1}{2}\int\eta^2\mathrm{d}Q(\eta), \beta = \frac{1}{3!}\int\eta^3\mathrm{d}Q(\eta)$. Note that the definition given here is not the same as that in Marriott (2002), and details can be found in Anaya-Izquierdo and Marriott (2005, submitted). In particular note that when $f_X(x\,\mu)$ is a natural exponential family, parametrized by the mean parameter then it follows that the terms $\frac{\partial^i}{\partial\mu^i}f_X(x|\mu)$ are Fisher orthogonal and linearly independent, thus we get a normal bundle structure. Furthermore, the family $f_X(x|\mu,\alpha,\beta)$ is identified in all its parameters, (K. Anaya-Izquierdo and P. Marriott (submitted)).

Since Example 1 satisfies the above conditions we can think of applying the local mixture model to the problem of marginal inference on μ. Thus as long as Q lies in the class of distributions where (4) is a good approximation, then the infinite dimensional marginalization (2) can be well approximated by a 2D marginalization

$$\int P(\mu|x_1,\ldots,x_n,Q)\mathrm{d}P(Q) \approx \int\left\{\prod_{i=1}^{n}f_X(x_i|\mu,\alpha,\beta)\right\}\mathrm{d}P(\alpha,\beta). \qquad (5)$$

This is clearly a much easier computation, but does need care over the region of (α,β) space where the approximation (4) makes sense. The ease of construction of the fibre in the affine space $\langle X_{\mathrm{Mix}}, V_{\mathrm{Mix}}, +\rangle$ comes at the price that it is a superset of the space of positive densities, and hence boundaries on parameter space have to be defined. There are in fact two types of boundary which need care in the analysis of Example 1. These are

Hard boundary: defined by condition that $f_X(x|\mu,\alpha,\beta) \geq 0$ for all x.
Soft boundary: defined by $f_X(x|\mu,\alpha,\beta)$ lying in convex hull of curve $f_X(x|\mu)$ in the mixture affine geometry.

The hard boundary ensures that we are dealing with real density functions, while the soft boundary ensures that the resultant approximation can be realised by an exact mixture model, see (K. Anaya-Izquierdo and P. Marriott (submitted)) for details.

In Example 1 the base family is $+1$-affine, while the fibres are -1-affine. The model is illustrated in Fig. 1, which represents a 3D manifold with boundary embedded in the affine space $\langle X_{\mathrm{Mix}}, V_{\mathrm{Mix}}, +\rangle$. It is the hard and soft boundaries discussed which give the illustrated boundaries, and note that these boundaries can have singularities. In general the geometry of the fibre is closer to that of a simplex than a manifold.

Recall that the question of interest for this paper is, *for a given data-set*, can a finite dimensional integral be constructed that gives an (arbitrarily) good

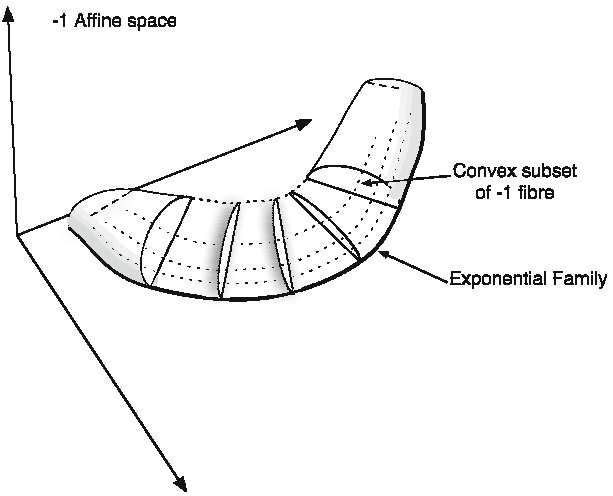

Fig. 1 Local mixture model embedded in −1 affine space

approximation to the infinite dimensional integral (2)? To investigate this consider the local approximation (4) in more detail. This approximation is used at a finite number of data points $x_i, i = 1, \ldots, n$ in (5) and can be thought of as a Taylor expansion inside the integral plus the regularity conditions which allow the operations of integration and differentiation to commute

$$\begin{aligned}\int f_X(x_i|\mu+\eta)\mathrm{d}Q(\eta) &= \int \{f_X(x_i|\mu) + \eta\frac{\partial}{\partial\mu}f_X(x_i|\mu) + \frac{\eta^2}{2}\frac{\partial^2}{\partial\mu^2}f_X(x_i|\mu) \\ &\quad + \frac{\eta^3}{3!}\frac{\partial^3}{\partial\mu^3}f_X(x_i|\mu) + R(x_i,\mu,\eta)\}\mathrm{d}Q(\eta) \\ &= f_X(x_i|\mu) + \alpha\frac{\partial^2}{\partial\mu^2}f_X(x_i|\mu) + \beta\frac{\partial^3}{\partial\mu^3}f_X(x|\mu) \\ &\quad + \int \{R(x_i,\mu,\eta)\}\,\mathrm{d}Q(\eta). \end{aligned} \tag{6}$$

Note that this explicitly uses the identification assumption that the expected value of Q is zero in the mean parametrization. Also it is easy to check, and shown in (K. Anaya-Izquierdo and P. Marriott (submitted)), that for many exponential families the second and higher derivatives in the mean parametrization satisfy orthogonality condition (3) i.e.

$$\int \frac{\frac{\partial^k}{\partial\mu^k}f_X(x|\mu)\frac{\partial}{\partial\mu}f_X(x|\mu)}{f_X(x|\mu)}\mathrm{d}x,$$

for $k \geq 2$.

Thus the approximation given by (4) will be useful when we know that the remainder is uniformly 'small' in i and the class of mixing distributions under consideration. Intuitively the local mixture approximation will be good for all mixing distributions Q which have most of their mass in a region where each of the n functions $f_X(x_i|\mu+\eta)$ can be well approximated by cubic functions of η.

Note that in the integral (2) it is required that the approximation

$$\prod_{i=1}^{n}\left\{f_X(x_i|\mu,\alpha,\beta)+\int R(x_i,\mu,\eta)\mathrm{d}Q(\eta)\right\}\approx\prod_{i=1}^{n}f_X(x_i|\mu,\alpha,\beta)$$

For this to be reasonable we must have that

$$\prod_{i=1}^{n}\left\{1+\frac{\int R(x_i,\mu,\eta)\mathrm{d}Q(\eta)}{f_X(x_i|\mu,\alpha,\beta)}\right\}\approx 1$$

Hence it is the relative error

$$\frac{\int R(x_i,\mu,\eta)\mathrm{d}Q(\eta)}{f_X(x_i|\mu,\alpha,\beta)} \tag{7}$$

which must remain small and this is impossible if the local mixture $f_X(x_i|\mu,\alpha,\beta)$ is zero, i.e. on the hard boundary,

2.2 Numerical illustration

To illustrate some of these issues a numerical example of Example 1 is explored here which shows some of the geometric points which arise. A data set has been generated from a discrete mixture of two well separated exponential distributions, one with mean 20 the other with mean 2 and a mixing proportion of 0.5. The data is plotted in Fig. 2, together with the maximum likelihood fit from the unmixed model family. Of note in the data is the very large observation which will turn out to be highly influential in a mixture model analysis. Also shown in Fig. 2 is the posterior distribution for the mean parameter μ under the assumption that the unmixed model holds. This is object which is of interest, in that we want to see how inference on μ is affected by different assumptions on possible mixing distributions. Note that knowing that this is the object of interest gives us a 'scale' relative to which we can measure how good an approximation is. Perturbations which only have a small effect on the μ-posterior can, will little inferential loss, be disregarded.

To illustrate the quality of the Taylor approximation given by (6) consider Fig. 3. In order to get good behaviour we need the approximation to be good at each of the observed points x_i uniformly in some region of $\mu+\eta$. Figure. 3 illustrates the approximation at $\mu=\widehat{\mu}$, the maximum likelihood estimate of the unmixed model. Furthermore, the function $f_X(x_i|\mu+\eta)$ is shown in a region of

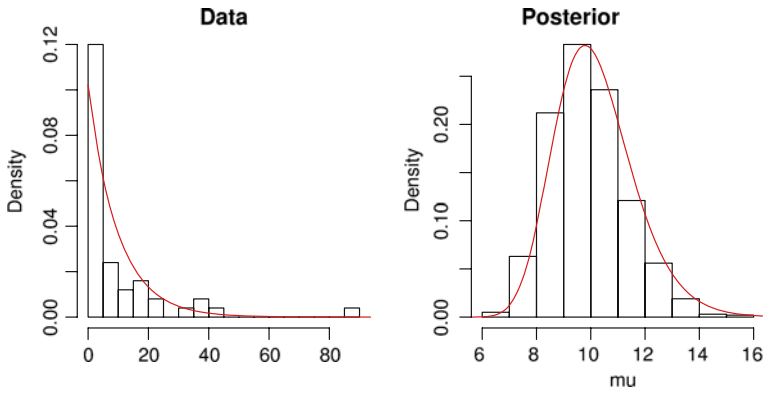

Fig. 2 Dataset with fitted exponential density

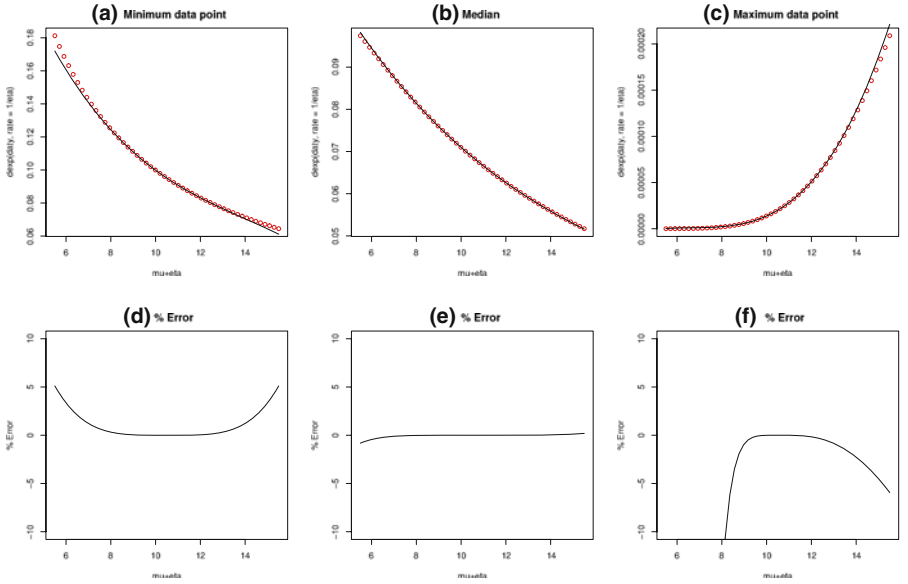

Fig. 3 Checking the quality of the polynomial approximation for Example 1

parameter space $\widehat{\mu} \pm 5$ which is reasonably large relative to the uncertainty in μ shown in Fig. 2. This figure illustrates the effect of looking at different observed data values by using the minimum, median and maximum value of the data respectively in panels (a), (b) and (c). The panels below each of these shows the relative error, on a percentage scale, of approximation (6), i.e. by plotting

$$\frac{R(x_i, \mu, \eta)}{f_X(x_i|\mu) + \eta \frac{\partial}{\partial \mu} f_X(x_i|\mu) + \frac{\eta^2}{2} \frac{\partial^2}{\partial \mu^2} f_X(x_i|\mu) + \frac{\eta^3}{3!} \frac{\partial^3}{\partial \mu^3} f_X(x_i|\mu)} \tag{8}$$

for values of $\eta \in [-5, 5]$. It is clear that if (8) is *uniformly* small then the important term (7) will be small for *all* mixing distributions which have support in $\mu \pm 5$. A good diagnostic is to see how large his relative error gets in the region of interest.

There are various issues to be considered. Firstly it is clear that the quality of the approximation is not uniform at each observed data point, rather at the extremes the worst approximation can be seen. This is quite consistent with the well known fact that with mixture models, unlike exponential families, inferential information is not uniformly spread across all data points, rather there can be highly influential points. Secondly it is clear from panels (c) and (f) why the polynomial approximation used in (6) is poor. The function $f_X(x_{\max}|\mu + \eta)$ approaches zero for large negative η exponentially fast, and so the polynomial approximation can not do well here. In fact it is clear that simply adding higher order derivatives to the approximation (6) will not help either since a finite polynomial can not approximate exponential decay.

3 Global extensions

The local mixture model is thus based on expanding each of the terms $f_X(x_i|\mu + \eta)$ in a linear space of basis functions of η such that a uniformly small remainder results. The polynomial set of basis functions which works well in the local expansion do not work well over larger possible values of η since the number of components needed very rapidly gets too large. Another way of selecting a functional basis needs to be found.

In Marriott and Vos (2004) a data based eigen-function approach was used to find approximate sufficient statistics for highly curved (in a +1-geometry sense) parametric families. Analogously we can consider using the −1-geometry version of this to construct the basis of functions using which we will make approximations.

The tool that is going to be used is a functional version of principle component analysis (PCA) constructed using the Karhunen–Loève (K–L) decomposition, see Papoulis (1984). In order to make this rigourous we first define a compact subsets C of parameter space, and a class of distributions defined on C. Consider the eigen-function equation

$$\int_C G(\mu_1, \mu_2) e_i(\mu_1) \mathrm{d}\mu_1 = \lambda e_i(\mu_2), \tag{9}$$

where the kernel can have the form

$$G_1(\mu_1, \mu_2) = \int \frac{(f_X(x|\mu_1) - f_X(x|\mu))\,(f_X(x|\mu_2) - f_X(x|\mu))}{f_X(x|\mu)} \mathrm{d}x \tag{10}$$

or alternatively a kernel more determined by the data-set in question

$$G_2(\mu_1, \mu_2) = \sum_{i=1}^{n} \left\{ \frac{(f_X(x_i|\mu_1) - f_X(x_i|\mu))\,(f_X(x_i|\mu_2) - f_X(x_i|\mu))}{f_X(x_i|\mu)} \right\}. \tag{11}$$

By spectral theory, Rudin (1973, pp. 305–311), if C is compact and G continuous there exists a countable set of eigenfunctions $\{e_i\}$ with eigenvalues $\{\lambda_i\}$ ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. These eigenfunctions can be chosen to form an orthonormal basis for the set of smooth functions from C to R with respect to the inner product defined by

$$\langle f, g \rangle = \int_C f(\mu) g(\mu) \mathrm{d}\mu.$$

The number of eigen-functions which are selected for the basis expansion of $f_X(x_i|\mu + \eta)$ is then chosen using the spectrum of the eigen-function equation (9) in the usual PCA way. The idea is to throw away only the eigen-functions with small eigen-values. If we do this, keeping a basis of K eigen-functions, then the corresponding version of (6) will be

$$\int_C \left\{ f_X(x_i|\mu) + \sum_{j=1}^{K} s_j(x_i) e_j(\mu + \eta) + R(x_i, \mu, \eta) \right\} \mathrm{d}Q(\eta). \tag{12}$$

Using this expansion the global mixture expansion will have $K + 1$ parameters and has the form

$$f_X(x_i|\mu, \alpha_1, \ldots, \alpha_K) = f_X(x_i|\mu) + \sum_{j=1}^{K} \alpha_j s_j(x_i),$$

where

$$\alpha_j = \int_C e_j(\mu + \eta) \mathrm{d}Q(\eta).$$

Thus the infinite dimensional marginalizing equation (2) has become a finite dimensional integral over K-dimensions. As before the computation has to be over a region which takes care of the hard and soft boundaries which ensures that:

[C1] : $f_X(x_i|\mu, \alpha_1, \ldots, \alpha_K) \geq 0$,
[C2] : $f_X(x_i|\mu, \alpha_1, \ldots, \alpha_K) \in$ convex hull of $\{f_X(x|\mu + \eta)|\mu + \eta \in C\}$.

In order to show that these functions $s_i(x)$ can define a normal bundle in the same way that the derivatives of $f_X(x|\mu)$ we need to check that they lie in the vector space V_{Mix}.

Theorem 1 *The function s_i defined in* (12) *are elements of V_{Mix}.*

Proof This follows from the definition of s_i in (12) and the othogonality of the eigenfunctions by

$$s_i(x) = \int_C \{f_X(x, \mu + \eta) - f_X(x, \mu)\}\, e_i(\eta) \mathrm{d}\eta.$$

Hence it follows that

$$\begin{aligned}
\int_S s_i(x)\mathrm{d}x &= \int_S \int_C \{f_X(x, \mu + \eta) - f_X(x, \mu)\}\, e_i(\eta)\mathrm{d}\eta\mathrm{d}x \\
&= \int_C \int_S \{f_X(x, \mu + \eta) - f_X(x, \mu)\}\, \mathrm{d}x\; e_i(\eta)\mathrm{d}\eta \\
&= \int_C 0 e_i(\mu)\mathrm{d}\mu \\
&= 0.
\end{aligned}$$

□

As shown in (K. Anaya-Izquierdo and P. Marriott (submitted)) when $f_X(x|\mu)$ is an exponential family in the mean parametrisation, then all terms in the Taylor approximation, after the first, are Fisher orthogonal to the score vectors $\frac{\partial}{\partial \mu} f_X(x|\mu)$, thus the normal bundle structure is automatic. This will not necessarily hold in the eigen-function expansions and so care needs to be taking with identification. In general the expansion is done in the subspace which is orthogonal to $\frac{\partial}{\partial \mu} f_X(x|\mu)$.

3.1 The error in the expansion

It is important to understand the nature of the remainder term in any of these expansions. The following calculations shows how knowledge of the spectrum of the eigen-function expansion (9) determines bounds on this remainder. In particular for approximations to be relevant to marginal posterior expressions such as (2) we need approximations which are uniform within the relevant class of mixing distributions $\mathcal{Q}$.

In this section the data driven kernel G_2 is considered, but similar results can be calculated for G_1 and other kernels. Throughout we will assume sufficient regularity on the family $f_X(x|\mu)$ so that all operations below are valid.

Denote the element of V_{Mix} which is centred at $f_X(x|\mu)$ by

$$\overline{f_X(x|\eta)} = f_X(x|\mu+\eta) - f_X(x|\mu),$$

and the remainder

$$R(x,\eta) = \overline{f_X(x|\eta)} - \sum_{i=1}^{K} s_j(x)e_j(\eta) = \sum_{j=K+1}^{\infty} s_j(x)e_j(\eta).$$

The following calculation is useful in showing the nature of the bound.

$$\begin{aligned}
\sum_{i=1}^{n} \frac{R(x_i,\mu)^2}{f_X(x_i|\mu)} &= \sum_{i=1}^{n} \frac{1}{f_X(x_i|\mu)} \left\{ \sum_{j=K+1}^{\infty} s_j(x_i)e_j(\eta) \right\}^2 \\
&= \sum_{i=1}^{n} \sum_{j,k>K} \frac{1}{f_X(x_i|\mu)} s_j(x_i)s_k(x_i)e_j(\eta)e_k(\eta) \\
&= \sum_{i=1}^{n} \sum_{j,k>K} \frac{1}{f_X(x_i|\mu)} \left\{ \int_C \overline{f_X(x_i|\eta)} e_j(\eta)\mathrm{d}\eta \right\} \\
&\quad \times \left\{ \int_C \overline{f_X(x_i|\eta)} e_k(\eta)\mathrm{d}\eta \right\} e_j(\eta)e_k(\eta) \\
&= \sum_{j,k>K} \left\{ \int_C \int_C \sum_{i=1}^{n} \left\{ \frac{\overline{f_X(x_i|\eta_1)}\ \overline{f_X(x_i|\eta_2)}}{f_X(x_i|\mu)} \right\} \right. \\
&\quad \left. \times e_j(\eta_1)e_k(\eta_2)\mathrm{d}\eta_1\mathrm{d}\eta_2 \right\} e_j(\eta)e_k(\eta) \\
&= \sum_{j,k>K} \int_C \left\{ \int_C G_2(\eta_1,\eta_2)e_j(\eta_1)\mathrm{d}\eta_1 \right\} e_k(\eta_2)\mathrm{d}\eta_2 e_j(\eta)e_k(\eta) \\
&= \sum_{j,k>K} \left\{ \int_C \lambda_j e_j(\eta_2)e_k(\eta_2)\mathrm{d}\eta_2 \right\} e_j(\eta)e_k(\eta) \\
&= \sum_{j,k>K} \lambda_j \delta_{jk} e_j(\eta)e_k(\eta) \\
&= \sum_{j>K} \lambda_j e_j^2(\eta),
\end{aligned}$$

where δ_{jk} is the delta function. Thus a bound on the tail of the expansion $\sum_{j>K} \lambda_j e_j^2(\eta)$ is a bound, for each observed x_i on the relative error

$$\frac{R(x_i, \mu)^2}{f_X(x_i|\mu)}.$$

Since, for each fibre $f_X(x_i|\mu)$ is known the tail of the expansion gives a direct bound on $R(x_i, \mu)$ which is uniform for all possible mixing distributions Q. Thus the smallness of the bound (7) will be assured as long as the term $f_X(x_i|\mu, \alpha, \beta)$ is bounded away from zero as discussed above.

3.2 Numerical illustration

In this section the eigen-function approach is applied to the data used in Sect. 2.2. All calculations are done numerically with a compact set C being chosen and then discretized uniformly into a vector of dimension 50 in each set. The eigen-functions from (9) are then calculated from the natural discretised version of this equation. In general this means that all calculations are done in a relatively high, but finite, dimensional space. There are various computational issues which need careful consideration tol allow numerical schemes to run at a suitable speed but these issues will only be sketched here.

The method requires that a compact set C is defined around μ in order to analyse the -1 fibre at μ. Here two choices are made with a small and large C being selected, again small and large are relative to the posterior plot in Fig. 2. Here small was taken as $\mu \pm 0.1$ while large was $\mu \pm 5$ in the mean parametrization. Once the compact set is chosen an eigen-analysis of (9) is undertaken given eigen-functions and eigen-values. For the small set the top eigen-functions are shown in Fig. 4 and for the large set in Fig. 5. It is clear from Fig. 4 that the polynomial approximation used by the local mixture model is reproduced by the eigen-functions in (6). However it is also clear from Fig. 5 that the basis

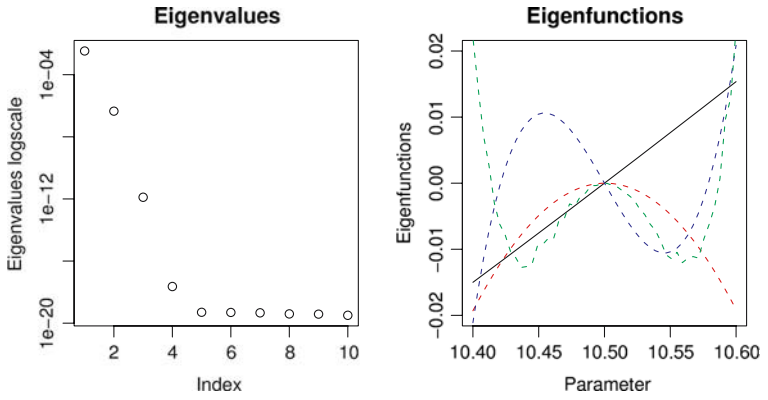

Fig. 4 Eigenvalues and vectors for very small set $\mu \pm 0.1$

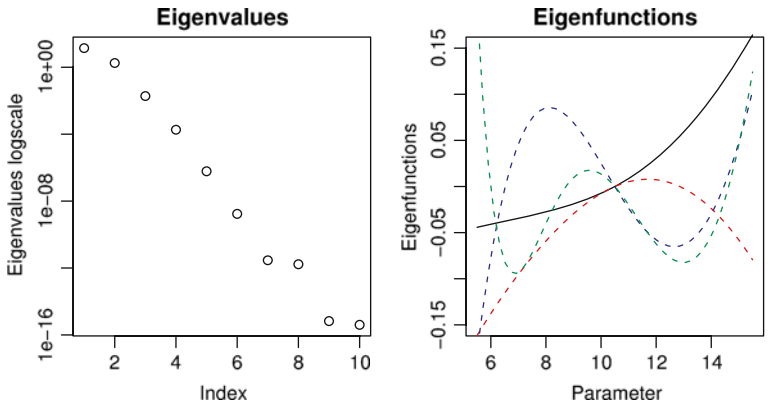

Fig. 5 Eigenvalues and vectors for large set $\mu \pm 5$

of polynomials do not do a good job for the larger C. Further the spectra of the two different cases shows that fewer eigen-functions are needed in the local analysis than the global. These spectra give us a way of selecting the number of terms needed, in the way familiar to PCA.

The key idea in the eigen-function approximation is that the approximation

$$f_X(x_i|\mu+\eta) \approx f_X(x_i+\mu) + \sum_{j=1}^{K} s_j(x_i) e_i(\mu+\eta) \tag{13}$$

is good for all observed data points x_i and for all η in the set C. If this holds then *any* integral $\int_C f_X(x_i|\mu+\eta)\mathrm{d}Q(\eta)$ can be approximated by $f_X(x_i|\mu) + \sum_{j=1}^{K} \alpha_j s_j(x_i)$ for some vector $\alpha_1, \ldots, \alpha_K$ where $\alpha_j := \int_C e_j(\mu+\eta)\mathrm{d}Q$.

The quality of the approximation (13) in the large C case is shown in Fig. 6 for a selection of different data values in the same way as Fig. 3. It can be seen that the (4D) approximation, given by the dots, gives an extremely good approximation to the actual likelihood, given by the solid line. The percentage relative error is plotted on the same scale as Fig. 3 and on this scale the error is negligible. The quality of this approximation can be formalised by analysing the spectrum of the eigen-function equation given by (9) and is shown in Figs. 4 and 5.

Having shown that the finite dimensional approximation $f_X(x_i|\mu, \alpha_1, \ldots, \alpha_K)$ is all that is needed to understand the marginal inference problem on μ we need to address the hard and soft boundaries for which this model is defined. The soft boundary is defined via a convex hull in the -1-affine space. One way to characterise this is to find its extremal point. This is done numerically, using discretisation, in Fig. 7 for the small C case. The points lie in a 3D affine space and the pairwise co-ordinates are shown. It is necessarily to sample from the posterior of this convex hull and this can be done by calculating the Delaunay triangulation using an algorithm such as Quick–Hull (see Barber et al. 1996) then rejection sampling can easily be done to sample from the posterior.

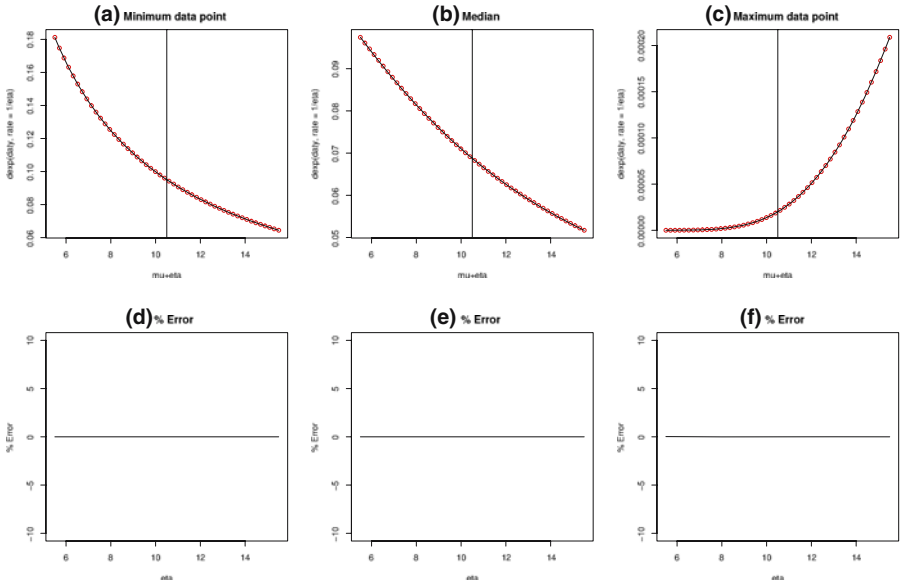

Fig. 6 Checking the quality of the eigenvalue approximation for Example 1

4 Discussion

This entire theory has been conditional on choosing the compact set C. This set defined the class of mixing distributions over which the marginalization has be done over. Loosely we have integrated over the class of $\mathcal{Q}_C$ such that (1) either Q has support in C , (2) or, and this is much more useful,

$$\int_C \mathrm{d}Q(\eta) \geq 1 - \epsilon$$

for some pre-selected small ϵ.

Also of interest is the effect on C on the number of components needed in the approximation. Let $K(C)$ be the number of components required for set C according to any well-defined criterion used in PCA, for example a percentage of the sum of the chosen eigen-values in terms of the total. Numerical experiments with Example 1 gives rise to the following conjecture.

Conjecture The number of components $K(C)$ will be bounded as $|C| \to \infty$, but the bound will be data dependent. The bound on the number of components needed tells us the maximum possible number of 'nuisance' parameters that are needed for marginal inference on μ for a given data-set. This has been described in the literature as the effective degrees of freedom for a problem. Note that unlike the derivation of this quantity in versions of information criteria such as AIC or BIC, the underlying geometric model here is not a manifold, but much more closely related to a simplex due to the convex hull condition.

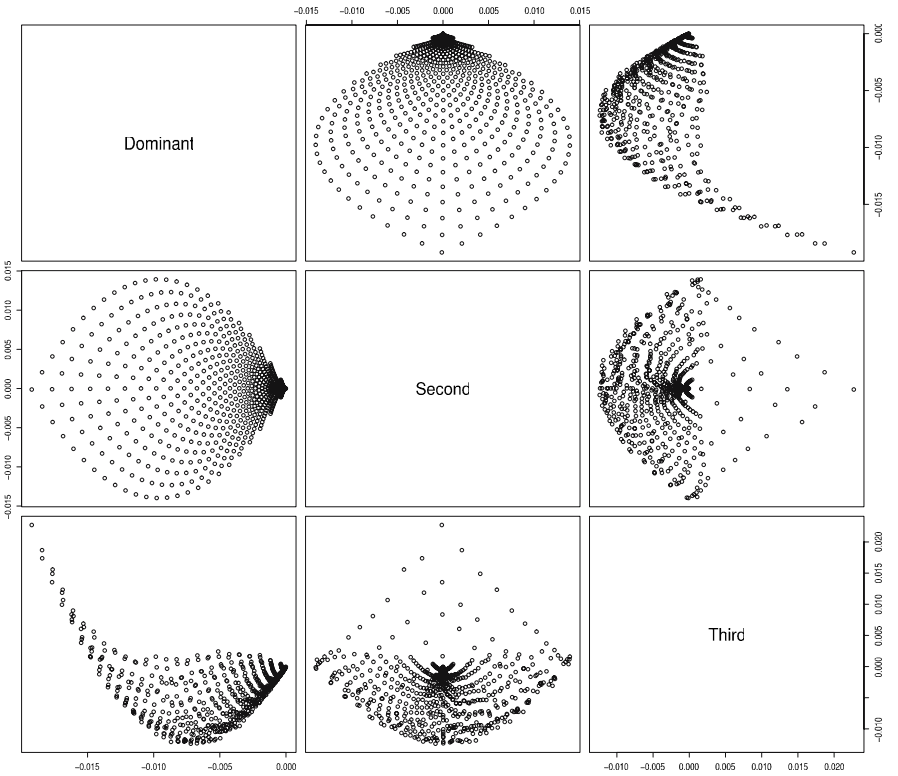

Fig. 7 Convex hull defined by extremal points in −1 fibre. In this plot, via discretization, a set of extremal points of the convex hull in 3D has been plotted. The convex hull is then well approximated by the convex hull of these points

References

Amari, S.-I. (1990). *Differential geometry methods in statistics.* Lecture Notes in Statistics, *28*. Berlin Heidelberg NewYork: Springer.

Anaya-Izquierdo, K., Marriott, P. (2005). Local mixtures of Natural Exponential families with quadratic variance function. In: *Proceeding of the 2nd International symposium on information geometry and its applications*, pp. 190–197.

Barber, C. B., Dobkin, D. P., Huhdanpaa, H. T. (1996). The Quickhull algorithm for convex hulls. *ACM Transactions Mathematical Software*, *22*(4), 469–483.

Critchley, F., Marriott, P. (2003). Data informed influence analysis. *Biometrika*, *91*(1), 125–140.

Eguchi, S. (2005). Tubular modelling approach to statistical method for observational studies. In: *Proceeding of the 2nd International symposium on information geometry and its applications*, pp. 1–8.

Marriott, P. (2002). On the local geometry of mixture models. *Biometrika*, *89*(1), 77–93.

Marriott, P. (2003). On the geometry of measurement error models. *Biometrika*, *90*(3).

Marriott P. (2005). Local and global mixture models. In: *Proceeding of the 2nd International symposium on information geometry and its applications*, pp. 82–88.

Marriott, P., Vos, P. (2004). On the global geometry of parametric models and information recovery. *Bernoulli*, *10*(2), 1–11.

Papoulis, A. (1984), *Probability, random variables, and stochastic processes.* London: McGraw-Hill.

Rudin, W. (1973). *Functional Analysis.* London: McGraw-Hill.