Local Mixture Models of Exponential Families

Karim Anaya-Izquierdo
Department of Statistics
The Open University
Walton Hall
Milton Keynes, MK7 6AA
UK

Paul Marriott*
Department of Statistics and Actuarial Science
University of Waterloo
200 University Avenue West
Waterloo,
Ontario, N2L 3G1
Canada

July 6, 2006

Running Title: Local mixture models
Key words: Affine geometry; Convex geometry; Differential geometry; Dispersion model; Exponential families; Mixture model; Statistical manifold.

Abstract

Exponential families and the related exponential dispersion models are the workhorses of parametric modelling theory. One reason for their popularity is their associated inference theory which is very clean both from a theoretical and a computational point of view. One way in which this set of tools can be enriched in a natural and interpretable way is through mixing. This paper applies and develops the idea of local mixture modelling to exponential families. It shows that the resulting highly interpretable and flexible models have enough structure to retain the attractive inferential properties of exponential families. In particular results on identification, parameter orthogonality and log-concavity of the likelihood are proved. The paper explores the differences between mixing with small exponential dispersion families and more general mixtures in a very geometry way. The paper also looks at the strong links with Amari's -1-fibre bundles and their associated geometry.

1 Introduction

Mixture models have many appealing properties for statistical modelling, see Titterington, Smith, and Makov (1985), Lindsay (1995) or McLachlan and Peel (2001). Marriott (2002) and Marriott (2003) consider a restricted form of mixing, called *local mixing*, which allows a considerable increase in inferential tractability when compared to general mixture models. This paper constructs a comprehensive theoretical framework for the theory of local mixture models for the exponential and exponential dispersion families. It introduces a new class of models called *true local mixtures* whose members are flexible, identifiable, and interpretable. Furthermore the class has a very tractable likelihood structure which means that these models are straightforward, if non-standard, inferentially.

Exponential models form the backbone of parametric modelling. They are, in terms of the natural parameter, of the form

$$f_X(x;\theta) = \exp\{\theta x - k_\nu(\theta)\}\,\nu(x)$$

with respect to the σ-finite measure ν on $\mathbf{R}$. An alternate and important parametrization is the expected parameter μ, where the transformation from the natural parameters is defined by

$$\mu = E_{f_X(x;\theta)}(X).$$

Throughout the regularity conditions on parametric families, given in Appendix A and similar to those in Amari (1985), is assumed. Important special cases of these models which are given special attention in this paper are families of the type called real natural exponential families with quadratic variance functions (NEF-QVF), see Morris (1982) and Letac (1992).

Generalisations of these families are given by exponential dispersion models. These have proved to be very successful at increasing modelling flexibility while keeping within a well understood inferential framework. An excellent treatment of their theory and application can be found in Jorgensen (1997). This paper there also considers local mixtures for models of the exponential dispersion form

$$f_X(x;\theta,\lambda) := \exp[\lambda\{x\theta - k_{\nu_\lambda}(\theta)\}]\nu_\lambda(x)$$

where $\nu_\lambda(x)$ is independent of θ, see Jorgensen (1997, page 72).

The paper has the following structure. Section 2 gives, in Definition 2, an alternative definition of the local mixture model to the one found in Marriott (2002). It also defines the true local mixture model with Theorem 1 giving a computationally attractive characterisation. Theorem 2 is the main tool for the identification and orthogonality results used in Theorem 3. Section 2.2 explores the parallels between local mixture models and the fibre bundle structure of Amari (1985), with Theorem 6 showing the simple, but inferentially very powerful result that the likelihood function is log-concave on the fibre. In contrast with this similarity to the likelihood for exponential families the sensitivity of the boundary

of the likelihood to single data points is also explored in this section. Section 3 looks at different asymptotic expansions which have interpretations as local mixture models with Theorem 7 and Section 4 giving the main results for the continuous and discrete mixing cases respectively. These results give an asymptotic interpretation to the parameters in the local mixture models. They are applied, mostly to the binomial case, in §4.2. In particular the way that different types of boundaries in the fibre affect inference are described in detail. Finally §5 and Theorem 9 explore how local mixing and adding a dispersion parameter both enrich exponential families in similar ways.

2 Local mixture models

DEFINITION 1 Consider the affine space defined by

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

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

$$X_{\text{Mix}} = \{f(x) | f \in L^2(\nu), \int f(x) d\nu = 1\},$$

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

$$V_{\text{Mix}} = \{f(x) | f \in L^2(\nu), \int f(x) d\nu = 0\}.$$

Finally the addition operator is the usual addition of functions.

The following definition gives a general definition of a local mixture model in the case of a regular exponential family.

DEFINITION 2 The *local mixture model* of a regular exponential family $f(x;\mu)$ is defined via its mean parameterisation as

$$g(x;\mu,\lambda) := f(x;\mu) + \sum_{i=2}^{r} \lambda_k f^{(k)}(x;\mu),$$

where

$$f^{(k)}(x;\mu) = \frac{\partial^k}{\partial \mu^k} f(x;\mu).$$

Here r is called the order of the local mixture model.

In order to see the link between Definitions 1 and 2 note that $f(x;\mu) \in X_{\text{Mix}}$ and furthermore, by the regularity conditions in Appendix A, all μ-derivatives of $f(x;\mu)$ are elements of V_{Mix}. Note also that elements of X_{Mix} are not restricted to be non-negative rather the space of regular density functions, $\mathcal{F}$, is a convex subset of the affine space $(X_{\text{Mix}}, V_{\text{Mix}}, +)$. It follows that restricting the family $g(x;\mu,\lambda)$ to $\mathcal{F}$ induces a boundary in the parameter space.

DEFINITION 3 The *hard boundary* of the local mixture model is defined as the subset of parameter space where

$$g(x;\mu,\lambda) = 0$$

for some x in the support of ν.

EXAMPLE 1 The local mixture model for the binomial family, of order 4, has a probability mass function of the form

$$g(x;\mu,\lambda_2,\lambda_3,\lambda_4) = \frac{n!\,\mu^z\,(n-\mu)^{n-z}}{z!\,(n-z)!\,n^n}\left\{1+\lambda_2 p_2(x,\mu)+\lambda_3 p_3(x,\mu)+\lambda_4 p_4(x,\mu)\right\} \qquad (1)$$

where p_i are polynomials which are given explicitly in Appendix C.

In §3 it is shown how local mixture models can be viewed as asymptotic approximations to actual mixtures. However the following example shows a way in which local mixtures can be qualitatively different from mixtures and motivates the definition of a true local mixture in Definition 4.

EXAMPLE 2 Consider the following example of local mixing over the normal family, $\phi(x;\mu,1)$, with known variance of 1. The local mixture model of order 4 is

$$\begin{aligned}\phi(x;\mu,1)\{1 \quad &+ \quad \lambda_2\left(-1+x^2-2\,x\mu+\mu^2\right)+\lambda_3\left(-3\,x+x^3-3\,x^2\mu+3\,x\mu^2+3\,\mu-\mu^3\right)\\ &+ \quad \lambda_4\left(3-6\,x^2+12\,x\mu-6\,\mu^2+x^4-4\,x^3\mu+6\,x^2\mu^2-4\,x\mu^3+\mu^4\right)\}.\end{aligned}$$

It is easy to show that the variance of $g(x;\mu,\lambda_2,\lambda_3,\lambda_4)$ is $1+2\lambda_2$. However while the model $g(x;\mu,-0.01,0,0.003)$ is a true density, since the parameter values satisfy the positivity condition

$$g(x;\mu,\lambda) > 0, \quad \forall x \in S,$$

its variance is less than 1. So the local mixture model has parameter values which result in a reduced variance when compared to the unmixed model $\phi(x;\mu,1)$. This runs counter to the well-known result that if mixed and unmixed models have the same mean then the variance should be increased by mixing, see for example Shaked (1980). The condition for being a true local mixture ensures that natural moment based inequalities for mixtures also hold for local mixtures. It also allows the parameters of the true local mixture model to have a natural interpretation in terms of possible mixing distributions.

DEFINITION 4 A local mixture model $g(x;\mu,\lambda)$, of order r, of the regular exponential family $f(x;\mu)$ is called *true* if and only if there exists a distribution $Q_{\mu,\lambda}$ and corresponding exact mixture

$$\int f(x;m)dQ_{\mu,\lambda}(m)$$

such that the first r moments of both distributions agree.

True local mixtures can be characterised in terms of convex hulls in finite dimensional affine spaces in the following way. Let X^r_{Mix} be the convex subset of X_{Mix} where the first r non-central moments exist, then define the r-moment mapping from X^r_{Mix} to an r-dimensional vector space via

$$M_r(f) = \left(E_f(X), E_f(X^2), \ldots, E_f(X^r)\right).$$

LEMMA 1 *The mapping $M_r : X^r_{\text{Mix}} \to \mathbf{R}^r$ is an affine map.*

PROOF: (i) Since all moments exist the result follows from the simple fact that

$$M_r(\rho f + (1-\rho)g) = \rho M_r(f) + (1-\rho)M_r(g),$$

for all $f, g \in X^r_{\text{Mix}}$. ■

THEOREM 1 *Let $f(x;\mu)$ be a regular exponential family, M a compact subset of the mean parameter space and let the order r local mixture of $f(x;\mu)$ be $g(x;\mu,\lambda)$.*

(i) If for each μ the moments $M_r(g(x;\mu,\lambda))$ lie in the convex hull of

$$\{M_r(f(x;\mu))|\mu \in M\} \subset \mathbf{R}^r,$$

then $g(x;\mu,\lambda)$ is a true local mixture model.

(ii) If $g(x;\mu,\lambda)$ has the same r-moment structure as $\int f(x;m)dQ_{\mu,\lambda}(m)$ where $Q_{\mu,\lambda}$ has support in M then the moments $M_r(g(x;\mu,\lambda))$ lie in the convex hull of

$$\{M_r(f(x;\mu))|\mu \in M\} \subset \mathbf{R}^r.$$

PROOF: First note that from the standard properties of exponential families all moments of $f(x;\mu)$ exist. Furthermore from the form of the derivatives of exponential families in Appendix D it is immediate that all moments of the local mixture $g(x;\mu,\lambda)$ also exists. Hence the local mixture model is mapped by M_r into $\mathbf{R}^r$.

(i) This result follows from Carathéodory's theorem, see Barvinok (2002, Theorem 2.3), since a point lies inside the convex hull of a set in a r-dimensional affine space if it can be represented as a convex combination of at most $r+1$ points of the set. Hence for each $i = 1, \ldots, r$ there exists a discrete distribution $Q_{\mu,\lambda}$ such that

$$\int x^i g(x;\mu,\lambda)dx = \int \left\{\int x^i f(x;m)dx\right\} dQ_{\mu,\lambda}(m) = \int x^i \left\{\int f(x;m)dQ_{\mu,\lambda}(m)\right\} dx.$$

Thus the r-moments of $g(x;\mu,\lambda)$ and $\int f(x;m)dQ_{\mu,\lambda}(m)$ agree and so $g(x;\mu,\lambda)$ is a true local mixture model.

(ii) By assumption when $i = 1,\ldots,r$

$$\int x^i g(x;\mu,\lambda)dx = \int x^i \int f(x;m)dQ_{\mu,\lambda}(m)\,dx.$$

Since $Q_{\mu,\lambda}$ has support in M it can be considered as the weak limit of a sequence Q_n of discrete distributions with support in M. It is immediate that for each n the point $M_r(\int f(x;m)dQ_n(m))$ lies in the convex hull. Since M is compact the corresponding convex hull is compact and hence closed, see Barvinok (2002, Corollary 2.4). Thus the limit $M_r(\int f(x;m)dQ_{\mu,\lambda}(m))$ also lies in the convex hull. ■

The proof of Theorem 1 (i) is similar in spirit to the existence of the non-parametric maximum likelihood estimate of the likelihood of a mixture given in Lindsay (1995). In both results there is a representation of a general mixture in terms of a finite mixture with a bounded number of components. This paper looks at such finite mixtures in §4.

The compactness assumption in Theorem 1 might seem restrictive but is necessary for a complete characterisation. The consequences of this constraint are considered in more detail in the rest of the paper.

It might be tempting to define a true local mixture in terms of the convex hull of a family inside the infinite dimensional affine space $(X_{\text{Mix}}, V_{\text{Mix}}, +)$ when this space is given enough topological structure for the Krein-Milman Theorem to hold, see Phelps (1966). It is surprising to note that there exists examples where the local mixture model does not lie in this larger convex hull unless $\lambda = 0$. For example in Anaya-Izquierdo and Marriott (2006) this is shown to be true for the negative exponential distribution. The restriction to the finite dimensional moment space is both more computational and of more practical value.

2.1 Identification

The definition of the local mixture model, Def. 2, has been given explicitly in terms of the mean parametrization of the base exponential family. Using the chain rule it is easy to see what form the local mixture model must take in an arbitrary reparametrization, see Marriott (2002) and Anaya-Izquierdo (2006). However, in §3 a more subtle aspect of reparametrisation comes into play when local mixture models are interpreted in terms of asymptotic expansions.

One major difference between Def. 2 and the treatment in Marriott (2002) is that there is no $\lambda_1 f^{(1)}(x;\mu)$ term in the expansion. This term is dropped for identification reasons, however as is shown in §3 there is no loss in generality in this when interpreting local mixtures in terms of asymptotic expansions of mixture models.

The following definition will be used throughout.

DEFINITION 5 If $f(x;\mu)$ is a natural exponential family in the mean parametrization then $V_f(\mu)$ defined by

$$V_f(\mu) := E[(X-\mu)^2] = \int (x-\mu)^2 f(x;\mu)\nu(dx)$$

is called the *variance function* of the natural exponential family, which together with the measure ν characterises the natural exponential family.

The derivatives of natural exponential families in both the θ and μ parametrizations are easy to calculcate directly and are given in Appendix D in terms of the variance function. This means that explicit forms of the local mixture model are easy to write out. Most of the computational issues are therefore concerned with calculating the hard boundary and checking convex hull conditions.

If the variance function $V_f(\mu)$ is quadratic then the corresponding exponential families have very attactive statistical properties. Examples include the normal, Poisson, gamma, binomial and negative binomial families which form the backbone of parametric statistical modelling. One example of the special properties is given by the following result.

THEOREM 2 *Let $f(x;\mu)$ be a regular natural exponential family with μ the mean parametrization and assume that the variance function $V_f(\mu)$ is a polynomial of degree at most 2. For each μ, the system of polynomials*

$$P_k(x;\mu) := V_f^k(\mu)\,\frac{f^{(k)}(x;\mu)}{f(x;\mu)}$$

for $k = 0, 1, \ldots$ is orthogonal with respect to $f(x;\mu)$. Moreover, $P_k(x;\mu)$ has exact degree k in both x and μ with leading term x^k.

PROOF: See Morris (1983). ■

The following result follows from direct computations similar to those shown in Appendix D and from Theorem 2.

LEMMA 2 *Let $f(x;\mu)$ be a regular natural exponential family and μ its mean parametrization. The local mixture model of order r can be written as*

$$g(x;\mu,\lambda) = f(x;\mu)\left[1 + \sum_{k=2}^{r} \lambda_k \frac{P_k(x;\mu)}{V_f^k(\mu)}\right] \tag{2}$$

where $\{P_k(x;\mu)\}$ is a polynomial of degree k in x.

Furthermore if the variance function $V_f(\mu)$ is a polynomial of degree at most 2, then $\{P_k(x;\mu)\}$ is the orthogonal system of polynomials described in Theorem 2.

From these results identification of the local mixture model follows.

THEOREM 3 *Let $f(x;\mu)$ be a regular natural exponential family and μ the mean parametrization, then the local mixture model $g(x;\mu,\lambda)$ is identified in all its parameters.*

Furthermore if the variance function $V_f(\mu)$ is a polynomial of degree at most 2, then the (μ,λ) parametrization is orthogonal at $\lambda = 0$.

PROOF: By repeatedly differentiating the identity

$$\int xf(x;\mu)dx = \mu$$

with respect to μ it is easy to see that

$$\int xf^{(k)}(x;\mu)dx = 0$$

for $k \geq 2$. Hence it follows immediately that for each μ the mean of $g(x;\mu,\lambda)$ is exactly μ.

It is sufficient to show that the λ-score vectors are linearly independent which follows from Lemma 2 since each polynomial is of a different degree.

The orthogonality result follows immediately from Theorem 2. ■

Note that it is clear from this result why it is advantageous to drop the first derivative in the local mixture expansion and hence the difference between the definition given here and that in Marriott (2002). Furthermore, in §3, it will be shown that in applications there is no loss of generality in this revised definition.

The next theorem gives a further advantage for working in the mean parametrization and a direct interpretation of the λ_2 parameter in many important cases of true local mixtures.

THEOREM 4 *Let $g(x;\mu,\lambda)$ be a true local mixture for the regular exponential family $f(x;\mu)$. (i) If $Q_{(\mu,\lambda)}$ is the mixing distribution defined in Definition 4 then the expected value of $M \sim Q_{(\mu,\lambda)}$ satisfies*

$$E_{Q_{(\mu,\lambda)}}(M) = \mu.$$

(ii) If it is further assumed that $f(x;\mu)$ has a quadratic variance function $V(\mu)$ such that $2 + V^{(2)}(\mu) > 0$ then $\lambda_2 \geq 0$.

PROOF: (i) As shown in Theorem 3 $E_{g(x;\mu,\lambda)}(X) = \mu$ for all λ. Furthermore since $g(x;\mu,\lambda)$ is a true local mixture $E_{g(x;\mu,\lambda)}(X)$ can be calculated using conditional expectations giving

$$E_{g(x;\mu,\lambda)}(X) = E_{Q_{(\mu,\lambda)}}(E_{f(x;M)}(X)|M)) = E_{Q_{(\mu,\lambda)}}(M),$$

giving the required result.

(ii) For any local mixture model

$$\int x^2 g(x;\mu,\lambda)dx = \int x^2 f(x;\mu)dx + \lambda_2 \int x^2 f^{(2)}(x;\mu)dx + \sum_{i=3}^{r} \lambda_k \int x^2 f^{(k)}(x;\mu)dx. \quad (3)$$

Furthermore, when $f(x;\mu)$ has variance function $V(\mu)$ its follows that

$$\int x^2 f(x;\mu)dx = \mu^2 + V(\mu),$$

and since $V(\mu)$ is quadratic it follows that

$$\int x^2 f^{(2)}(x;\mu)dx = 2 + V^{(2)}(\mu) > 0, \int x^2 f^{(k)}(x;\mu)dx = 0,$$

for all $k \geq 3$. Hence substituting into (3) gives

$$\int (x-\mu)^2 g(x;\mu,\lambda)dx = V(\mu) + \lambda_2(2 + V^{(2)}(\mu)).$$

Since $g(x;\mu,\lambda)$ is a true local mixture the effect of mixing must be to increase the variance, hence $\lambda_2 \geq 0$. ■

From Morris (1982, Table 1) the condition on the variance function in Theorem 4 holds for the normal, Poisson, gamma, negative binomial and binomial (for size > 1) families.

2.2 -1 Normal bundles

Amari (1985) showed the statistical importance of normal bundles both for inference in curved exponential families and undertaking inference on an interest parameter in the presence of nuisance parameters. One way of understanding these structures is to consider when $f(x;\mu)$ is a parametric family of density functions embedded inside a larger family $f(x;\mu,\xi)$ such that (i) $f(x;\mu,0)$ equals $f(x;\mu)$ for all μ, (ii) for each fixed μ_0 the family $f(x;\mu_0,\xi)$ is Fisher orthogonal to $f(x;\mu)$ at μ_0 and (iii) the family $f(x;\mu_0,\xi)$ has zero -1-curvature either at $(\mu_0,0)$ or globally. The -1-curvature is defined via the ∇^{-1}-connection, see Amari (1985). A family $f(x;\mu,\xi)$ which satisfies conditions (i), (ii) and (iii) is 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. Normal bundles arise naturally when $f(x;\mu,\xi)$ is a full exponential family and the fibres are defined by the auxiliary family associated with the maximum likelihood estimate of the curved exponential family $f(x;\mu)$.

THEOREM 5 *(i) The local mixture model $g(x;\mu,\lambda)$ for a regular exponential family $f(x;\mu)$ given by Definition 2 is a -1-affine fibre bundle in $(X_{\text{Mix}}, V_{\text{Mix}}, +)$.*
(ii) If $f(x;\mu)$ is an exponential family with quadratic variance function then the -1-fibres are Fisher orthogonal to the family $f(x;\mu)$ at $\lambda = 0$.

PROOF: (i) The fibre at μ_0 is parametrised by the $r-1$ dimensional vector $\lambda = (\lambda_2, \ldots, \lambda_r)$ and it is immediate that when $\lambda = 0$ then $g(x; \mu_0, \lambda) = f(x; \mu_0)$. It is required to show that the fibre is -1-affine which follows from a direct calculation using the linearity of the λ parameter in $g(x; \mu_0, \lambda)$.

(ii) The orthogonality follows from the definition of the Fisher metric in the -1 representation as, for $k \geq 2$

$$\begin{aligned} \int \frac{\frac{\partial g(x;\mu,0)}{\partial \mu}|_{\mu=\mu_0} \frac{\partial g(x;\mu_0,\lambda)}{\partial \lambda_k}|_{\lambda=0}}{f(x;\mu_0)} dx &= \int \frac{\frac{\partial f(x;\mu)}{\partial \mu}|_{\mu=\mu_0} f^{(k)}(x;\mu_0)}{f(x;\mu_0)} dx \\ &= 0, \end{aligned}$$

from Theorem 2. ∎

When restricted to the case where the local mixture model is a density the family is no-longer an affine space but rather a convex subset of the affine space. Despite the slight abuse of notation it is this convex set that will be referred to as the fibre rather than the full subspace.

The excellent statistical properties of -1-fibre bundles discussed in Amari (1985) also extend to the local mixture models discussed here. The following theorem shows, for example, that the log-concavity of the likelihood function, one of the most important properties of natural exponential families, is paralleled in the fibres of local mixture models.

THEOREM 6 *The log-likelihood function for λ for a fixed, known μ_0, based on the density function $g(x; \mu_0, \lambda)$ and the random sample $x_1, \ldots, x_n$ is concave.*

PROOF: Consider first a 1-dimensional affine subspace of $(X_{\text{Mix}}, V_{\text{Mix}}, +)$ which can be written as $f(x) + \lambda v(x)$, where $f(x) \in X_{\text{Mix}}, v(x) \in V_{\text{Mix}}$. The corresponding log-likelihood, defined on the convex subset of densities, is

$$\ell(\lambda) = \sum_{i=1}^{n} \log \{f(x_i) + \lambda v(x_i)\}$$

and so

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\sum_{i=1}^{n} \frac{v(x_i)^2}{(f(x_i) + \lambda v(x_i))^2} < 0,$$

hence is concave.

In general consider any two points f_1, f_2 in the fibre at μ_0 which are density functions. The convex combination of f_1 and f_2 is a one dimensional affine space in the fibre hence the corresponding log-likelihood is concave. It follows that

$$\ell(\rho f_1 + (1-\rho) f_2) \geq \rho \ell(f_1) + (1-\rho)\ell(f_2),$$

for $0 \leq \rho \leq 1$. The log-concavity for the fibres of the local mixture model inside the hard boundary therefore follows immediately. ■

Thus there is a clear parallel between the shape of the log-likelihood on a fibre and that on an exponential family. One difference between these two cases is that in the fibre there ca
illu

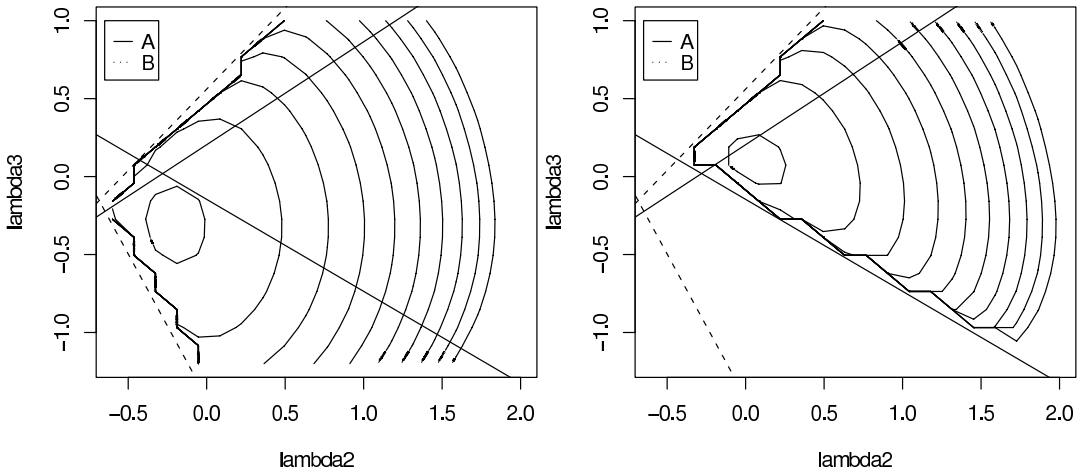

Figure 1: The log-likelihood function on a fibre of the local mixture of a binomial model. The hard boundary is shown by the solid lines (A) while the singularity in the log-likelihood occurs at the dotted lines (B). Just one point in the sample has changed between the two plots

Example 1 (Revisited) The local mixture for a binomial model has a hard boundary in a fibre which is defined, as is normal with local mixture models, as the intersection of half-spaces in the parameter space. For example the fibre of the local mixture model of order 3, $g(x; \mu, \lambda_2, \lambda_3)$, has a parameter space which is a subset of $\mathbf{R}^2$, shown in Fig. 1. The hard boundary is defined by the intersection of half-spaces of the form

$$\{(\lambda_2, \lambda_3) \,|\, 1 + \lambda_2 q_2(x_i, \mu_0) + \lambda_3 q_3(x_i, \mu_0) > 0\}\,,$$

where $x_i \in \{0, \ldots, n\}$ and q_2 and q_3 are polynomials in x.

In Fig. 1 the log-likelihood for this fibre is shown as a contour plot in a case where $n = 10$. In both panels the hard boundary simplifies as

$$\{(\lambda_2, \lambda_3) \,|\, 1 + \lambda_2 q_2(10, \mu_0) + \lambda_3 q_3(10, \mu_0) > 0\} \bigcap \{(\lambda_2, \lambda_3) \,|\, 1 + \lambda_2 q_2(0, \mu_0) + \lambda_3 q_3(0, \mu_0) > 0\}$$

and the hard boundary is shown as solid lines.

The log-concavity of the likelihood is clear in both plots, and the singularities in the log-likelihood can also be seen. In the left hand panel the sample size is 50 and singularities can be seen along the dotted lines defined by

$$1 + \lambda_2 q_2(x_m, \mu_0) + \lambda_3 q_3(x_m, \mu_0) = 0$$

where x_m is the maximum (minimum) observed value in the dataset which happens to be 8 (1).

In the right hand panel the log-likelihood for the fibre is shown with the same data except that one of the observations which was 8 has been changed to 10. The singularity has jumped and now lies on the hard boundary. Thus it can be seen that, unlike exponential families, the log-likelihood in local mixtures can be very sensitive to a single data-point, and is especially sensitive to large or small observations.

3 Applications

The previous results give an abstract framework in the space $(X_{\text{Mix}}, V_{\text{Mix}}, +)$ and the subset of densities. In this section the theory is applied to different modelling situations.

3.1 Asymptotic expansions

In Marriott (2002) local mixture models were partially motivated by the idea of a modelling situation where a mixture model is used, but the mixing is only responsible for a relatively small amount of the variation in the model.

In more detail consider the Q-mixture density defined by

$$g(x; Q) = \int_{\Theta} f(x; \theta) dQ(\theta), \tag{4}$$

where Q is a distribution over the parameter space Θ. If the mixing distribution is unknown (4) appears to define an infinite dimensional family over which inference would seem to be difficult. In fact some inferential results can be found, for example the existence of a maximum likelihood estimate for Q, see Lindsay (1995). Moreover for many practical applications the unmixed model $f(x; \theta)$ is chosen as a regular family which explains most of the variation and the mixing only adds a small component to improve the fit of the baseline model. Examples of such applications include frailty modelling in lifetime data analysis, random effects modelling in generalised linear models and measurement error modelling in regression. Local mixture models use modelling assumptions on the 'smallness' of the mixture to approximate the class of models given by (4) by a finite dimensional parametric family where the parameters decompose into the interest parameter θ and a small number of well-defined and interpretable nuisance parameters.

When the mixing distribution is continuous one sensible and useful interpretation of smallness is that the mixing distribution is close to a degenerate delta function, i.e. it is close to the case of no mixing. In such an example a Laplace expansion gives an asymptotic tool which enables us to construct the local mixing family, see for example Wong (2001). The basic results on the existence of local mixture models can be found in Marriott (2002). This section gives a more advanced treatment which uses an asymptotically equivalent definition but one which (i) proposes a new parametrization which gives identified and interpretable nuisance parameters, (ii) deals with the case when the parameter space over which there is mixing has boundaries and (iii) has a log-likelihood structure in the new parametrization which has excellent convexity properties enabling inference to be done in a very clean way.

The relationship between the parametrization and the topology of the space over which there is mixing needs careful consideration. For example the space for the mean parameter in the normal distribution case is $\mathbf{R}$, while for the mean parameter for a Poisson model it is $\mathbf{R}^+$. When considering the 'smallness' of the mixing the existence of boundaries needs to be considered, see §3.2. In light of the following theorem the results of Marriott (2002) can be generalised to include the class of *proper dispersion model* which are defined in Jorgensen (1997), see also Appendix B for details.

One important, but subtle, feature of the following result is that the proper dispersion mixing model $dQ(\theta, \vartheta, \epsilon)$ and the Laplace expansion are most naturally centred at ϑ, which is sometimes a mode, while, as is shown in Theorem 4, the previous theory has been centred at the mean of the mixing distribution.

THEOREM 7 *Let $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ be a regular family and also let*

$$\mathcal{Q} = \{dQ(\theta;\vartheta,\epsilon) : \vartheta \in \Theta, \epsilon > 0\}$$

be a family of proper dispersion models, defined on Θ. The $\mathcal{Q}$-mixture of $\mathcal{F}$ has the following asymptotic expansion:

$$\begin{aligned} g(x;Q(\theta,\vartheta,\epsilon)) &= \frac{\displaystyle\int_\Theta f(x;\theta)V^{-1/2}(\theta)\exp\left(-\frac{d(\theta,\vartheta)}{2\epsilon}\right)d\theta}{\displaystyle\int_\Theta V^{-1/2}(\theta)\exp\left(-\frac{d(\theta,\vartheta)}{2\epsilon}\right)d\theta} \\ &\sim f(x;\vartheta) + \sum_{i=1}^{2r} A_i(\vartheta,\epsilon)f^{(i)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^{r+1}) \end{aligned} \tag{5}$$

as $\epsilon \to 0$, for fixed $\vartheta \in \Theta$ and x and for functions A_i such that

$$\begin{aligned} A_i(\vartheta,\epsilon) &= O_\vartheta(\epsilon^{u(i)}) \\ \frac{E_Q[(\theta-\vartheta)^i]}{i!} &\sim A_i(\vartheta,\epsilon) + O_\vartheta(\epsilon^{r+1}), \qquad i = 1,2,\ldots,2r, \end{aligned}$$

where $u(i) = \lfloor (i+1)/2 \rfloor$.

The following alternative expansion is also valid

$$g(x; Q(\theta, \vartheta, \epsilon)) \sim f(x; M_1(\vartheta, \epsilon)) + \sum_{i=2}^{2r} M_i(\vartheta, \epsilon) f^{(i)}(x; M_1(\vartheta, \epsilon)) + O_{x,\vartheta}(\epsilon^{r+1}), \qquad (6)$$

for functions M_i *such that,*

$$\begin{aligned} E_Q[\theta] &\sim M_1(\vartheta, \epsilon) = \vartheta + A_1(\vartheta, \epsilon) + O_\vartheta(\epsilon^3) \\ M_i(\vartheta, \epsilon) &= O_\vartheta(\epsilon^{u(i)}) \\ \frac{E_Q[(\theta - E_Q[\theta])^i]}{i!} &\sim M_i(\vartheta, \epsilon) + O_\vartheta(\epsilon^{r+1}), \qquad i = 2, \ldots, 2r. \end{aligned}$$

If the density $f(x; \theta)$ *and all its derivatives are bounded then the statement will be uniform in* x.

PROOF: See Appendix B ∎

The first form of the expansion, (5) is be referred to as the *ϑ-centred* expansion and the second expansion, (6), as the *mean-centred* expansion. Note that actually this latter expansion is not centred at the exact mean but at the function $M_1(\vartheta, \epsilon)$ which is very close to the exact mean when ϵ is small. The function $M_1(\vartheta, \epsilon)$ is referred to as the *pseudo-mean* of the proper dispersion model, and the functions $M_i(\vartheta, \epsilon)$ as the central *pseudo-moments* as they behave like the exact moments to order ϵ^{r+1}.

It follows immediately that expression (6) is of the form given in Definition 2, and is therefore a local mixture, after truncating the remainder term. The form (5) can be thought of either as a direct Laplace expansion, as it was in Marriott (2002), or as a simple reparametrisation of (6). This fact shows the generality of Definition 2.

It also follows from Lemma 5 in the Appendix that if a local mixture model given by Definition 2 is of order $r = 4$ then the remainder term is $O_\vartheta(\epsilon^3)$. However the simpler looking $r = 3$ expansion is not well defined asymptotically since both $E_Q[(\theta - E_Q[\theta])^3]$ and $E_Q[(\theta - E_Q[\theta])^4]$ are of the same asymptotic order.

The basic idea of a local mixture model is to truncate one of the expansions given above and treat the coefficients A_i or M_i as unknown parameters to be estimated. Theorem 7 enables the accuracy of a local mixture approximation to be assessed, and furthermore it gives clear interpretations of the meaning of the new parameters in terms of the (pseudo-) moments of the mixing distribution.

3.2 Boundary cases

In order to be able to treat the coefficients in the above expansions as parameters the dependence of $A_i(\vartheta, \epsilon)$ and $M_i(\vartheta, \epsilon)$ on the point of expansion must be clarified, in particular

when the parameter space has boundaries. This is done in the following corollaries of Theorem 7. Here special cases of the exponential dispersion models, called local-scale and local dispersion exponential dispersion models are used, see Appendix B.

COROLLARY 1 *Let $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta = \mathbf{R}^+\}$ be a regular family and consider scale dispersion mixtures $\mathcal{Q}_{sd} = \{Q(\theta;\vartheta,\epsilon) : \vartheta \in \Theta, \epsilon > 0\}$. The $\mathcal{Q}_{sd}$-mixture of $\mathcal{F}$ has the following expansion:*

$$g(x;Q(\theta,\vartheta,\epsilon)) \quad \sim \quad f(x;\vartheta) + \sum_{i=1}^{2r} \vartheta^i A_i^*(\epsilon) f^{(i)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^{r+1}),$$

as $\epsilon \to 0$ for fixed $\vartheta \in \Theta$ and x and for functions A_i^* such that

$$\begin{aligned} A_i^*(\epsilon) &= O(\epsilon^{u(i)}) \\ \frac{E_Q[(\theta-\vartheta)^i]}{i!} &\sim \vartheta^i (A_i^*(\epsilon) + O(\epsilon^{r+1})), \qquad i = 1,2,\ldots,2r, \end{aligned}$$

where $u(i) = \lfloor (i+1)/2 \rfloor$. The following alternative expansion is also valid

$$g(x;Q(\theta,\vartheta,\epsilon)) \sim f(x;M_1^*(\vartheta,\epsilon)) + \sum_{i=2}^{2r} \vartheta^i M_i^*(\epsilon) f^{(i)}(x;M_1^*(\vartheta,\epsilon)) + O_{x,\vartheta}(\epsilon^{r+1}),$$

for functions M_i^* such that

$$\begin{aligned} E_Q[\theta] &\sim M_1^*(\vartheta,\epsilon) = \vartheta[1 + A_1^*(\epsilon) + O(\epsilon^3)], \\ M_i^*(\epsilon) &= O(\epsilon^{u(i)}), \\ \frac{E_Q[(\theta - E_Q[\theta])^i]}{i!} &\sim \vartheta^i [M_i^*(\epsilon) + O(\epsilon^{r+1})], \qquad i = 2,\ldots,2r. \end{aligned}$$

PROOF: See Appendix B. ∎

COROLLARY 2 *Let $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta = \mathbf{R}\}$ be a regular family and also let $\mathcal{Q}_{ld} = \{Q(\theta;\vartheta,\epsilon) : \vartheta \in \Theta, \epsilon > 0\}$ be a location dispersion model defined on Θ. The $\mathcal{Q}_{ld}$-mixture of $\mathcal{F}$ has the following expansion:*

$$g(x;Q(\theta,\vartheta,\epsilon)) \quad \sim \quad f(x;\vartheta) + \sum_{i=1}^{2r} A_i^*(\epsilon) f^{(i)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^{r+1}),$$

as $\epsilon \to 0$ for fixed $\vartheta \in \Theta$ and x and for functions A_i^* such that

$$\begin{aligned} A_i^*(\epsilon) &= O(\epsilon^{u(i)}) \\ \frac{E_Q[(\theta-\vartheta)^i]}{i!} &\sim A_i^*(\epsilon) + O(\epsilon^{r+1}), \qquad i = 1,2,\ldots,2r, \end{aligned}$$

where $u(i) = \lfloor (i+1)/2 \rfloor$.

The following alternative expansion is also valid

$$g(x; Q(\theta, \vartheta, \epsilon)) \sim f(x; M_1^*(\vartheta, \epsilon)) + \sum_{i=2}^{2r} M_i^*(\epsilon) f^{(i)}(x; M_1^*(\vartheta, \epsilon)) + O_{x,\vartheta}(\epsilon^{r+1}),$$

for functions M_i^* *such that*

$$\begin{aligned} E_Q[\theta] &\sim M_1^*(\vartheta, \epsilon) = \vartheta + A_1^*(\epsilon) + O(\epsilon^3), \\ M_i^*(\epsilon) &= O(\epsilon^{u(i)}), \\ \frac{E_Q[(\theta - E_Q[\theta])^i]}{i!} &\sim M_i^*(\epsilon) + O(\epsilon^{r+1}), \qquad i = 2, \ldots, 2r. \end{aligned}$$

PROOF: See Appendix B. ∎

From these results the following definitions are natural.

DEFINITION 6 Let $\mathcal{F} = \{f(x;\mu) : \mu \in M = \mathbf{R}^+\}$ be a regular exponential family. The *local scale mixture model* of order r of the family $\mathcal{F}$ is defined as the parametric family

$$\mathcal{G}_{\mathcal{F}}^{scale} = \left\{ g(x;\mu,\gamma) = f(x;\mu) + \sum_{k=2}^{r} \frac{\mu^k \gamma_k}{k!} f^{(k)}(x;\mu) \, : \, \mu \in M \right\}.$$

DEFINITION 7 Let $\mathcal{F} = \{f(x;\mu) : \mu \in M = \mathbf{R}\}$ be a regular exponential family. The *local location mixture model* of order r of the family $\mathcal{F}$ is defined as the parametric family

$$\mathcal{G}_{\mathcal{F}}^{loc} = \left\{ g(x;\mu,\gamma) = f(x;\mu) + \sum_{k=2}^{r} \frac{\gamma_k}{k!} f^{(k)}(x;\mu) \, : \, \mu \in M \right\}.$$

The form for these models in any other parametrization is given by the chain rule, see Anaya-Izquierdo (2006).

Note these definitions are restricted to the cases where the parameter space for $\mathcal{F}$ is $\mathbf{R}^+$ (scale) and $\mathbf{R}$ (location). Here γ is considered constant as function of μ but for each $\mu \in M$ the parameter space for γ depends on μ. This is just a reflection of the fact that, for us, a mixing distribution is going to be small not only when ϵ is small but also relative to the family $\mathcal{F}$.

To illustrate this consider the case where $\mathcal{F}$ is a Poisson family. If the mixing distribution is a scale dispersion model with position m and dispersion parameter ϵ sufficiently small such that the usual normal approximation holds, Jorgensen (1997), then the distribution of μ is close to a normal with mean m and variance m^2. For the mixing distribution to be local it is reasonable to require that $\epsilon\, m^2 < m$, otherwise the mixing distribution

will have more variability around m than X. In the special case where $\mathcal{F}$ is a scale family with scale parameter $\mu \in \mathbf{R}^+$ then it is straightforward to check that μ is still a scale parameter for the family $\mathcal{G}_{\mathcal{F}}^{scale}$ and therefore the parameter space for γ does not depend on θ. Similarly, when $\mathcal{F}$ is a location family with location parameter $\mu \in \mathbf{R}$ then μ is still a location parameter for the family $\mathcal{G}_{\mathcal{F}}^{loc}$ and again the parameter space for γ does not depend on θ. So, in these special cases the smallness of the mixing distribution depends only on the dispersion parameter ϵ.

The interpretation of the local scale and local location models is clear. When we have a regular family $\mathcal{F}$ is parametrized in such a way that $\Theta = \mathbf{R}^+$ ($\Theta = \mathbf{R}$) and the mixing distribution on that parametrization is a scale (location) dispersion model with small ϵ, then the above models mimic the behaviour of the mean-centred expansions in Corollaries 1 and 2 respectively. Moreover, the parameters γ_i play the role of the pseudo normalized moments in the scale case and the role of the pseudo moments in the location case. This justifies the introduction of the factorials in the previous definitions.

4 Discrete mixing and marginal inference

To see the relationship between discrete mixture models and local mixtures consider a family of discrete finite distributions which shrink around their common mean μ

$$Q(\theta; \mu, \epsilon) = \sum_{i=1}^{n} \rho_i I\left\{\theta \leq \theta_i(\epsilon)\right\}$$

where $|\mu - \theta_i(\epsilon)| = O(\epsilon^{1/2})$, $\sum_{i=1}^n \rho_i \theta_i(\epsilon) = \mu$, $\sum_{i=1}^n \rho_i = 1$, $\rho_i \geq 0$, and I is the indicator function. The mixture over such a finite distribution has the form

$$f(x; \mu, Q(\theta; \mu, \epsilon)) = \sum_{i=1}^{n} \rho_i f(x; \theta_i(\epsilon)). \tag{7}$$

This has the asymptotic expansion

$$f(x; \mu) + \sum_{j=2}^{r} M_j(Q) f^{(j)}(x; \mu) + R(x, \mu, Q), \tag{8}$$

where

$$M_j(Q) = \sum_{i=1}^{n} \rho_i \frac{(\theta_i(\epsilon) - \mu)^j}{j!} = O(\epsilon^{j/2}),$$

and $R(x, \mu, Q) = O(\epsilon^{(j+1)/2})$. This is a natural generalisation of the Expansion (6) in Theorem 7.

DEFINITION 8 Following Expansion (8), define the function Φ by

$$\Phi(Q) = (M_2(Q), \ldots, M_r(Q)) .$$

Thus it follows that

$$\int f(x;m)dQ(m;\mu,\epsilon) - g(x;\mu,\Phi(Q)) = R(x,\mu,Q). \tag{9}$$

A comparison of expansion (8), and those in Theorem 7 reveals interesting differences. In expansion (6) the fibre is centred at the pseudo-mean M_1 and the order of the terms is $u(i) = \lfloor (i+1)/2 \rfloor$, while in (8) the asymptotic order is the 'more natural' $i/2$ and the expansion is around the exact mean. One reason for these differences is the requirements for a valid asymptotic expansion in Theorem 7 imposed by Watson's Lemma, Wong (2001), that the tail of the (continuous) mixing distribution must have exponentially decreasing tails. There are, of course no such restrictions for discrete mixtures as long as the number of components is known or bounded.

Related to this difference is the idea of the smallness of the mixing. In Theorem 7 the idea of the mixture being close to the unmixed model was captured by the small variance of the mixing distribution. There is however a quite different notion to that expressed above of what it means for mixing to be small in the discrete case. The simplest example of this is given by a two component finite mixture

$$\rho f(x;\theta_1) + (1-\rho) f(x;\theta_0) = f(x;\theta_0) + \rho \left\{ f(x;\theta_1) - f(x;\theta_0) \right\}, \tag{10}$$

for any regular family $f(x;\theta)$. The form on the right hand side shows the natural way that this mixture lies inside the affine space $(\mathcal{X}_{\text{Mix}}, \mathcal{V}_{\text{Mix}}, +)$, since $\int f(x;\theta_0)\,\nu(dx) = 1$ and $\int \{f(x;\theta_1) - f(x;\theta_0)\}\,\nu(dx) = 0$. The new interpretation of when (10) is 'close' to the model $f(x;\theta_0)$ is when ρ is small, rather than θ_1 being close to θ_0.

The simple observations that there are mixtures which are arbitrary close to an unmixed model, $f(x;\theta_0)$, which can have components, $f(x;\theta_1)$, which are far from being local shows that the interpretation of local mixture models in terms of Laplace expansions is not exhaustive.

4.1 Marginal inference on μ

DEFINITION 9 For a regular exponential family $f(x;\mu)$ let $\{M(\mu)\}$ be a family of compact subsets of the mean parameter space such that $\mu \in M(\mu)$. Define $\mathcal{Q}_{M(\mu)}$ to be the set of distributions which have support on $M(\mu)$ and have expected value μ. Furthermore let $\mathcal{Q}^{dis}_{M(\mu)}$ be the subset of $\mathcal{Q}_{M(\mu)}$ defined by the finite mixtures. Since each $M(\mu)$ is compact its length can be defined by $|M(\mu)| = \max M(\mu) - \min M(\mu)$.

Suppose that the local mixture model $g(x;\mu,\lambda)$ is to be used for marginal inference on μ. Interpreting this in a Bayesian sense means it is of interest to know if marginalising over

some subset of the parameter space for λ is equivalence to marginalising over a set of mixing distributions. The marginal posterior defined over some class of mixing distributions $\mathcal{Q}(\mu)$, each with mean μ, has the form

$$\int_{\mathcal{Q}(\mu)} p(\mu, Q|x_1, \ldots, x_n) dP(Q), \tag{11}$$

where $p(\mu, Q|x_1, \ldots, x_n)$ is the joint posterior over μ and $Q \in \mathcal{Q}(\mu)$ and $dP(Q)$ is a measure over $\mathcal{Q}(\mu)$. On the other hand the marginal distribution over the local mixture has the form

$$\int_{\lambda \in \Lambda(\mu)} p(\mu, \lambda|x_1, \ldots, x_n) dP(\lambda), \tag{12}$$

where $p(\mu, \lambda|x_1, \ldots, x_n)$ is the posterior over μ, λ in the local mixture model and $\Lambda(\mu)$ is the set of parameters corresponding to distributions in $\mathcal{Q}(\mu)$.

In (11) the posterior is proportional to

$$\prod_{i=1}^{n} \int f(x_i; m) dQ(m) \times \pi(\mu, Q) \tag{13}$$

for some prior $\pi(\mu, Q)$, while in (12) the corresponding posterior is

$$\prod_{i=1}^{n} g(x_i; \mu, \lambda) \times \pi(\mu, \lambda) \tag{14}$$

again for a prior $\pi(\mu, \lambda)$.

LEMMA 3 *If $Q \in \mathcal{Q}^{dis}_{M(\mu)}$ and $|M(\mu)| = O(\epsilon^{1/2})$ then from (9) it follows that*

$$\int_{M(\mu)} f(x; m) dQ(m) - g(x; \mu, \Phi(Q)) = R(x, \mu, Q) = O(\epsilon^{r+1/2}).$$

in particular there exists a bound $\delta(x, \mu)$ on $R(x, \mu, Q)$ which is uniform for all mixing distributions in $\mathcal{Q}^{dis}_{M(\mu)}$.

PROOF: The remainder term $R(x, \mu, Q)$ can be expressed, using Taylor's theorem, as $M_{r+1} f^{(r+1)}(x, \mu^*)$ for some $\mu^* \in M(\mu)$. Since $M(\mu)$ is compact there is a uniform bound for both $f^{(r+1)}$ and the M_{r+1} term for all $Q \in \mathcal{Q}^{dis}_{M(\mu)}$. Thus the result follows immediately. ■

The following result shows that marginal inference for μ over a local mixture model is asymptotically equivalent to that over all distributions with compact support as long as the parameter space is bounded away from possible singularities in the log-likelihood function.

THEOREM 8 *Let $f(x;\mu)$ be a regular exponential family and $g(x;\mu,\lambda)$ the corresponding local mixture model of order r. Also assume that the compact covering $\{M(\mu)\}$ satisfies $|M(\mu)| = O(\epsilon^{1/2})$. For each μ let $\Lambda(M(\mu))$ be defined by*

$$\Lambda(M(\mu)) := \left\{\Phi(Q)|Q \in \mathcal{Q}^{dis}_{M(\mu)}\right\}.$$

Suppose further that for all $Q \in \mathcal{Q}^{dis}_{M(\mu)}$

$$g(x_i;\mu,\Phi(Q)) \geq C > 0$$

for every observed data point x_i.

Under these assumptions there exists a prior $\pi(\mu,\lambda)$, depending $\pi(\mu,Q)$, such that

$$\left|\int_{\mathcal{Q}^{dis}_{M(\mu)}} \left\{\prod_{i=1}^{n}\int_{M(\mu)} f(x_i;m)dQ(m)\times\pi(\mu,Q)\right\}dP(Q) - \int_{\Lambda(M(\mu))}\left\{\prod_{i=1}^{n} g(x_i;\mu,\lambda)\times\pi(\mu,\lambda)\right\}dP(\lambda)\right| \leq R_2(\epsilon),$$

where $R_2(\epsilon) = O(\epsilon^{(r+1)/2})$.

PROOF: By direct computation it follows that the marginal posterior for μ over $\mathcal{Q}^{dis}_{M(\mu)}$

$$\begin{aligned}
\int_{\mathcal{Q}^{dis}_{M(\mu)}} \left\{\prod_{i=1}^{n}\int_{M(\mu)} f(x_i;m)dQ(m)\times\pi(\mu,Q)\right\}dP(Q) &= \int_{\mathcal{Q}^{dis}_{M(\mu)}} \left\{\prod_{i=1}^{n}\left\{g(x_i;\mu,\Phi(Q)) - R(x_i,\mu,Q)\right\}\pi(\mu,Q)\right\}dP(Q)\\
&= \int_{\mathcal{Q}^{dis}_{M(\mu)}} \left\{\prod_{i=1}^{n} g(x_i;\mu,\Phi(Q))\left\{1-\frac{R(x_i,\mu,Q)}{g(x_i;\mu,\lambda)}\right\}\pi(\mu,Q)\right\}dP(Q)\\
&= \int_{\mathcal{Q}^{dis}_{M(\mu)}} \left\{\prod_{i=1}^{n} g(x_i;\mu,\Phi(Q))\prod_{i=1}^{n}\left\{1-\frac{R(x_i,\mu,Q)}{g(x_i;\mu,\lambda)}\right\}\pi(\mu,Q)\right\}dP(Q).
\end{aligned}$$

The assumptions of the theorem and the results of Lemma 3 give that

$$\left|\prod_{i=1}^{n}\left\{1-\frac{R(x,\mu,Q)}{g(x_i;\mu,\lambda)}\right\}\right| \leq 1 + R_3(\epsilon)$$

where the bound $R_3(\epsilon)$ is uniform in Q and of order $\epsilon^{(r+1)/2}$.

Thus it follows that there exists a prior $\pi(\mu,\lambda) = \int_{\{\Phi(Q)=\lambda\}}\pi(\mu,Q)dP(Q)$ for which

$$\left|\int_{\mathcal{Q}^{dis}_{M(\mu)}} \left\{\prod_{i=1}^{n}\int_{M(\mu)} f(x_i;m)dQ(m)\times\pi(\mu,Q)\right\}dP(Q) - \int_{\Lambda(M(\mu))}\left\{\prod_{i=1}^{n} g(x_i;\mu,\lambda)\times\pi(\mu,\lambda)\right\}dP(\lambda)\right| \leq R_2(\epsilon).$$

■

Theorem 8 has the following interpretation. If $\Lambda(M(\mu))$, the set of λ-values of interest, is bounded away from any of the possible singularities in the log-likelihood then, for a sufficiently small compact cover $\{M(\mu)\}$, there is little loss in undertaking marginal inference on μ with the local mixture model, as compared to the set of all finite mixing distributions, $\mathcal{Q}^{dis}_{M(\mu)}$. By weak convergence this result extends to the space $\mathcal{Q}_{M(\mu)}$, i.e. all mixing distributions with support in the compact cover.

Since many important mixing distributions do not have compact support, this result still might seem somewhat restrictive. Note however that as far as the contribution to the posterior is concerned, since

$$\int f(x_i;m)dQ(m) = \int_{m\in M(\mu)} f(x_i;m)dQ(m) + \int_{m\notin M(\mu)} f(x_i;m)dQ(m),$$

there is be only a small loss in extending to distributions with small 'tail probability' that $m \notin M(\mu)$.

4.2 Geometry of the fibre

In order to illustrate the previous results consider the fibre defined for a local mixture family. In Theorem 8 the mapping from a set of mixing distributions defines a subset of the fibre, $\Lambda(M(\mu))$. It is immediate that this is a convex subset of the fibre and it is illuminating to consider its convex geometry.

By interpreting local mixtures as asymptotic expansions, either in the Laplace form of (6) or the discrete form (8), then the λ parameter of a local mixture model has a direct interpretation, at least to high asymptotic order, as a moment of the mixing distribution. By looking at attainable values of these moments a characterisation of which points in the fibre correspond to true local mixture models can be given.

EXAMPLE 1 (Revisited) The binomial case has the simplifying advantage that the parameter space for $\mu = n\pi$ is compact, but it also has boundaries as discussed in §3.2. The following result of the binomial family is a corollary of Theorem 1.

COROLLARY 3 *Let $g(x;\mu,\lambda)$ be a local mixture model of order $r = 4$ for $f(x;\mu)$ the binomial family $Bin(n,\pi)$. Following Theorem 1 it follows that $g(x;\mu,\lambda)$ is a true local mixture if and only if $M_4\left(g(x;\mu,\lambda)\right)$ lies in the convex hull of $\{M_4(f(x;\mu))|\mu \in M\} \subset \mathbf{R}^4$.*

PROOF: The result follows immediately by setting the compact set $M = [0,n]$ in Theorem 1 (ii). ∎

Following Teuscher and Guiard (1995) any distribution with mean μ can be approximated as a mixture of discrete distributions with two support points of the form

$$Q(m) = \rho I(m \le \mu_1) + (1-\rho)I(m \le \mu_2) \tag{15}$$

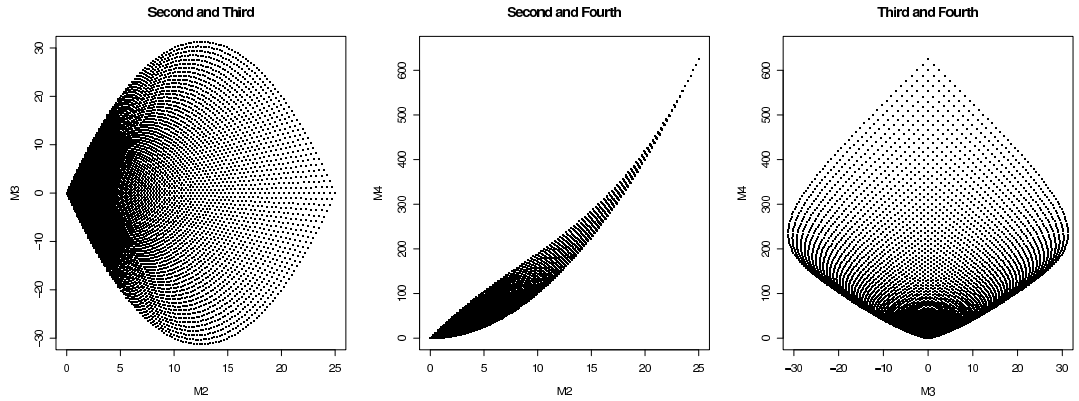

Figure 2: The convex hull in the fibre for a binomial model

where $\rho\mu_1 + (1-\rho)\mu_2 = \mu$ and $0 \leq \rho \leq 1$. Such mixing distributions are the extremal points of the convex hull and give a convenient way to characterise it.

An example of such a set of points is illustrated in Fig. 2. In this plot, for clarity, the central moments

$$\left(E((X-\mu)^2), E((X-\mu)^3), E((X-\mu)^4)\right)$$

are plotted for mixtures of $Bin(n,\pi)$ over distributions of the form (15). These mixtures lie in the fibre defined by $E(X) = \mu$. For fixed μ the central moments are a linear transformation of the non-central ones, thus extremal points are preserved. By Corollary 3 $g(x;\mu,\lambda)$ will be a true local mixture model if and only if its central moments lie in this convex hull.

For inferential purposes it is required to have a test which can be implemented computationally which checks if a point lies in a convex hull. Discretizing the compact set M and the interval $[0,1]$ gives triples of points $(\mu_{1i}, \mu_{2i}, \rho_i)$ which define mixing distributions of the form (15). The corresponding moments of the binomial, after mixing by such distributions, defines a set of extremal points in $\mathbf{R}^3$. To check that a point lies in the convex hull of a set of points in $\mathbf{R}^3$ is a simple exercise in linear programming.

The previous methodology also holds for any choice of compact cover as shown by the following corollary.

COROLLARY 4 *Let $f(x;\mu)$ be a binomial model with local mixture model, of order 4, $g(x;\mu,\lambda)$. If $\lambda = \Phi(Q)$ for $Q \in \mathcal{Q}^{dis}_{M(\mu)}$ then*

$$M_4\left(g(x;\mu,\lambda)\right) = M_4\left(\int_{M(\mu)} f(x;m)dQ(m)\right).$$

PROOF: From Theorem 2 the polynomials $x-\mu, p_2(x,\mu), p_3(x,\mu), p_4(x,\mu)$ defined in Ap-

pendix C from the derivatives of $f(x;\mu)$ are orthogonal and hence span the space of polynomials of degree less than or equal to 4.

The remainder term $R(x,\mu,Q)$ defined in (9) can be expressed as a linear combination of terms $f^{(k)}(x;\mu)$ for $k \geq 4$. By orthogonality of Theorem 2 these terms satisfy

$$\int x^i f^{(k)}(x;\mu)dx = 0$$

for $i = 1,2,3,4$ and $k \geq 4$. Thus it follows that the term $R(x,\mu,Q)$ does not effect the first four moments, and hence from (9) it is immediate that

$$M_4\left(g(x;\mu,\lambda)\right) = M_4\left(\int_{M(\mu)} f(x;m)dQ(m)\right).$$

■

From Corollary 4 it follows that characterising the set $\Lambda(M(\mu))$ is essentially the same type of problem as characterising true local mixtures. In particular consider $\{M(\mu)\}$ of length $O(\epsilon^{1/2})$ as used in Theorem 8, then the set $\Lambda(M(\mu))$ is a convex hull which can be characterised by extremal points. These points are defined by the moments of a binomial random variable after mixing over distributions of the form (15) but now with the added constraint that $\mu_1, \mu_2 \in M(\mu)$. If μ is far from the boundary of the parameter space and $M(\mu)$ is symmetric around μ then the corresponding hull looks very similar to Fig. 2. If the point μ is near one of the boundaries then a more asymmetric set of points is generated. In such a case the interpretation of the asymptotic expansion should be through the models in §3.2.

As another example of how a restriction on the space of mixing distributions affects the allowable subset of the -1-fibre consider Fig. 3. Since the space of proper dispersion models is not convex then the subset of allowable moments is also not convex. It is then not possible to characterise this set using extremal points. Rather the subset is explored numerically by sampling where allowed central moments are plotted in Fig. 3. In this case, since the parameter space is compact, points are included if the corresponding mixing distribution has a log-concave density. It can be seen that the subset shown in Fig. 3 is not convex and has an interesting geometry for very small variances. The space has a cusp and is for very small exponential mixing distributions almost two dimensional. The consequences of this unusual geometry on inference are currently being explored by the authors.

By Theorem 8 is it clear that local mixtures can be used for marginal inference as long as a small ϵ is assumed and the singularities in the log-likelihood, shown in Fig. 1, are bounded away from $\Lambda(\mu)$. It can be shown for the binomial case that the hard boundary is always bounded away from the convex hull shown in Fig. 2. For computations with true local mixture models the Markov chain Monte Carlo algorithm can be used, however this

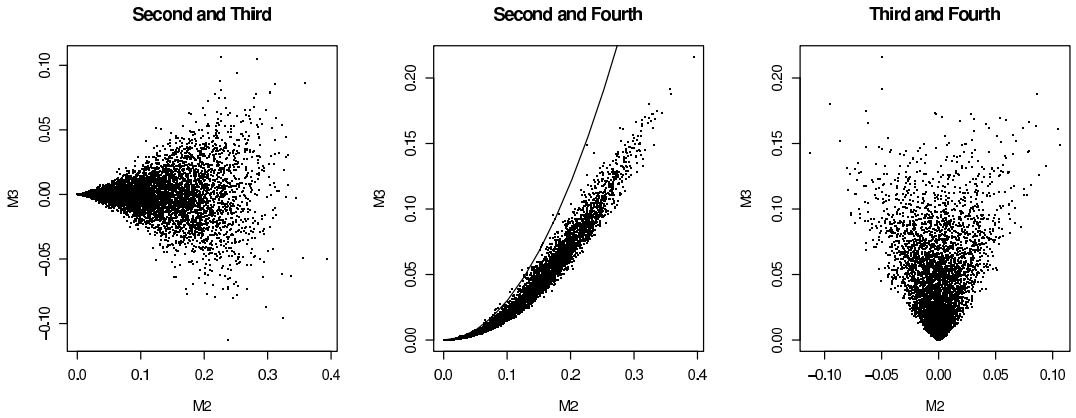

Figure 3: The space of moments for proper dispersion models

is not the only possible approach, for example Marriott (2003) uses a simple closed form geometrically based estimator, while Critchley and Marriott (2004) use moment based methods. However the richness of the output of Markov chain Monte Carlo gives the clearest illustration of the power of the local mixture approach. Furthermore Theorem 6 indicates that it is to be expected that the performance of Markov chain Monte Carlo will be extremely good due to the log-concavity of the posterior on the fibre. The boundaries, which are a fundamental part of the geometry, are included in the Monte Carlo algorithm by a simple rejection step.

5 Local mixtures and exponential dispersion families

One natural generalisation of exponential model are exponential dispersion families of the form

$$f(x;\theta,\phi) = \exp\left[\frac{x\theta - k(\theta)}{\phi}\right]\nu_\phi(x), \tag{16}$$

see for example Jorgensen (1997). These are a rich and highly applicable set of modelling families of great importance in the theory of generalised linear models. The moment generating function of (16) is

$$M_X(t) = \exp\left(\frac{k(\theta+\phi t) - k(\theta)}{\phi}\right).$$

Hence the mean parameter is given by $k'(\theta)$ which does not depend on the dispersion parameter ϕ. However the variance is $\phi k''(\theta)$ which depends on ϕ multiplicatively, as do the higher moments. Thus the dispersion gives a way of controlling the variance structure independently of the mean. There is a clear parallel here with the structure of true local

mixture models where the vector value parameter λ controls the higher moment structure in a way which is asymptotically interpretable as the result of mixing.

The additional variation in dispersion models corresponds to scaling, or adding, on the random variable scale. This is different to mixing but still changes the moment structure hence it is interesting to compare dispersion models with local mixtures.

The following result shows when these different ways of thinking about controlling the higher moment structures give formally different sets of models with the interesting special cases of the Normal and Poisson models which are looked at separately.

THEOREM 9 *Let $f(x;\mu,\phi)$ be an exponential dispersion model of the form (16) such that for fixed and known ϕ the corresponding exponential family is a regular natural exponential family. Let μ be the mean parametrization and assume that the variance function $V_f(\mu)$ is a polynomial of degree at most 2. Let $g(x;\mu,\phi,\lambda)$ be a local mixture model based on (16) defined by*

$$g(x;\mu,\phi,\lambda) = f(x;\mu,\phi) + \sum_{i=2}^{r} \lambda_i \frac{\partial^i f}{\partial \mu^i}(x;\mu,\phi).$$

Finally assume that

$$\frac{\partial}{\partial \phi} \log \nu_\phi(x) \tag{17}$$

is not a polynomial in x. Then the local mixture model $g(x;\mu,\phi,\lambda)$ is, locally to $\lambda = 0$, identified in all its parameters.

PROOF: The result is immediate since the scores with respect to the λ parameters are spanned by polynomials by Theorem 2. ■

The condition in Theorem 9 on the normalising factor $\nu_\phi(x)$ is quite general and applies to the binomial, gamma, negative binomial and many other families, see Jorgensen (1997, pp 85-91). However there are two very important special cases for which this condition does not hold. These are the normal and Poisson families. These cases are treated here separately as they require a rather more detailed analysis.

EXAMPLE 2 (Revisited) The well-known fact that a mixture of normal densities $\phi(x;\mu,\sigma^2)$ can itself be normal, in particular that

$$\int \phi(z;\mu+\eta,\sigma_1^2)\phi(\eta;0,\sigma_2^2)d\eta = \phi(z;\mu,\sigma_1^2+\sigma_2^2) \tag{18}$$

makes one suspect that the identification issue for this family must be quite delicate. Direct calculations show that the condition on $\nu_\phi(x)$ in Theorem 9 does not hold. Figure 3 throws light on this non-identification issue. The solid line plotted in the middle panel corresponds to the moment structure of the normal family with a fixed mean and varying the dispersion parameter. It can be seen that this non-identified case seems to form a boundary in the

set of parameters which can be achieved as true local mixtures of the form of (6). Thus normal models are, in some sense, close to being identified with the difficult case given by (18) being a boundary case.

Jorgensen (1997, p 90) shows how to write the Poisson family as an additive exponential dispersion model,

$$\frac{\lambda^z}{z!}\exp(\theta z - \lambda\exp(\theta)),$$

but also shows that this is the only family for which θ and λ are not identified, since $\mu = \lambda\exp(\theta)$. Hence this family falls outside the regularity conditions of Theorem 9.

Acknowledgments

Part of this work was undertaken while the authors were visiting the Institute of Statistics and Decision Sciences, Duke University. The authors would also like to thank Frank Critchley and Paul Vos for many helpful discussions. Also the authors would like to thank the Universidad Nacional Autnoma de México for financial support.

References

Amari, S.-I. (1985). *Lecture notes in statistics-28: Differential-geometrical methods in statistics*. Springer-Verlag Inc.

Anaya-Izquierdo, K. (2006). *Statistical and Geometric Analysis of Local Mixture Models and a Proposal of some New Tests of Fit for Censored Data*. Ph. D. thesis, Universidad Nacional Autónoma de México.

Anaya-Izquierdo, K. and P. Marriott (2006). Local mixtures of negative exponential distribution. *submiited to AISM*.

Barvinok, A. (2002). *A course in convexity*. Graduate studies in mathematics Vol. 54. AMS.

Critchley, F. and P. Marriott (2004). Data-informed influence analysis. *Biometrika 91*(1), 125–140.

Jorgensen, B. (1997). *The theory of dispersion models*. Chapman & Hall Ltd.

Letac, G. (1992). *Monografias de Matemtica No. 50: Lectures on natural exponential families and their variance functions*. Instituto de Matematica Pura e Eplicada, Rio de Janeiro.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. Institute of Mathematical Statistics.

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), 567–576.

McLachlan, G. J. and D. Peel (2001). *Finite mixture models*. John Wiley & Sons.

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

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

Phelps, R. (1966). *Lectures on Choquet's theorem*. D van Nostrand Inc.

Shaked, M. (1980). On mixtures from exponential families. *Journal of the Royal Statistical Society, Series B, Methodological 42*, 192–198.

Teuscher, F. and V. Guiard (1995). Sharp inequalities between skewness and kurtosis for unimodal distributions. *Statistics & Probability Letters 22*, 257–260.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical analysis of finite mixture distributions*. John Wiley & Sons.

Wong, R. (2001). *Asymptotic approximations of integrals*. Classics in applied mathematics. SIAM.

Wood, G. R. (1999). Binomial mixtures: geometric estimation of the mixing distribution. *The Annals of Statistics 27*(5), 1706–1721.

Appendix A: Regularity Conditions

The following conditions on $\mathcal{F}$ are basically taken from Amari (1985, p. 16).

DEFINITION 10 We will say that the parametric family $\mathcal{F}$ is *regular* if it satisfies the following conditions

(A1) There exists a measure ν on $\mathbf{R}$ such that the measures generated by the members of $\mathcal{F}$ are equivalent to ν. This implies that all measures in $\mathcal{F}$ have common support, so they are mutually absolutely continuous.

(A2) Θ is an open subset of $\mathbf{R}$.

(A3) Every density $f(x;\theta)$ is smooth as a function of θ a.e.$[\nu]$ and partial derivatives $\frac{\partial^i}{\partial\theta^i}$ and integration with respect to the measure ν always commute.

(A4) For all $\theta \in \Theta$, the random variables

$$\frac{\dfrac{\partial^k f(x;\theta)}{\partial\theta^k}}{f(x;\theta)}\,, \qquad k = 1, 2, \ldots$$

are square integrable with respect to the measure $f(x;\theta)\nu(dx)$

(A5) For each $\theta_0 \in \Theta$ there exist ν integrable functions $h_k(x,\theta_0)$ for $k = 1, 2, \ldots$, such that

$$\left|\frac{d^k f(x;\theta)}{d\theta^k}\right| \leq h_k(x,\theta_0)$$

and

$$\int h_k(x,\theta_0) f(x;\theta)\, d\nu(x) < \infty\,,$$

holds in a neighbourhood N_{θ_0} of θ_0 a.e. $[\nu]$.

Appendix B: Proofs

Throughout this discussion let $d : \mathbf{R} \to \mathbf{R}$ be a nonnegative twice continuously differentiable function satisfying $d(0) = 0$, $d(\mu) > 0$ for $\mu \neq 0$ and $d''(0) > 0$.

DEFINITION 11 A model is a *proper dispersion model* if the unit deviance d is regular and the density takes the form

$$q(\mu; m, \epsilon) = a(\epsilon) V^{-1/2}(\mu) \exp\left\{-\frac{1}{2\epsilon} d(\mu; m)\right\} \tag{19}$$

for μ, m in the parameter space, a suitable function a and $V(\mu)$ the unit variance function defined as $V(\mu) = 2\left(\frac{\partial^2 d}{\partial \mu^2}(\mu, \mu)\right)^{-1}$, for details see Jorgensen (1997).

DEFINITION 12 If the integral

$$\frac{1}{a(\epsilon)} = \int \exp\left\{-\frac{1}{2\epsilon} d(\mu - m)\right\} d\mu$$

is finite for $\epsilon \in (0, \epsilon_0)$ for some $\epsilon_0 > 0$ then

$$q(\mu; m, \epsilon) = a(\epsilon) \exp\left\{-\frac{1}{2\epsilon} d(\mu - m)\right\} \tag{20}$$

is defined to be a *location-dispersion model* with location parameter m. $\mathcal{Q}_{ld}$ is then defined as being the set of all *location-dispersion models*.

DEFINITION 13 If the integral

$$\frac{1}{c(\epsilon)} = \int \exp\left\{-\frac{1}{2\epsilon} d(\mu/m)\right\} d\mu$$

is finite for $\epsilon \in (0, \epsilon_0)$ for some $\epsilon_0 > 0$ then

$$q(\mu; m, \epsilon) = c(\epsilon) \exp\left\{-\frac{1}{2\epsilon} d(\mu/m)\right\} d\mu \tag{21}$$

is defined to be a *scale-dispersion model* with scale parameter m. $\mathcal{Q}_{sd}$ is then defined as being the set of all *scale-dispersion* models.

LEMMA 4 *The first four moments (centered at ϑ) of a proper dispersion model $q(\theta;\vartheta,\epsilon)$ have the following asymptotic expansions:*

$$\begin{aligned}
E_Q[\theta-\vartheta] &\sim B_1(\vartheta)\,\epsilon + B_2(\vartheta)\,\epsilon^2 + O_\vartheta(\epsilon^3)\,,\\
E_Q[(\theta-\vartheta)^2] &\sim 2\,C_1(\vartheta)\,\epsilon + 2\,C_2(\vartheta)\,\epsilon^2 + O_\vartheta(\epsilon^3)\\
E_Q[(\theta-\vartheta)^3] &\sim 6D_1(\vartheta)\epsilon^2 + O_\vartheta(\epsilon^3)\\
E_Q[(\theta-\vartheta)^4] &\sim 24E_1(\vartheta)\,\epsilon^2 + O_\vartheta(\epsilon^3)\,, \qquad \forall\,\vartheta\in\Theta
\end{aligned} \tag{22}$$

as $\epsilon\to 0$, where

$$\begin{aligned}
B_1(\vartheta) &= -\left[\frac{V^2 d_3 + 2V'}{4}\right]\\
B_2(\vartheta) &= -\frac{3(V')^3}{4V} + V'V'' - \frac{V}{16}\left\{4V''' + 5d_3(V')^2\right\} + \frac{V^2}{8}\left\{d_4V' + 2d_3V''\right\}\\
&\quad -\frac{V^3}{16}\left\{d_5 + 2V'd_3^2\right\} + \frac{V^4}{6}d_3d_4 - \frac{5V^5d_3^3}{64}\\
C_1(\vartheta) &= V/2
\end{aligned}$$

$$\begin{aligned}
C_2(\vartheta) &= \frac{3(V')^2}{8} - \frac{VV''}{4} + \frac{V^2V'd_3}{4} - \frac{V^3d_4}{8} + \frac{5V^4d_3^2}{32}\\
D_1(\vartheta) &= -\left[\frac{VV'}{4} + \frac{5V^3d_3}{24}\right]\\
E_1(\vartheta) &= \frac{[V(\vartheta)]^2}{8}
\end{aligned}$$

Also V, V', V'', V''' are the variance function and its derivatives evaluated at $\theta=\vartheta$ and

$$d_i = d_i(\vartheta,\vartheta) := \frac{\partial^i}{\partial\theta^i}\,d(\theta,\vartheta)\Big|_{\theta=\vartheta} \qquad i=3,4,5.$$

PROOF: This follow from direct computation. ∎

LEMMA 5 *The second, third and fourth centered (at the mean) moments of a proper dispersion model have the following asymptotic expansions as* $\epsilon \to 0$

$$\begin{aligned} E_Q[(\theta - E_Q[\theta])^2] &\sim V(\vartheta)\epsilon + [2C_2(\vartheta) - B_1^2(\vartheta)]\epsilon^2 + O_\vartheta(\epsilon^3) \\ E_Q[(\theta - E_Q[\theta])^3] &\sim 6[D_1(\vartheta) - B_1(\vartheta)C_1(\vartheta)]\epsilon^2 + O_\vartheta(\epsilon^3) \\ E_Q[(\theta - E_Q[\theta])^4] &\sim 24E_1(\vartheta)\epsilon^3 + O_\vartheta(\epsilon^3) \end{aligned}$$

PROOF: The results results from Lemma 4 and that

$$\begin{aligned} E_Q[(\theta - E_Q[\theta])^2] &= E_Q[(\theta - \vartheta)^2] - (E_Q[\theta - \vartheta])^2 \\ &\sim V(\vartheta)\epsilon + [2C_2(\vartheta) - B_1^2(\vartheta)]\epsilon^2 + O_\vartheta(\epsilon^3) \\ E_Q[(\theta - E_Q[\theta])^3] &= E_Q[(\theta - \vartheta)^3] + 2(E_Q[\theta - \vartheta])^3 - 3E_Q[\theta - \vartheta]E_Q[(\theta - \vartheta)^2] \\ &\sim 6[D_1(\vartheta) - B_1(\vartheta)C_1(\vartheta)]\epsilon^2 + O_\vartheta(\epsilon^3) \\ E_Q[(\theta - E_Q[\theta])^4] &= E_Q[(\theta - \vartheta)^4] - 4E_Q[(\theta - \vartheta)^3]E_Q[(\theta - \vartheta)] \\ &\quad +6E_Q[(\theta - \vartheta)^2](E_Q[(\theta - \vartheta)])^2 - 3(E_Q[(\theta - \vartheta)])^4 \\ &\sim E_Q[(\theta - \vartheta)^4] + O_\vartheta(\epsilon^3) = 24E_1(\vartheta)\epsilon^2 + O_\vartheta(\epsilon^3) \end{aligned}$$

■

Theorem 7 PROOF: For clarity in the exposition, only the proof in the case $r = 2$ which is enough for our purposes. Extensions to higher r are straightforward. Using Laplace's method in both numerator and denominator of $g(x; Q(\theta, \vartheta, \epsilon))$ and then dividing the two series, one obtains (after a considerable amount of algebra),

$$f(x;\vartheta) + \sum_{i=1}^{4} A_i(\vartheta, \epsilon) f^{(i)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^3)$$

as $\epsilon \to 0$, where

$$\begin{aligned}
A_1(\vartheta,\epsilon) &= B_1(\vartheta)\,\epsilon + B_2(\vartheta)\,\epsilon^2 \\
A_2(\vartheta,\epsilon) &= C_1(\vartheta)\,\epsilon + C_2(\vartheta)\,\epsilon^2 \\
A_3(\vartheta,\epsilon) &= D_1(\vartheta)\epsilon^2 \\
A_4(\vartheta,\epsilon) &= E_1(\vartheta)\epsilon^2 \\
R_a(x;\vartheta,\epsilon) &= \sum_{k=5}^{\infty} R_i(\vartheta,\epsilon) f^{(k)}(x;\vartheta) = O_{x,\vartheta}(\epsilon^3)\,,
\end{aligned}$$

for some functions R_i for $i \geq 5$ and $B_1, B_2, C_1, C_2, D_1, E_1$ are defined in the proof of Lemma 4. One obtains the same result by making use (in the denominator) of the saddlepoint approximation of a proper dispersion model, see Jorgensen (1997), which states that

$$a(\epsilon) \sim \sqrt{2\pi\epsilon}$$

as $\epsilon \to 0$. Use of the formulae given in Lemma 4 states the first form of the expansion stated in this theorem. Now define

$$\begin{aligned}
M_1(\vartheta,\epsilon) &:= \vartheta + A_1(\vartheta,\epsilon) + O_\vartheta(\epsilon^3) \\
M_2(\vartheta,\epsilon) &:= \frac{2C_1(\vartheta)\epsilon + [2C_2(\vartheta) - B_1^2(\vartheta)]\epsilon^2}{2} \\
M_3(\vartheta,\epsilon) &:= \epsilon^2[D_1(\vartheta) - B_1(\vartheta)C_1(\vartheta)] \\
M_4(\vartheta,\epsilon) &:= E_1(\vartheta)\epsilon^2\,.
\end{aligned}$$

Using Taylor's Theorem (with $\delta_\vartheta(\epsilon) := A_1(\vartheta,\epsilon)+O_\vartheta(\epsilon^3)$ as the increment) on $f(x;M_1(\vartheta,\epsilon))$ and $f^{(i)}(x;M_1(\vartheta,\epsilon))$ for $i=2,3,4$ we obtain

$$\begin{aligned}
& f(x;M_1(\vartheta,\epsilon)) + \sum_{i=2}^{4} M_i(\vartheta,\epsilon) f^{(i)}(x;M_1(\vartheta,\epsilon)) \\
= \;& f(x;\vartheta+\delta_\vartheta(\epsilon)) + \sum_{i=2}^{4} M_i(\vartheta,\epsilon) f^{(i)}(x;\vartheta+\delta_\vartheta(\epsilon)) \\
\sim \;& f(x;\vartheta) + [B_1(\vartheta)\epsilon + B_2(\vartheta)\epsilon^2] f^{(1)}(x;\vartheta) + [C_1(\vartheta)\epsilon + C_2(\vartheta)\epsilon^2] f^{(2)}(x;\vartheta) \\
& + D_1(\vartheta)\epsilon^2 f^{(3)}(x;\vartheta) + \frac{V^2(\vartheta)\epsilon^2}{8} f^{(4)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^3) \\
= \;& f(x;\vartheta) + \sum_{i=1}^{4} A_i(\vartheta,\epsilon) f^{(i)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^3)
\end{aligned}$$

as $\epsilon \to 0$. By making use of the formulae given in Lemma 5 we can state the second form of the expansion in the theorem. The bounded derivatives assumption implies uniformity in x as described in Marriott (2002). ■

Corollary 1

PROOF: Proceeding as in the proof of Theorem 7 we have that as $\epsilon \to 0$

$$g(x;Q(\theta;\vartheta,\epsilon)) \sim f(x;\vartheta) + \sum_{i=1}^{4} \vartheta^i A_i^*(\epsilon) f^{(i)}(x;\vartheta) + O_{x,\vartheta}(\epsilon^3)$$

where

$$\begin{aligned}
A_1^*(\epsilon) &= -\epsilon\left(1+\frac{d_0^{(3)}}{4}\right) \\
&\quad +\epsilon^2\left(\frac{d_0^{(3)}d_0^{(4)}}{6} - 2 + \frac{d_0^{(4)}}{4} - \frac{d_0^{(5)}}{16} - \frac{3d_0^{(3)}}{4} - \frac{5[d_0^{(3)}]^3}{64} - \frac{[d_0^{(3)}]^2}{4}\right) \\
A_2^*(\epsilon) &= \frac{\epsilon}{2} + \epsilon^2\left(\frac{5[d_0^{(3)}]^2}{32} + 1 - \frac{d_0^{(4)}}{8} + \frac{d_0^{(3)}}{2}\right) \\
A_3^*(\epsilon) &= -\epsilon^2\left(\frac{1}{2} + \frac{5d_0^{(3)}}{24}\right) \\
A_4^*(\epsilon) &= \frac{\epsilon^2}{8}
\end{aligned}$$

where now $d_0(u)$ is a function with absolute minimum at $u = 1$ and $d_0^{(i)}$ for $i = 3, 4, 5$ are its third to fifth derivatives evaluated at that minimum. The rest of the proof follows as in the proof of Theorem 7 and therefore will be omitted. The formula for the M_i^* functions are

$$\begin{aligned} M_1^*(\epsilon) &= A_1^*(\epsilon) \\ M_2^*(\epsilon) &= \frac{\epsilon}{2} + \epsilon^2 \left[\frac{[d_0^{(3)}]^2}{8} + \frac{1}{2} - \frac{d_0^{(4)}}{8} + \frac{d_0^{(3)}}{4} \right] \\ M_3^*(\epsilon) &= -\frac{\epsilon^2 d_0^{(3)}}{12} \\ M_4^*(\epsilon) &= \frac{\epsilon^2}{8} \end{aligned}$$

The formulae for the functions A_i^* and M_i^* are

$$\begin{aligned} A_1^*(\epsilon) &= -\epsilon \left(\frac{d_0^{(3)}}{4} \right) + \epsilon^2 \left(\frac{d_0^{(3)} d_0^{(4)}}{6} - \frac{d_0^{(5)}}{16} - \frac{5[d_0^{(3)}]^3}{64} \right) \\ A_2^*(\epsilon) &= \frac{\epsilon}{2} + \epsilon^2 \left(\frac{5[d_0^{(3)}]^2}{32} - \frac{d_0^{(4)}}{8} \right) \\ A_3^*(\epsilon) &= -\epsilon^2 \left(\frac{5 d_0^{(3)}}{24} \right) \\ A_4^*(\epsilon) &= \frac{\epsilon^2}{8} \end{aligned}$$

where now $d_0(u)$ is a function with absolute minimum at $u = 0$ and $d_0^{(i)}$ for $i = 3, 4, 5$ are its third to fifth derivatives evaluated at that minimum. Also we have

$$\begin{aligned} M_1^*(\epsilon) &= A_1^*(\epsilon) \\ M_2^*(\epsilon) &= \frac{\epsilon}{2} + \epsilon^2 \left[\frac{[d_0^{(3)}]^2}{8} - \frac{d_0^{(4)}}{8} \right] \\ M_3^*(\epsilon) &= -\frac{\epsilon^2 d_0^{(3)}}{12} \\ M_4^*(\epsilon) &= \frac{\epsilon^2}{8} \end{aligned}$$

■

Appendix C

The polynomials in the local mixture expansion of the Binomial model are given by

$$
\begin{aligned}
p_2(x,\mu) &= \frac{1}{\mu^2\,(n-\mu)^2}\left\{x^2n^2+n\,(2\,\mu-2\,\mu\,n-n)\,x+n\left(\mu^2n-\mu^2\right)\right\}\\
p_3(x,\mu) &= \frac{n}{\mu^3(n-\mu)^3}\left\{x^3n^2+\left(-3\,\mu\,n^2+6\,\mu\,n-3\,n^2\right)x^2+\right.\\
&\quad \left.\left(-6\,\mu\,n+3\,\mu^2n^2-9\,\mu^2n+6\,\mu^2+3\,\mu\,n^2+2\,n^2\right)x-2\,\mu^3-\mu^3n^2+3\,\mu^3n\right\}\\
p_4(x,\mu) &= \frac{n}{\mu^4(n-\mu)^4}\{x^4n^3+\left(-6\,n^3+12\,\mu\,n^2-4\,\mu\,n^3\right)x^3+\\
&\quad \left(-36\,\mu\,n^2-30\,\mu^2n^2+36\,\mu^2n+11\,n^3+12\,\mu\,n^3+6\,\mu^2n^3\right)x^2+\\
&\quad \left(-36\,\mu^2n+24\,\mu^3n^2+24\,\mu^3+30\,\mu^2n^2-6\,n^3-4\,\mu^3n^3-8\,\mu\,n^3-44\,\mu^3n-6\,\mu^2n^3+24\,\mu\,n^2\right)x-\\
&\quad 6\,\mu^4n^2+11\,\mu^4n-6\,\mu^4+\mu^4n^3\}
\end{aligned}
$$

Appendix D

Then the derivatives of the exponential densities have a simple form

$$
\begin{aligned}
\frac{df(x;\theta)}{d\theta} &= f(x;\theta)\left[x-k'_\nu\right] \qquad (23)\\
\frac{d^2f(x;\theta)}{d\theta^2} &= f(x;\theta)\left[(x-k'_\nu)^2-k''_\nu\right]\\
\frac{d^3f(x;\theta)}{d\theta^3} &= f(x;\theta)\left[(x-k'_\nu)^3-3(x-k'_\nu)k''_\nu-k'''_\nu\right]\\
\frac{d^4f(x;\theta)}{d\theta^4} &= f(x;\theta)\left[(x-k'_\nu)^4-6(x-k'_\nu)^2k''_\nu-4(x-k'_\nu)k'''_\nu+3(k''_\nu)^2-k''''_\nu\right]
\end{aligned}
$$

and so on. Here $k'_\nu, k''_\nu, k'''_\nu, k''''_\nu$ are just the first four derivatives of k_ν evaluated at θ. If the mean parametrization $\mu = k'_\nu(\theta)$ is used and denote the variance function by $V_f(\mu)$ then

these expressions turn into

$$\begin{aligned}
\frac{df(x;\mu)}{d\mu} &= f(x;\mu)\left[\frac{x-\mu}{V_f(\mu)}\right] \\
\frac{d^2f(x;\mu)}{d\mu^2} &= f(x;\mu)\left[\frac{(x-\mu)^2-(x-\mu)V_f'(\mu)-V_f(\mu)}{V_f^2(\mu)}\right] \\
\frac{d^3f(x;\mu)}{d\mu^3} &= f(x;\mu)\left[\frac{(x-\mu)^3-3(x-\mu)^2V_f'(\mu)}{V_f^3(\mu)}\right. \\
&\quad \left.+\frac{-(x-\mu)\left[3V_f(\mu)+V_f(\mu)V_f''(\mu)-2[V_f'(\mu)]^2\right]+2V_f(\mu)V_f'(\mu)}{V_f^3(\mu)}\right] \\
\frac{d^4f(x;\mu)}{d\mu^4} &= \frac{f(x;\mu)}{V_f^4(\mu)}\left\{(x-\mu)^4-6(x-\mu)^3V_f'(\mu)\right. \\
&\quad -(x-\mu)^2[6V_f(\mu)+4V_f(\mu)V_f''(\mu)-11[V_f'(\mu)]^2] \\
&\quad +(x-\mu)[14V_f(\mu)V_f'(\mu)+6V_f(\mu)V_f'(\mu)V_f''(\mu)-6[V_f'(\mu)]^3-V_f(\mu)^2V_f'''(\mu)] \\
&\quad \left.-6V_f(\mu)[V_f'(\mu)]^2+3V_f^2(\mu)V_f''(\mu)+3V_f^2(\mu)\right\}.
\end{aligned}$$